META-ANALYSIS OF THE RELATION BETWEEN MENTAL HEALTH PROFESSIONALS' CLINICAL AND EDUCATIONAL EXPERIENCE AND JUDGMENT ACCURACY: REVIEW OF CLINICAL JUDGMENT RESEARCH FROM 1997 TO 2010

A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

BY

LOIS A. PILIPIS

DISSERTATION ADVISOR: DR. PAUL M. SPENGLER

BALL STATE UNIVERSITY

MUNCIE, INDIANA

JULY 2010

META-ANALYSIS OF THE RELATION BETWEEN MENTAL HEALTH PROFESSIONALS' CLINICAL AND EDUCATIONAL EXPERIENCE AND JUDGMENT ACCURACY: REVIEW OF CLINICAL JUDGMENT RESEARCH FROM 1997 TO 2010

A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

BY

LOIS A. PILIPIS

Approved by:

_____	_____
Committee/Dissertation Chair	Date
_____	_____
Committee Member	Date
_____	_____
Committee Member	Date
_____	_____
Committee Member	Date

BALL STATE UNIVERSITY

MUNCIE, INDIANA

JULY 2010

Acknowledgements

I would like to thank everyone who helped support me throughout the dissertation writing process, namely my doctoral committee members, Dr. Paul Spengler, Dr. W. Holmes Finch, Dr. Carrie Ball, and Dr. Annette Leitze. I would also like to thank my School Psychology professors who have been available to guide and support me throughout my program.

Additionally, I would like to thank my family and friends who have patiently awaited the completion of my degree. Without their love and support, none of this would have been possible.

Table of Contents

Abstract

Researchers have addressed many clinician and client attributes in relation to the accuracy of judgments made by mental health professionals. One such moderator addressed clinicians' judgment accuracy in relation to experience. Contrary to what many clinicians expect, a number of studies have failed to demonstrate a positive correlation between judgment accuracy and experience (e.g., Berman & Berman, 1984; Ruscio & Stern, 2005; Schinka & Sines, 1974). In Spengler et al. (2009), the relationship between judgment accuracy and experience was assessed via a large-scale meta-analysis that examined studies of clinical judgment and experience from 1970 to 1996. The result was a small but reliable, homogeneous effect demonstrating a positive correlation between judgment accuracy and experience. The Spengler et al. meta-analysis found relatively few significant moderator effects influencing the experience-accuracy effect, namely the type of judgment made by clinicians, the criterion validity of accuracy measures used, and publication source. In the present study, results from clinical judgment and experience studies from 1997 to 2010 were combined in a meta-analysis. An update and extension allowed for cross-validation of the Spengler et al. meta-analysis with more recent research as well as an exploration of additional moderator variables, such as profession type and inclusion of non-mental health participants. The overall effect was .16, with a 95 percent confidence interval that was above zero (CI = .05 to .26). This overall effect indicated experience significantly impacted judgment accuracy, consistent with expectations. The overall effect was shown to be heterogeneous, indicating the *Q* statistic was sufficiently large to reject the null hypothesis regarding homogeneity of the effect size distribution. Exploratory analyses revealed the presence

of two significant moderator variables, namely judgment type and publication source. Limitations included lack of variability of judgment type and difficulty with or complete inability to assess other potential moderators of interest, such as feedback and utilization of test protocols for the stimulus measure. Other limitations included utilization of a less exhaustive search strategy, in which some relevant studies may have been missed. Despite limitations, the results of the present meta-analysis largely replicated those of the Spengler et al. meta-analysis.

Chapter 1: Introduction

It is generally believed by clinicians in psychology as well as their clients that clinical judgments evolve and improve as a result of increased experience. These experiences can consist of more face-to-face time with clients, further training in specific skill areas, increased supervision, or even continuing education courses, for example. Regardless of the types of experiences obtained, many clinicians and their clients would assume these experiences have had a direct impact on clinicians' abilities to make accurate and well-informed clinical judgments, whether those are decisions related to diagnosis, prognosis, or intervention choice. However, clinical judgment research has not been completely supportive of a positive correlation between judgment accuracy and experience (e.g., Berman & Berman, 1984; Ruscio & Stern, 2005; Schinka & Sines, 1974). In these studies, judgment accuracy has not been shown to improve with experience, and in some, judgment accuracy has even been shown to worsen with experience. By contrast, other studies have shown a positive correlation between experience and judgment accuracy (e.g., Berven, 1985; Brammer, 2002; Rerick, 1999; Wilkin, 2001).

Purpose of the present study

The present study aimed to clarify the relationship between clinical judgment accuracy and experience through an update and extension of the Spengler et al. (2009) meta-analysis. The Spengler et al. meta-analysis examined clinical judgment studies between the years 1970 and 1996 to determine whether or not more experience is correlated with improved judgment accuracy. The Spengler et al. meta-analysis was part of a more comprehensive project known as the Meta-Analysis of Clinical Judgment

project (MACJ; Spengler et al.) in which a comprehensive search strategy was performed locating every clinical judgment study within the given time span and entering each study into a database for examination. The Spengler et al. meta-analysis revealed an overall effect of .12, indicating a modest, positive relationship between experience and judgment accuracy. The corresponding confidence interval had a lower limit that was greater than zero, indicating more experienced clinicians were more accurate in their judgments. The effect in the Spengler et al. meta-analysis was shown to be homogeneous after the removal of one outlier study (Garcia, 1993). Exploratory analyses were conducted to assess the impact of moderator variables. There were very few moderator variables shown to have a statistically significant effect, including type of judgment made by clinicians, the criterion validity of accuracy measures used, and publication source. A replication and extension of the Spengler et al. meta-analysis allowed for analysis of the experience-accuracy effect with more recent research as well as an examination of additional moderator variables. Additionally, the present meta-analysis allowed for a test of the robustness of the Spengler et al. findings. In order to gain perspective on the history of clinical judgment research as well as the state of the current research base, however, it will be necessary to examine various factors related to the judgment accuracy-experience debate, such as how experience and expertise differ, the role of cognitive biases and heuristics in clinical judgment, and the specific performance differences assumed to be present when novice clinicians are compared to more experienced clinicians. A discussion of the aforementioned factors will be necessary to fully understand the complex relationship between clinical judgment and experience.

Finally, this chapter provides a rationale for the present study and discusses implications of the present study's findings for the mental health field.

Experience-expertise distinction

In order to fully understand the relationship between clinical judgment and experience, one must first understand the concept of expertise. Frensch and Sternberg (1989) defined expertise as "the ability, acquired by practice and experience, to perform qualitatively well in a particular task domain" (p. 189). This definition cites practice and experience as the catalysts for achieving expert status. However, an experienced counselor should not necessarily be considered an expert based on having worked a certain number of years in the field (Eells, Lombart, Kendjelic, Turner, & Lucas, 2005). Moreover, increased experience does not necessarily lead to expertise (Eells et al.; Glaser & Chi, 1988; Sedlmeier, 2005).

Although there has been some empirical support for the quality of expert judgments being better than those of novices, it has also been noted that experts perform tasks differently than novices (Bereiter & Scardamalia, 1986; Chan, 2006; Chi, Glaser, & Rees, 1982; Ericsson, 2005; Glaser & Chi, 1988; Haerem & Rau, 2007; Hinsley, Hayes, & Simon, 1978; Leon & Perez, 2001; Lesgold et al., 1988; Nickerson, Perkins, & Smith, 1985; O'Reilly, Parlette, & Bloom, 1980; Polanyi, 1962). Some of these expert differences include spending greater amounts of time understanding or analyzing problems before attempting solutions, displaying a larger knowledge base, displaying a better-organized knowledge base, having the ability to perceive large, meaningful patterns in their domain, having greater automatic processing, displaying superior long- and short-term memory, having the ability to perceive and represent a problem in their

domain at a meaningful level, and displaying more advanced self-monitoring skills (Chan; Chi et al.,; Glaser & Chi; Ericsson; Haerem & Rau; Hinsley et al.; Leon & Perez; Lesgold et al.; Nickerson et al.; O'Reilly et al.; Polanyi; Sedlmeier, 2005). According to Frensch and Sternberg (1989), however, not all of these differences result in better quality judgments for experts. For example, they indicated the increased ability to automatize knowledge and the possession of a larger and more structured knowledge base has prevented many experts from outperforming novices on certain tasks. For example, expert clinicians may be able to form cohesive case conceptualizations more rapidly at intake; however, the costs of these more organized and automatic processes as compared to novices are sometimes reflected in the under emphasis of or failure to identify important client data. Frensch and Sternberg asserted that expert performance was likely to be poorer than that of novices when basic-level information or nonintegrated information had to be retrieved, when experts were called upon to restructure their existing knowledge, and when existing knowledge had to be deliberately or consciously selected. Due to the somewhat subjective and abstract nature of clinical decision-making, these demands are likely to occur frequently. Furthermore, the variability in individual clients and their presenting problems make it likely that clinicians will be called upon to restructure old knowledge bases and make conscious, newly constructed, and deliberate treatment decisions. The aforementioned cognitive shortcomings of expert clinicians noted by Frensch and Sternberg are prevalent in non-mental health professions to varying degrees. The following is a more in-depth investigation of some of the mechanisms underlying mental health clinicians' faulty judgments, specifically biases based on client variables and cognitive heuristics.

Bias interference on clinical judgment

Without adding the variable of experience into the equation, research has shown that clinicians operate with a degree of personal and professional bias (Agell & Rothblum, 1991; Arkes, Wortmann, Saville, & Harkness, 1981; Biaggio, Roades, Staffelbach, Cardinali, and Duffy, 2000; Bowers & Bieschke, 2005; Clavelle & Turner, 1980; Coutinho, Oswald, Best, & Forness, 2002; Dailey, 1952; Davis-Coelho, Waltz, & Davis-Coelho, 2000; Fernbach, Winstead, & Derlega, 1989; Friedlander & Stockman, 1983; Gauron & Dickinson, 1966; Hansen & Reekie, 1990; Herbert, Nelson, & Herbert, 1988; Langer & Abelson, 1974; Lewis, Croft-Jeffreys, & David, 1990; McNiel & Binder, 1995; Mohr, Israel, & Sedlacek, 2001; Neighbors, Trierweiler, Ford, & Muroff, 2003; Oskamp, 1965; Pfeiffer, Whelan, & Martin, 2000; Raines & Rohrer, 1955; Richards & Wierzbicki, 1990; Rosenhan, 1973; Sandifer, Hordern, & Green, 1970; Snyder, 1977; Spengler, 2000; Spengler, Strohmer, Dixon & Shivy, 1995; Strohmer & Leierer, 2000; Temerlin, 1968; Teri, 1982; Trachtman, 1971; Wilson, 2000; Whaley, 2001). In some of these studies, clinicians' judgments of the degree and type of pathology displayed by clients were significantly affected by the clients receiving a label of "patient" or "client." In other words, individuals labeled as patients or clients were often viewed as exhibiting more pathology than individuals labeled as "normal" with the exact same behaviors or characteristics (e.g., Herbert et al.; Langer & Abelson; Snyder, 1977). Additionally, clients' socioeconomic statuses have been shown to influence clinicians' judgments. Trachtman, for example, found that social class had a significant impact on clinicians' assessment of clients' Rorschach responses. The clinicians' judgments were based upon two identical Rorschach protocols with only the social class varied. Specifically, clients

of lower socioeconomic status were rated more negatively by clinicians as compared to clients of middle socioeconomic status. Furthermore, Lopez (1989) found that clinicians were negatively biased towards those diagnosed with developmental disabilities and members of relatively low socioeconomic status when clinicians were given identical sets of client information with only developmental disability and socioeconomic statuses varied.

Other potential sources of bias include client gender, sex-role, race, and age. For example, in some studies, clinicians gave poorer prognoses to females when identical case histories were used for both genders (Agell & Rothblum, 1991; Fernbach et al., 1989; Hansen & Reekie, 1990; Teri, 1982). Gender bias was especially prevalent when clinicians were asked to predict the occurrence of violence. In other words, violence was more likely to be predicted with male clients than female clients (Lewis et al., 1990; Lidz, Mulvey, & Gardner, 1993; McNiel & Binder, 1995). These gender bias studies involved providing clinicians with identical case vignettes with only the gender of the client varied. This finding is sometimes dependent, however, on the gender of the clinician. Elbogen, Williams, Kim, Tomkins, and Scalora (2001) found that female clinicians predicted male psychiatric patients as more dangerous than did male clinicians. When it comes to bias based on clients' sex-role, clinicians were sometimes found to predict that clients who possessed stereotypical traits of their sex had a better prognosis for treatment response as compared to clients with identical information who did not possess stereotypical sex-role traits. For example, Rosenthal (1982) found that lesbian clients with more stereotypically masculine traits were rated to have poorer treatment prognoses than were clients with more stereotypical sex-role traits. Bias based on race has been

especially prevalent in the area of violence prediction. African-American psychiatric inpatients are often predicted by clinicians to be more violent than their Caucasian counterparts, for example (Garb, 1998; Hoptman et al., 1999). Finally, age seemed to play a role in clinicians' judgments of clients. For example, clinicians were often shown to give poorer prognoses to elderly clients compared to middle-aged or younger clients with the same case histories (Ford & Sbordone, 1980; Hansen & Reekie, 1990; Hillman, Stricker, & Zweig, 1997; James & Haley, 1995; Meeks, 1990; Ray, McKinney, & Ford, 1987; Ray, Raciti, & Ford, 1985; Settin, 1982; Wrobel, 1993).

Bias studies such as those described in this section often employ some method of establishing a baseline for the particular clinicians' judgment tendencies. For example, identical case vignettes are often utilized with only the variable of interest manipulated (Abramowitz & Dockeki, 1977). This method allows for comparisons between conditions but often does not allow the researcher to conclude whether or not judgment errors have occurred. One of the reasons for this limitation is due to differences in population base rates for certain events or conditions. For example, in studies of gender bias for judgments of violence likelihood, researchers are often not able to determine whether or not the higher observed likelihood ratings for male clients are due to true bias or clinician consideration of higher violence base rates for males (Lewis et al., 1990; Lidz, Mulvey, & Gardner, 1993; McNiel & Binder, 1995). This limitation occurs in many studies regarding client variable biases due to the observed differences in base rates across populations as well as the lack of a clear-cut standard by which to evaluate clinicians' judgments.

In addition to exhibiting bias based on client and clinician demographic or personality characteristics, bias has been demonstrated in the form of faulty functioning in clinicians' cognitive processes. These types of biases, generally referred to as cognitive heuristics, include confirmatory bias (Wood & Nezworski, 2005), hindsight bias (Belknap, 2000), anchoring and adjustment heuristics (Cioffi, 2001), representativeness heuristics (Nisbett, Krantz, Jepson, Kunda, 1983; Tversky & Kahneman, 1974), and the availability heuristic (Nisbett & Ross, 1980). A more recent addition to the previous set of commonly studied heuristics is known as the affect heuristic (Slovic, Finucane, Peters, & MacGregor, 2002), which is associated with the vividness bias identified by Nisbett and Ross.

The first type of heuristic mentioned, confirmatory bias, refers to "the tendency to seek out information that is consistent with a belief or hypothesis and to ignore or overlook information that is potentially inconsistent" (Davies, 2003, p. 736). One way in which confirmatory bias is tested in research studies is through use of the positive test strategy. The positive test strategy involves testing a hypothesis by searching for instances in which the hypothesized characteristics can be found rather than searching for instances in which the hypothesized characteristics would be absent. For example, Kunda, Fong, Sanitioso, and Reber (1993) asked respondents questions such as, "Are you happy with your social life?" or "Are you extraverted?" Respondents tended to reply with examples that confirmed rather than contradicted the hypothesized characteristic. In addition, more respondents tended to claim they possessed the hypothesized characteristic than respondents who were asked the opposite question. Confirmatory bias has been shown to have an effect on mental health clinicians' judgments regarding their

clients. Haverkamp (1993) examined 65 counseling trainees' hypothesis testing strategies in response to a videotaped counseling session. A tendency to exhibit confirmatory bias was found among the trainees, with a mean of 64 percent of their responses falling in the confirmatory category and only 15 percent labeled as disconfirmatory. Likewise, Strohmer, Shivy, and Chiodo (1990) found confirmatory bias to be present in the way clinicians remembered clinical information. When presented with a written report and later asked to remember and select information, clinicians tended to remember and select more confirmatory than disconfirmatory information. This finding held even when the written report had contained more disconfirmatory information.

Hindsight bias, on the other hand, involves making after-the-fact assessments of a particular outcome. Individuals engaging in hindsight bias have post-hoc knowledge of the particular outcomes and claim to have had this knowledge before the outcomes occurred. Research has shown hindsight bias at work in a variety of real-world situations, for example in the ability to judge sporting events (Bonds-Raacke, Fryer, Nicks, & Durr, 2001). Hindsight bias has been examined in psychology research and was found to be robust across a wide variety of task environments (Ash, 2009). Slovic and Fischhoff (1977), for example, examined hindsight bias in relation to whether or not debiasing strategies would improve clinicians' judgments of the results of a research study. Specifically, they were forced to consider alternatives to the actual results of the study. The clinicians were assigned to either a hindsight or a foresight condition depending on whether they were asked to consider alternatives to a *hypothetical* outcome (foresight) or to an *already existing* outcome (hindsight). Slovic and Fischhoff found that

asking clinicians to consider alternatives resulted in a decrease in the hindsight bias, regardless of whether or not they were asked prior to or after the particular outcome.

Hindsight bias has been partially explained by the use of anchoring and adjustment heuristics (Hawkins & Hastie, 1990). Anchoring and adjustment heuristics involve individuals making changes to their initial assessments based on knowledge of the true outcomes. Two types of anchoring and adjustment heuristics have been proposed. The first, expectation-based adjustments, describes how individuals assess how surprising they found a given outcome and make adjustments to their retrospective judgments in accordance with this information (Müller & Stahlberg, 2007). Research has shown that the more surprising an individual judges an outcome to be, the less likely he or she will claim to have been able to judge the outcome in the first place, leading to a reverse hindsight bias effect. However, if an outcome is judged by an individual to be predictable, that individual will be at risk of employing hindsight bias. Therefore, surprise has been shown to be negatively correlated with hindsight bias (Ash, 2009). The second type of anchoring and adjustment heuristic, known as the experience-based adjustment, describes how individuals adjust their retrospective assessments by using their beliefs about their levels of expertise in the given domains. More specifically, if an individual judges himself or herself to be an expert in a particular domain, the individual will make a rather small adjustment from the outcome. Individuals who believe they lack experience or expertise in a given domain, however, will most likely make a relatively larger adjustment from the outcome. Therefore, the experience-based adjustment is more situation-general than the expectation-based adjustment, which can be described as situation-specific (Ash).

Representativeness heuristics are yet another type of cognitive bias individuals utilize when making judgments. Representativeness heuristics involve individuals making judgments about an object or person by comparing the object or person to another and unknowingly invoking relevant schemata. Representativeness heuristics involve the failure to take into account appropriate base rates for the condition or event being addressed (Gilovich, Griffin, & Kahneman, 2002). According to Ashcraft (2002), "The representativeness heuristic is a judgment rule in which an estimate of the probability of an event is determined by one of two features: how similar the event is to the population of events it came from or whether the event seems similar to the process that produced it" (p. 468). According to Ashcraft's description, base rates are either not known or dismissed by the clinician. Mental health clinicians, for example, may employ the representativeness heuristic when assigning diagnoses based upon comparisons of the clients in question with the prototypical client pertaining to that specific diagnosis without taking the appropriate base rates into account. Garb (1996) tested the representativeness heuristic among psychologists and interns by providing them with case histories and asking them to rate the likelihood of disorder, confidence in their ratings, and how similar they believed the case to be in comparison with the "typical" case corresponding with the particular disorder they chose. Findings revealed a positive correlation between likelihood ratings and similarity ratings, indicating clinicians seemed to judge the particular cases based on how similar those cases were to the perceived prototype of the disorder in question. According to Garb (1996), clinicians in this condition relied upon the representativeness heuristic to form their judgments rather than adhering to base rates or criteria based upon the *Diagnostic and Statistical Manual of*

Mental Disorders, Fourth Edition, Text Revision (DSM-IV-TR; American Psychological Association, 2000). Representative heuristics are especially pertinent to the mental health field due to the illusiveness and covertness of decision-making processes involved (Garb, 1996).

The availability heuristic refers to the tendency to attend to the data that are most readily available (Nisbett & Ross, 1980; Kahneman & Tversky, 1973). When clinicians utilize the availability heuristic, they form judgments of clients based upon the most readily available clinical data, which is usually the most salient and/or memorable data. Much of what we find salient or memorable is influenced by training and/or theoretical orientation as well as personal factors. For example, psychoanalytical clinicians will find clients' early social and emotional development as well as information pertaining to the clients' families of origin to be more useful and salient in case conceptualization than would more behaviorally-oriented clinicians. Psychoanalytically-trained clinicians, in comparison with behaviorally-trained clinicians, will also most likely have attended to this type of data in a more diligent and overt manner, thus increasing the likelihood of remembering the information later. In addition to differences in the utilization of the availability heuristic based upon training and theoretical orientation, clinicians utilize their previously established personal schemata for understanding their clients. Specifically, clinicians often find client data to be more salient and more memorable when the client data are vivid, emotional, amusing, or shocking. Client data that are routine or pallid are often underemphasized or ignored, even though this type of information may have been more diagnostically relevant than the more vivid client data (Dumont, 1993). The availability heuristic often plays a role in vocational

overshadowing. Vocational overshadowing occurs when clinicians underemphasize or ignore clinically relevant data regarding clients' professional problems in favor of what the clinicians perceive to be more interesting problems in the clients' personal lives (Spengler, 2000). In relation to the availability heuristic, clinicians may remember clients' personal problems more vividly than vocational problems, thus ignoring potentially important vocational data.

Finally, the affect heuristic, introduced by Slovic et al. (2002), refers to reliance on emotions and feeling-states to guide decision-making. Affect heuristics, like the previously mentioned heuristics, operate at varying levels of consciousness. According to dual-process theories of cognition, individuals utilize two fundamentally different methods of comprehending reality. The first is a reliance on analytical, deliberative, and verbal means. The second involves relying on intuitive, experiential, and emotional means. Both can be quite irrational and have the potential of leading to inaccurate appraisals and judgments (Epstein, 1994). In studies of relative risk as perceived by the general public, feelings of dread were shown to determine how risky a particular event would be. For example, judges associated radiation exposure from nuclear power plants with higher dread and rated this event as far riskier than radiation from medical x-rays. However, the likelihood of becoming exposed to radiation from a nuclear power plant is much less than that of exposure to radiation from medical x-rays. Most individuals undergo x-rays at some point in their lives and many receive x-rays repeatedly, thus increasing their risk for radiation exposure. However, the participants in Epstein who associated radiation exposure from nuclear power plants with higher feelings of dread judged that type of radiation exposure as far riskier than radiation exposure from medical

x-rays. Assigning greater risk to events inducing more feelings of dread may lead to an overestimation of the actual risk of those events. In the area of clinical judgment, the affect heuristic is believed by Garb (2005) to be of particular importance. Although not yet extensively studied in the clinical judgment research, the affect heuristic has the potential to bias clinicians' initial liking or disliking of clients (especially when based on client variables such as race, gender, age), their appraisals of client malingering, and attributions of client problems. The affect heuristic has relevance for vocational overshadowing, specifically related to clinicians' tendencies to be less empathetic towards vocational problems (Hill, Tanney, & Leonard, 1977; Melnick, 1975). Vocational overshadowing is related to both the affect heuristic as well as the previously discussed availability heuristic, which tends to be interrelated. Since the availability heuristic often involves individuals emphasizing or remembering salient information, it is logical that salient information tends to be more emotion-laden in nature (Nisbett & Ross, 1980). In terms of vocational overshadowing, clinicians may find personal problems to be more salient due to the clinicians' positive affect and empathy towards personal problems as compared to vocational problems (Borresen, 1965). In Spengler, Blustein, and Strohmer (1990), counseling psychologists were found to differ in their preferences for working with personal problems in comparison to vocational problems. Findings revealed that clinicians with greater preferences for addressing personal problems were less likely to assess, diagnose, and treat vocational problems. These findings provided support for the connection between affect and clinical judgment, with clinician preferences reflecting, in part, clinicians' empathy toward and/or positive feelings

towards working with personal problems. The automatic and instinctive processing involved in the affective heuristic may lead to inaccurate judgments.

Biases based on faulty cognitive processes have long been the subjects of research studies involving non-clinicians (Hogarth & Karelaia, 2007). However, the examination of how certain types of heuristics influence judgment is especially pertinent to the mental health field, where the use of heuristics and the presence of bias can have a dramatic effect on the course and outcome of treatment (Garb, 1999). Mental health clinician biases can also occur as a result of other types of errors, such as those based on invalid or unreliable data, invalid or unreliable instruments, and failing to recognize regression towards the mean (Lichtenberg, 1997). Biases are especially pertinent to this study due to their impact on clinical judgment accuracy. Researchers have long been interested in how clinicians are affected by various biases and heuristics and how clinicians' judgment accuracy can be skewed as a result (Garb, 1999). Although biases and heuristics should not be viewed as inherently negative and do not necessarily lead to inaccurate clinical judgments, it is generally believed that clinicians are at risk of making inaccurate judgments if they are not diligent in developing critical thinking and self-monitoring skills. The automatic nature of most biases and heuristics may prevent clinicians from utilizing self-awareness in order to effectively gauge the accuracy of their judgments (Hogarth & Karelaia). The use of biases and heuristics is of special concern in regards to more experienced clinicians, who have had more time to cement and automatize their clinical judgment practices (Garb & Grove, 2005). In the next section, the clinical judgments of clinicians will be examined as a function of varying levels of experience.

Differences in clinical judgment in relation to experience

In terms of the effects of experience on clinical judgment, much research has found novice and experienced clinicians to perform similarly (e.g., Boland, 2002; Chandler, 1970; Faust et al., 1988; Garb & Boyle, 2003; Gaudette, 1992; Goldberg, 1959; Hickling, Blanchard, Mundy, & Galovski, 2002; Kim & Ahn, 2002; Walker & Lewine, 1990; Yeo et al., 2001). Goldberg compared staff psychologists, psychology trainees, and untrained secretaries regarding their ability to diagnose brain damage on the basis of the Bender-Gestalt Test. The groups were not found to differ in the accuracy of their judgments. Similarly, Yeo et al. examined generally-trained nurses and psychiatrically-trained nurses in their diagnostic decision-making of depression, mania, and schizophrenia case vignettes. Judgment accuracy was not shown to differ as a function of whether or not nurses had additional psychiatric training.

Other research has found that more experienced clinicians showed a significant improvement in judgment accuracy (e.g., Arkell, 1976; Berven, 1985; Brammer, 2002; Garcia, 1993; Rerick, 1999; Wilkin, 2001). For example, Arkell asked clinicians to classify the human figure drawings of functionally normal and maladjusted children. Clinicians with more experience were found to more accurate classify the figure human drawings than those with less experience. Wilkin, on the other hand, asked general practitioners, pediatricians, and psychologists to provide accurate diagnoses for case vignettes portraying clients with Attention-Deficit/Hyperactivity Disorder. Psychologists were shown to significantly outperform the other two professional groups, displaying a 79 percent hit rate, compared to 66 percent and 64 percent for pediatricians and general practitioners, respectively.

