INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Confidence in psychodiagnosis:

A study of clinicians' judgment confidence in a psychological assessment task as a function of reliance on four inferential heuristics and clinical experience

J. David Smith

Department of Educational and Counselling Psychology

McGill University, Montreal

Thesis submitted in partial fulfillment of the requirements for the degree of

PhD in Counselling Psychology

Table of Contents

List of Tables

Abstract

Research in several domains has revealed that when individuals are asked to estimate the probability that their judgments are correct, they reveal an overconfidence effect. Judgments produced in decision environments such as psychodiagnosis, which are by their nature ambiguous and complex, appear to be most vulnerable to overconfidence. By implication, this phenomenon threatens the validity of clinical judgment and subjects clients to risks of flawed diagnoses and unsuitable treatments.

In an effort to identify variables implicated in judgment confidence and overconfidence, this study examined the relationship between four different inferential biases (dispositionalism, confirmationism, truncated data search, and narrow problem formulation) and diagnostic confidence in the context of a psychological assessment task. A second aspect of this study examined the effect of clinical experience on psychodiagnostic confidence. Thirty-six clinicians (18 experienced professionals and 18 clinical trainees) were individually presented a written client casefile, which was segmented and serially presented, to read and clinically interpret aloud. Analyses of participants' verbal protocols revealed that one of the four inferential biases studied (i.e., dispositionalism) accounted for a significant proportion of the variance in psychodiagnostic confidence scores. The author concludes that other clinician variables likely moderate the relationship

between particular heuristics and judgment confidence. Regarding the second hypothesis, the data revealed no difference between experienced clinicians and clinical trainees in the degrees of psychodiagnostic confidence manifested in their verbal protocols.

The author proposes that effective remedies to overconfidence begin in training programs that lead students through problem-solving experiences that can invalidate facile, premature, and dubious diagnostic judgments. The author delineates a number of strategies that may be used by educators to achieve this end.

Résumé

La recherche dans divers milieux a démontré la présence d'un effet de surconfiance parmis les individus lorsqu'ils estiment la probabilité de l'exactitude de leurs jugements. Les jugements produits dans l'environnement décisionnel de la psychodiagnose, sont de par leur nature ambigus et complexes. De ce fait, ils semblent les plus vulnérables à l'effet de surconfiance. Ceci implique donc que ce phénomène menace la validité des jugements des cliniciens et soumet les clients à des risques de faux diagnostics et de traitements non appropriés.

Cette étude tente d'identifier les variables impliquées dans la confiance et la surconfiance des jugements en examinant la relation entre quatre différents biais inférentiels (le dispositionalisme, le confirmationisme, la recherche tronquée de données et la formulation étroite des problèmes) et la confiance du diagnostic lors d'une évaluation psychologique. Le chercheur demanda à trente-six cliniciens (18 professionnels et 18 stagiaires) de lire et d'interpréter individuellement le dossier d'un client. Ce dossier écrit était segmenté et présenté en série. L'analyse des propos verbaux des participants révéla que seulement un des quatre biais inférentiels étudiés (le dispositionalisme) expliquait une importante variance du niveau de confiance psychodiagnostique. L'auteur suggère que d'autres variables pertinentes au clinicien pourraient modérer la relation entre les heuristiques de jugement populaires et la confiance dans le jugement. Concernant la deuxième

hypothèse, les données ne révélèrent aucune différence entre les cliniciens professionnels et les stagiaires concernant le niveau de confiance psychodiagnostique manifestée par leur propos verbaux.

Pour remédier effectivement au problème de la surconfiance, l'auteur propose des programmes de formation qui amèneraient l'étudiant à passer au travers du processus de résolution des problèmes et qui rendraient invalide les jugements diagnostics faciles, prématurés et douteux. L'auteur décrit plusieurs stratégies que les éducateurs pourraient utiliser pour réaliser cet objectif.

Preface: Contributions to Knowledge

Recent studies in the areas of social and clinical judgment have attempted to explain the frequently observed overconfidence effect in human judgment and decision making. This study advances these efforts by investigating a suspected source of overconfidence, namely inferential heuristics and biases, which presumably are used to simplify complex judgment tasks. Specifically, this study examined the relationship between clinicians' reliance on inferential heuristics to aggregate and interpret a clinical casefile and the degrees of psychodiagnostic confidence they manifested in their verbal reports. The findings do not support a direct, linear relationship between reliance on heuristics and judgment confidence, and it appears that, at least in a clinical context, other variables may moderate the relationship between particular heuristics and psychodiagnostic confidence.

Another contribution of this study to knowledge is the development of a procedure to measure linguistic expressions of psychodiagnostic confidence. This procedure to quantify verbally expressed confidence in clinicians' verbal protocols was developed and validated specifically for this study