Yet other studies have shown judgment accuracy to worsen with more experience (e.g., Berman & Berman, 1984; Falvey, Bray, & Hebert, 2005; Ruscio & Stern, 2005; Schinka & Sines, 1974). For example, Falvey et al. examined varying levels of experience in relation to diagnostic accuracy as well as comparing their case conceptualization and treatment planning skills. Twenty practicing clinicians provided diagnostic judgments for a standardized case simulation of Attention-Deficit/Hyperactivity Disorder as well as follow-up explanations of their judgment processes. Not only were more experienced clinicians shown to have superior case conceptualization and treatment planning skills as judged by an expert panel, but more experienced clinicians provided more accurate diagnoses.

An examination of the individual clinical judgment studies makes apparent the lack of a definitive or clear relation between experience and judgment accuracy. Due to the wealth of clinical judgment accuracy research, it has become apparent to some scholars that a different approach is needed to gain a more accurate and complete understanding of the experience-accuracy effect (Falvey et al., 2005). Some have conducted traditional narrative reviews in order to determine whether or not there is a positive correlation between experience and clinical judgment. Many narrative reviews have concluded that experience does not facilitate better clinical judgments. For example, Wiggins (1973) found there was "little empirical evidence that justifies the granting of 'expert' status to the clinician on the basis of training, experience, or information processing ability" (p. 131). Similarly, Watts (1980) stated, "There are many studies...suggesting that the clinical judgment of psychologists is no better than that of, say physical scientists; and that psychologists with clinical training have no better

judgment than those without" (p. 95). Perhaps one of the most important reviews had been conducted by the American Psychological Association (1982), from which the conclusion was drawn that neither professional training nor experience was related to professional competence. Other reviews have repeated this conclusion (Faust, 1986; Faust and Ziskin, 1988; Garb, 1998; Highlen & Hill, 1984; Lichtenberg, 1997).

Explanations for clinical judgment findings

It is important to understand not only the relationship between clinical judgment and experience but also the reasons why many studies have shown that clinical judgment does not improve with increased experience. There are many reasons why this may be in addition to the cognitive biases discussed earlier. Dawes (1994) discussed some of these reasons. First, he asserted that clinicians often treat the process of gaining clinical experience as if it were identical to the process of gaining other types of experiences, such as learning motor skills. Many types of skills are learned somewhat automatically through repetitive practice. However, clinical judgment requires complex cognitive processes. Perhaps the most crucial difference is the lack of immediate feedback when clinicians make clinical decisions as opposed to the often naturalistic feedback that occurs with learning other skills. Without this feedback, clinicians run the risk of committing subsequent errors, a costly occurrence given the importance of accurate clinical judgment (Garb, 2005). Clinical skill learning can be categorized as experiential learning, in which the clinician learns by doing. In order for experiential learning to be successful, however, two conditions must be met: a) a clear understanding of what constitutes an incorrect response or error in judgment; and b) immediate, unambiguous, and consistent feedback when such errors are made. Unfortunately, these two conditions

are rarely met in the actual practice of psychology (Dawes; Garb, 2005; Lichtenberg, 1997; Zeldow, 2009).

Differences in performance of experienced and novice clinicians

Some research has supported the claim that novices and experienced clinicians perform differently. For example, experienced clinicians have been shown to differ from novice clinicians on a number of important cognitive dimensions, such as the complexity of knowledge structures, short- and long-term memory, efficiency in client conceptualization, number of concepts generated, flexibility in therapeutic response, and the quality of their cognitive schemata regarding case material (Brammer, 2002; Cummings, Hallberg, Martin, Slemon, & Hiebert, 1990; Eells et al., 2005; Kim & Ahn, 2002; Kivlighan & Quigley, 1991; Martin, Slemon, Hiebert, Hallberg, & Cummings, 1989; Mayfield, Kardash, & Kivlighan, 1999; Mumma & Mooney, 2007; O'Byrne & Goodyear, 1997; Tracey, Hays, Malone, & Herman, 1988). Additionally, experienced clinicians have been shown to be able to employ statistical heuristics more effectively when statistical heuristics are deemed as being important by the clinicians (Nisbett et al., 1983). Although it may be reassuring to some in the field that novice and experienced clinicians have been shown to differ on these cognitive dimensions, researchers have warned against reaching positive conclusions. For example, Martin et al. (1989) examined how novice and experienced clinicians differed on the extensiveness of their therapeutic conceptualizations and found, "Experienced clinicians conceptualized the specific problems of their individual clients in relation to their conceptual structures for counseling in general" (p. 399). In other words, more experienced clinicians tended to rely on their previously established schemata regarding the general counseling process

with minor adjustments according to the needs of each case. Novices, on the other hand, were found to require considerably more client-specific concepts in order to form their conceptualizations of the individual clients and their problems. The more experienced clinicians' greater automaticity in case conceptualization resulted in greater error in clinical judgment, consistent with experience-based adjustment models (Schwarz & Stahlberg, 2003).

The previous paragraphs outlined the variability in clinical judgment accuracy findings in relation to experience. Through an examination of individual studies as well as traditional narrative reviews in this research area, it is apparent that the relation between judgment accuracy and experience is not clear. This is especially alarming given the frequent demands on clinicians to make accurate judgments, for example when conducting psychological assessments, establishing appropriate course of treatment, and the reporting of diagnostic impressions to third parties for reimbursement. In addition, the variability in experience-accuracy findings is worrisome given the amount of time, effort, and money contributing to training of clinicians both in their early careers as well as throughout (Spengler et al., 2009).

Purpose and rationale for present study

Until recently, individual studies as well as traditional narrative reviews have comprised the research base regarding clinical judgment and experience. Individual studies have revealed mixed findings and traditional narrative reviews, for the most part, have found clinical judgment does not improve with experience. Recently, however, the large-scale Spengler et al. (2009) meta-analysis synthesized the clinical judgment and experience research from the years 1970 to 1996. Spengler et al. combined results from

75 clinical judgment studies including the judgments of 4,607 clinicians and found a small, reliable effect, demonstrating a positive correlation between experience and judgment accuracy. They noted a 13 percent increase (d = .12) in clinicians' decision-making accuracy with more experience, with few significant effects from moderator variables. Spengler et al. noted that even though the overall effect found was modest, it was not trivial. They asserted the overall effect was reliable since the confidence interval did not cross zero and few moderator variables significantly impacted the overall effect. When compared with meta-analysis findings analyzing the relation between experience and client outcome (Lambert & Ogles, 2004), the Spengler et al. experience-accuracy effect is meaningful. As Spengler et al. stated, "meta-analytic reviews of psychotherapy find little evidence for a relation between experience and client outcome" p. 26). In a reaction to Spengler et al., Ridley and Shaw-Ridley described the meta-analysis findings as "sobering and instructive" (p. 402, 2009). They addressed the Spengler et al. findings in light of the effort and funds contributed to the establishment of professional training and safeguards, specifically addressing the utilization of rigorous accreditation programs designed to promote clinician competency and effectiveness. According to Ridley and Ridley-Shaw, the marginal improvement in clinical judgment accuracy as a function of greater experience has serious implications for clients, namely the adherence to premature and/or inaccurate case formulations, the consequent selection of inappropriate or ineffective intervention strategies, and the inaccurate appraisal of client outcomes.

The Spengler et al. (2009) meta-analysis provided an in-depth, comprehensive analysis of the relation between experience and judgment accuracy, examining various moderator variables that may have impacted the overall effect. The purpose of the

present study is to conduct a meta-analysis in order to investigate whether or not clinical judgment improves with experience, updating and expanding upon the work of Spengler et al. The Spengler et al. meta-analysis addressed studies conducted between 1970 and 1996. An update and extension will allow for cross-validation of the Spengler et al. meta-analysis and an examination of whether or not findings vary as a function of using current research. Due to the presence of only one available meta-analysis examining the experience-judgment accuracy effect, cross-validation of this meta-analysis will be a beneficial contribution to the clinical judgment research. An update will also provide a test of robustness of the Spengler et al. findings using current research.

In the present study, like in the Spengler et al. (2009) meta-analysis, experience encompassed both clinical experience and educational training. Clinical experience was defined as the time clinicians spend directly providing services to clients. Educational training, in contrast, represented the level of graduate training clinicians have reached as well as training in specific skill areas or the receipt of supervision. Clinical judgments in the present study encompassed a variety of decisions commonly made by clinicians in professional practice. For example, clinicians are often called upon to make determinations of diagnosis, prognosis, behavior prediction, and intervention choice. The different methods of assessing judgment accuracy reflected varying degrees of criterion validity. Some methods of assessing judgment accuracy included a priori validation of stimulus materials, comparison of judgments with observable client behaviors, expert consensus, and comparisons of judgments with psychological test scores. Brammer (2002), for example, provided clinicians with a computer-based case vignette and subsequently asked the clinicians to provide a DSM-IV diagnosis. Judgment accuracy

was then determined by a panel of "expert" judges, consisting of four psychologists with an average of 18 years of licensed, clinical experience. The judges were provided with a list of 30 possible diagnoses and were asked to rate the clinicians' diagnoses on a 4-point Likert-type scale ranging from *unlikely* to *definite*. Consistent with the Spengler et al. (2009) meta-analysis, studies in which judgment accuracy was determined via various types of professional consensus were considered to have low criterion validity due to the relatively subjective nature of the method for determining judgment accuracy.

It is interesting to note that although included studies addressed whether or not experience was correlated with greater judgment accuracy, some of the studies employed "experienced" or "expert" clinicians as assessors of judgment accuracy. The use of clinicians with more experience as judges for determining other clinicians' judgment accuracy casts doubt on the ability of these more experienced judges to make accurate assessments, especially considering the varied research findings regarding the experience-accuracy effect. Kirk and Hsieh (2004) alluded to the problem of using more experienced judges' assessments as validation for less experienced clinicians. Kirk and Hsieh found that when clinicians of varying levels of experience were asked to provide a diagnosis for a case vignette, there was a 50-50 split between the selection of two different diagnoses. They noted, "This is troubling because the judgments of experienced clinicians, such as those in this study, are used in many reliability studies as the 'validating diagnosis' or the 'gold standard' against which other diagnostic methods are compared for accuracy" (p. 8). Therefore, using professional consensus as a method of validating clinicians' judgment accuracy presents a major problem in clinical judgment research.

More valid ways of determining judgment accuracy than using professional consensus have been found in the clinical judgment research. For example, higher criterion validity can be found when the accuracy of clinicians' judgments is assessed based on such measures as those used in Ogloff and Daffern (2006). In this study, psychiatric nurses' predictions of patients' violence were compared to the patients' actual violence as recorded on the Overt Aggression Scale (OAS) by a second set of nurses. Steps were taken to enhance the accurate recording of ratings on the OAS, such as providing training and support for the second set of nurses and requiring them to rate incidents of violence either right after it occurred or at the end of their shifts. This type of standard by which to measure judgment accuracy reflects a higher level of criterion validity than does professional consensus; however, error can still occur. For example, error in the Ogloff and Daffern study may have occurred in the form of variability in nurses' observations and reporting of violence. The studies included in the present meta-analysis employed a variety of assessment methods for determining clinicians' judgment accuracy. Criterion validities (high versus low) were determined for each assessment method.

Clinical judgment studies in which judgment accuracy cannot be determined will be excluded. This includes many studies of clinician bias. Although bias often plays a role in clinicians' judgment accuracy, it will be impossible to determine accuracy in some cases. In other words, clinicians may display bias towards members of a given race, age group, or gender but may or may not be able to formulate accurate judgments with those groups. It is often difficult to determine judgment accuracy in these cases due to a lack of a standard by which to evaluate the judges' decisions. For example, Mohr, Weiner,

Chopp, and Wong (2009) presented mental health clinicians of varying experience levels with identical case vignettes involving clients of either heterosexual, bisexual, or homosexual orientation. The purpose of the study was to determine whether clinician bias exists in relation to client differences in sexual orientation. Moreover, the authors were interested in determining the conditions under which the most bias occurred, specifically examining whether or not there was an increase in clinician bias when the client vignettes included stereotypes about bisexuality or homosexuality. The clinicians were asked to rate the clients' GAF scores based on the information provided in the vignettes. Results revealed bias in the way clinicians viewed clients based on their sexuality, especially when the vignettes included stereotypical information about sexuality. Although the results revealed bias as measured by comparisons of clinicians' ratings across sexual orientation conditions, the authors noted they could not determine judgment accuracy because there was no clear-cut standard by which to judge the clinicians' ratings. As explained in the study, "One interpretation of the results is that the effect of client bisexuality on clinical judgment may partly reflect therapists' accurate beliefs about differences among bisexual, gay, and heterosexual men" (p. 173). Therefore, in studies pertaining to bias in clinical judgment, it is oftentimes impossible to determine a standard for accuracy.

Moderator variables were chosen on the basis of prior research findings related to clinical judgment and experience. Other moderator variables were chosen because they are commonly addressed in meta-analyses (e.g., study quality). The present study examined the moderating effects of the following variables, the majority based on the Spengler et al. (2009) meta-analysis and some added for the present meta-analysis: (a)

experience type, (b) experience breadth, (c) judgment type, (d) criterion validity for accuracy dependent measure, (e) provision of feedback, (f) publication source, (g) ecological validity of the method of study, (h), ecological validity of stimulus, (i) relation of experience to the research design, (j) experience as a major variable, (k) study quality, (l) profession type, (m) inclusion of non-mental health participants, and (n) publication year. Whereas most moderator variables were drawn from the Spengler et al. meta-analysis to allow for comparison, profession type and inclusion of non-mental health participants were added for the present meta-analysis. It is hypothesized that moderator variables will reveal few significant findings, as was found in the Spengler et al. meta-analysis. A more detailed discussion of the present study's hypotheses will be outlined in Chapter 2.

Implications for findings

The findings of the present meta-analysis will have important implications for the field of psychology. It has been long assumed that clinical judgment improves with experience. Clients and clinicians as well as the general public assume that more experienced clinicians yield more accurate judgments. If this assumption were shown to be untrue, the perceived benefits of receiving mental health services from more experienced clinicians would be called into question. If clients perceive that even a small increase in judgment accuracy will have important personal or professional implications, however, they may nonetheless choose relatively more experienced clinicians (Spengler et al., 2009). Training programs would be forced to make integral changes in order to improve their curricula. Even worse, clients and the general public may experience less faith in the counseling process overall. Also, psychotherapy is likely to be ineffective if

clinicians' decisions are inaccurate (Wolfgang et al., 2006). Additionally, client drop-out becomes more of a problem with inaccurate judgments (Epperson, Bushway, & Warman, 1983). The Spengler et al. meta-analysis revealed judgment accuracy slightly improved with increased clinical experience. However, an update and expansion is needed to cross-validate and test the robustness of the Spengler et al. findings with more recent research as well as test additional moderator variables. For these reasons, the present meta-analysis seeks to investigate the relationship between clinical judgment and experience of studies published from 1997 to 2010.

Chapter 2: Literature Review

The purpose of this chapter is to provide a comprehensive review of the literature regarding the relationship between clinical judgment and experience. Additionally, this chapter will provide a rationale for updating the Spengler et al. (2009) meta-analysis. The purpose of the present study is to conduct a meta-analysis of studies between the years 1997 and 2010 that address whether or not clinical judgment accuracy improves with experience. The Spengler et al. meta-analysis synthesized studies published between 1970 and 1996 and found a small, but reliable effect of $d = .12$. There was little variability in the obtained results, suggesting that experience marginally improves judgment accuracy regardless of most other variables. The present study is needed to capture the potential impact of using current research as well as assess new moderator variables in addition to those drawn from the Spengler et al. meta-analysis.

Relationship between experience and expertise

Contrary to what is sometimes assumed experience and expertise seem to be related, but separate concepts (Eells et al., 2005; Glaser & Chi, 1988; Sedlmeier, 2005). Hayes (1985), for example, distinguished experience from expertise and quantified the amount of experience needed to gain expertise. He found that experts in various fields acquired an expert level of performance after approximately 10 years of practice. Similarly, Eells et al., in a study of the differences in the quality of psychotherapy case formulations, distinguished experienced clinicians from expert clinicians. In their study, 65 clinicians were divided into three separate groups. The first group consisted of 24 novice clinicians with fewer than 1,500 hours of supervised psychotherapy experience. The second group consisted of 19 experienced clinicians with at least 10 years of

experience. The third group, in contrast, consisted of 22 expert clinicians. The expert group was identified as "meeting one or more of three criteria: (a) developed a method of psychotherapy case formulation; (b) led one or more workshops for professionals on how to construct case formulations; or (c) published one or more scientific articles, books, or book chapters on the topic of psychotherapy case formulation" (Eells et al., p. 581). After listening to six standardized vignettes via audiotape, the clinicians were asked to provide verbal case formulations. Case formulations were evaluated on multiple dimensions, such as comprehensiveness, complexity, and formulation elaboration.

Eells et al. (2005) found that expert clinicians performed better than novice and experienced clinicians when considering many dimensions of case formulation quality, for example demonstrating superior comprehensiveness, elaboration, and complexity of case formulations as compared to those of novices or experienced clinicians. In addition, expert clinicians were found to provide more elaborate treatment plans that were better fitted to the clients' needs as compared to the other groups of clinicians. Moreover, the expert clinicians' formulations across the six vignettes displayed more consistency and structure compared to those of the other groups. This finding may allude to the hypothesis that experts' judgments have become somewhat systematic or automatized. Eells et al. attributed the superiority of expert case formulations to the hypothesized tendency of expert clinicians to produce responses based on a priori cognitive schema, whereas novice and experienced clinicians were believed to have less structured or well-defined schema. Although Eells et al. found that expert clinicians provided the highest quality case formulations, they found that novice clinicians outperformed experienced clinicians when comparing these two groups with each other. Eells et al. hypothesized

that both expert and novice clinicians diligently and consistently worked to ensure they kept to a high standard of case formulation, whereas experienced clinicians often worked in the field for years without calibrating their work with professional standards. They concluded that acquiring a certain number of years of experience did not always guarantee an expert level of performance, as evidenced by novices outperforming experienced clinicians. Instead, Eells et al. explained that novice clinicians often have opportunities to work closely with expert clinicians through their studies and practica. Experts, as mentioned previously, were believed by Eells et al. to calibrate their performance via cognitive tools, such as highly structured and well developed schema. Eells et al. believed both novices and experts tended to engage more frequently in self-monitoring, whereas experienced clinicians were believed to work in a more isolated setting less conducive to self-monitoring and calibration with expert performance.

For the present study, it is important to understand the distinction between experience and expertise. If expertise is not automatically gained with a certain number of years of experience, one can easily see how some clinicians never reach an expert level of performance. Frensch and Sternberg (1989) defined expertise as "the ability, acquired by practice and experience, to perform qualitatively well in a particular task domain" (p. 189). The definition provided above, however, assumes practice and experience is the cause of gaining expertise. Perhaps this was because it has commonly been assumed that expertise should be gained through increased practice and experience rather than from some innate talent or proclivity. Additionally, the definition addressed the construct of quality, rather than judgment accuracy, decision-making speed or some other measure, as being the central indicator of expertise status. Moreover, the definition assumed that

expertise would be gained in a particular domain, rather than assuming one could gain global expertise status across a wide variety of tasks. In other words, clinicians may be experts in some aspects of clinical practice, such as providing quality case formulations, but not perform at expert levels in other aspects. Eells et al. recognized this distinction when they stated, "Thus, one would not necessarily expect the experts in the present study to excel in other aspects of the practice of their profession, including perhaps applying the formulation in therapeutic interventions" (p. 587). According to Frensch and Sternberg, expertise seemed to be domain-dependent.

While many reviewers in the mental health field agree with the Frensch and Sternberg (1989) definition of expertise (Lichtenberg, 1997; Sedlmeier, 2005), others have described expertise as if it were some mystical, innate, and domain-independent construct that a clinician either has or does not have (Zeldow, 2009). In this sense, novice clinicians can be sure to feel some degree of hopelessness as they struggle to discover whether or not they were blessed with this talent. Additionally, they may never know whether or not they have gained expertise status even through years of practice and experience (Lichtenberg). According to Witteman and van den Bercken (2007), when experts are questioned about a decision they have made, they will often deny knowing how they arrived at a particular decision. This unawareness or underutilization of meta-cognitive and self-reflective processes may prove to be dangerous for clinicians and clients alike. Not only can faulty cognitive processes occur beyond the awareness of the clinician, but this lack of awareness has the potential to present problems in the training of future clinicians. One can imagine how "expert" clinicians, with limited awareness of

how they arrived at particular decisions, would find it difficult to explicitly teach these skills to trainees in a formal, unambiguous manner.

The second perspective on expertise, that of expertise as an elusive, innate quality, is by far a grimmer one for inexperienced and experienced clinicians alike. In other words, both inexperienced and experienced clinicians would find it difficult to achieve expert status and know when they have reached it due to the emphasis on an abstract, ill-constructed definition of expertise. They then may be forced to rely on alternative methods of identification, such as peer nomination (Elman et al., 2005; Kahneman & Klein, 2009; Sonnentag, 1998). Peer nomination allows for expert identification by relying on the evaluations of the so-called expert's peers. According to Kahneman and Klein, peer nomination techniques have the potential to distinguish between expert, experienced, and novice performers. For example, Sonnentag conducted a study in which software designers were observed by their peers in a software design task. The peers were required to rate the performance of the software designers and identify "high performers" (p. 703). The selection of these high performers was further evaluated with regards to the software designers' objective performance on the software design task. In this case, peer nomination was shown to be a valid method of expert identification. According to Elman et al., peer nomination offers one method of expert identification; however, peer nomination, in addition to the attainment of licensure and certification, the receipt of professional awards, and the publication of journal articles, does not encompass the full spectrum of what an expert clinician represents. Peer nomination techniques may prove to be less helpful in the identification of expert mental health clinicians due to the variety of ways in which expert clinicians approach tasks

consistent with their theoretical orientations. For example, it may be more difficult for a mental health clinician who subscribes to behaviorist theories and techniques to judge the expertness of another mental health clinician who employs psychodynamic approaches.

Although there is some degree of variability in how experts are identified and/or defined, research has attempted to extrapolate the differences in how experts perform compared to novice clinicians. Experts have been found to perform qualitatively different from novices (Bereiter & Scardamalia, 1986; Chan, 2006; Chi et al., 1982; Ericsson, 2005; Glaser & Chi, 1988; Haerem & Rau, 2007; Hinsley et al., 1978; Leon & Perez, 2001; Lesgold et al., 1988; Nickerson et al., 1985; O'Reilly et al., 1980; Polanyi, 1962). Experts, for example, have demonstrated the ability to make broader inferences due to their knowledge being organized into broad and complex memory structures (Ericsson). Additionally, they have been shown to be able to make connections between seemingly irreconcilable concepts (Hinsley et al.). Expert judgments have been shown to be more sophisticated and rely on more advanced critical reasoning skills (Chi et al.; Polanyi). Some other differences include having the ability to perceive large, meaningful patterns, displaying superior long- and short-term memory, and displaying more advanced self-monitoring skills (Chan; Chi et al.; Ericsson; Glaser & Chi; Haerem & Rau; Hinsley et al.; Leon & Perez; Lesgold et al.; Nickerson et al.; O'Reilly et al.; Polanyi; Sedlmeier, 2005). Perhaps one of the most easily recognizable differences in expert performance is the greater automatic processing of information (Glaser & Chi).

One of the most important differences in how experts in any field approach tasks may be the difference in their initial perceptions of a task (Chan, 2006; Day & Lord, 1992; Ericsson & Charness, 1994; O'Reilly et al., 1980). Not only do experts differ in

the depth and breadth of their knowledge bases, but they also differ in how they comprehend aspects of a given problem or situation. This difference in perception may be one of the causes for why it is difficult to improve the performances of novices by a simple transfer of knowledge. Experts may perceive problems as less complex than non-experts due to being more familiar with the basic principles underlying those tasks. As a result, experts view problems as more manageable than do non-experts. Specifically, experts tend to view problems according to their fundamental principles or deep underlying structures. Novices, on the other hand, tend to view problems according to their superficial characteristics or surface structures (Chi, Feltovich, & Glaser, 1981; Ericsson & Charness, 1994; Haerem & Rau, 2007). Surface structures, in this case, refer to objects, keywords, or physical configurations involving the interaction of several objects. Deep structures, on the other hand, refer to the underlying principles presented in a given problem. A study by Chi et al. demonstrated this difference. Subjects in this study were asked to assess problems in physics. The surface structures mentioned in the physics problems included physical descriptors of the objects, such as "spring" or "inclined plane" and/or keywords contained in the problem, such as "rotation" or "velocity problems." The deep structures involved in the physics problems were related to underlying principles of physics, such as Newton's laws. Experts were shown to utilize abstracts principles of physics when analyzing a problem, whereas novices focused on the problem's surface structures, including physical descriptors of the objects and keywords contained in the problems. As Haerem and Rau noted, expert-novice differences in task perception also involve task complexity. For example, experts were found to view tasks as less variable and more analyzable than non-experts due to experts'

tendencies to focus on the underlying principles of the task. Haerem and Rau's findings supported differences in how experts and non-experts perceive tasks; therefore, they emphasized the importance of future research in expert-novice differences to examine not only how these different groups perform but also how their initial perceptions of the tasks differed. Haerem and Rau concluded that although their findings supported differences in how experts and non-experts perceive tasks, these differences in perceptions do not always result in the differences in performance we would expect. Several other factors may be involved in measuring expert-novice performance differences, for example, task complexity and motivation.

In bringing the discussion of expertise and characteristics of experts more specifically to the field of psychology, the frameworks of Hollon and Kriss (1984) as well as Wierzbicki (1993) are useful. Expertise in counseling, as it relates to clinical judgment, can be analyzed in terms of three interrelated constructs: cognitive structures, processes, and products. Cognitive structures have been defined as organizational bodies that contain clinicians' knowledge, beliefs, and assumptions about themselves, their clients, and their world (Hollon & Kriss). Cognitive structures are believed to be accessed when the clinician attempts to label and explain incoming information as well as search for additional information (Wierzbicki). Cognitive processes, on the other hand, pertain to the methods used when clinicians combine incoming information with existing knowledge structures in order to form judgments. Finally, cognitive products have been defined as the results of information processing (Hollon & Kriss). Errors may occur within any of the three aspects of decision-making. Often, errors in one lead to errors in another due to the interrelated nature of these cognitive domains. For example, clinicians

who utilize faulty cognitive structures will most likely produce inaccurate judgments. Although one would expect and hope that functioning in all three of these cognitive domains would improve with increased clinical practice and training, a growing body of research suggests otherwise.

Influence of bias on clinical judgment

When discussing novice-expert differences and attempting to understand how novices can achieve expert performance levels, a review of judgment bias is essential. According to Gambrill (2005), "Bias is a systematic 'leaning to one side' that distorts the accuracy of results" (p. 328). Although distinct from judgment error, bias has long been considered an important obstacle preventing many clinicians from arriving at accurate judgments (Garb, 1998). In terms of Hollon and Kris's framework, judgment biases based on client and clinician variables relate to functioning in the cognitive structures domain, where clinicians hold their attitudes, beliefs, and knowledge, and assumptions about themselves and others (Wierzbicki, 1993). Bias comes in many forms, ranging from bias related to client characteristics (e.g., socioeconomic status, age, sex-role, race, and gender; Abramowitz & Dokecki, 1977; Garb, 1998) to bias related to various characteristics of the clinician (Garb, 1998; Raines & Rohrer, 1955). Garb's influential book, *Studying the Clinician: Judgment Research and Psychological Assessment*, was published in 1998. In it Garb provided a comprehensive review of studies pertaining to clinical judgmental biases in the mental health field.

An overview of biases based on race, social class, gender, age, and sex role revealed mixed empirical results. For example, Garb (1998) cited studies in which African-American psychiatric inpatients were predicted to be more violent than their

Caucasian counterparts (Lewis et al., 1990; McNiel & Binder, 1995). On the other hand, this type of race bias did not seem to be present in the clinical prediction of violence taking place in the community (Lewis et al.; Lidz et al., 1993). Similarly, Garb (1998) found mixed results for studies addressing gender bias. Some studies were cited in which gender bias was demonstrated in clinicians' ratings of prognosis (Agell & Rothblum, 1991; Fernbach et al., 1989; Hansen & Reekie, 1990; Teri, 1982). In most studies of prognosis, however, gender bias was not shown to be present (Adams & Betz, 1993; Bernstein & LeComte, 1982; Billingsley, 1977; Dailey, 1980; Elovitz & Salvia, 1982; Fischer, Dulaney, Fazio, Hudak, & Zivotofsky, 1976; Foon, 1989; Hardy & Johnson, 1992; Lewis et al.; Lopez, Smith, Wolkenstein, & Charlin, 1993; Rabinowitz & Lukoff, 1995; Schwartz & Abramowitz, 1975; Settin, 1982; Stearns, Penner, & Kimmel, 1980; Wrobel, 1993; Zygmond & Denton, 1988).

For the purposes of this study, it is important to note that assessing clinician bias is a multifaceted issue. In Garb's (1998) comprehensive review, biases of many types were found to be present in many studies, but only under certain conditions. Although it is sometimes difficult to capture particular client variable biases at work in the mental health setting, scholars believe these biases do pertain to clinicians of varying experience levels and that they have the potential to negatively impact judgment accuracy. In a discussion of the rates of Borderline Personality Disorder (BPD) diagnosis among lesbian, gay, and bisexual clients, for example, Eubanks-Carter and Goldfried (2006) stated, "The association of male homosexuality and BPD may also be related to the strong association between BPD and female gender" (p. 753). Eubanks-Carter and Goldfried were interested in examining the relationship between clients' gender and

sexual orientation and clinician-provided diagnoses of BPD. They explained that although homosexuality was no longer an official diagnosis in the *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition, Text Revision (DSM-IV-TR; American Psychological Association, 2000) and should not be viewed as pathological, clinicians still displayed tendencies to overpathologize gay and bisexual men. The risk of clinicians misdiagnosing gay and bisexual men with BPD is especially high when considering certain borderline-like traits that are common among gay and bisexual men experiencing a crisis in sexual identity, including identity disturbance, affective instability, and self-mutilation (Gonsiorek, 1982).

The discussion regarding BPD diagnoses and sexual orientation highlights the complexity of the relationship between bias and judgment error. Clinicians with negative biases towards gay and bisexual clients are at risk of overpathologizing the clients' reported problems, which oftentimes leads to clinicians assigning diagnoses that are inappropriate or that are overly severe. In clinical judgment research, judgment error is oftentimes relatively easy to measure and detect due to its necessary reliance on some agreed upon standard by which to rate individual judgments. Judgment bias, however, often goes unnoticed and is more difficult to detect due to researchers' difficulties establishing a baseline for comparison.

In addition to biases based on client variables, clinicians display bias based on faulty cognitive processes. These types of biases interfere with the way clinicians integrate novel information with preexisting knowledge (Wierzbicki, 1993). Some of these biases include confirmatory bias, hindsight bias, anchoring and adjustment heuristics, representativeness heuristics, availability heuristics, and the newly defined

affective heuristic (Garb, 2005; Garb, 1998; Tversky & Kahneman, 1974). Confirmatory bias exists when clinicians display a tendency to seek out information that confirms or supports their prior beliefs and assumptions about their clients. Disconfirmatory evidence is de-emphasized or, in the worst cases, ignored (Davies, 2003). Many studies have revealed confirmatory bias at work for clinicians and laypeople (Davies; Einhorn & Hogarth, 1978; Elstein, Shulman, & Sprafka, 1978; Mahoney, 1976; O'Brien, 2009; Snyder, 1981; Snyder & Swann, 1978; Snyder, Tanke, & Berscheid, 1977; Strohmer, Moilanen, & Barry, 1988; Strohmer, Shivy, & Chiodo, 1990).

Hindsight bias has been defined as the tendency to believe, once the outcome is revealed, that the outcome could have been predicted more easily than actually could in reality. Hindsight bias occurs when individuals try to make sense of certain outcomes by elaborating causal relations between antecedent conditions and the outcome (Blank & Nestler, 2007). Hindsight bias has been demonstrated in a variety of domains, such as historical events (Fischhoff, 1975), medical diagnoses (Arkes, Faust, Guilmette, & Hart, 1988), sporting events (Pezzo, 2003; Roese & Maniar, 1997), and political elections (Blank, Fischer, & Erdfelder, 2003). Additionally, hindsight bias has been demonstrated in the area of clinical judgment (Arkes, Wortmann, Saville, & Harkness, 1981; Fischhoff), and has been associated with inaccurate judgments.

Judgmental heuristics are cognitive shortcuts commonly used in decision-making by people in general and including mental health clinicians (Kahneman, Slovic, & Tversky, 1982; Nisbett & Ross, 1980). The anchoring heuristic is described as a phenomenon in which individuals are presented with information in a sequential manner and emphasize information received first. These individuals tend to make initial

judgments based on the first information presented and then make insufficient adjustments away from that initial judgment as additional information is presented. Even though information presented later in the sequence may be disconfirming, or have greater importance to the judgment task at hand, it is ignored or given inadequate attention (Carlson, 1990). Research has shown the anchoring heuristic to cause an effect on clinical decision-making (Clavelle & Turner, 1980; Dailey, 1952; Friedlander & Phillips, 1984; Friedlander & Stockman, 1983; Gauron & Dickinson, 1966; Meehl, 1960; Oskamp, 1965; Richards & Wierzbicki, 1990; Sandifer et al., 1970). For example, Friedlander and Phillips asked clinicians to rate two clients after reading summaries of five therapy sessions for each client. Client summaries included information about the presence of suicidal ideation or anorexia at either the first or fourth therapy session. Those clinicians who were presented with this information first often rated the clients as more maladaptive than those who received the information at the fourth therapy session. The variability in the timing of information during sessions and/or intake is one way in which clinicians commonly risk engaging in bias, which subsequently may lead to judgment error.

The representativeness heuristic can be described as the tendency for individuals to judge the likelihood that a person has some characteristic based on the degree to which the person is similar to the class of persons with that characteristic without taking base rates into consideration (Kahneman et al., 1982; Nisbett & Ross, 1980; Wierzbicki, 1993). When base rates are ignored, clinicians risk failing to recognize that some clinical predictions are unlikely because of the infrequency of the diagnosis or event (Garb, 1996). Garb (1998) noted three types of biases based on the representativeness heuristic.

The first involves the clinician comparing his or her client to others who clearly possess the trait in question. The clients who clearly possess the trait are referred to as *exemplars*. The second type involves the clinician exhibiting bias based on his or her concept of the "typical" person possessing the trait in question. In this case the bias is based on *stereotype*. The third type refers to the clinician exhibiting bias based on comparison to a theoretical standard that represents the specific trait. This type involves clinician bias based on *prototype*. Garb (1998) stated prototypes were more often the focus of theory and research than were exemplars and stereotypes. In the area of clinical judgment, clinicians relying on the representativeness heuristic are at risk for error in various ways. For example, a clinician evaluating a client for Bipolar Disorder may compare that client to one in the past who is believed by the clinician to be the exemplar of Bipolar Disorder. The clinician's accuracy in his or her evaluation of the present client depends largely upon whether or not the client in the past represented an accurate depiction of an individual with Bipolar Disorder. The representativeness heuristic can also cause judgment error in behavioral prediction. Poole, Lindsay, Memon, & Bull (1995) asked clinicians to list client indicators that suggest previous childhood sexual abuse. Clinicians most frequently reported "adult sexual dysfunction" (p. 430) as an indicator. However, the research field is lacking in empirical support for a connection between childhood sexual abuse and adult sexual dysfunction. Poole et al. noted that the relation between childhood sexual abuse and adult sexual dysfunction is at risk of being overestimated by clinicians, which may lead them to produce inaccurate conceptualizations of their clients' histories. In addition, one can imagine that if a clinician evaluating a client with diagnosed sexual dysfunction suspects childhood sexual

abuse, medical causes for the dysfunction may be overlooked and/or underemphasized. Some studies have demonstrated these types of errors in progress in the mental health field (Dawes, 1994; Kahneman & Tversky, 1973; Meehl, 1973; Meehl & Rosen, 1955; Tversky & Kahneman, 1974).

Other types of errors and bias beyond the scope of the present review include errors based on invalid or unreliable data and/or instruments and errors based on failing to recognize regression toward the mean (Kahneman & Tversky, 1973). There are a multitude of areas and ways in which error can occur to influence clinicians' judgment, including biases based on client characteristics, clinician characteristics, faulty cognitive processes, invalid or unreliable data and/or instruments, and a lack of understanding of regression toward the mean. When the experience level of the clinician is assessed in terms of its effect on judgment accuracy, one can easily understand how clinicians unknowingly develop tendencies to make these errors over time, and how it is possible that the more experience clinicians have, the more time they have had to practice and make these types of errors automatic.

Effects of experience on clinical judgment

It is because of these pervasive biases and errors outlined above that some scholars have concluded experienced clinicians should not outperform novices (Dawes, 1994; Faust, 1984, 2006; Garb, 1998, 2005; Lichtenberg, 1997; Ruscio, 2006; Sternberg, Roediger, & Halpern, 2007; Wiggins, 1973). Some research suggests that novice clinicians perform as well as, or even better than, more experienced clinicians (e.g., Blashfield, Sprock, Pinkston, & Hodgin, 1985; Boland, 2002; Garb & Boyle, 2003; Gaudette, 1992; Hickling et al., 2002; Kim & Ahn, 2002; Leon & Perez, 2001;

Rodriguez, 2002; Ruscio & Stern, 2005; Zozula, 2001). According to Garb (1998), "Overall, results on presumed expertise, experience, training, and validity are disappointing" (p. 17). Garb (1998) concluded in his review that experts are no more accurate than less experienced clinicians and clinicians were not found to be more accurate than graduate students. In his estimation, the only uplifting results were that clinicians may be more accurate than beginning graduate students and advanced graduate students were found to be more accurate than beginning graduate students. Furthermore, clinicians were found to be more accurate than lay judges. In a recent study, Witteman and van den Bercken stated that the gains in judgment accuracy associated with more experience that have been found in medical research have not been realized in the mental health field.

By contrast, other research has demonstrated a positive correlation between judgment accuracy and experience (e.g., Arkell, 1976; Berven, 1985; Brammer, 2002; Garcia, 1993; Rerick, 1999; Wilkin, 2001). Brammer, for example, asked psychology graduate students and psychologists to provide diagnoses after participating in a computerized case simulation in which a clinical interview was reproduced. Based upon questions asked by the clinicians, the simulated client provided appropriate, pregenerated responses. According to Brammer, the ability of the clinicians to arrive at a correct diagnosis was associated with the clinicians' level of training and years of experience. In other words, clinicians with higher levels of training and greater years of experience were relatively more accurate in their diagnostic decisions.

Other research has attempted to further describe and explain differences in clinical judgment with novices compared to more experienced clinicians (Brammer,

2002; Cummings et al., 1990; Eells et al., 2005; Kim & Ahn, 2002; Kivlighan & Quigley, 1991; Martin et al., 1989; Mayfield et al., 1999; Mumma & Mooney, 2007; Nisbett et al., 1983; O'Byrne & Goodyear, 1997; Tracey et al., 1988). As stated in Chapter 1, however, these differences do not always result in experienced clinicians outperforming novices on measures of judgment accuracy. Martin et al., for example, found that novice and experienced clinicians differed on the extensiveness of their therapeutic conceptualizations. As an added inquiry, Martin et al. addressed whether or not differences existed in the conceptualizations used by novice versus experienced counselors regarding a) the counseling process in general, and b) the specific problems of individual clients. They found that the more experienced the clinician, the more he or she conceptualized the client problems in relation to deep, underlying structures regarding the counseling process in general. Less experienced clinicians, in contrast, tended to focus on surface elements, such as the specific problems of the individual clients. This is consistent with previous findings of the differences in how novices and more experienced clinicians conceptualize client problems (Chi et al., 1981; Ericsson & Charness, 1994; Haerem & Rau, 2007). Martin et al. noted, "Experienced clinicians conceptualized the specific problems of their individual clients in relation to their conceptual structures for counseling in general" (p. 399). When addressing specific client problems, more experienced clinicians tended to rely on their previously established schemata regarding the general counseling process, with minor adjustments. The novice counselors, in contrast, were shown to require considerably more client-specific concepts in order to form their conceptualizations of the individual clients and their problems. While the schemata employed by experienced clinicians was likely more time- and energy-efficient

and led to a greater feeling of confidence by the seasoned counselors, one could also envision how having more automatic and inflexible schemata could lead to a "blinder" (Martin et al.) effect. Martin et al. concluded that while greater automization of decision-making is generally perceived as a beneficial acquisition in such fields as mathematics, physics, computer programming, and medicine, the subjective nature of clinical decision-making in the field of psychology means greater automization has the potential to hinder accuracy.

Traditional narrative reviewers investigating the relationship between clinical experience and judgment accuracy have typically claimed that gaining clinical experience does not necessarily lead to better judgment accuracy (Faust, 1986; Faust & Ziskin, 1988; Garb, 1998; Highlen & Hill, 1984; Lichtenberg, 1997; Watts, 1980; Wiggins, 1973; Ziskin, 1981). A related report of great importance was conducted by the American Psychological Association in 1982, which clearly stated no evidence had been found of a positive relationship between professional competence and years of professional experience. It was further suggested that it was "important, perhaps, imperative, that psychology begin to assemble a body of persuasive evidence bearing on the value of specific educational and training experience" (p. 2). Although traditional narrative reviews are not without their limitations, they have served to add to the experience-judgment accuracy debate by highlighting the lack of a definitive and clear-cut positive relationship between clinician experience and judgment accuracy. In order to better understand the complexities of the debate, it is important to address the possible reasons why judgment accuracy may not automatically improve with more clinical experience.

Explanations for clinical judgment research findings

There are many reasons why clinical judgment has not been shown to improve with experience in much research. First, clinical judgment involves complex, cognitive processes that are difficult to replicate and teach. Second, clinicians are often forced to make decisions in the absence of immediate feedback, increasing risk for error. Third, clinical skill learning can be categorized as experiential learning, in which the clinician learns by doing. The two conditions for experiential learning, which involve a clear understanding of what constitutes an incorrect response and immediate, unambiguous, and consistent feedback, are rarely met in the practice of psychology (Dawes, 1994).

In addition to the lack of clear-cut guidelines as to what constitutes a correct clinical response and the absence of useful feedback, Dawes (1994) highlighted other reasons clinical judgment does not necessarily improve with experience. One of these reasons pertains less to clinician characteristics and more to the nature of the clinical problem itself. He stated that clinical judgment problems were often "ill-structured" (Dawes, p. 232). Ill-structured problems lack clear parameters or constraints. For example, clinicians are often faced with decision-making regarding appropriate interventions for specific clients. Given the plethora of various interventions and the ever-changing views on what constitutes appropriate and/or ethical interventions, clinicians are forced to make treatment decisions based on ill-defined constraints. Furthermore, intervention choice is often a function of non-clinical constraints, such as limits of reimbursement by insurance carriers. Yet another issue of debate is related to establishing proper end-points and solutions for clients. For example, success in counseling can be measured by various constructs and taken from the perspective of

various individuals. If a client were to attend career counseling, one clinician might view the client applying for various jobs as a successful outcome, while another might only consider the client obtaining stable employment as a success. There are so many ways to measure effective solutions that it is often difficult to judge whether or not clinicians have demonstrated the ability to treat clients effectively (Eells, et al., 2005).

Although the nature of clinical problems as ill-structured adds even more difficulty to the clinical judgment debate, most clinicians maintain that experts should be better able than novices to transform these problems into solvable problems, even if that means imposing artificial parameters for the sake of manageability. Due to the subjective and socially constructed nature of clinical problems, however, one can easily reach the conclusion that expertness is also a socially constructed concept (Kahneman & Klein, 2009). Especially in the field of psychology, where most, if not all clinical problems can be described as ill-defined, "expert" may be defined by the clinician who is able to reach a consensus with his or her client, colleagues, and community as to whether or not an outcome was successful. According to Zeldow (2009), "Clinical practice is not a science that aspires to truth and the development of replicable and standardized interventions. Rather, it is an interpretive or narrative activity whose objective is the reduction of suffering in particular individuals" (p. 3). Additionally, it has been said that clinical judgment often relies on practical reasoning, or the act of making the most appropriate and well-informed treatment decisions under uncertain circumstances (Montgomery, 2006). For the most part, there are no clear-cut guidelines upon which clinicians can judge their outcomes as there are in other fields. Much of clinicians' responsibility is to create persuasive arguments in support of their treatment decisions. In fact, effective

case-making and argumentation may even be viewed as more important than accurate clinical judgment when it comes to gauging the quality of clinicians' judgments.

Given the long-standing debate and variable findings regarding whether or not judgment accuracy improves with more experience, a comprehensive assessment of the research is needed. In the next sections, rationales are provided for the use of meta-analysis in assessing the experience-accuracy relation as well as for the importance of updating and expanding Spengler et al.'s (2009) comprehensive meta-analysis.

Limitations of traditional narrative reviews

According to Cook and Leviton (1980), traditional narrative reviews have been criticized based on three important points: a) a simple box count is used to tally the number of studies of statistical significance regardless of effect size; b) the group of studies for which the review is conducted is often biased; and c) a simple box count ignores important statistical interactions. The first point addresses the idea that traditional narrative reviews ignore information about the direction and magnitude of relationships (Light & Smith, 1971). The result is that findings are overly conservative due to results of statistical non-significance being counted as failures regardless of their direction. This is especially problematic in the social sciences, where studies often rely on small sample sizes, which in turn, lead to low statistical power and non-significance (Cook & Leviton). Meta-analysis, with its inherent synthesis of effect sizes, and other estimates of magnitudes less dependent on sample size, avoids some of these pitfalls of the traditional narrative review (Light & Smith; Smith & Glass, 1977).

The second major problem of traditional narrative reviews, that of bias in the sample of studies used, can manifest in too narrow of a literature search, studies being

excluded based on methodological issues, and studies excluded based on unrelated theoretical constructs (Cook & Leviton, 1980). Meta-analysts also risk applying too rigid of exclusionary criteria and excluding relevant and informative studies (Smith & Glass). According to Glass (1978), when conducting a meta-analysis a wide net should be cast to include any relevant studies regardless of variability in how the variables of importance were defined. Instead of excluding studies based upon perceived variability, post-hoc comparisons should be performed afterwards to identify subsets of the data. In this way, meta-analysis has the potential to diminish subjective exclusion of articles based upon the meta-analysts' preliminary reviews of the included studies. The third problem with traditional narrative reviews pertains to the idea of the box count ignoring important interactions. According to Cook and Leviton, many traditional narrative reviews intend to test simple main effects only and ignore other interactions deemed irrelevant by the reviewers. Due to this exclusion, traditional narrative reviews risk oversimplifying phenomena and failing to recognize complex, important interactions. As Cook and Leviton state, "The best narrative reviews identify and explain contradictory and unexpected data patterns for which no specific boxes were initially set up" (p. 463).

Rationale for the use of meta-analysis

Due to the limitations of traditional narrative reviews, and the importance of understanding the relation between experience and judgment accuracy, a review of this research will be conducted using meta-analysis techniques. The statistical foundations of meta-analysis were laid out by William Gemmell Cochran, who in 1937 discussed a method of combining effect sizes across independent studies. He was also a forerunner in laying out much of the statistical techniques that modern meta-analysis is built upon,

such as inverse variance weighting and homogeneity testing. Although Cochran's methods laid the statistical foundation for meta-analysis, little attention was paid to meta-analysis until the 1950s and not until the 1970s in the social sciences (Hunt, 1997). In 1976, meta-analysis received its name by Gene Glass in his presentation at a San Francisco conference. Glass' paper presented five basic phases of the meta-analytic process. In Glass' (1976) words,

> Meta-analysis refers to the analysis of the analyses. I use it to refer to the statistical analysis of a large collection of results from individual studies for the purpose of integrating the findings. It connotes a rigorous alternative to the casual, narrative discussions of research studies which typify our attempts to make sense of the rapidly expanding research literature (p. 3).

Perhaps one of the most influential periods in the history of meta-analysis, and definitely one of the most controversial, included Smith and Glass' (1977) meta-analysis of psychotherapy outcomes, entitled "The benefits of psychotherapy." Smith and Glass' study arose in response to Hans Eysenck's (1952) review in which nineteen studies, covering over seven thousand cases related to both psychoanalytic and eclectic types of treatment, were analyzed. The findings of Eysenck's review failed to support the hypothesis that psychotherapy facilitated recovery from neurotic disorder. Many scholars, including Smith and Glass, were quick to critique Eysenck's review and defend psychotherapists' livelihoods. In response, Smith and Glass conducted the first ever social science meta-analyses in an effort to assess the effectiveness of psychotherapy. They combined studies involving a variety of therapeutic orientations as applied to a variety of populations. According to their findings, psychotherapy works and all types

work approximately equally well (Hunt). Several additional meta-analyses of psychotherapy effects have been conducted since Smith and Glass, and meta-analysis has been recently utilized for a wide variety of research topics, for example hindsight bias (Guilbault, Bryant, Brockway, & Posavac, 2004), cognitive-behavior therapy for anxiety disorders (Belleville, Cousineau, Levrier, & St. Pierre-Delorme, & Machand, 2010), and internalizing problems in children (Reijntjes, Kamphuis, Prinzie, & Telch, 2010).

Rationale for present study

To date, only one meta-analysis has been conducted evaluating the relationship of experience and judgment accuracy (Spengler et al., 2009). According to Spengler et al., "No other area of clinical judgment research has been synthesized by meta-analytic techniques" (p. 7) aside from two meta-analyses evaluating clinical versus statistical prediction (Ægisdóttir et al., 2006; Grove, Zald, Lebox, Snitz, & Nelson, 2000). In the Spengler et al. study, the effects from 75 clinical judgment studies spanning the years 1970 to 1996 were meta-analyzed. Spengler et al. found a small, but reliable effect demonstrating a positive correlation between experience and judgment accuracy. They noted a 13 percent increase (d = .12) in clinicians' decision-making accuracy with more experience, regardless of other factors. While this finding was more hopeful than what was previously found in narrative reviews of experience and judgment accuracy, it did not reflect an overwhelming improvement in judgment accuracy as a function of experience. Spengler et al. stated, "Given the amount of time, money, effort, and training required for clinicians, the present findings suggest that they do not receive much 'payoff' or benefit for their cost" (p. 35).

The discussion of the explicit teaching of clinical judgment skills has continued since the Spengler et al. (1995) article. Grove (2001) commented on an American Psychological Association Division 12 Task Force study, in which the importance of teaching clinical judgment in training programs was second only to the teaching of ethical and legal standards. Lilienfeld et al. (2003) have stated that in order to receive accreditation by the American Psychological Association, graduate training programs should be required to provide formal education in the area of clinical decision-making. Although there has been steady publication of literature emphasizing the importance of sound clinical judgment skills, it seems the impact on training programs and clinicians' actual judgment-making behaviors has been scarce. In an attempt to investigate how much academic time is actually dedicated to the explicit teaching of clinical judgment skills, however, Harding (2007) examined courses from APA accredited programs as well as surveys completed by academic training directors. The findings from the Harding study are disappointing given the assertions about the importance of the teaching of clinical judgment skills. Although academic training directors agreed upon the importance of possessing effective clinical decision-making skills, greater than 70 percent admitted their programs were in need of more explicit instruction on the topic. Harding also found that actual course content and program literature indicated less focus on the teaching of clinical judgment and decision-making skills than was suggested by survey responses from the academic training directors. In general, it seems the calls for reform in the way clinical judgment skills are taught and enacted in actual practice have been largely ignored.

An update and extension of the Spengler et al. (2009) meta-analysis will allow for the examination of additional moderator variables as compared to those drawn from the Spengler et al. study. In individual studies allowing for examination of the relation between experience and judgment accuracy, few studies included analyses of the moderating or mediating effects of additional variables. Oftentimes the experience-accuracy relation was presented in supplementary or post-hoc analyses, decreasing the likelihood of in-depth examination of moderating or mediating variables of the effect. In other studies, the experience-accuracy relation was presented as the primary analysis, but likewise relatively few impacting variables were assessed. For example, in Brammer (2002), the number of diagnostic questions asked by the clinicians was assessed as a mediating variable of the experience-accuracy relation. In Witteman and van den Bercken (2007), the experience-accuracy effect was assessed in relation to the impact of the type of case presented (i.e., child, adult) and the experience level of the clinicians (i.e., novice, intermediate, experienced). In the present study, two moderator variables are assessed as compared to the Spengler et al. meta-analysis. Profession type was assessed as a moderator variable in the present meta-analysis in order to explore the possibility of its impact on the experience-accuracy effect. Profession type referred to the subfield of mental health from which the clinician sample was drawn, for example social work, psychiatry, and psychology. Although the Spengler et al. meta-analysis did not analyze profession type as a moderator, it is reasonable to examine whether or not various mental health professionals learn differently from experience. Profession type has implications for the type of training mental health professionals undergo in relation to the formulation of clinical judgments. Specifically, various mental health professions

may emphasize different epistemologies in both training and clinical practice. These various epistemologies may influence how the clinicians accumulate knowledge and improve their judgment-making skills throughout their careers.

In addition to profession type, the inclusion of non-mental health participants was newly addressed as a moderator variable. Spengler et al. (2009) addressed studies in which mental health clinicians were compared to each other in relation to their judgment accuracy; therefore, the lowest level of training in the sample pertained to first-year graduate students in the mental health field. As stated by Lambert and Ogles (2004), psychotherapy meta-analyses often address studies in which the relative levels of experience between untrained and trained psychotherapists are quite small. Relatively fewer studies, such as Goldberg (1959), explored greater experience ranges in relation to judgment accuracy. Goldberg, in particular, asked staff psychologists, psychology trainees, and hospital secretaries to make diagnostic judgments of brain damage based upon Bender-Gestalt protocols of 15 organic and 15 nonorganic patients. In his review, Garb (1998) commented that experts are no more accurate than less experienced clinicians and clinicians are not found to be more accurate than graduate students; however, clinicians may be more accurate than beginning graduate students and also more accurate than laypeople. To demonstrate the effect of training and experience on judgment accuracy utilizing a relatively wide experience range, Lambert and Wertheimer compared undergraduates, graduate students, and paraprofessionals in their ability to accurately diagnose psychopathology based on client case histories. These groups were divided into three groups according to their education level (i.e., no education, low education, and moderate education) as well as three experience groups (i.e., no

experience, low experience, and moderate experience). Case histories were constructed with adherence to DSM-IV criteria. Results revealed positive experience-accuracy effects, with those in the low education group performing significantly better than those in the no education group. In addition, those in the moderate education performed significantly better than the low and no education groups. In regards to the experience groups, those in the low experience group performed slightly better than the no experience group. However, those in the moderate experience group performed significantly better than the low education and no education groups. Lambert and Wertheimer concluded that the inclusion of lower levels of training and experience resulted in the larger, positive experience-accuracy effect. These findings and comments suggest that in order to maximize the finding of a larger, statistically significant positive correlation between experience and accuracy, or performance in the case of the psychotherapy research cited by Lambert and Ogles, there needs to be a relatively large range in the experience levels of the participants. Lambert & Wertheimer (1988) addressed the problem of restricted range in clinical judgment research, stating "Even if there is a strong relationship between training or experience and diagnostic accuracy, that relationship may not be detectable if all the participants in a study are selected from a group that already has a substantial amount of training and experience" (p. 50). The exclusion of studies in which undergraduates and non-mental health professionals were compared to mental health clinicians in Spengler et al. may have resulted in difficulty capturing the larger overall effect. Unlike the Spengler et al. meta-analysis, the present meta-analysis included studies in which non-mental health clinicians are compared to mental health clinicians (i.e., those with at least some graduate level training in a mental

health field). It is expected that this inclusion will maximize the finding of the positive experience-accuracy effect.

Moreover, an update and expansion of the Spengler et al. (2009) meta-analysis will allow for cross-validation of the Spengler meta-analysis. Due to the presence of only one available meta-analysis examining the experience-judgment accuracy effect, cross-validation of this meta-analysis will be a beneficial contribution to the clinical judgment research. An update will provide a test of robustness of the Spengler et al. findings using current research. It is hypothesized that the present meta-analysis will reveal a modest, positive correlation between judgment accuracy and experience based upon the findings of the Spengler et al. meta-analysis as well as an inclusion of additional moderator variables.

Spengler et al. (2009) examined several moderator variables, which will also be analyzed in the present study in order to test the robustness of their findings. Original moderators based upon the Spengler et al. meta-analysis included experience type, experience breadth, type of judgment, criterion validity, provision of feedback, publication source, ecological validity of method of study, ecological validity of stimulus, study quality, and age of study. Additional moderator variables include profession type (psychology, psychiatry, nursing, social work, or combination) and inclusion of non-mental health participants (yes, no).

Experience type was divided into three categories: clinical, educational, or both. According to Spengler et al. (2009), experience type was not found to make a significant impact on judgment accuracy. This is consistent with the multitude of findings demonstrating nonsignificant and negative findings regarding the relationship between

judgment accuracy and experience (e.g., Goldberg, 1959; Highlen & Hill, 1984; Kirk & Hsieh, 2004; Silverman, 1959; Witteman & van den Bercken, 2007). These studies employed a variety of operational definitions of experience, including clinical and educational. For the present study, it is therefore hypothesized that experience type will not have a statistically significant impact on the experience-accuracy effect.

Experience breadth was divided into three categories: general, specific, and both. According to Spengler et al. (2009), specificity of experience did not impact the experience-accuracy effect. Popular opinion holds that having specific or specialized training or experience in a particular area would result in better judgment making in that area. Some research has supported this claim (Fairman, Drevetz, Kreisman, & Teitelbaum, 1998; Goldstein, Deysach, & Kleinknecht, 1973). However, other research has suggested having specialized training or experience merely leads to more confidence in related tasks. In addition, research on experts in psychology has found that experts often form judgments based on a priori schemata regarding the presenting problem and that while their judgments are often more automatic and made with greater confidence, they are not always more accurate (Eells et al., 2005; Goldberg, 1959; Witteman & van den Bercken, 2007). For the present study, it is hypothesized that experience breadth will not make a statistically significant impact on the experience-accuracy effect.

Judgment type made refers to the kind of judgment participants were required to make and could be defined as problem type, hit rate, treatment, severity, prognosis, problem recall, other, or combined. In the Spengler et al. (2009) meta-analysis, it was found that more experienced clinicians showed greater diagnostic accuracy, were more accurate at formulating appropriate treatment recommendations, and more accurately

recalled problems. Additionally, the "other" category revealed a moderate effect with more experienced clinicians making more accurate judgments. One could speculate that these categories offer more clear-cut guidelines for decision-making than would judgments of severity and prognosis. Therefore, it is hypothesized that the present study will find statistically significant moderator effects for type of judgment made, with more experienced clinicians showing better accuracy in diagnosis, formulating treatment recommendations, and recalling problems.

Criterion validity could be categorized as low, high, or both and refers to how the standard for judgment accuracy was established. For example, Yeo et al. (2001) asked participants to provide ratings of clients' problem type based upon clinical vignettes. They were provided with several problem type choices, including stress, depression, schizophrenia/paranoid schizophrenia, mania, anxiety, physical weakness, mental weakness, being possessed, and 'other.' Due to the lack of standardized diagnostic choices as well as the lack of *a priori* validation methods (i.e., pilot study to validate clinical vignettes), this included study was considered to have low criterion validity. Contrary to expectations, Spengler et al. (2009) found that studies with low criterion validity resulted in higher effect sizes. This is consistent with findings that experts' knowledge is organized into broader and complex memory structures, allowing them to make connections between seemingly irreconcilable concepts (Hinsley et al., 1978). Additionally, it has been found that experts view problems according to their fundamental principles or underlying structures, as opposed to novices who focus more on superficial characteristics (Chi et al., 1981; Ericsson & Charness, 1994; Haerem & Rau, 2007). These characteristics could aid more experienced clinicians in their

assessments of problems with "fuzzy" criteria. Novices, on the other hand, may demonstrate poorer performance with low criterion validity tasks due to their tendencies to focus more on the details. Therefore, it is hypothesized that the present study will find a statistically significant moderator effect, with more experienced clinicians outperforming less experienced clinicians when the judgment tasks reflect low criterion validity.

Provision of feedback was categorized dichotomously (i.e., yes or no) and refers to whether or not feedback was provided to participants relevant to the judgment task. The lack of direct feedback in psychological practice has long been emphasized by scholars and researchers as one of the major reasons judgment accuracy has not been shown to improve with greater experience (Dawes et al., 1989; Garb & Boyle, 2003; Lichtenberg, 1997). However, it is also noted that not all feedback is equally valid and useful to the clinician making the judgment. Oftentimes, the feedback clinicians receive is not based on objective assessments of their performance but involves bias and pseudoscience perspectives on the part of the individual providing the feedback (Lilienfeld, et al., 2003). In the Spengler et al. (2009) meta-analysis, only two studies were found that addressed feedback (deMesquita, 1992; Horner, Guyer, & Kalter, 1993). They concluded that feedback had not played a significant role in the experience-accuracy effect. Based on initial reviews of studies to be included in the present meta-analysis, it is expected that the ability to test this hypothesis will be limited based on the few studies found that provided feedback to the participants. However, it is hypothesized that feedback will not show a statistically significant impact on the experience-accuracy effect.

Publication source could be categorized as APA journal, another psychology journal, psychiatry or medical journal, or dissertation. It has been noted that publication bias exists and can skew the results of meta-analysis if not accounted for. Specifically, larger effect sizes are found for published journal articles in comparison to unpublished studies, theses, or dissertations (Kurosawa, 1984; Light & Pillemer, 1984; Peters, Sutton, Jones, Abrams, & Rushton, 2006; Rosnow & Rosenthal, 1989). Spengler et al. (2009) tested this moderator and found that studies published in non-APA psychology journals reported much smaller effects than those found in APA journals. It is therefore hypothesized that studies published in APA journals will reveal greater effects than those in non-APA journals.

Ecological validity of method of study could be categorized as analogue, archival, or *in vivo* and refers to the way in which material was presented to the participants. According to the Spengler et al. (2009) meta-analysis, the ecological validity did not make a significant impact on the experience-accuracy effect. Likewise, it is hypothesized that the present study's findings will not reveal a statistically significant impact on the experience-accuracy effect.

Ecological validity of stimulus could be categorized as direct, indirect, or both and refers to the method used for the stimulus presentation. The Spengler et al. (2009) meta-analysis reported that neither direct nor indirect presentation of the stimulus played a significant role in judgment accuracy. It is hypothesized that the present study will reveal similar findings.

Relation of experience to the research design could be categorized as not in design, in primary design, in supplementary analysis, and multiple, and referred to

whether or not experience was included as a component of the primary research design. In quantitative research, supplementary analyses often capitalize on chance, inflating the size of the effect (Keppel, Saufley, & Tokunaga, 1992). They are considered to be opportunistic and, if not controlled for through the use of techniques such as the Tukey test (Tukey, 1953) or the Scheffé test (Scheffé, 1953), can lead to distortion of or overemphasis on unplanned significant effects. Studies could be coded as not in the research design when the researchers did not include an analysis of experience in either the primary design, supplementary analysis, or in multiple parts of the design. These studies required extrapolation of the data from tables or from descriptions as well as reorganization of the data to fit the present meta-analysis format. In addition, "not in design" studies could be coded for studies whose authors required contact to obtain the necessary experience comparisons and data. For example, Hannan et al. (2005) examined clinicians' abilities to accurately detect treatment failure. One author from this study was contacted to obtain the necessary data for the experience variable due to the experience variable not being included or reported in the research design. Spengler et al. (2009) found no statistically significant impact of relation of experience to the research design on the experience-accuracy effect. Likewise, it is expected that the present study will find similar results.

Experience as a major variable could be categorized dichotomously (i.e., yes or no) and referred to the inclusion of experience as a conceptual or theoretical variable of importance in the original plan of the study. When not conceived as theoretical variables of importance, the examined variables in a given study may result in larger, more spurious effects (Keppel et al., 1992). As found in the Spengler et al. (2009) meta-

analysis, it is hypothesized that the present study's findings will not find a statistically significant difference based on whether or not experience was included as a major variable of importance.

Study quality could be categorized either as acceptable, good, or excellent and refers to the subjective rating of overall methods and analyses. The impact of study quality on meta-analysis is important in that meta-analysis cannot correct for studies with serious flaws (Eysenck, 1994). In the Spengler et al. (2009) meta-analysis, study quality was not found to be a significant moderating variable. It is hypothesized that the present study will reveal similar findings.

Profession type could be categorized as psychology, psychiatry, nursing, social work, or a combination. Within the mental health profession, there may be significant differences in training and experiences based on the specific profession addressed that affect the clinicians' abilities to learn from experience. At a basic level, various profession types emphasize different etiologies of pathology as well as varying aspects of presenting problems. Although the profession types of the clinicians may impact their judgment accuracy throughout their careers, empirical support was not found for a significant relation between certain profession types and judgment accuracy. Therefore, no specific hypothesis was formed regarding profession type and its impact on the experience-accuracy effect.

Inclusion of non-mental health participants could be coded dichotomously as "yes" or "no." The Spengler et al. (2009) meta-analysis excluded studies based upon comparisons between non-mental health participants and mental health clinicians if analysis solely within the mental health clinician group was not conducted in the original

study. However, the inclusion of non-mental health participants may achieve a greater range of experience and will allow for comparisons to be made when the baseline level of experience for participants is significantly lower than that of the mental health clinicians. According to Lambert and Wertheimer (1988), the failure of many studies to reveal a positive experience-accuracy effect may in fact be due to the efficacy of graduate-level training. In other words, after reaching a certain level of graduate status, these less experienced students may perform at a similar accuracy rate as compared to more experienced clinicians. Unless studies incorporate relatively large experience ranges (i.e., undergraduates versus experienced clinicians), the experience-accuracy effect may be diminished. In his review, Garb (1998) noted the greatest differences in judgment accuracy seemed to be found when the experience range of the participants in question was considerably large, such as when the judgment accuracy of laypeople or first-year graduate students was compared to that of highly experienced clinicians. In Spengler et al., the typical study included in the meta-analysis focused on small to moderates ranges of experience. In fact, Lambert and Wertheimer was included in the Spengler et al. meta-analysis and revealed one of the largest experience-accuracy effects. The problem of restricted range is common in behavioral psychology research, resulting in underestimations of relations between variables. For example, if a researcher aims to examine the correlation between intelligence and political affiliations, he or she will encounter the problem of restricted range if the sample includes only undergraduates. It is likely that individuals in a university setting will have average to above average IQ scores, precluding the researcher from addressing individuals with lower IQ in relation to their political affiliations. In clinical judgment research, it is common for studies to

incorporate restricted ranges of experience (Lambert and Wertheimer), precluding researchers from generalizing findings beyond the particular experience range examined in the study. It is hypothesized that studies in which non-mental health participants are included will show significantly larger experience-accuracy effects.

Publication year will be categorized as the exact year of publication of each study. Publication year was addressed in the present meta-analysis in order to examine possible differences in the experience-accuracy effect based on when the study was published. As discussed previously in this chapter, important contributions in the clinical judgment literature have occurred since the time of the Spengler et al. (2009) meta-analysis, specifically Garb's (1998) review of clinician factors impacting clinical judgment as well as APA's calls for more explicit emphasis on fostering clinical judgment skills in training programs (Grove, 2001). However, it is difficult to gauge the actual impact these contributions have had in clinical practice through research studies, especially when the utilization of survey methods introduce social desirability bias (Harding, 2007). As found in Spengler et al., it is hypothesized that publication year will not have a significant impact on the experience-accuracy effect.

Chapter Summary

This chapter discussed the empirical findings in the areas of research related to clinical judgment and experience research. These areas included the distinction between experience and expertise, unique characteristics of expert performance, investigations of client variable biases and heuristics, and a summary of the research evaluating the relationship between clinical experience and judgment accuracy. Additionally, this section provided an overview on the limitations of traditional narrative reviews. Finally,

a rationale was provided as to why an update and expansion of the Spengler et al. (2009) meta-analysis is important. Similar to the findings of the Spengler et al. meta-analysis, it is hypothesized that the present study will reveal a small, positive effect, indicating a slight increase in judgment accuracy in relation to experience. With the exceptions of type of judgment, criterion validity, publication source, and inclusion of non-mental health participants, it is hypothesized that the remaining moderator variables will not significantly impact the experience-accuracy effect. In the following chapter, research methodology is discussed that will allow the present study to address these hypotheses.

Chapter 3: Methods

The purpose of this chapter is to outline the research methods that were used for the present study. In this chapter, study search and selection are discussed and a description of the coding procedures utilized is provided. Also, important terms and variables are defined and operationalized, such as the main variables of judgment accuracy and experience as well as the moderator variables. Finally, meta-analysis methodology is discussed in greater detail, with a focus on effect size as the unit of measurement and the use of the random effects model for the purposes of the present study.

Study search

As part of the more comprehensive MACJ project, Spengler et al. (2009) evaluated studies from 1970 to 1996 that addressed the relationship between experience and judgment accuracy. A study search was conducted using electronic databases, including PsychINFO, ERIC, Dissertation Abstracts, MEDLINE, and Social Science Index. For the MACJ project, 35,000 studies were initially identified and reviewed for inclusion. Subsequent to review, 4,617 studies were chosen that addressed some type of judgment, either distinctively or possibly mental health-related or mental health-related. Upon coding these 4,617 studies for their content, 1,135 studies met the initial inclusion criteria described later in this chapter. Only 316 of the 1,135 identified could be coded for the experience variable. Moreover, only 106 of the 316 studies established a standard for judgment accuracy. Finally, 75 studies remained that included sufficient statistical data necessary to calculate effect sizes.

The present study used the same electronic databases as in Spengler et al. (2009) with the exclusion of BRS, which had dissolved by the time of the present meta-analysis search. Moreover, the same search terms were used as in the Spengler et al. meta-analysis and applied to studies from 1997 to 2010 (see Appendix A). For each database, limiters were set to prevent the inclusion of editorials, comments, replies, book chapters, and studies conducted with non-human subjects in the search retrieval. When necessary, customer support for the databases was sought to confirm the utilization of the specific limiters in question for the purposes of the present meta-analysis. These limiters greatly reduced the number of articles retrieved for each search term to those that involved empirical studies of varying research designs. At two stages, the search was expanded by searching for related studies in reference sections as well as searching for studies cited by the original studies found (i.e., forward and backward cross-referencing). Forward referencing was carried out with the 75 included studies of the Spengler et al. meta-analysis. Additionally, forward and backward cross-referencing was implemented with the final 37 studies selected for the present meta-analysis. This resulted in the return of many of the same studies that were previously reviewed in the first stages of the search process.

For the present meta-analysis, 7,789 studies were initially identified using the search terms in the Spengler et al. (2009) meta-analysis as well as the search limiters (see Appendix B). Of these 7,789 studies, 862 were chosen that were believed to address some type of mental health judgment accuracy task (e.g., diagnostic decision-making, violence risk assessment, prediction of treatment failure) as identified in their titles and further confirmed with their abstracts. This screening for judgment accuracy studies

resulted in the exclusion of studies that clearly examined judgments for which there was no standard for accuracy, for example studies of judgment bias, judgment processes, and studies in which mental health clinicians were surveyed regarding their attitudes towards various client populations and problems. The remaining 862 studies were located and obtained through various methods, which included downloading electronic copies, ordering electronic copies through Interlibrary Loan, seeking out hard copies of articles and dissertations, and purchasing electronic copies of dissertations. In the next step of the search process, studies were reviewed to ensure they included at least one group of mental health professionals in their participant sample. This step resulted in the exclusion of studies whose sole participant base were undergraduates, medical professionals (i.e., those without mental health training), or those in unrelated professions (e.g., chemists, school principals, police detectives). The remaining 302 studies were further subjected to the inclusion criteria outlined later in this section. Upon meeting inclusion criteria, 85 studies were then coded for their content. From these 85 studies, 50 were confirmed as establishing a standard for judgment accuracy. Only 37 of these studies, however, included sufficient statistical data to calculate effect sizes. In some cases, the authors of the studies were contacted to obtain the necessary statistical data for the included studies.

Study selection

In the included studies for both the Spengler et al. (2009) meta-analysis as well as the present meta-analysis, experience was defined as either clinical, educational, or both. Studies focused on mental health issues and involved clinical judgment, clinical judgment bias, or clinical versus statistical prediction. The types of judgment being made in the

studies were related to problem type, hit rate, treatment, problem severity, prognosis, or problem recall. Moreover, studies that included some measure of judgment accuracy were included, although the measures, or benchmarks, utilized for the judgment accuracy task reflected varying degrees of criterion validity. Additionally, included studies were required to investigate at least two groups for comparison of experience and judgment accuracy relationships or a first-order correlation of judgment accuracy and experience. The mental health judges included professionals working in various fields, such as psychology, psychiatry, social work, counseling (mental health, school, rehabilitation, and pastoral), and psychiatric nursing. Studies evaluating graduate students in these fields were also included. Unlike the Spengler et al. meta-analysis, studies including undergraduate students and/or professionals in other fields were also admitted as long as these groups were compared to at least one of the previously mentioned groups of mental health clinicians. Included studies were also required to provide the data necessary to calculate effect sizes. If a study did not report sufficient data to calculate the necessary effect size, the authors were contacted and asked to provide the necessary data.

Coding

Studies were coded for variables related to study characteristics and statistics. An original coding form created for the present study was utilized that allowed for the recording of moderator variables (see Appendix C). In addition, a coding form from the MACJ Project (Spengler et al., 2009) meta-analysis was utilized that allowed for the recording of statistical data (see Appendix D). Since the present study examined only one of the sub-analyses of the MACJ (i.e., effect of experience on judgment accuracy), fewer variables were coded in comparison to Spengler et al.

In terms of study characteristics, the Moderator Coding Sheet (see Appendix C) was used to code for variables related to the experience type, experience breadth, judgment type, criterion validity, provision of feedback, publication source, method of study, validity of stimulus, relation of experience to design, experience as a major variable, study quality, profession type, inclusion of non-mental health participants, and publication year. In order to ensure coding was performed with adherence to predetermined guidelines and criteria, two raters selected and coded all included studies using the Moderator Coding Sheet. The second rater was a third-year graduate student in school psychology who received prior training from the author on the use of the coding sheets. Cohen's kappa was calculated to determine interrater reliability since it is a relatively conservative measure of agreement, especially when used in tasks with many categorical variables (Cohen, 1960). Coding discrepancies were discussed with the doctoral committee chairperson for further clarification and resolution.

Cohen's kappa was calculated for 13 categories across all 37 included studies. Kappa ratings ranged from .71 to .96, indicating substantial agreement to almost perfect agreement according to Landis and Koch (1977). While these subjective labels are by no means universally agreed-upon, they assist in providing a more informative examination of the individual kappa ratings. Table 1 displays the kappa ratings achieved for each coding category using the Moderator Coding Sheet.

In relation to statistics, studies were coded using the Metrics Coding Sheet, created for the Spengler et al. meta-analysis (see Appendix D). The Metrics Coding Sheet allowed for coding of the statistical relationship between the dependent and independent variables, whether that was in the form of means and standard deviations,

correlation coefficients, hit rate percentages, odds ratios, *F* and *t* distributions, or chi-square distributions. For the dependent variables, the more general terms were recorded first (e.g., problem type) with more specific definitions provided in parentheses (e.g., judgment of neurological impairment versus no impairment). Likewise for the independent variable, more specific definitions (e.g., years of clinical experience) followed the more general term (e.g., experience). Once the statistical relationship between judgment accuracy and experience was recorded, it was necessary to determine the direction of the effect, if possible. Additionally, the rater was prompted to code level of confidence in the rating of accuracy (low versus high). Finally, a global rating of methods and analyses employed by the individual study was coded (poor, adequate, or excellent). The final two items of metrics coding, level of confidence in rating of accuracy and global rating of methods/analyses, called for subjective judgments to be made on the part of the rater. The level of confidence in rating of accuracy item involved judging whether or not the individual study employed standards for accuracy that were clearly defined, objective, and validated in some way (e.g., a priori predictions of violence compared with actual violence recorded on a standardized measure). The global rating of methods/analyses item involved use of Cook and Campbell's (1979) discussion of threats to internal and external validity.

Definitions of terms

Independent measure: Experience. For the present study, the term experience was operationalized similarly to the Spengler et al. (2009) meta-analysis. Experience encompassed both clinical and educational. Clinical experience referred to number of clients seen, number of tests administered, time of counseling experience, job setting, or

others. Educational experience referred to number of graduate courses taken, year in graduate training, level of training, amount of face-to-face clinical supervision, training intervention, or others. In addition, experience could be general, specific, or both. General experience referred to the type of experience gained as part of core training and common tasks as a mental health clinician. Specific experience, in contrast, referred to the type of experience gained that would be of specific use for the judgment task in question. These differences in types of experience were reflected in the included studies. For example, Falvey et al. (2005) operationalized experience as number of years the clinicians have worked in the field as well as their exposure to clients with a diagnosis of Attention-Deficit/Hyperactivity Disorder. The first type was considered general clinical experience and the second was considered specific clinical experience. In Monsen and Frederickson (2002), on the other hand, experience was defined as a specific training intervention in interviewing and case formulation techniques for school psychologist trainees. This was considered specific educational experience.

Dependent measure: Judgment accuracy. Judgment accuracy, likewise, was defined and operationalized in many different ways. For the present study, like in the Spengler et al. (2009) meta-analysis, judgment accuracy referred to the validity of clinicians' judgments related to various constructs, such as problem type, hit rate, treatment, severity, prognosis, problem recall, or others. Studies varied in the measures and in the quality of measures used to evaluate judgment accuracy. For example, in Garb and Boyle (2003), 25 neuropsychologists were provided with two sets of written case material and asked to make judgments of neurological impairment and dementia if applicable. The written case material included neuropsychological test protocols based

on the average scores for community dwelling 38-year-old and 74-year old individuals. Since the test scores were strictly based upon the average performances of a 38-year-old and 74-year-old, it would have been difficult for clinicians to justify making judgments of neurological impairment or dementia. Any judgments made of neurological impairment or dementia could confidently be said to reflect errors in judgment. While this method provided a relatively objective measure of clinicians' judgment accuracy, other studies employed methods in which the standards for accuracy were more difficult to determine. Monsen and Frederickson (2002) assessed school psychologist trainees both before and after a specialized training session in the use of accessible reasoning techniques during school consultations with teachers. One of the outcomes on which school psychologist trainees were assessed, judgment accuracy, was defined as "the degree to which the aspects identified by the participants were based upon the facts of the case, as opposed to including errors such as over-generalization and speculative inference" (Monsen & Frederickson, p. 204). The trainees' judgment accuracy was assessed by the first author and cross-validated by two tutors "experienced in assessing written problem understanding" (p. 204). In this study, the standard for judgment accuracy was relatively difficult to determine, as it relied on the subjective judgments of the study's author as well as two others rather than comparison to some type of objective standard. Judgment accuracy, however operationalized, was evaluated for criterion validity as in the Spengler et al. study.

Moderator variables

Experience type reflected the specific kind of experience addressed in the study, whether that was clinical, educational, or both. Clinical experience could refer to number

of clients seen, number of tests administered, time of counseling experience, job setting, or others. Educational experience could refer to number of graduate courses taken, year in graduate training, level of training, amount of face-to-face clinical supervision, training intervention, or others. Brammer (2002) included years of experience as well as level of training (i.e., master's, doctorate) into the estimation of a participant's amount of experience. In the case of the Brammer study, the focus on the level of training the participants had received would fall into the category of educational experience. Brammer also addressed the clinical experience of the participants by recording the number of years the participants had worked in the field. It was possible for studies, as was the case in Brammer, to include both educational and clinical types of experiences.

Experience breadth referred to whether the study addressed general and/or specific experience. General experience could be addressed by studies reporting on clinicians' general level of training or years of experience in the mental health field, for example. Specific experience, in contrast, could be addressed by studies reporting on participants' specific experiences or training related to the type of judgment task being studied. For example, Falvey et al. (2005) reported on clinicians' previous exposure to similar cases as the ones presented to them in the study. However, other researchers simply reported on clinicians' years of experience working in the mental health field (e.g., Witteman & van den Bercken, 2007).

Judgment type referred to the kind of judgment participants were required to make, such as problem type, hit rate, treatment severity, prognosis, problem recall, other, or combined. Brammer (2002), for example, asked clinicians to make diagnostic decisions after they utilized an artificial intelligence program that simulated a clinical

interview. This type of judgment fell under the category of problem type. Problem recall pertained to studies in which clinicians were asked to recall clients' presenting problems. Dawson, Zeitz, and Wright (1989), for example, asked novice and experienced clinicians to make judgments based upon multiple observations of children's behaviors with varying degrees of aggression. One specific judgment type addressed by this particular study was problem recall, in which the clinicians were rated on the amount of information they correctly recalled from the behavior observations.

Criterion validity referred to the rating of high, low, or both based on how the standard for judgment accuracy was established. High criterion validity was noted when the researchers included standards for accuracy that were highly valid or objective. For example, Witteman and van den Bercken (2007) required participants to make diagnostic decisions based on preexisting, standardized case studies included in the DSM-IV Case Book. Members of the APA DSM Task Force and Work Group as well as groups of advisers and consultants wrote the cases. Less objective methods of judging accuracy, for example utilizing a panel of "expert" judges (Brammer, 2002), reflected low criterion validity since there was much more room for disagreement among judges who had no part in the writing of the case studies and had not already been members of an organized professional group.

Provision of feedback referred to whether or not feedback was provided to the participants of the study at some point in their judgment processes. This moderator was coded only if judgment accuracy was measured before and after the provision of feedback or if various feedback conditions were measured between groups in relation to judgment accuracy.

Publication source referred to the original location of the study, in other words an APA journal, another psychology journal, a psychiatry or medical journal, or if the study was a dissertation.

Ecological validity of method of study referred to the way in which material was presented to the participants to obtain judgments and could be categorized as analogue, archival, or *in vivo*. For example, in some studies participants were presented with case studies of fictitious clients (e.g., Brammer, 2002), while in other studies, live client scenarios were used or simulated (e.g., Hickling et al., 2002). In the former study, the analogue method was used while an *in vivo* method was used in the latter.

Ecological validity of stimulus referred to the method used for the presentation of material and could be categorized as directly experienced, indirectly experienced, or both. In other words, participants could experience the material directly, for example via videotape, role-playing, or live presentation, or the participants could experience the material indirectly, for example through written case studies or test protocols.

Relation of experience to the research design referred to whether or not experience was not in the design, part of the primary design, part of the supplementary analysis, or in multiple areas.

Experience as a major variable referred to whether or not experience was conceived as a major theoretical variable of importance. Experience as a major variable was also coded if the study made specific *a priori* hypotheses in regards to the effects of experience.

Study quality referred to the overall quality assigned to each study. Studies were judged as acceptable, good, or excellent. Decisions about study quality included the

study's design, execution, and analysis. Cook and Campbell's (1979) review of research methods and design offered guidelines for judging overall study quality.

Profession type referred to the specific profession of the mental health clinician, whether that is psychology, psychiatry, nursing, social work, or a combination. For example, Jopp (2001) included clinicians in the field of psychology and was therefore coded as, "psychology." This moderator reflected *only* the mental health clinicians, meaning those with at least graduate-level experience.

Inclusion of non-mental health participants referred to the inclusion of participants not in the mental health field (i.e., the inclusion of undergraduates or other non-mental health professionals) in comparison to mental health clinicians in terms of judgment accuracy. Non-mental health participants included undergraduates, other professionals, or a combination. Studies were coded as "yes" or "no" depending on whether or not non-mental health participants were compared to participants with at least some graduate training in a mental health field.

Publication year referred to the exact year of publication.

Chapter 4: Results

From the initial database of 7,789 studies identified using the search terms, 85 met the criteria for study selection and were coded using the Moderator Coding Sheet (Appendix C) and Metrics Coding Sheet (Appendix D). However, only 37 contained sufficient statistical data to warrant inclusion in the meta-analysis. Analysis of the data included several steps. First, a random effects model was chosen based upon previously established goals and limitations of the present study's findings. Following this model, the meta-analysis was conducted with the effect size data from the individual studies. After examining the overall weighted mean effect size, tests for homogeneity of the overall effect were conducted using Hedges' (1982) *Q*. Moderator analyses were then performed to explore the impact of the following variables on the experience-accuracy effect: (a) experience type, (b) experience breadth, (c) judgment type, (d) criterion validity for accuracy dependent measure, (e) provision of feedback, (f) publication source, (g) ecological validity of the method of study, (h), ecological validity of stimulus, (i) relation of experience to the research design, (j) experience as a major variable, (k) study quality, (l) profession type, (m) inclusion of non-mental health participants, and (n) publication year.

Random effects model

A random effects model was used for the meta-analysis. According to Sánchez-Meca and Marín-Martinez (2008), the random effects model has been favored in recent years over the fixed effects model due to its more realistic assumptions of how correlations occur. Specifically, the random effects model explicitly accounts for heterogeneity because the model assumes a different underlying effect for each study and

takes this into consideration as an additional source of variation. The weights in a random effects model are more evenly distributed than those in a fixed effects model. Additionally, the confidence intervals in a random effects model are wider than those in a fixed effects model (Hunter & Schmidt, 2000). Moreover, a random effects model was used because it assumes there is a population of effect sizes from which the studies drawn are a random sample (Hedges & Vevea, 1998). A random effects model was used for the present study due to the high amount of variation observed in how variables of importance were measured, specifically the measurement of experience and judgment accuracy. In addition, this model was chosen based upon the assumption that the studies identified were only a sample of the population of relevant studies. In Spengler et al. (2009), 35,000 studies were reviewed over a period of several years to find relevant studies for inclusion. The research team ensured every relevant study was identified by examining entire articles for experience-accuracy analyses. In contrast, the present meta-analysis employed a search strategy in which two raters reviewed the initial 7,789 studies identified via their titles and abstracts. If abstracts did not identify analyses of mental health judgments in which standards for accuracy could reasonably be inferred, further review of the studies was not conducted. Due to the longer time frame and utilization of a large team of raters to search for and locate relevant studies, the Spengler et al. meta-analysis utilized a fixed effects model. Since it could be confidently stated that every clinical judgment study was located, a fixed effects model allowed Spengler et al. to make conditional inferences based upon the located studies.

Calculation of effect sizes

A total of 190 separate effect sizes were inputted into *Comprehensive Meta-Analysis*, Version 2 (Borenstein, Hedges, Higgins, & Rothstein, 2005) to assist in conducting the meta-analysis. This software allows the researcher to input the individual data from each study in over 100 formats (i.e., means and standard deviations, odds ratios, effect sizes) and is able to transform the data into a standardized metric. For each study, an effect size (*d*) was calculated using the reported means and standard deviations (Glass, McGaw, & Smith, 1981). However, if the means and standard deviations were not reported in a study, other statistics were used to calculate *d*, for example *F* or *t* (Hedges & Olkin, 1985). In studies providing multiple outcome measures, the results were pooled by weighting effect sizes to ensure each study contributes only one effect size (Lipsey & Wilson, 2001; Marín-Martinez & Sánchez-Meca, 1999). An average of the effect estimates from each individual study was calculated by weighting each by its inverse variance. As a result, a confidence interval was obtained (Sánchez-Meca & Marín-Martinez, 2008). Positive effect sizes indicated studies in which judgment accuracy was shown to improve with experience, whereas negative effect sizes indicated studies reporting the opposite effect. Zero effects indicated no statistically significant differences in judgment accuracy based on experience. As in the Spengler et al. (2009) meta-analysis, zero effects were employed when it was not possible to calculate effect sizes from the statistical information given or when studies simply stated that the relationship between judgment accuracy and experience was not statistically significant. As Spengler et al. noted, this process produced a conservative estimate of the effect sizes, or a bias toward the null hypothesis; therefore, significant effort was made to contact the

authors of these studies to obtain the necessary statistical data. Only one study resulted in the utilization of a zero effect conversion (Leon & Perez, 2001).

The overall effect of the present meta-analysis based on a random effects model from 37 individual studies was .17, with a 95 percent confidence interval that was above zero (CI = .06 to .28). Because the confidence interval did not include zero, it was possible to assume the effect was different from zero. This overall effect indicated experience significantly impacted judgment accuracy, consistent with expectations. Table 2 provides a summary of the average weighted effect size estimates for the individual studies. After removal of an outlier study, explained in the subsequent paragraph, the overall effect was .16, with a confidence interval of .05 to .26. Removing the outlier study maintained the positive, small effect size with a confidence interval that was above zero, indicating judgment accuracy shows a slight improvement with experience.

Homogeneity testing

Hedges' (1982) homogeneity statistic, Q, was used to assess the overall sample of studies in order to give support for testing moderator variables as well as identify possible outliers. According to Hedges & Olkin (1985), Q is significant when the variability among the studies cannot be due to chance. The Q statistic was sufficiently large to reject the null hypothesis regarding homogeneity of the effect size distribution, Q (36) = 59.00, p = .009. However, one outlier study was identified (Rerick, 1999) via the Extreme Studentized Deviate (ESD) method (Grubbs, 1969). The ESD method allows for calculation of the maximum deviation from the mean value and subsequent comparison with a critical value. If the maximum deviation is greater than the critical

value, the maximum deviation is then considered to be an outlier and considered for removal. With the present effect size data, the ESD method revealed Rerick as an outlier study, with significance at the .05 level. After reviewing the outlier study to ensure no inputting errors were committed upon inclusion into the meta-analysis program, it was decided that the study be removed due to its relatively large effect size (d_{i+} = 1.86). The Q statistic was still sufficiently large to reject the null hypothesis of homogeneity after the outlier was removed, Q (35) = 52.37, p = .030. The removal of the outlier study resulted in an overall experience-accuracy effect of .16, with a 95 percent confidence interval of .05 to .26. Previously identified moderators were subsequently analyzed in order to gain a more precise understanding of the moderating effects of particular variables in relation to the experience-accuracy effect. According to Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella (2006), exploratory moderator analyses should be conducted when homogeneity testing reveals more heterogeneity that can be explained by chance. In addition, moderator analyses were conducted to allow for a more in-depth comparison with the analyses in Spengler et al. (2009).

Hypothesis testing

Moderator variables were analyzed using Hedges and Olkin's (1985) counterpart to the analysis of variance (ANOVA) statistic. Moderators were chosen based on the Spengler et al. (2009) meta-analysis as well as additional moderators chosen based on a review of the clinical judgment literature. Moderators included (a) experience type, (b) experience breadth, (c) judgment type, (d) criterion validity for accuracy dependent measure, (e) provision of feedback, (f) publication source, (g) ecological validity of the method of study, (h), ecological validity of stimulus, (i) relation of experience to the

research design, (j) experience as a major variable, (k) study quality, (l) profession type, (m) inclusion of non-mental health participants, and (n) publication year. The following outlines the impact of moderator variables in relation to the experience-accuracy effect. For moderators categorized in multiple levels (i.e., analogue, archival, and *in vivo*), the between-class effect was used (Q_B) in addition to the reporting of the 95 percent confidence interval. Continuous variables, however, were analyzed using Rosenthal's (1991) focused comparison of effect size (z). For categorical moderator variables in which fewer than 3 studies pertained to any particular category, that category was removed from the moderator analysis. The limit of 3 studies per moderator category and the exclusion of those categories not meeting this requirement has been the methodology utilized in various meta-analyses (e.g., Ægisdóttir et al., 2006; Hall, Coats, & Smith LeBeau, 2005; Hoeve et al., 2009) as well as recommended by meta-analysts (Lipsey & Wilson, 2001). Table 2 displays the results of the main effects for each study and the effect sizes for the particular judgment type used. Table 3 displays the categorical variable codes for each study. Table 4 displays the results of the moderator analyses. Table 5 displays the corresponding stem-and-leaf plot.

Experience type will not have a significant impact on the experience-accuracy effect. Experience type was comprised of three categories: clinical, educational, or both. It was hypothesized that experience type would not have a moderating effect on judgment accuracy, which was found to be true, $Q_B(2) = .94$, $p = .627$.

Experience breadth will not have a significant impact on the experience-accuracy effect. Experience breadth could be coded as general, specific, or both. The

experience-accuracy effect was not hypothesized to differ based upon experience breadth, which was found to be true, $Q_B(2) = .63$, $p = .729$.

Judgment type will have a significant impact on the experience-accuracy effect, with more experienced clinicians showing better accuracy in diagnosis, formulating treatment recommendations, and recalling problems. Type of judgment made was categorized as problem type, hit rate, treatment, severity, prognosis, problem recall, other, or combined. Consistent with expectations, the type of judgment made had a significant impact on the experience-accuracy effect, $Q_B(3) = 8.27$, $p = .041$. More experienced clinicians were better at assessing clients' problems ($d_{i+} = .29$) and were more accurate in studies employing a combination of judgment tasks ($d_{i+} = .29$). However, only one study was identified that involved clinicians making prognostic judgments. Other types of judgments, such as treatment, severity prognosis, and problem recall were not interpretable due to none or too few of the included studies utilizing these specific types. The findings of the present meta-analysis are comparable with those of the Spengler et al. (2009) meta-analysis, in which more experienced clinicians were found to perform significantly better on problem type judgment tasks. In addition, the Spengler et al. meta-analysis found more experienced clinicians to be significantly better at forming appropriate treatment recommendations, accurately recalling client problems, and other judgment types.

Criterion validity for the accuracy dependent measure will have a significant impact on the experience-accuracy effect, with more experienced clinicians outperforming less experienced clinicians when the judgment tasks reflect low criterion validity. Criterion validity could be categorized as low, high, or both and

contrary to expectations, did not have a significant moderating effect on experience and judgment accuracy, $Q_B(2) = 2.29$, $p = .318$.

Provision of feedback will not have a significant impact on the experience-accuracy effect. The provision of feedback was coded dichotomously as "yes" or "no" and did not reveal a significant impact on the experience-accuracy effect, as hypothesized, $Q_B(1) = .02$, $p = .883$. However, only three studies were identified that incorporated feedback as part of the judgment task (Hickling et al., 2002; Kim & Ahn, 2002; Wood, 2004).

Studies published in APA journals will reveal greater effects for experience and judgment accuracy than those in non-APA sources. Publication sources were coded as APA, other psychology journal, psychiatric or medical journal, or dissertation. Consistent with expectations and the findings of Spengler et al. (2009), publication source had a significant impact on the experience-accuracy effect, $Q_B(3) = 16.48$, $p = .001$. Studies published in APA journals reported the largest experience-accuracy effects ($d_{i+} = .54$). This phenomenon has been well-researched (Kurosawa, 1984; Light & Pillemer, 1984; Peters et al., 2006; Rosnow & Rosenthal, 1989) and has been explained in terms of the competitiveness of major journals such as APA journals and the subsequent tendencies of studies with larger effect sizes to be accepted for publication. In addition, studies published in psychiatric or medical journals included greater experience-accuracy effects ($d_{i+} = .23$). Finally, the experience-accuracy effects found in dissertations tended to be significantly larger as well ($d_{i+} = .16$).

Ecological validity of method of study will not have a significant impact on the experience-accuracy effect. Ecological validity of method of study was categorized

as analogue, archival, or *in vivo*. As expected, there was no significant moderating effect of the relation between experience and judgment accuracy, $Q_B(1) = .00$, $p = .989$.

Ecological validity of stimulus will not have a significant impact on the experience-accuracy effect. Ecological validity of stimulus was categorized as directly experienced, indirectly experienced, or both. As expected, ecological validity of stimulus did not have a significant impact on the experience-accuracy effect, $Q_B(1) = .86$, $p = .353$.

Relation of experience to the research design will not have a significant impact on the experience-accuracy effect. Relation of experience to the research design was categorized as not in design, in primary design, in supplementary analyses, and in multiple formats. As expected, no significant moderating effect was found for this variable, $Q_B(2) = 4.70$, $p = .095$.

Experience as a major variable will not have a significant impact on the experience-accuracy effect. Experience as a major variable was coded dichotomously as "yes" or "no" and was not shown to significantly moderate the experience-accuracy effect, as expected, $Q_B(1) = .02$, $p = .882$. Most studies cited research pertaining to the impact of experience on clinical judgment in the introduction sections and/or rationales. Also, many included hypotheses regarding the relation between experience and judgment accuracy.

Study quality will not have a significant impact on the experience-accuracy effect. Study quality was categorized as acceptable, good, or excellent and was not shown to make a significant impact on the relation between experience and judgment accuracy, $Q_B(2) = .94$, $p = .625$. The majority of studies were coded as "good" and

reported adequate sampling methods, study design, and appropriate measures for establishing internal and external validity.

No specific hypothesis was formed regarding the impact of profession type on the experience-accuracy effect. Profession type was categorized as psychology, psychiatry, nursing, social work, or combination. Profession type did not have a significant moderating impact on the experience-accuracy effect, $Q_B(1) = 1.38$, $p = .239$. The vast majority of mental health clinicians practiced in the field of psychology as opposed to psychiatry, nursing, or social work, with only one study examining the judgments of psychiatrists alone. In fact, the categories of psychiatry, psychiatric nursing, and social work were not interpretable due to too few studies coded in these categories.

Inclusion of non-mental health participants will have a significant impact on the experience-accuracy effect, with larger effects found within studies including comparisons between non-mental health participants and mental health clinicians. Inclusion of non-mental health participants was coded "yes" or "no" and did not reveal a statistically significant moderating effect, contrary to expectations, $Q_B(1) = 2.37$, $p = .124$.

Publication year will have a significant impact on the experience-accuracy effect, with more recent studies showing larger experience-accuracy effects. Study age was recorded as the year of publication and was not shown to have a statistically significant impact on the experience-accuracy effect, $z = -.71$, $p = .476$.

Fail-safe analysis

As in Spengler et al. (2009), a "fail-safe" analysis was conducted to address the problem of publication bias (Rosenthal, 1991). According to Rosenthal (1979), "journals are filled with the 5% of the studies that show Type I errors, while the file drawers back at the lab are filled with the 95% of the studies that show nonsignificant (e.g., $p > .05$) results" (p. 638). Results revealed 79 studies with zero effects would be needed to decrease the overall experience-accuracy effect to a level of unreliability, for which the 95 percent confidence interval would include zero. In order to combat the problem of publication bias, every effort was made to locate a representative sample of studies via the standardized search process.

Chapter 5: Discussion

The present meta-analysis found an overall experience-accuracy effect of .16 after the removal of one outlier study, a small effect according to Cohen's (1998) guidelines. The corresponding 95 percent confidence interval was above zero (.05 to .26), indicating there was a reliable difference in judgment accuracy as a function of experience. In comparing the findings from the present meta-analysis to the Spengler et al. (2009) meta-analysis, it was apparent that the corresponding 95 percent confidence intervals overlapped significantly and that both were above zero. In the present study, homogeneity testing revealed the null hypothesis of homogeneity of effect sizes was rejected. However, an analysis of outliers revealed a study with a relatively large effect size (Rerick, 1999). This outlier was removed for the overall analyses as well as the moderator analyses. After removing this outlier the null hypothesis regarding homogeneity of the effect sizes was still rejected. In order to provide a more in-depth analysis of the heterogeneity, exploratory moderator analyses were conducted (Eagly & Wood, 1994). In addition, the moderator analyses allowed for further comparison of the present study's findings with those of Spengler et al.'s meta-analysis. Spengler et al., in contrast, explored moderator variables despite achieving homogeneity of the effect size, enabling Spengler et al. to explore previously unexamined variables.

Interpretation of the overall effect

Although the overall experience-accuracy effect in the present meta-analysis was small according to Cohen's (1998) guidelines, the interpretation, meaning, and importance are relative to several factors dependent upon the reader. Mental health clinicians, for example, may interpret the overall effect in various ways. For example,

clinicians may feel satisfied that there was a positive, reliable experience-accuracy effect found, even though the effect was modest. In light of the continued skepticism and critical stance towards the validity of clinical judgment (e.g., Lichtenberg, 2009; Lilienfeld, Lynn, & Lohr, 2003), as well as the ongoing debate over the utilization and benefits of statistical methods in comparison to unaided clinical judgment (see Ægisdóttir et al., 2006, for a meta-analytic review), it is expected that many mental health clinicians will welcome the positive finding. Westen and Weinberger (2004), for example, expressed concern over the skepticism towards clinical judgment, stating clinicians themselves have become susceptible to what they call "clinicism" (p. 601) referring to the cynical attitude towards or negative stereotype of mental health clinicians. Westen and Weinberger (2004) cited misunderstandings of the clinical-statistical literature as some of the causes for the skeptical attitude towards clinical judgment. Although the overall effect found in the present study was positive and reliable, the strength of the relation between judgment accuracy and experience did not seem to reflect the lofty claims commonly made by many mental health clinicians regarding the positive effects of experience on their professional growth. Many mental health clinicians take pride in their years of training and experience and frequently cite these as the main source of their skills and knowledge (Westen & Weinberger, 2005). According to Jennings, Hanson, Skovholt, and Grier (2005), "The emphasis on hard work (e.g., thousands and thousands of hours) and motivation versus inborn characteristics suggests that we (i.e., mental health practitioners) too can become experts if we keep at this work" (p. 28). Unfortunately, this perspective may be somewhat naïve, especially when examined within the context of clinical judgment research that highlights the tendency for more

experienced clinicians to demonstrate more *confidence*, but not necessarily more accuracy or competence (e.g., Goldberg, 1959; Holsopple & Phelan, 1954; Oskamp, 1965; Twaites, 1974). In fact, Arkes et al. (1981) emphasized overconfidence as one of the main factors that hindered clinicians' ability to learn from experience. Mental health clinicians, therefore, may be humbled by the findings of the present meta-analysis as well as those of the Spengler et al. meta-analysis. When considering the almost 40-year time span between the two meta-analyses (1970-2010), the small, stable experience-accuracy effect may be somewhat disappointing and surprising for some clinicians.

Individuals seeking mental health treatment, however, may also have varying reactions and may reach different conclusions regarding the small, positive experience-accuracy effect that was found both in the present study and in Spengler et al. (2009). In debates about the validity of clinical judgment, it is commonly cited that clients place a great deal of trust in their mental health providers and that this trust may or may not reflect the clinicians' true competence levels (e.g., Garb, 1998). As discussed in Spengler et al., individuals seeking mental health services may reevaluate the relative importance they place on obtaining 'expert' care in accordance with their immediate and long-term needs. In other words, if individuals feel there are high-stakes decisions to be made, of which the clinicians will play a major role, they may be likely to choose more experienced clinicians even though the gains in judgment accuracy associated with experience may only be modest. For instance, clients frequently require the assistance of mental health clinicians when they are involved in high-conflict custody cases (McCurley, Murphy, & Gould, 2005). In these situations, the clinicians may play the role of custody evaluator, guardian *ad litem*, parenting coordinator, or mental health therapist

to name a few. The choice in clinician, which is oftentimes determined by outside parties, can be integral for the outcome of the case. Many judges rely heavily upon the assessments of the clinicians, especially due to the clinicians' relatively greater time spent with the clients and greater familiarity with the clients' situations (Galatzer-Levy & Kraus, 1990). However, it should be noted that greater familiarity with the clients' situations and greater time spent with the clients involved does not automatically result in clinicians making more accurate judgments. In fact, child custody evaluators are among one of the most visible groups of mental health clinicians criticized for their perceived failure to adhere to scientifically-based methods and ethical codes (Tippins & Wittmann, 2000). Other high-stakes areas in which clients may place relatively greater importance in obtaining the services of an experienced clinician, regardless of the modest experience-accuracy relation, include competency evaluations, divorce mediation, and expert testimony. Whether or not clients accept the modest gain in judgment accuracy in relation to clinicians' experience will most likely be driven by the clients' perceptions of the consequences or payoff associated with the clinicians' decisions.

Interpretation of moderator variables

In terms of the overall experience-accuracy effect, few significant moderating variables were discovered in the present meta-analysis, namely judgment type and publication source. None of the other moderators were found to be significant, including experience type, experience breadth, criterion validity, provision of feedback, method of study, validity of stimulus, relation of experience to the research design, experience as a major variable, study quality, profession type, and inclusion of non-mental health participants. Considering the large number of moderator variables chosen for both

studies, it is noteworthy that so few were found to make a significant impact on the experience-accuracy effect. Present findings revealed more experienced clinicians demonstrated more accurate judgments in studies involving a combination of judgment types. Findings revealed more experienced clinicians were less accurate in their prognostic ratings; however, this effect was only based upon one study.

Regarding judgment type, results of the present meta-analysis revealed more experienced clinicians demonstrated significantly better accuracy with problem type judgments, consistent with Spengler et al. (2009). However, the Spengler et al. meta-analysis also revealed more experienced clinicians were shown to perform significantly better than less experienced clinicians on tasks for which they were required to formulate treatment recommendations, accurately recall clients' problems, and on tasks categorized as "other." The problem type category included judgments of problem(s), diagnose(s), or symptom(s). These results suggest that experienced clinicians perform better when assessing clients' problems when compared to novices. It may be that the relatively explicit graduate and post-graduate training in assessing clients' problems, as compared to other judgment tasks, give clinicians of at least a graduate-level education an advantage in problem type assessments. According to Westen and Weinberger (2005), clinical training often emphasizes teaching diagnostic and assessment skills over skills related to the prediction of future client behavior and/or outcome. They noted that even though clinicians make implicit predictions as part of their daily practice, assessments of clients' problems make up the vast majority of clinical judgment tasks. Moreover, practicing clinicians are often required to make problem type judgments, specifically diagnostic judgments, with adherence to previously specified, explicit criteria. This type

of assessment frequently calls for reliance upon structured diagnostic measures as well as the sharing of that information with third parties, for example supervisors, insurance companies, and government funding agencies. Moreover, experienced clinicians often come to understand and appreciate the very practical ramifications for lack of adherence to preselected criteria, for example difficulties in receiving third party reimbursement. According to Tetlock and Mitchell (2008), the impact of accountability on clinical judgment is often overlooked in clinical judgment research. Tetlock and Mitchell argued that even those judgment processes purported to be implicit, for example underlying prejudices and cognitive heuristics, are often made implicit in situations where clinicians will be held accountable for their decisions. In these situations, clinicians' self-awareness is heightened and they pay more deliberate attention to outward manifestations of their judgment processes and beliefs. In the case of underlying prejudices, Tetlock and Mitchell discussed the often ignored policy-related and legal ramifications that moderate the judgment processes and decisions of clinicians in actual practice settings. In regards to the present study's findings that experienced clinicians made more accurate problem type assessments, Tetlock's (2000) perspective on accountability may contribute to an appropriate explanation. It is likely that more experienced clinicians have had greater exposure to the professional ramifications and accountability underlying clinical judgment, and therefore more directly adhere to practice standards (Levant, 2005; Tetlock).

An additional moderating variable of the present study's experience-accuracy effect was publication source, which was categorized as APA journal, other psychology journal, psychiatric or medical journal, or dissertation. Present findings, consistent with

those of Spengler et al. (2009), revealed publication source as a significant moderator variable. The largest experience-accuracy effects were included in studies published in APA journals. The examination of publication source has been preferred practice in meta-analysis due to the tendency of major journals, such as APA journals, to accept studies in which relatively larger, statistically significant results have been found. Due to the competitive nature of these major journals and the tendency for studies accepted by them to reveal larger, statistically significant findings, meta-analysts must make efforts to seek out other sources for studies in order to provide the most accurate, unbiased assessment possible of the research base (Kurosawa, 1984; Light & Pillemer, 1984; Peters et al., 2006; Rosnow & Rosenthal, 1989). The present meta-analysis findings revealed studies published in psychiatry or medical journals as well as dissertations displayed significant, positive experience-accuracy effects.

The present meta-analysis allowed for the examination of two previously unexamined moderator variables, profession type and inclusion of non-mental health participants. Profession type referred to the type of mental health profession to which the participants belonged. The studies included in the present meta-analysis overwhelmingly utilized clinicians in psychology as opposed to clinicians of other mental health professions, with 25 of the original 37 studies incorporating clinicians in psychology. The second largest group included a combination of several mental health profession types, with a total of 7 studies in this category. The treatment, severity, prognosis, and problem recall categories were not interpretable due to too few studies coded in these categories. The paucity of studies involving other mental health professionals, such as psychiatrists, psychiatric nurses, and social workers may have prevented the discovery of

a larger moderating effect for profession type. Overall, profession type was not found to be a significant moderator.

The second moderator added for the present meta-analysis, inclusion of non-mental health participants, referred to whether or not studies incorporated a "zero mental health experience" condition by including comparisons of mental health participants with other groups, such as undergraduates and/or other professionals. Studies were coded as "yes" if they allowed for examinations of judgment accuracy in undergraduates and/or other professionals as compared to mental health participants. In contrast, studies were coded as "no" when all the participants were at least graduate-level clinicians in a mental health field. This moderator was included in the present study in order to address the problem of restricted range of the experience variable, as discussed in Spengler et al. (2009) and other works (e.g., Lambert & Ogles, 2004; Lambert & Wertheimer, 1988). In Spengler et al., the lowest level of experience in the experience-accuracy comparison was restricted to at least graduate-level in a mental health field. The present study, in contrast, included undergraduates and other professionals in order to capture the lower end of the experience continuum. This categorization was utilized due to the hypothesized initial gain in competency and judgment accuracy as a result of graduate-level studies (Garb, 1998). It should be noted that undergraduates and other professionals were included together in categorizing this particular moderator variable and that this introduced heterogeneity within the "zero-mental health experience" group. Clearly, some types of professionals possess specific skills and receive training in areas that would aid them in clinical judgment accuracy tasks. For example, police officers' training in investigative techniques, lie detection, and criminal profiling may far

outweigh the training and experiences possessed by undergraduates and may even rival the training and experiences of mental health clinicians in regards to certain judgment tasks, especially those that are commonly conducted by forensic psychologists (Kocsis, 2003). Hazelwood, Ressler, Depue, and Douglas (1995), in regards to the task of criminal profiling, stated "no amount of education can replace the experience of having investigated crimes" (p. 119). Particularly in the area of forensic psychology judgment, it is often difficult to determine which type of professions *should* demonstrate the best performance (Kocsis, 2003). The overall effect was not shown to differ based upon this moderator variable.

Other moderator analyses were under assessed (i.e., feedback) due to a scarcity of clinical judgment research examining these variables. The provision of feedback and its potential moderating role on the experience-accuracy effect was understudied due to the presence of only three studies examining its impact. This was comparable to the Spengler et al. (2009) meta-analysis, in which only two studies were identified as including the provision of feedback as part of the research design. Neither found feedback to be a significant moderating variable. The lack of studies examining feedback and its relation with experience and/or judgment accuracy is puzzling, especially when the lack of feedback is often provided as one of the main reasons mental health clinicians do not adequately learn from experience (e.g., Ericsson & Lehmann, 1996; Garb, 1998; Kahneman & Klein, 2009; Lichtenberg, 1997; Slobogin, 2003). As discussed previously, feedback can be beneficial in learning from experience under certain conditions, specifically when clinicians clearly understand what constitutes an incorrect response as well as when immediate, unambiguous, and consistent feedback is given upon each

incorrect response. In the mental health field, these conditions are rarely met (Dawes, 1994; Garb, 2005; Lichtenberg, 1997; Zeldow, 2009). It could be argued, however, that clinicians do receive various types of feedback in actual practice. For example, nonverbal cues (i.e., facial expression, posture) can often provide a great amount of feedback during therapy sessions. In addition, clinicians often receive verbal feedback from clients and clients' families. Although oftentimes more indirect or covert, this naturalistic feedback provides almost a constant stream of information to the clinicians. The difficulty lies in sorting through and interpreting the feedback, however. Since it has already been demonstrated that mental health clinicians are vulnerable to various cognitive heuristics and biases (e.g., Nisbett & Ross), it is likely these 'shortcuts' would also permeate the clinicians' utilization of feedback.

Another potential moderator not analyzed in the present meta-analysis included the utilization of test data for the judgment task. The lack of studies including test data as part of the stimulus measure precluded the inclusion of this potential moderator in the present study as well as in Spengler et al. (2009). In very few studies, clinicians were provided with test protocols as part of the stimulus measure. In Garb and Boyle (2003), for example, neuropsychologists were provided with various neuropsychological test protocols and asked to provide ratings of neurological impairment. In this particular study, the test protocols were based upon the average scores of community-dwelling 38- and 74-year olds. None of the clinicians were found to make errors in judgments. It appeared that the inclusion of test protocols greatly decreased the challenge level of the judgment task, so much so that Garb and Boyle were not able to complete intended supplementary analyses of differences in judgment accuracy based upon clinician

variables. Due to the longstanding debate about, and continued research support for actuarial methods, it was disappointing and surprising that so few studies were found that included test data as part of the stimulus materials for the judgment tasks. Test cutoff scores provide a relatively easy and familiar method of incorporating statistical methods (Bury & Bagby, 2002; Graham, Watts, & Timbrook, 1991; Stein et al., 1999). As described in Ægisdóttir et al. (2006), test cutoff scores are often more readily available and easier to construct than are statistical formulas. Given that some of the resistance towards statistical methods has been the complex nature of statistical formulas, it seems that clinicians would more readily assimilate test cutoff scores as a way to bridge the gap between unaided clinical judgment and pure statistical methods. Due to the high stakes nature of many of the decisions clinicians are required to make (i.e., predicting suicide, estimating violent reoffense risk), even relatively small increases in judgment accuracy are often important.

Implications of present findings

The present study's findings have serious implications for the field of psychology in general, but especially for practicing clinicians. The present meta-analysis was conducted in order to update and expand upon the Spengler et al. (2009) meta-analysis, addressing experience-accuracy research from 1997 to 2010. Despite calls for better training in and greater focus on critical thinking skills and evidence-based practice, the present findings reveal little change in the direction and/or magnitude of study findings since the time period covered by Spengler et al., 1970 to 1996. Overall, the state of clinical judgment research seems to be remarkably similar to that of the time frame assessed in the Spengler et al. meta-analysis. The same moderators Spengler et al. had

intended to assess but could not based upon lack of relevant research studies (e.g., feedback, expertness) likewise precluded assessment in the present meta-analysis. Although clinical judgment research continues to address the relation between judgment accuracy and experience, the methodology utilized and the moderators assessed in the studies have remained relatively constant. Overall, the present study revealed remarkably similar findings as the Spengler et al. meta-analysis, in terms of the small, positive effect as well as the few impacting moderator variables discovered. The present study sought to cross-validate and test the robustness of the Spengler et al. meta-analysis, and it seems the present findings largely replicated those of the former.

The present study findings have important implications for training in psychology. Without a significantly large increase in judgment accuracy with greater experience (including educational), it will be difficult for training programs to justify the amount of time and resources utilized. As was found by Harding (2007), it is quite possible that training programs are lacking in the explicit teaching of critical thinking and clinical judgment skills and that this deficit in training could hamper clinicians' ability to learn from their experiences. One way in which training programs may foster the development of critical thinking and sound clinical judgment skills is through explicit instruction of cognitive biases and heuristics (Kahneman, 2003). The clinical judgment research continues to emphasize the role cognitive biases and heuristics play in clinical decision-making; however, further steps are needed to translate those results to inform actual practice and training programs. For example, the newly defined affect heuristic (Slovic et al., 2002) addresses clinicians' tendencies to utilize emotions and feeling states to guide judgment. This type of heuristic has relevance for many types of clinical

decisions. Awareness and examination of the affect heuristic has the potential to aid future clinicians in maintaining a balanced, critical perspective.

Finally the present study findings provide further support for clinicians learning to work in tandem with technological and statistical aids, for example assisting clinicians in making suicide predictions based on statistical calculations (Gustafson, Greist, Stauss, Erdman, & Laughren, 1977). Although the present study revealed an overall positive effect for experience and judgment accuracy, it is a smaller effect than what most would probably suspect or hope for, especially those who believe their years of experience have resulted in a significant increase in their abilities to make accurate decisions. Mainstream journals, such as the *Journal of Clinical and Consulting Psychology and Psychotherapy: Theory, Research, Practice, Training*, have printed special issues outlining specific technological developments in the field of psychology and how they are transforming clinical practice (Caspar, 2004). One area in which statistical models are gaining popularity and utility is in the prediction of violence (Ægisdóttir et al., 2006). It is likely that with the continued development of technological and statistical methods of clinical judgment, the positive effects of experience will be even more overshadowed (Grove et al., 2000; Monahan, Steadman, Silver, Appelbaum, & Robbins, 2001). However, if clinicians can learn to work effectively with technological and statistical aids, clinical judgment may be greatly enhanced.

Limitations of the present meta-analysis

The results of the present meta-analysis should not be interpreted without an examination of the present study's limitations. First, the study search was initially conducted via review of titles and abstracts in comparison to the Spengler et al. (2009)

meta-analysis, in which many more full articles and/or dissertations were sought early in the search process. When reviewing the titles and abstracts for the present meta-analysis, studies were screened first for inclusion of some type of clinical judgment task. It is possible, therefore, that some studies allowing for experience-accuracy analyses may have been excluded. If the studies included experience-accuracy examinations in supplemental or post-hoc analyses, it is logical that these studies may be prematurely excluded based upon the titles and abstracts not alluding to the experience-accuracy analyses. In addition, forward and/or backward referencing was only conducted with the 75 studies included in the Spengler et al. meta-analysis as well as the final 37 studies included in the present meta-analysis. If forward and backward cross-referencing had occurred throughout the study search process, as in the Spengler et al. meta-analysis, it is possible that more relevant studies would be found. However, it should be noted that the majority of the relevant studies found as a result of the forward and backward cross-referencing had also been located initially through the database search process using the pre-selected search terms. In order to account for possible missing studies as a result of the relatively more limited search process, a random effects model was chosen for the present study in comparison to the fixed effects model utilized in the Spengler et al. meta-analysis. The random effects model assumed between- as well as within-study variability. In the Spengler et al. meta-analysis, the fixed effects model assumed variability only within each included study due to the assumption that there is one underlying effect. Since it was reported in the Spengler et al. meta-analysis that every relevant study was included, the fixed effects model allowed for conditional inferences to be made upon the pool of selected studies.

Another limitation of the present meta-analysis related to the limited variability in the problem type moderator variable. As noted previously in the discussion of important moderator findings, experienced clinicians were found to make significantly more accurate problem type judgments. Problem type studies, such as Parmley (2006), often utilized case vignettes as the stimulus measure for the judgment task. In Parmley, for example, clinicians were asked to read and provide diagnoses for two case vignettes describing clients with either a psychotic disorder or anxiety disorder. Out of 37 originally included studies in the present meta-analysis, 22 utilized case vignettes. Case vignettes have been frequently employed in the clinical judgment research (e.g., Finlayson & Koocher, 1991; Kalichman & Craig, 1991; Warner-Rogers, Hansen, & Spieth, 1996; Zellman, 1990). One benefit of utilizing the case vignette method is that it often allows the researcher to have greater control over the systematic manipulation of variables. For example, vignettes may be constructed with strict adherence to DSM-IV-TR criteria. In addition, client variables, such as age, race, and problem type can easily be manipulated across the study's conditions. Although case vignettes potentially offer convenience and control for the researcher as well as a variety of other advantages, they are never able to capture or replicate real-life judgment tasks. In actual clinical practice, clinicians are faced with a variety of confounding factors that may hinder their ability to make accurate client appraisals. These *potential* obstacles may include clients' tendencies to portray themselves in an overly positive or negative manner, family members' subjective input regarding the clients' problems, and third party demands for clinical performance (i.e., managed care requirements, company policies).

However well-constructed, case vignettes often neglect to account for these additional factors, transforming what is typically a 'fuzzy' decision-making task into a more manageable and simplified one (Hansen et al., 1997). In the present meta-analysis, the abundance of case vignette studies resulted in an oversimplification of judgment tasks. This restriction most likely prevented an examination of more difficult judgment tasks, such as those in which clinicians are asked to make judgments *in vivo*, including all naturally occurring confounds. One study example of *in vivo* methods is Hannan et al. (2005), in which 48 therapists at a university outpatient clinic were asked to predict their clients' progress and outcomes during a 3-week period, specifically whether or not they would drop out of treatment prematurely. Hannan et al. found that only one clinician predicted accurately and that results did not differ based upon clinicians' knowledge of client deterioration base rates. As in Spengler et al. (2009), studies that incorporated *in vivo* methods of study were relatively rare. It is possible, therefore, that the inclusion of so many case vignette studies where the judgment tasks were relatively simpler and more contrived inflated the overall experience-accuracy effect. Alternatively, it is somewhat surprising that a larger experience-accuracy effect was not found due to the abundance of rather simplistic, straightforward judgment tasks.

Overall, many of the moderator analyses had low power to detect potential moderating effects. For example, in judgment type, four judgment types were excluded from the analysis due to having fewer than three studies in each category. It seems the relative uniformity of judgment accuracy research, specifically the preponderance of analogue research utilizing case vignettes, prevented a thorough analysis of moderator variables. It was extremely difficult to accurately gauge the impact of potential

moderator variables when so much of the research assessed the experience-accuracy relation in similar ways, oftentimes utilizing contrived situations in which participants are given a restricted amount of information. The tendency of judgment accuracy studies to assess the experience-accuracy relation in similar ways, therefore, resulted in the exclusion of several moderator categories and even in the complete prevention of accurate assessments of particular moderator variables, for example the provision of feedback variable. The problem of low power is common in meta-analyses, especially when assessing moderator variables across several categories (Lipsey & Wilson, 2003).

Suggestions for future research

Future studies involving the experience-accuracy relation should attempt to create more precise guidelines for the operationalization of the experience variable. The definitions of experience varied greatly from one study to the next. In future studies, maintaining experience as a continuous variable will allow for more precise analyses of the experience-accuracy effect. In some of the included studies, subjective categories were created based upon certain loosely defined variables related to experience (e.g., Witteman & van den Bercken, 2007). For example, years of clinical experience was sometimes transformed into a categorical measure of experience and analyzed in terms of two levels (i.e., novice versus expert). Although the creation of the two levels often relied upon straightforward cutoff criteria based on years of experience, researchers sometimes seemed to confound other variables of experience within the categories, such as degree type and level of training. In future studies, researchers can address this challenge by creating more clear-cut guidelines for the experience variable.

In addition, future researchers addressing the experience-accuracy effect may wish to employ participants from various subfields of mental health as well as non-mental health participants. Meta-analysts in the future will therefore be able to achieve a more comprehensive and accurate perspective on the moderating effects of profession type and the inclusion of non-mental health participants. In order for this to occur, however, it will be necessary for researchers to report data for each subsample (e.g., social workers, undergraduates) instead of reporting data for the overall sample.

Finally, future researchers of the relation between experience and accuracy may wish to expand their study methods to incorporate those with varying degrees of ecological validity. In the present meta-analysis, 22 out of 37 studies utilized clinical case vignettes to present judgment tasks to participants. These vignettes were oftentimes constructed based upon rigid adherence to DSM-IV-TR criteria, thus decreasing the likelihood that participants would make incorrect responses or inaccurate judgments (e.g., Brammer, 2002). Other studies employed relatively contrived situations in order to provide a standard for accuracy, which led to a decrease in the generalizability of the results (e.g., Garb, 2006; Ruscio & Stern, 2005). According to Ruscio and Stern (2005), for example, "The judgment task in this research was the simplest holistic task that we could conceive, and the failure of participants to perform well under these conditions casts serious doubt on the efficacy of holistic judgment more broadly" (p. 62). In real-world practice, however, clinicians are often called upon to make "fuzzy" decisions for which the "correct" answer is difficult, if not impossible, to determine (Dawson et al., 1989; Lichtenberg, 2009). Future clinical judgment researchers may wish to explore a

variety of methods for examining judgment accuracy rather than relying on case vignettes for their clarity and standardized nature.

Future clinical judgment research should continue to *critically* examine the presence and role of feedback in learning from experience. While it has often been stated that mental health clinicians are not able to learn from experience in part due to the lack of unambiguous feedback (e.g., Westen & Weinberger, 2005), it could also be said that mental health clinicians *should* be trained in sorting through and interpreting this specific type of ambiguous feedback. Given the difficulties clinicians have had in establishing a direct, consistent relation between experience and improvements in accuracy and competence, it may be time to explore other methods of training as well as explore additional moderating and mediating variables.

Conclusion

In conclusion, the present meta-analysis found a small, but positive, reliable effect between experience and judgment accuracy, with judgment type and publication source as moderating factors. The results of this meta-analysis are comparable to those found in Spengler et al. (2009) in which an examination of studies from 1970 to 1996 revealed a small, positive experience-accuracy effect. Similar to the Spengler et al. meta-analysis, the results of the present meta-analysis revealed little impact of theoretically important variables (e.g., experience breadth, experience type) on the experience-accuracy effect. Given the limitations of the present study and the relatively smaller sample of studies, the fact that the findings of the present meta-analysis largely reflected the findings of the Spengler et al. meta-analysis is significant and noteworthy. Although the meta-analyses spanned two separate time periods and included a distinct set of selected studies, it seems

the experience-accuracy effect remains a small, positive effect. In addition, it is worth mentioning that the various moderator variables assessed only revealed a few significant effects in each meta-analysis, indicating the overall effect is relatively stable. Although the present meta-analysis and the Spengler et al. meta-analysis utilized two different models (i.e., random, fixed), the overall effect sizes in each study were virtually identical. The results of the present meta-analysis are humbling given the popular belief and desire amongst professionals for a large, clear-cut payoff for training and experience. These results provide support for critical examination of training programs in search of more effective methods of teaching sound clinical judgment and critical thinking skills as well as an empirically exploration of additional moderating and mediating variables. In summation, given the findings of the present meta-analysis as well as those of Spengler et al., there needs to be a field-wide acknowledgement of the difficulties more experienced clinicians face in making accurate judgments as well as an emphasis on investigating the particular circumstances in which clinical judgment accuracy *is* shown to improve with experience.

References

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

Abramowitz, C., & Dokecki, R. (1977). The politics of clinical judgment: Early empirical returns. *Psychological Bulletin*, *84*, 460-476.

Adams, E. V., & Betz, N. E. (1993). Gender differences in counselors' attitudes toward and attributions about incest. *Journal of Counseling Psychology*, *40(2)*, 210-216.

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Freels, G. R. & Rush, J. D. (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.

Agell, G. & Rothblum, E. D. (1991). Effects of clients' obesity and gender on the therapy judgments of psychologists. *Professional Psychology: Research and Practice*, *22(3)*, 223-229.

*Akehurst, L., Bull, R., Vrij, A., Köhnken, G. (2004). The effects of training professional groups and lay persons to use criteria-based content analysis to detect deception. *Applied Cognitive Psychology*, 18, 877-891.

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

American Psychological Association Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist*, *61*, 271-285.

American Psychological Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC.

American Psychological Association (1982). Report of the task force on the evaluation of education, training, and service in psychology. *American Psychological Association*, Washington, D.C.

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

Arkes, H. R., Faust, D., Guilmette, T. J., & Hart, K. (1988). Eliminating the hindsight bias. *Journal of Applied Psychology*, *73*, 305-307.

Arkes, H. R., Wortmann, R. L., Saville, P. D., & Harkness, A. R. (1981). Hindsight bias among physicians weighing the likelihood of diagnoses. *Journal of Applied Psychology*, *66*, 252-254.

Ash, I. K. (2009). Surprise, memory, and retrospective judgment making: Testing cognitive reconstruction theories of the hindsight bias effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2009, *35(4)*, 916-933.

Ashcraft, M. (2002). Math anxiety: Personal, educational, and cognitive consequences. *Current Directions in Psychological Science*, *11(5),* 181-185.

Bagby, R. M., & Bury, A. S. (2002). The detection of feigned uncoached and coached posttraumatic stress disorder with the MMPI-2 in a sample of workplace accident victims. *Psychological Assessment*, *14(4)*, 472-84.

Belknap, D. D. (2000). Tarasoff and the hindsight bias: After-the-fact evaluations of the reasonableness of therapists' judgments. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, *61(3-B)*, 1693.

Belleville, G., Cousineau, H., Levrier, K., St.-Pierre-Delorme, M. E., & Marchand, A. (2010). The impact of cognitive-behavior therapy for anxiety disorders on concomitant sleep disturbances: A meta-analysis. *Journal of Anxiety Disorders*, *24(4)*, 379-86.

Bereiter, C. & Scardamalia, M. (1986). Educational relevance of the study of expertise. *Interchange*, *17(2)*, 10-19.

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.

Bernstein, B. L., & LeComte, C. (1982). Therapist expectancies: Client gender, and therapist gender, profession, and level of training. *Journal of Clinical Psychology*, *38(4)*, 744-754.

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

Biaggio, M., Roades, L. A., Staffelbach, D., Cardinali, J., & Duffy, R. (2000). Clinical evaluations: Impact of sexual orientation, gender, and gender role. *Journal of Applied Social Psychology*, *30(8)*, 1657-1669.

Billingsley, D. (1977). Sex bias in psychotherapy: An examination of the effects of client sex, pathology, and therapist sex on treatment planning. *Journal of Consulting and Clinical Psychology*, *45(2)*, 250-256.

Blank, H., Fischer, V., & Erdfelder, E. (2003). Hindsight bias in political elections. *Memory*, *11*, 491-504.

Blank, H., & Nestler, S. (2007). Cognitive process models of hindsight bias. *Social Cognition*, *25*, 132-146.

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

*Boland, K. (2002). Ethical decision-making among hospital social workers. Unpublished doctoral dissertation, Marywood University, Bernville.

Bonds-Raacke, J. M., Fryer, L. S., Nicks, S., & Durr, R. T. (2001). Hindsight bias demonstrated in the prediction of a sporting event. *The Journal of Social Psychology*, *141(3)*, 349-352.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2005). Comprehensive Meta-Analysis (Version 2) [Computer software]. Englewood, NJ: Biostat.

Borreson, A. M. (1965). Counselor influence on diagnostic classification of client problems. *Journal of Counseling Psychology*, 12, 252-258.

Bowers, A. M. V. & Bieschke, K. J. (2005). Psychologists' clinical evaluations and attitudes: An examination of the influence of gender and sexual orientation. *Professional Psychology: Research and Practice*, *36(1)*, 97-103.

*Brammer, R. (2002). Effects of experience and training on diagnostic accuracy. *Psychological Assessment*, *14(1)*, 110-113.

*Butte, D. G. (1998). *Accuracy of Nurses Assessing Patients Using American Psychiatric Association Criteria.* Unpublished doctoral dissertation, University of Utah, Salt Lake City.

Caldwell, T. (1998). *Beliefs about interventions and outcomes for schizophrenia and depression: Mental health nurses within the professional and public context*. Unpublished thesis, Australian National University, Canberra.

Carlson, B. W. (1990). Anchoring and adjustment in judgments under risk. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16(4)*, 665-676.

Caspar, F. (2004). Technological developments and applications in clinical psychology and psychotherapy: Introduction. *Journal of Clinical Psychology*, *60*, 221-238.

Chan, D. (2006). Interactive effects of situational judgment effectiveness and proactive personality on work perceptions and work outcomes. *Journal of Applied Psychology*, *91*, 475-481.

Chandler, M. J. (1970). Self-awareness and its relation to other parameters of the clinical inference process. *Journal of Consulting and Clinical Psychology*, *35(2)*, 258-264.

Chi, M., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5(2)*, 121-151.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. Sternberg (Ed.), *Advances in the Psychology of Human Intelligence* (pp. 1-75). Hillsdale, NJ: Erlbaum.

Cioffi, J. (2001). A study of the use of past experiences in clinical decision making in emergency situations. *International Journal of Nursing Studies*, *38*(5), 591-599.

Clavelle, P., & Turner, A. (1980). Clinical decision-making among professional and paraprofessionals. *Journal of Clinical Psychology*, *36*, 833-838.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20(1)*, 37-46.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis for field settings. Boston, MA: Houghton-Mifflin.

Cook, T. D., & Leviton, L. C. (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. *Journal of Personality*, *48(4)*, 449-472.

Coutinho, M. J., Oswald, D. P., Best, A. M., & Forness, Steven, R. (2002). Gender and sociodemographic factors and the disproportionate identification of culturally and linguistically diverse students with emotional disturbance. *Behavioral Disorders*, *27(2)*, 109-125.

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.

Dailey, R. C. (1980). Relationship between locus on control, task characteristics and work attitudes. *Psychological Reports*, *47(3 Pt 1)*, 855-861.

Dailey, C. (1952). The effects of premature conclusions upon the acquisition of understanding of a person. *Journal of Psychology*, *33*, 133-152.

Davies, M. (2003). Confirmatory bias in the evaluation of personality descriptions: Positive test strategies and output inference. *Journal of Personality and Social Psychology*, *85(4)*, 736-744.

Davis-Coelho, K., Waltz, J., & Davis-Coelho, B. (2000). Awareness and prevention of bias against fat clients in psychotherapy. *Professional Psychology: Research and Practice*, *31(6)*, 682-684.

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

Day, D. V., & Lord, R. G. (1992). Expertise and problem categorization: The role of expert processing in organizational sense-making. *Journal of Management Studies*, *29*, 35-47.

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

Dumont, F. (1993). Inferential heuristics in clinical problem formulation: Selective review of their strengths and weaknesses. *Professional Psychology: Research and Practice*, *24(2)*, 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: Sage.

*Ebling, R., & Levenson, R. (2003). Who are the marital experts? *Journal of Marriage and Family*, 65, 130-142.

Eells, T., Lombart, K. G., Kendjelic, E. M., Turner, L. C., & Lucas, C. P. (2005). The quality of psychotherapy case formulations: A comparison of expert, experienced, and novice cognitive-behavioral and psychodynamic therapies. *Journal of Consulting and Clinical Psychology*, *73(4)*, 579-589.

Einhorn, H., & Hogarth, R. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, *85*, 395-416.

*Ekman, P., O'Sullivan, M., & Frank, M.G. (1999). A few can catch a liar. *Psychological Science*, *10(3)*, 263-266.

Elbogen, E. B., Williams, A. L., Kim, D., Tomkins, A. J., Scalora M. J. (2001). Gender and perceptions of dangerousness in civil psychiatric patients. *Legal and Criminology Psychology. 6*: 215–28.

Elman, N. S., Illfelder-Kaye, J., & Robiner, W. N. (2005). Professional development: Training for professionalism as a foundation for competent practice in psychology. *Professional Psychology: Research and Practice*, *36(4)*, 367-375.

Elovitz, G. P., & Salvia, J. (1982). Attractiveness as a biasing factor in the judgments of school psychologists. *Journal of School Psychology*, *20(4)*, 339-345.

Elstein, A., Shulman, L., & Sprafka, S. (1978). *Medical problem solving: An analysis of clinical reasoning*. Cambridge, MA: Harvard University Press.

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.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*, *49*, 709–724.

Ericsson, K. A. (2005). Recent advances in expertise research: A commentary on the contributions to the special issue. *Applied Cognitive Psychology*, *19(2)*, 223-241.

Ericsson, K. A., Charness, N. (1994). Expert-performance-Its structure and acquisition. *American Psychologist*, *49*, 725-747.

Ericsson, K A, & Lehmann, A C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology*, *47*, 273 -306.

Eubanks-Carter, C., & Goldfried, M. R. (2006). The impact of client sexual orientation and gender on clinical judgments and diagnosis of borderline personality disorder. *Journal of Clinical Psychology*, *62*, 751-770.

Eysenck, H. J. (1994). Systematic Review: Metaanalysis and its problems. *British Medical Journal*, *309*, 789-792.

Eysenck, H. J. (1952). The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology*, *16*, 319-324.

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., Bray, T. E., & Hebert, D. J. (2005). Case conceptualization and treatment planning: Investigation of problem-solving and clinical judgment. *Journal of Mental Health Counseling*, *27(4)*, 348-372.

Faust, D. (2006). Decision research can increase the accuracy of clinical judgment and thereby improve patient care. In S. Lilienfeld & W. O'Donohue (Eds.). The great ideas of clinical science: 17 principles that every mental health professional should understand (pp. 29-48). New York, NY: Routledge.

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, MA: Allyn & Bacon.

Faust, D. (1986). Research on human judgment and its application to clinical practice. *Professional Psychology: Research and Practice*, *17(5)*, 420-430.

Faust, D. (1984). *Limits of scientific reasoning*. Minneapolis, MN: University of Minnesota Press.

Faust, D., Guilmette, T. J., Hart, K. J., 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.

Fernbach, B. E., Winstead, B. A., & Derlega, V. J. (1989). Sex differences in diagnosis and treatment recommendations for antisocial personality and somatization disorders. *Journal of Social and Clinical Psychology*, 238-255.

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.

Fischer, J., Dulaney, D. D., Fazio, R. T., Hudak, M. X., & Zivotofsky, E. (1976). Are social workers sexist? A replication. *Social*, *25*, 46-50.

Fischhoff, B. (1975). Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1*, 288-299.

Foon, A. E. (1989). Mediators of clinical judgment: An exploration of the effect of therapists' locus of control on clinical expectations. *Genetic, Social, & General Psychology Monographs*, *115(2)*, 243-266.

Ford, C. V., & Sbordone, R. J. (1980). Attitudes of psychiatrists toward elderly patients. *American Journal of Psychiatry*, *137(5)*, 571–575.

Frensch, R. A., & Sternberg, R. J. (1989). Expertise and intelligent thinking: When is it worse to know better? In Sternberg, R. J. (Ed.), *Advance in the psychology of human intelligence* (pp. 157-188). Hillsdale, NJ: Erlbaum.

Friedlander, M., & Phillips, S. (1984). Preventing anchoring errors in clinical judgment. *Journal of Consulting and Clinical Psychology*, *52*, 366-371.

Friedlander, M., & Stockman, S. (1983). Anchoring and publicity effects in clinical judgment. *Journal of Clinical Psychology*, *39*, 637-643.

Gadol, I. (1969). The incremental and predictive validity of the Rorschach test in personality assessments of normal, neurotic and psychotic subjects. *Dissertation Abstracts International*, *29(9-B)*, 3482-3483.

Galatzer-Levy, R., Kraus, L., & Galatzer-Levy, J. (eds.) (2009, in press) *The Scientific Basis of Child Custody Decisions.* Second Edition. John Wiley & Sons.

Gambrill, E. (2005). *Critical thinking in clinical practice: Improving the quality of judgments and decisions*. Hoboken, NJ: John Wiley & Sons, Inc.

*Garb, H. N. (2006). The conjunction effect and clinical judgment. *Journal of Social and Clinical Psychology*, *25(9)*, 1048-1056.

Garb, H. N. (2005). Clinical judgment and decision making. *Annual Review of Clinical Psychology*, *1(1)*, 67-89.

Garb, H. N. (1999). Call for a moratorium on the use of the Rorschach Inklot Test in clinical and forensic settings. *Assessment*, *6(4)*, 313-317.

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

Garb, H. N. (1996). The representativeness and past-behavior heuristics in clinical judgment. *Professional Psychology: Research and Practice*, *27(3)*, 272-277.

*Garb, H. N., & Boyle, P. A. (2003). The diagnosis of neurological disorders in older adults. *Assessment*, *10(2)*, 129-134.

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

Garb, H. N., & Grove, W. M. (2005). On the merits of clinical judgment: Comment. *American Psychologist*, *60(6)*, 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.

Garner, A. M. & Smith, G. M. (1976). An experimental videotape technique for evaluating trainee approaches to clinical judging. *Journal of Consulting and Clinical Psychology*, *44(6)*, 945-950.

Gaudette, M. D. (1992). Clinical decision-making in neuropsychology: Bootstrapping the neuropsychologist utilizing Brunswik's lens model. *Dissertation Abstracts International*, *53(4-B)*, 2059.

Gauron, E., & Dickinson, J. (1966). Diagnostic decision-making in psychiatry. 2. Diagnostic styles. *Archives of General Psychiatry*, *14*, 233-237.

*Gerbe, N. (2007). *School psychologists' knowledge of Asperger's Disorder, its differential diagnosis, and treatment recommendations*. Unpublished doctoral dissertation, St. John's University, New York.

Gilovich, T., Griffin, D. W., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. New York, NJ: Cambridge University Press.

Glaser, R., & Chi, M. T. H. (1988). Overview. In Chi, M.T. H., Glaser, R., & Farr, M. J. (Eds.), *The nature of expertise* (pp. xv-xxvii). Hillsdale, NY: Erlbaum.

Glass, G.V (1978). Integrating findings: The meta-analysis of research. *Review of Research in Education*, *5*, 351-379.

Glass, G. V (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*, 3-8.

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

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

Goldstein, E. G. (2007) 'Social work education and clinical learning: yesterday, today, and tomorrow', *Clinical Social Work Journal*, *35*, 15–23.

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.

Gonsiorek, J.C. (1982). The use of diagnostic concepts in working with gay and lesbian populations. *Journal of Homosexuality*, *7*, 9–20.

Graham, J. R. (1967). A Q-sort study on the accuracy of clinical descriptions based on the MMPI. *Journal of Psychiatric Research*, *5(4)*, 297-305.

Grigg, A. E. (1958). Experience of clinicians, and speech characteristics and statements of clients as variables in clinical judgment. *Journal of Consulting Psychology*, *22(4)*, 315-319.

Groth-Marnat, G. (2000). Visions of clinical assessment: Then, now, and a brief history of the future. *Journal of Clinical Psychology*, *56*, 349-365.

Grove, W. M. (2001). Recommendations of the Division 12 task force: "Assessment for the century: A model curriculum." *Clinical Science*, 8.

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.

Grubbs, F. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11, 1-21.

Guilbault, R. L., Bryant, F. B., Brockway, J. H., & Posavac, E. J. (2004). A meta-analysis of research on hindsight bias. *Basic and Applied Social Psychology, 26,*103–117.

Gustafson, D., Greist, J., Stauss, F., Erdman, H., & Laughren, T. (1977). A probabilistic system for identifying suicide attemptors. *Computers And Biomedical Research*, *An International Journal*, *10(2)*, 83-89

Haerem, T. & Rau, D. (2007). The influence of degree of expertise and objective task complexity on perceived task complexity and performance. *Journal of Applied Psychology*, *92(5)*, 1320-1331.

Hall, J. A., Coats, E. J., & Smith LeBeau, L. (2005). Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological Bulletin*, *131(6)*, 898-924.

*Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., & Sutton, S. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology*, *61(2)*, 155-163.

*Hansen, D. J., Bumby, K. M., Lundquist, L. M., Chandler, R. M., Le, P. T., & Futa, K. T. (1997) The influence of case and professional variables on the identification and reporting of child maltreatment: A case study of licensed psychologists and certified masters social workers. *Journal of Family Violence*, *12(3)*, 313-332.

Hansen, F. J. & Reekie, L. (1990). Sex differences in clinical judgments of male and female therapists. *Sex Roles*, *23(1-2)*, 51-64.

Hanson, R. K., & Morton-Bourgon, K., Canada. (2004). *Predictors of sexual recidivism: An updated meta-analysis*. Ottawa, Ont.: Public Works and Government Services Canada.

Harding, T. P. (2007). Clinical decision-making: How prepared are we? *Training and Education in Professional Psychology*, *1(2)*, 95-104.

Hardy, D. M., & Johnson, M. E. (1992). Influence of therapist gender and client gender, socioeconomic status and alcoholic status on clinical judgments. *Journal of Alcohol and Drug Education*, *37(2)*, 94-102.

Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin*, *197*, 311-327.

Hayes, J. R. (1985). Three problems in teaching general skills. In S. F. Chipman & J. W. Segal, & R. Glaser (Eds.), *Thinking and learning skills: Research and open questions*. Hillsdale, NJ: Erlbaum.

Hazelwood, R. R., Ressler, R. K., Depue, R. L., & Douglas, J. E. (1995). Criminal investigative analysis:An overview. In R. R. Hazelwood & A. W. Burgess (Eds.), *Practical aspects of rape investigation:A multidisciplinary approach*(2nd ed., pp. 115-126). Boca Raton, FL:CRC Press.

Heaton, R. K., Smith, H. H., Lehman, R. A., & Vogt, A. T. (1978). Prospects for faking believable deficits on neuropsychological testing. *Journal of Consulting and Clinical Psychology*, *46(5)*, 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.

Hedges, L. V., Vevea J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3(4),* 486-504.

Herbert, D., Nelson, R., & Herbert, J. (1988). Effects of psychodiagnostic labels, depression severity, and instructions on assessment. *Professional Psychology: Research and Practice*, 19(5), 496-502.

*Hickling, E. J., Blanchard, E. B., Mundy, E., & Galovski, T. (2002). Detection of malingered MVA related posttraumatic stress disorder: An investigation of the ability to detect professional actors by experienced clinicians, psychological tests and psychophysiological assessment. *Journal of Forensic Psychology Practice*, *2(1)*, 33-53.

Highlen, P. S. & Hill, C. E. (1984). Factors affecting client change in individual counseling: Current status and theoretical speculations. In S. D. Brown & R. W. Lent (Eds.), *Handbook of counseling psychology* (pp. 334-396). New York, NJ: Wiley-Interscience.

Hill, C. E., Tanney, M. F., & Leonard, M. M. (1977). Counselor reactions to female clients: Type of problem, age of client, and sex of counselor. *Journal of Counseling Psychology*, *24*, 60-65.

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

*Hillman, J. L., Stricker, G., & Zweig, R. A. (1997). Clinical psychologists' judgments of older patients with character pathology: Implications for practice. *Professional Psychology: Research and Practice*, *28(2)*, 179–183.

Hilton, N. Z., Harris, G. T., Rice, M. E. (2006). Sixty-Six Years of Research on the Clinical Versus Actuarial Prediction of Violence . *The Counseling Psychologist, 34(3),* 400-409.

Hinsley, D.A., Hayes, J.R., & Simon, H.A. (1978). From words to equations: Meaning and representation in algebra word problems. In P.A. Carpenter & M.A. Just (Eds.) Cognitive processes in comprehension (pp. 89-106). Hillsdale, NJ: Erlbaum.

Hogarth, R. M., & Karelaia, N. (2007). Heuristic and linear models of judgment: Matching rules and environments. *Psychological Review*, *114(3)*, 733-758.

Hollon, S. D., & Kriss, M. R. (1984). Cognitive factors in clinical research and practice. *Clinical Psychology Review*, *4*, 35-76.

Holsopple, J. Q., Phelan, J. (1954) The skills of clinicians in analysis of projective tests. *Journal of Clinical Psychology*, *10*, 307-320.

Hoptman, M. J., Yates, K. F., Patalinjug, M. B., Wack, R. C., & Convit, A. (1999). Clinical prediction of assaultive behavior among male psychiatric patients at a maximum-security forensic facility. *Psychiatric Services*, *50*, 1461-1466.

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.

Hsieh, D. K., & Kirk, S. A. (2003). The effect of social context on psychiatrists' judgments of adolescent antisocial behavior. *Journal of Child Psychology and Psychiatry*, *44(6)*, 877-887.

Huedo-Medina, T., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). *Assessing heterogeneity in meta-analysis: Q statistic or I2 index?* Retrieved July 5, 2010, from http://digitalcommons.uconn.edu/chip docs/19.

*Huffaker, S. (2008). Factors predicting case formulation proficiency in traumatic brain injury: Experience and knowledge. Unpublished doctoral dissertation, Alliant International University, San Diego.

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

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment*, *8(4)*, 275-292.

Imig, C., Krauskopf, C. J., & William, J. L. (1967). Clinical prediction and immediate feedback training. *Journal of Counseling Psychology*, *14(2)*, 180-186.

*Jackson, R. L., Rogers, R., & Shuman, D. W. (2004). The adequacy and accuracy of sexually violent predator evaluations: Contextualized risk assessment in clinical practice. *International Journal of Forensic Mental Health*, 3(2), 115-129.

James, J. W., & Haley, W. E. (1995). Age and health bias in practicing clinical psychologists. *Psychology and Aging*, *10(4)*, 601–616.

Jennings, L., Hanson, M., Skovholt, T. M., & Grier, T. (2005). Searching for mastery. *Journal of Mental Health Counseling*, *27(1)*, 19-31.

Jorm, A. F., Korten, A. E., Rodgers, B., & Pollitt, P. (1997). Beliefs about the helpfulness of interventions for mental disorders: A comparison of general practitioners, psychiatrists, and clinical psychologists. *Australian New Zealand Journal of Psychiatry*, *31*, 844-851.

*Jopp, D. (2001). An examination of the diagnostic overshadowing bias. Unpublished doctoral dissertation, University of Illinois, Chicago.

Kalichman, S. C., & Craig, M. E. (1991). Professional psychologists' decisions to report suspected child abuse: Clinician and situation factors. *Professional Psychology: Research and practice*, *22*, 84-89.

Kahneman, D. (2003). *Maps of bounded rationality: Psychology for behavioral economics*. American Economic Review, 93(2), 162-168.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64(6)*, 515-526.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237-251.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, London: Cambridge University Press.

*Kellner, L. J. (2001). *Are school psychologists knowledgeable about adolescent suicide?* Unpublished doctoral dissertation, Fairleigh Dickinson University, Madison.

Keppel, G., Saufley, W. H. Jr., & Tokunaga, H. (1992). *Introduction to design and analysis: A student's handbook*. (2nd ed.). New York, NY: W. H. Freeman and Company.

*Kim, N. S. & Ahn, W. (2002). Clinical psychologists' theory-based representations of mental disorders predict their diagnostic reasoning and memory. *Journal of Experimental Psychology: General*, *131(4)*, 451-476.

Kirk, S. A. & Hsieh, D. K. (2004). Diagnostic consistency in assessing conduct disorder: An experiment on the effect of social context. *American journal of Orthopsychiatry*, *74(1)*, 43-55.

*Kitamura, T., Kitamura, F., Ito, A., Okazaki, Y., Okuda, N., Mitsuhashi, T., & Katoh, H. (1999). Image of psychiatric patients' competency to give informed consent to treatment in Japan II. A case vignette study of competency judgements. *International Journal of Law and Psychiatry*, *22(2)*, 133-142.

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

*Kocsis, R. N. (2003). Criminal Psychological Profiling: Validities and Abilities. *International Journal of Offender Therapy and Comparative Criminology. 47(2),* 126-144.

Kurosawa, K. (1984). Meta-analysis and selective publication bias. *American Psychologist*, *39(1)*, 73-74.

Kunda, Z., Fong, G. T., Sanitioso, R., & Reber, E. (1993). Directional questions direct self-conceptions. *Journal of Experimental Social Psychology, 29,* 63– 86.

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., & Ogles, B. M. (2004). The efficacy and effectiveness of psychotherapy (pp. 139 - 193). In Michael J. Lambert (Ed.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (5th ed.). NY: John Wiley & Sons, Inc.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Langer, E., & Abelson, R. (1974). A patient by any other name...: Clinical group differences in labeling bias. *Journal of Consulting and Clinical Psychology*, *42*, 4-9.

*Leon, J. A. & Perez, O. (2001). The influence of prior knowledge on the time course of clinical diagnosis inferences: A comparison of experts and novices. *Discourse Processes*, *31(2)*, 187-213.

Lesgold, A.M., Rubinson, H., Feltovich, P. J., Glaser, R., Klopfer, D., & Wang, Y. (1988). Expertise in a complex skill: Diagnosing x-ray pictures. In: Chi, M. T. H., Glaser, R., Farr, M. J. (Eds.), *The nature of expertise* (pp. 311-342). Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Lewis, G., Croft-Jeffreys, C., & David, A. (1990). Are British psychiatrists racist? *British Journal of Psychiatry*, *157*, 410-415.

Lichtenberg, J. W. (2009). Comment: Effects of experience on judgment accuracy. *The Counseling Psychologist*, *37*, 410-415.

Lichtenberg, J. W. (1997). Expertise in counseling psychology: A concept in search of support. *Educational Psychology Review*, *9(3)*, 221-238.

Lidz, C. W., Mulvey, E. P., & Gardner, W. (1993). The accuracy of predictions of violence to others. *The Journal of the American Medical Association*, *269(8)*, 1007-11.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing the research*. Cambridge, MA: Harvard University Press.

Light, R. J., & Smith, P. V. (1971). Statistical issues in social allocation models in intelligence: A review and a response. *Review of Educational Research*, *41(4)*, 351-367.

Lilienfeld, S. O., Lynn, S. J., & Lohr, J. M. (2003). *Science and pseudoscience in clinical psychology*. New York, NY: Guilford Press.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications, Inc.

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

Lopez, S. R., Smith, A., Wolkenstein, B. H., & Charlin, V. (1993). Gender bias in clinical judgment: An assessment of the analogue method's transparency and social desirability. *Sex Roles*, *28(1-2)*, 35-45.

*Lubman, D. I., Hides, L., Jorm, A. F., & Morgan, A. J. (2007). Health professionals' recognition of co-occurring alcohol and depressive disorders in youth: A survey of Australian general practitioners, psychiatrists, psychologists and mental health nurses using case vignettes. *Australian and New Zealand Journal of Psychiatry*, *41*, 830-835.

McCurley, M. J., Murphy, K. J., & Gould, J. W. (2005). Protecting children from incompetent forensic evaluations and expert testimony. *Forensic Evaluations*, *19*, 277-319.

McNiel, D. E., & Binder, R. L. (1995). Correlates of accuracy in the assessment of psychiatric inpatients' risk of violence. *American Journal of Psychiatry*, *152*, 901-906.

Mahoney, M. (1976). *Scientist as subject: The psychological imperative*. Cambridge, MA: Ballinger.

Marín-Martínez, F., & Sánchez-Meca, J. (1999). Averaging dependent effect sizes in meta-analysis: A cautionary note about procedures. *The 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(4)*, 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.

Meehl, P. (1986). Causes and Effects of My Disturbing Little Book. *Journal of Personality Assessment*, *50*(3), 370.

Meehl, P. (1973). Why I do not attend case conferences. In P. Meehl (Ed.), *Psychodiagnosis: Selected papers* (pp. 225-302), Minneapolis, MN: University of Minnesota Press.

Meehl, P. (1960). The cognitive activity of the clinician. *American Psychology*, *15*, 19-27.

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

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

Meeks, S. (1990). Age bias in the diagnostic decision-making behavior of clinicians. *Professional Psychology: Research and Practice*, *21(3)*, 279–2.

Melnick, R. R. (1975). Counseling responses as a function of method of problem presentation and type of problem. *Journal of Counseling Psychology*, *22*, 108-112.

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

Mohr, J. J., Israel, T., & Sedlacek, W. E. (2001). Counselors' attitudes regarding bisexuality as predictors of counselors' clinical responses: An analogue study of a female bisexual client. *Journal of Counseling Psychology*, *48(2)*, 212-222.

Mohr, J. J., Weiner, J. L., Chopp, R. M., & Wong, S. J. (2009). Effects of client bisexuality on clinical judgment: When is bias most likely to occur? *Journal of Counseling Psychology*, *56(1)*, 164-175.

Monahan, J., Steadman, H. J., Silver, E., Appelbaum, P. S., & Robbins, P. C. (2001). *Rethinking risk assessment: The MacArthur study of mental disorder and violence*. New York: Oxford University Press.

Monsen, J. J., & Frederickson, N. (2002). Consultant problem understanding as a function of training in interviewing to promote accessible reasoning. *Journal of School Psychology*, *40(3)*, 197-212.

Montgomery, K. (2006). *How doctors think: Clinical judgment and the practice of medicine*, New York, NY: Oxford University Press.

Morey, L. C., & Ochoa, E. S. (1989). An investigation of adherence to diagnostic criteria: Clinical diagnosis of the DSM-III personality disorders. *Journal of Personality Disorders*, *3*, 180-192.

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

Müller, P. A., Stahlberg, D. (2007). The role of surprise in hindsight bias: A metacognitive model of reduced and reversed hindsight bias. *Social Cognition*, *25*, 32-47.

Mumma, G. H., & Mooney, S. R. (2007). Comparing the validity of alternative cognitive case formulations: A latent variable, multivariate time series approach. *Cognitive Therapy and Research*, *31(4)*, 451-481.

Neighbors, H. W., Trierweiler, S. J., Ford, B. C., & Muroff, J. R. (2003). Racial differences in DSM diagnosis using a semi-structured instrument: The importance of clinical judgment in the diagnosis of African Americans. *Journal of Health and Social Behavior*, *44(3)*, 237-256.

Newell, R., & Gournay, K. (2000). *Mental health nursing: An evidence-based approach*. Edinburgh: Churchill Livingstone.

Nickerson, R., Perkins, D., & Smith, E. (1985). *The teaching of thinking*. Hillsdale, NJ : Erlbaum.

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.

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

O'Brien, B. (2009). Prime suspect: An examination of factors that aggravate and counteract confirmation bias in criminal investigations. *Public Policy and Law*, *15(4)*, 315-334.

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.

O'Reilly, Jr., C. A., Parlette, G. N., & Bloom, J. R. (1980). Perceptual measures of task characteristics: The biasing effects of differing frames of reference and job attitudes. *Academy of Management Journal*, *23*, 118-131.

Ogloff, J. R. P., Daffern, M. (2006). The dynamic appraisal of situational aggression: An instrument to assess risk for imminent aggression in psychiatric inpatients. *Behavioral Sciences and the Law*, *24*, 799-813.

Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of Consulting Psychology*, *29*, 261-265.

Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs: General Applications*, *76(547)*, 1-28.

*Parker, G., Mahendran, R., Yeo, S. G., Loh, M. I., & Jorm, A. F. Diagnosis and treatment of mental disorders: a survey of Singapore mental health professionals. *Social Psychiatry and Psychiatric Epidemiology*, *34*, 555-563.

*Parmley, M. C. (2006). *The effects of the confirmation bias on diagnostic decision making*. Unpublished doctoral dissertation, Drexel University, Philadelphia.

*Persons, J. B., & Bertagnolli, A. (1999). Inter-rater reliability of cognitive-behavioral case formulations of depression: A replication. *Cognitive Therapy and Research*, *23(3)*, 271-283.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. *Journal of the American Medical Association*, *295(6)*, 676-680.

Pezzo, M. V. (2003). Surprise, defense, or making sense: What removes the hindsight bias? *Memory*, *11(4/5)*, 421-441.

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(4)*, 429-436.

Polanyi, M. (1962). *Personal Knowledge*, Chicago, IL: The University of Chicago Press.

Poole, D. A., Lindsay, D. S., Memon, A., & Bull, R. (1995). Psychotherapy and the recovery of memories of childhood sexual abuse: U.S. and British practitioners' opinions, practices, and experiences. *Journal of Consulting and Clinical Psychology*, *63*, 426-437.

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

Rabinowitz, J., & Lukoff, I. (1995). Clinical decision making of short- versus long-term treatment. *Research on Social Work Practice*, *5(1)*, 62-79.

Raines, G., & Rohrer, J. (1955). The operational matrix of psychiatric practice. II. Variability in psychiatric impression and the projection hypothesis. *American Journal of Psychiatry*, *117*, 133-139.

Ray, D. C., McKinney, K. A., & Ford, C. V. (1987). Differences in psychologists' ratings of older and younger clients. *The Gerontologist*, *27(1)*, 82–86.

Ray, D. C., Raciti, M. A., & Ford, C. V. (1985). Ageism in psychiatrists: Associations with gender, certification, and theoretical orientation. *The Gerontologist*, *25*, 497-500.

Reijntjes, A., Kamphuis, J. H., Prinzie, P., Telch, M. J. (2010). Peer victimization and internalizing problems in children: A meta-analysis of longitudinal studies. *Child Abuse and Neglect, 34(4),* 244-252.

*Rerick, K. E. (1999). *Improving counselors' attitudes and clinical judgement towards dual diagnosis*. Unpublished doctoral dissertation, University of South Dakota, Vermillion.

Richards, M., & Wierzbicki, M. (1990). Anchoring errors in clinical-like judgments. *Journal of Clinical Psychology*, *46*, 358-365.

*Rieffel, L. M. (2005). *The moderating effects of clinician cognitive complexity on the accuracy of neuropsychological clinical judgments for Hispanic patients*. Unpublished doctoral dissertation, Fielding Graduate University, Santa Barbara.

*Rodriguez, C. M. (2002). Professionals' attitudes and accuracy on child abuse reporting decisions in New Zealand. *Journal of Interpersonal Violence*, *17*, 320-342.

Roese, N. J., & Maniar, S. D. (1997). Perceptions of purple: Counterfactual and hindsight judgments at Northwestern Wildcats football games. *Personality and Social Psychology Bulletin*, *23*, 1245-1253.

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

Rosenthal, A., (1982), Heterosexism and Clinical Assessment. *Smith College Studies in Social Work*, *82(2)*, 143-53.

Rosnow, R. L., & Rosenthal, R. (1989). Definition and interpretation of interaction effects. *Psychological Bulletin*, *105*, 143-146.

Ruscio, J. (2006). The clinician as subject: Practitioners are prone to the same judgment errors as everyone else. In S. Lilienfeld & W. O'Donohue (Eds.), The great ideas of clinical science: 17 principles that every mental health professional should understand (pp. 29-48). New York, NJ: Routledge.

*Ruscio, J., & Stern, A. (2005). The consistency and accuracy of holistic judgment: Clinical decision making with a minimally complex task. *The Scientific Review of Mental Health Practice*, 4(2), 52-65.

Sánchez-Meca, J., & Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods*, *13(1)*, 31-48.

Sánchez-Sachs, I. (2003). The mediating effects of prototype learning, experience and case typicality on clinical judgments. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, *64(4-B)*, 1915.

Sandifer, M., Hordern, A., & Green, L. (1970). The psychiatric interview: The impact of the first three minutes. *American Journal of Psychiatry*, *126*, 968-973.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, *40*, 87-104.

Schinka, J. A. & Sines, J. O. (1974). Correlates of accuracy in personality assessment. *Journal of Clinical Psychology*, *30(3)*, 374-377.

Schwartz, J. M., & Abramowitz, S. I. (1975). Value-related effects on psychiatric judgment. *Archives of General Psychiatry*, *32(12)*, 1525-1529.

Schwarz, S., & Stahlber, D. (2003). Strength of hindsight bias as a consequence of meta-cognitions. *Memory*, *33*, 395-410.

Sedlmeier, P. (2005). *From associations to intuitive judgment and decision making: Implicitly learning from experience* (pp. 83-99). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Settin, J. M. (1982). Clinical judgment in neuropsychology practice. *Psychotherapy: Theory, Research, & Practice*, *19(4)*, 397-404.

*Shumaker, K. R. (1999). *Measured professional competence between and among different mental health disciplines when evaluating and making recommendations in cases of suspected child sexual abuse*. Unpublished doctoral dissertation, United States International University, San Diego.

Silverman, L. H. (1959). A Q-sort study of the validity of evaluations made from projective techniques. *Psychological Monographs*, *73(7, Whole No. 477)*, 28.

Simpson, G., Williams, J., Segall, A. (2007). Social Work Education and Clinical Learning. *Clinical Social Work Journal, 35(1),* 3-14.

Slobogin, C. (2003). Pragmatic Forensic Psychology: A Means of "Scientizing" Expert Testimony Form Mental Health Professionals? *Psychology Public Policy and Law*, *9(3)*, 275-300.

Slovic, P., & Fischhoff, B. (1977). On the psychology of experimental surprises. *Journal of Experimental Psychology*, *3*, 544-551.

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2002). The affect heuristic. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 397-420). New York, NJ: Cambridge University Press.

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

Snyder, M. (1981). “Seek and ye shall find.” In Higgins, E., Herman, C., & Zanna, M. (Eds.), *Social cognition: The Ontario symposium on personality and social psychology* (pp. 277-303), Hillsdale, NJ, Erlbaum.

Snyder, M. (1977). “A patient by any other name” revisited: Maladjustment or attributional locus of problem? Journal of Consulting and Clinical Psychology, 45(1), 101-103.

Snyder, M., & Swann, W. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology*, *36*, 1202-1212.

Snyder, M., Tanke, E., & Berscheid, E. (1977). Social perception and interpersonal behavior: on the self-fulfilling nature of social stereotypes. *Journal of Personality and Social Psychology*, *35*, 656-666.

Sonnentag, S. (1998). Expertise in professional software design: A process study. *Journal of Applied Psychology*, *83(5)*, 703-715.

Soskin, W. F. (1954). Frames of reference in personality assessment. *Journal of Clinical Psychology*, *1954*, 107-114.

Spengler, P. M. (2000). Does vocational overshadowing even exist? A test of the robustness of the vocational overshadowing bias. *Journal of Counseling Psychology*, *47(3)*, 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(4), 372-381.

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(3), 506-534.

Spengler, P. M., White, M. J., Ægisdóttir, S., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G. R., & Rush, J. D. (2009). *The meta-analysis of clinical judgment project: Effects of experience on judgment accuracy*. *The Counseling Psychologist, 37(3),* 350-399.

Stearns, B. C., Penner, L. A., & Kimmel, E. (1980). Sexism among psychotherapists: A case not yet proven. *Journal of Consulting and Clinical Psychology*, *48(4)*, 548-550.

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

Sternberg, R. J., Roediger, H. L., & Halpern, D. F. (Eds.). (2007). *Critical thinking psychology*. New York, NJ: Cambridge University Press.

*Strain, B. A. (2002). Influence of gender bias on the diagnosis of borderline personality disorder. Unpublished doctoral dissertation, Alliant International University, Fresno.

Strohmer, D. C., & Leierer, S. J. (2000). Modeling rehabilitation counselor clinical judgment. *Rehabilitation Counseling Bulletin*, *44(1)*, 3-9.

Strohmer, D. C., Moilanen, D. L., & Barry, L. J. (1988). Personal hypothesis testing: The role of consistency and self-schema. *Journal of Counseling Psychology*, *35(1)*, 56-65.

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(4)*, 465-472.

Temerlin, M. K. (1968). Suggestion effects in psychiatric diagnosis. *Journal of Nervous and Mental Disease*, *147*, 349-353.

Teri, L. (1982). Effects of sex and sex-role style on clinical judgment. *Sex Roles*, *8(6)*, 639-649.

Tetlock, P. E. (2000). Cognitive biases and organizational correctives: Do both disease and cure depend on the politics of the beholder? *Administrative Science Quarterly*, *45*, 293-326.

Tetlock, P. E., & Mitchell, G. (2008). Calibrating prejudice in milliseconds. *Social Psychology Quarterly*, *71(1)*, 12-16.

Tippins, T. M., & Wittman, J. P. (2005). Empirical and ethical problems with custody recommendations: A call for clinical humility and judicial vigilance. *Family Court Review*, *193*, 266-299.

Tracey, T. F., Hays, K. A., Malone, J., & Herman, B. (1988). Changes in counselor response as a function of experience. *Journal of Counseling Psychology*, *35*, 119-126.

Trachtman, J. P. (1971). Socio-economic class bias in Rorschach diagnosis: contributing psychosocial attributes of the clinician. *Journal of Personality Assessment, 35(3)*, 229-40.

Tukey, J. (1953). The problem of multiple comparisons. Unpublished manuscript, Princeton University.

Turk, D. C., & Salovey, P. (1988). *Reasoning, inference, and judgment in clinical psychology*. New York, NY: Free Press.

Turner, D. R. (1966). Predictive efficiency as a function of amount of information and level of professional experience. *Journal of Projective Techniques & Personality Assessment*, *30(1)*, 4-11.

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, Minneapolis.

*Walker, B. S. (1999). *The diagnosis and treatment of major depression in Aids patients: effect of counselor experience and attitude toward people with Aids*. Unpublished doctoral dissertation, Ball State University, Muncie.

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.

Warner-Rogers, J. E., Hansen, D. J., & Spieth, L. E. (1996). The influence of case and professional variables on identification and reporting of physical abuse: A study with medical students. *Child Abuse and Neglect*, *20*, 851-866.

Watts, F. (1980). Clinical judgment and clinical training. *British Journal of Medical Psychology*, *53*, 95-108.

Watts, D., Timbrook, R. E., & Graham, J. R. (1991). Detecting Fake-Goodand Fake-Bad MMPI-2 Profiles. *Journal of Personality Assessment*, *57(2)*, 264 277.

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

Westen, D., Weinberger, J. (2004). When clinical description becomes statistical prediction. *American Psychologist*, *59(7)*, 595-613.

Whaley, A. L. (2001). Cultural mistrust and the clinical diagnosis of paranoid schizophrenia in African American patients. *Journal of Psychopathology and Behavioral Assessment*, *23(2)*, 93-100.

Wierzbicki, M. (1993). *Issues in clinical psychology: Subjective versus objective approaches*. Boston, MA: Allyn & Bacon.

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

*Wilkin Bloch, S. (2001). *The diagnosis of Attention-Deficit/Hyperactivity Disorder (ADHD) by psychologists, pediatricians, and general practitioners*. Unpublished doctoral dissertation, University of Windsor, Windsor.

Wilson, K. B. (2000). Predicting vocational rehabilitation acceptance based on race, education, work status, and source of support at application. *Rehabilitation Counseling Bulletin*, *43(2)*, 97-105.

*Witteman, C. L. M. & van den Bercken, J. H. L. (2007). Intermediate effects in psychodiagnostic classification. *European Journal of Psychological Assessment*, *23(1)*, 56-61.

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.

*Wood, D. S. (2004). An intervention for diagnostic overshadowing. Unpublished doctoral dissertation, Arizona State University, Tempe.

Wood, J. M., & Nezworski, M. T. (2005). Science as a history of corrected mistakes: Comment. *American Psychologist*, *60(6)*, 657-658.

Wrobel, N. H. (1993). Effect of patient age and gender on clinical decisions. *Professional Psychology: Research and Practice*, *24(2)*, 206-212.

*Yeo, S. G., Parker, G., Mahendran, R., Jorm, A. F., Yap, H. L., Lee, C., & Loh, M. I. (2001). Mental health literacy survey of psychiatrically and generally trained nurses employed in a Singapore psychiatric hospital. *International Journal of Nursing Practice*, 7, 414-421.

Zeldow, P. B. (2009). In defense of clinical judgment, credentialed clinicians, and reflective practice. *Psychotherapy: Theory, Research, Practice, Training*, *46(1)*, 1-10.

Zellman, G. L. (1990). Report decision-making patterns among mandated child abuse reporters. *Child Abuse and Neglect*, *14*, 325-336.

Ziskin, J. (1981). *Coping with psychiatric and psychological testimony* (3rd ed., two vols). Venice, CA: Law and Psychology Press.

*Zozula, L. J. (2001). *The assessment process of psychologists as a function of clinical experience*. Unpublished doctoral dissertation, McGill University, Montreal.

Zygmond, M. J., & Denton, W. (1988). Gender bias in marital therapy: A multidimensional scaling analysis. *American Journal of Family Therapy*, *16(3)*, 262-272.

Appendix A

Electronic Data Base Search Terms

Clinical [judg(e)ment(s)] [bias(es)] [judg(e)ment(al)(s) bias(es)] [decision(s) (making)] [judge(s)] [assessment(s)] [prediction(s)]

Counselor [judg(e)ment(s)] [bias(es)] [judg(e)ment(al)(s) bias(es)] [decision(s) (making)] [assessment(s)] [prediction(s)]

Medical [judg(e)ment(s)] [bias(es)] [judg(e)ment(al)(s) bias(es)] [decision(s) (making)] [judge(s)] [assessment(s)] [prediction(s)]

Psychiatric [judg(e)ment(s)] [bias(es)] [judg(e)ment(al)(s) bias(es)] [decision(s) (making)] [judge(s)] [assessment(s)] [prediction(s)]

Psychological assessment(s)

Diagnostic [judg(e)ment(s) [decision(s) (making)] (accuracy)

Treatment [judg(e)ment(s)] [decision(s) (making)]

Intervention [judg(e)ment(s)] [decision(s) (making)]

Judg(e)ment(al) [heuristic(s)] [bias(es)] (accuracy)

Anchoring [heuristic(s)] [effect(s)] [bias(es)]

Representative(ness) [heuristic(s)] [effect(s)] [bias(es)]

Availability [heuristic(s)] [effect(s)] [bias(es)]

Salience(y) [heuristic(s)] [effect(s)] [bias(es)]

Vividness [heuristic(s)] [effect(s)] [bias(es)]

Illusory correlation(s)

Fundamental attribution error(s)

(Diagnostic) (Treatment) overshadowing

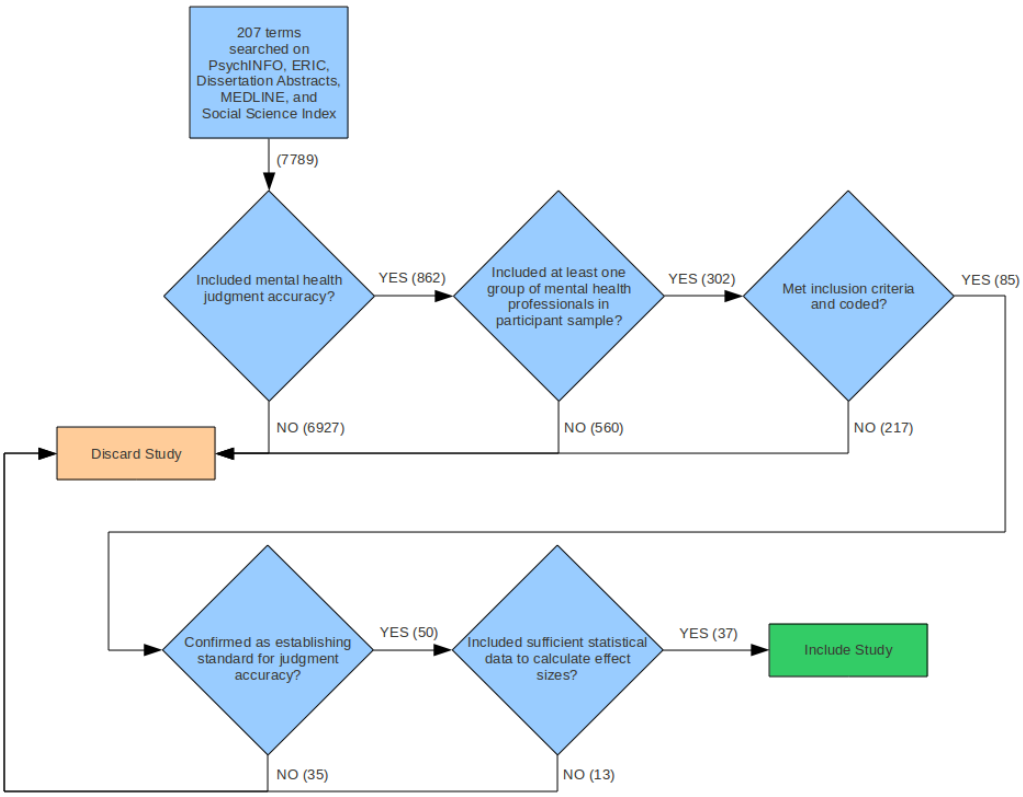
(Under)(Over)diagnosing

(Gender) [Race(ial)] (Socioeconomic) (Age) bias

(Sex)(Rac)ism

Appendix B

Search Strategy Flowchart



Appendix C

Moderator Coding Sheet

Title:

Author(s):

JN: **VOL:** **PGS:** **PY:**

Accept:	Reject:	Unsure:

Experience Type: Clinical, Educational, Both

Experience Breadth: General, Specific, Both

Judgment Type: Problem Type, Hit Rate, Treatment, Severity, Prognosis, Problem Recall, Other, Combined

Criterion Validity: Low, High, Both

Provision of Feedback: Yes, No

Publication Source: APA, Other Psychology, Medical, Dissertation

Method of Study: Analogue, Archival, In Vivo

Validity of Stimulus: Direct, Indirect, Both

Relation of Experience to Design: Not in Design, In primary design, Supplementary, Multiple

Experience Major Variable: Yes, No

Study Quality: Acceptable, Good, Excellent

Profession Type: Psychology, Psychiatry, Psychiatric Nursing, Social Work, Combination

Inclusion of Non-Mental Health Participants: Yes, No

Appendix D

Metric Coding Sheet

Rater_____ Date_____ Author(s): _____

Year: _____ Journal: _____

Independent Variable: _____

Dependent Variable: _____

Group 1 (name): _____ Group 2 (name): _____

Mean: _____ SD: _____ n: _____Mean: _____ SD: _____ n: _____

Test: _____ df error: _____

Test value: _____ df effect: _____

p-value: _____ effect size: _____

Which judgment direction or category is more accurate? _____ Can't determine _____

Your (the rater's) confidence in rating of accuracy is ( ) Low ( ) High

Dependent Variable: _____

Group 1 (name): _____Group 2 (name): _____

Mean: _____ SD: _____ n: _____Mean: _____ SD: _____ n: _____

Test: _____ df error: _____

Test value: _____ df effect: _____

p-value: _____ effect size: _____

Which judgment direction or category is more accurate? _____ Can't determine _____

Your (the rater's) confidence in rating of accuracy is ( ) Low ( ) High

Subjects (specify):

Global Rating Methods/Analyses:	1 POOR	2 ADEQUATE	3 EXCELLENT

Table 1

Interrater Agreement for Moderator Coding Sheet

Moderator	Kappa Rating	Level of Agreement
Experience Type	.74	Substantial Agreement
Experience Breadth	.71	Substantial Agreement
Judgment Type	.74	Substantial Agreement
Criterion Validity	.73	Substantial Agreement
Provision of Feedback	.84	Almost Perfect Agreement
Publication Source	.96	Almost Perfect Agreement
Method of Study	.88	Almost Perfect Agreement
Validity of Stimulus	.81	Almost Perfect Agreement
Relation of Experience to Design	.88	Almost Perfect Agreement
Experience Major Variable	.79	Substantial Agreement
Study Quality	.79	Substantial Agreement
Profession Type	.85	Almost Perfect Agreement
Inclusion of Non-Mental Health Participants	.96	Almost Perfect Agreement

Table 2

Corrected Effect Sizes Between Experience and Accuracy

Study	n	Study effect size (*d*)c	Lower	Upper	Public-ation Year	Problem Type	Hit Rate	Treat-ment	Sever-ity	Prog-nosis	Problem Recall	Other	Comb-ined
Akehurst et al. (2004)	58	0.21	-0.75	1.17	2004							0.21	
Boland[b] (2002)	239	-0.00	-0.26	0.25	2002							-0.00	
Brammer[b] (2002)	138	0.81	0.45	1.18	2002	0.81							
Butte[b] (1998)	10	-0.17	-2.15	1.82	1998		-0.17						
Ebling & Levenson[b] (2003)	101	-0.04	-0.64	0.56	2003								-0.04
Ekman et al. (1999)	627	-0.04	-0.39	0.30	1999							-0.04	
Garb (2006)	40	0.30	-0.48	1.08	2006							0.30	
Garb & Boyle (2003)	25	0.00	-0.84	0.84	2003	0.00							
Gerbe[b] (2007)	168	0.53	0.24	0.82	2007	0.53							

Study	n	Study effect	Lower	Upper	Public-ation	Problem Type	Hit Rate	Treat -	Sever -	Prog -	Problem Recall	Other	Comb-ined
Hannan et al. (2005)	48	0.54	-1.25	2.33	2005								0.54
Hansen et al.[b] (1997)	210	0.33	-0.07	0.73	1997								0.33
Hickling et al. (2002)	6	0.00	-2.26	2.26	2002		0.00						
Hillman et al.[b] (1997)	183	0.42	0.12	0.72	1997								0.42
Huffaker[b] (2008)	125	0.06	-0.30	0.42	2001	0.06							
Jackson et al. (2004)	222	-0.11	-0.41	0.19	2004		-0.11						
Jopp[b] (2001)	89	0.21	-0.22	0.64	2001	0.21							
Kellner[b] (2001)	109	-0.03	-0.62	0.56	2001								-0.03
Kim & Ahn[a] (2002)	20	-0.06	-1.01	0.89	2002		-0.06						
Kitamura[b] (1999)	406	0.23	-0.28	0.74	1999							0.23	
Kocsis (2003)	444	0.11	-0.41	0.62	2003							0.11	

Study	n	Study effect	Lower	Upper	Public-ation	Problem Type	Hit Rate	Treat -	Sever -	Prog -	Problem Recall	Other	Comb-ined
Leon & Perez[ab] (2001)	132	0.00	-0.64	0.64	2001	0.00							
Lubman et al.[b] (2007)	1230	0.16	-0.14	0.47	2007	0.16							
Parker et al.[b] (1999)	299	0.78	-0.02	1.57	1999	0.78							
Parmley[b] (2006)	102	0.15	-0.73	1.03	2006	0.15							
Persons & Bertagnolli (1999)	38	0.72	0.02	1.43	1999	0.72							
Rerick (1999)	27	1.86	0.57	3.86	1999	1.86							
Rieffel (2005)	47	0.04	-0.55	0.64	2005	0.04							

Study	n	Study effect	Lower	Upper	Public-ation	Problem Type	Hit Rate	Treat-ment	Sever-ity	Prog-nosis	Problem Recall	Other	Comb-ined
Rodriguez[b] (2002)	253	-0.33	-0.70	0.03	2002			-0.33					
Ruscio & Stern[b] (2005)	124	-0.46	-0.91	-0.00	2005					-0.46			
Shumaker[b] (1999)	204	0.45	-0.77	1.68	1999	0.45							
Strain[b] (2002)	52	-0.02	-0.58	0.54	2002	-0.02							
Walker[b] (1999)	281	-0.06	-0.69	0.57	1999	-0.06							
Wilkin[b] (2001)	120	0.45	-0.14	1.04	2001		0.45						

Study	n	Study effect	Lower	Upper	Public-ation	Problem Type	Hit Rate	Treat-	Sever-	Prog-	Problem Recall	Other	Comb-ined
Wittemann & van den Bercken (2007)	41	0.32	-0.46	1.11	2007		0.32						
Wood[b] (2004)	210	0.20	-0.37	0.78	2004	0.20							
Yeo et al.[b] (2001)	230	0.19	-0.19	0.57	2001	0.19							
Zozula[b] (2001)	27	-0.27	-1.77	1.22	2001		-0.27						
Overall *n* or *d*	6685	0.16	0.04	0.27		5.59	0.22	-0.33		-0.46		0.81	1.22
n of studies	37					17	7	1		1		6	5

Note. Overall *d* is the average of *d*s for studies as units of judgment accuracy corrected for sample size. Between subjects designs are used in all studies.

[a]Zero effect is inferred. The study reports statistically non-significant 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 3
Categorical Variables for Study Coding

Study	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Akehurst et al. (2004)	2	3	2	2	2	1	2	2	1	1	4	1	0	7
Boland (2002)	3	3	1	2	4	1	2	2	1	1	4	2	6	7
Brammer (2002)	3	1	1	2	1	1	1	2	1	1	1	2	3	3
Butte (1998)	3	1	1	2	4	3	1	2	1	1	3	1	3	3
Ebling & Levenson (2003)	3	3	3	2	2	1	1	2	1	3	1	1	4	3
Ekman et al. (1999)	3	3	1	2	2	1	1	2	1	1	1	1	0	6

Study	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Garb (2006)	2	1	2	2	2	1	2	3	2	2	1	2	0	3
Garb & Boyle (2003)	1	2	2	2	2	1	2	1	2	1	1	2	3	0
Gerbe (2007)	3	3	1	2	4	1	2	2	1	2	1	2	6	7
Hannan et al. (2005)	2	1	3	2	2	3	1	1	2	3	1	2	0	3
Hansen et al. (1997)	2	2	1	2	2	1	2	3	1	2	5	2	0	3
Hickling et al. (2002)	1	2	1	1	2	3	3	2	1	1	1	2	5	0
Hillman et al. (1997)	2	2	3	2	1	1	2	3	1	2	1	2	0	6

Study	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Huffaker (2008)	3	3	1	2	4	1	1	4	1	3	1	2	3	7
Jackson et al. (2004)	2	1	2	2	2	1	2	2	1	3	1	2	0	3
Jopp (2001)	1	2	1	2	4	1	2	3	1	3	1	2	5	0
Kellner (2001)	3	3	1	2	4	1	2	2	1	2	1	2	6	3
Kim & Ahn (2002)	1	1	1	1	1	1	2	2	1	1	1	2	3	0
Kitamura (1999)	2	1	2	2	3	2	2	2	1	2	2	1	0	6
Kocsis (2003)	2	1	1	2	2	1	2	2	1	1	1	1	0	6

Study	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Leon & Perez (2001)	2	1	2	2	2	1	2	4	1	2	1	1	0	3
Lubman et al. (2007)	2	1	1	2	3	1	2	2	1	2	5	1	0	6
Parker et al. (1999)	2	1	1	2	3	1	2	2	1	1	5	1	0	6
Parmley (2006)	2	2	2	2	4	1	2	2	1	3	1	2	0	5
Persons & Bertagnolli (1999)	3	3	1	2	2	1	1	2	1	2	5	2	3	7
Rerick (1999)	2	2	2	2	4	1	2	2	1	2	1	2	0	5
Rieffel (2005)	1	3	2	2	4	2	2	2	1	1	1	2	6	0

Study	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Rodriguez (2002)	2	1	1	2	2	1	2	2	1	2	1	1	0	6
Ruscio & Stern (2005)	2	1	2	2	2	1	2	4	1	2	1	1	0	7
Shumaker (1999)	3	1	2	2	4	1	2	3	1	2	5	1	6	7
Strain (2002)	3	1	2	2	4	1	2	2	1	2	5	2	3	3
Walker (1999)	1	2	2	2	4	1	2	2	1	2	1	2	6	0
Wilkin (2001)	2	1	2	2	4	1	2	2	1	2	1	1	0	6
Witteman & van den Bercken (2007)	1	1	2	2	2	1	2	2	1	2	1	2	3	0

Study	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Wood (2004)	2	3	2	1	4	1	2	3	1	2	5	1	0	3
Yeo et al. (2001)	2	2	1	2	3	1	2	2	1	2	3	1	0	6
Zozula (2001)	2	1	1	2	4	1	1	2	1	2	1	2	0	3

Note. Categorical variables and codes: A-experience type (1 = clinical, 2 = educational, 3 = both), B-experience breadth (1 = general, 2 = specific, 3 = both), C-accuracy criterion validity (1 = low, 2 = high, 3 = both), D-accuracy feedback (1 = yes, 2 = no), E-publication source (1 = APA journal, 2 = other psychology journal 3 = psychiatry or medicine journal, 4 = dissertation), F-method of study (1 = analogue, 2 = archival, 3 = *in vivo*), G-ecology of stimulus presentation (1 = directly experience, 2 = indirectly experienced, 3 = both), H-relation of experience to design (1 = not in design, 2 = in primary design, 3 = supplementary analysis, 4 = multiple), I-experience as a major variable (1 = yes, 2 = no), J-study quality (1 = acceptable, 2 = good, 3 = excellent), K-profession type (1 = psychology, 2 = psychiatry, 3 = nursing, 4 = social work, 5 = combination), L-inclusion of non-mental health participants (1 = yes, 2 = no), M-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 N-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).

Table 4

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]	Lower	Upper
Experience Type	0.94				
Clinical		7	0.09	-0.21	0.40
Educational		18	0.12	-0.03	0.27
Both		11	0.23*	0.05	0.41
Experience Breadth	0.63				
General		18	0.12	-0.04	0.28
Specific		8	0.23*	0.01	0.46
Both		10	0.15	-0.04	0.34
Judgment Type	8.27*				
Problem type		16	0.29**	0.17	0.41
Hit rate		7	0.02	-0.22	0.26
Treatment		(1)	(-0.34)	(-0.70)	(0.03)
Severity		(0)			
Prognosis		(1)	(-0.46*)	(-0.91)	(-0.00)
Problem recall		(0)			
Other		6	0.05	-0.12	0.22
Combined		5	0.29**	0.08	0.50
Criterion Validity	2.29				
Low		18	0.20**	0.06	0.33
High		15	0.05	-0.13	0.23
Both		3	0.30	-0.06	0.66
Provision of Feedback	0.02				
Yes		3	0.12	-0.43	0.66
No		33	0.16**	0.05	0.27
Publication Source	16.48**				
APA		3	0.54**	0.01	0.78
Other psychology		15	-0.02	-0.16	0.12
Medical		4	0.23*	0.01	0.45
Dissertation		14	0.15*	0.03	0.30
Ecological Validity of Method of Study	0.00				
Analogue		31	0.16**	0.04	0.27

Variable and levels	Between-class effect (Q_B)	k	Mean weighted effect size (d_{i+})[a]	Lower	Upper
Archival		(2)	(0.15)	(-0.32)	(0.61)
In vivo		3	0.17	-1.00	1.33
Ecological Validity of Stimulus	0.86				
Direct		8	0.26*	0.02	0.51
Indirect		27	0.13*	0.01	0.25
Both		(1)	(0.00)	(-2.29)	(2.29)
Relation of Experience to the Research Design	4.70				
Not in design		(2)	(0.11)	(-.69)	(0.3)
In primary design		25	0.16*	0.04	0.28
Supplementary		6	0.32**	0.08	0.56
Multiple		3	-0.13	-0.45	0.20
Experience as a Major Variable	0.02				
Yes		33	0.15**	0.05	0.26
No		3	0.20	-0.39	0.79
Study Quality	0.94				
Acceptable		11	0.20	-0.01	0.41
Good		19	0.17*	0.03	0.32
Excellent		6	0.04	-0.21	0.30
Profession Type	1.38				
Psychology		24	0.12	-0.02	0.26
Psychiatry		(1)	(0.23)	(-0.41)	(0.87)
Psychiatric Nursing		(2)	(0.16)	(-0.36)	(0.59)
Social Work		(2)	(0.04)	(-0.39)	(0.46)
Combination		7	0.30*	0.04	0.56
Inclusion of Non-Mental Health Participants	2.37				
Yes		15	0.06	-0.10	0.22
No		21	0.22**	0.09	0.36

*p < .05

**p < .01

Table 5

Stem-and-Leaf Plot of Combined Effect Sizes for Experience and Overall Accuracy Effects with Outlier Removed

Effect Sizes (*d*)		Summary Statistics	
Stem	Leaf		
1.8	6 (outlier removed)	Maximum	0.81
1.7			
1.6		Quartile 3	0.33
1.5			
1.4		Median	0.13
1.3			
1.2		Quartile 1	-0.04
1.1			
1.0		Minimum	-0.46
.9			
.8	1	SD	0.30
.7	2,8		
.6		Mean (weighted for sample n)	0.16
.5	3,4		
.4	2,5,5	*N*	36
.3	0,2,3		
.2	0,1,1,3	*Proportion with positive sign	0.67
.1	1,5,6,9		
.0	0,0,0,4,6		
-.0	0,2,3,4,4,6,6		
-.1	1,7		
-.2	7		
-.3	3		
-.4	6		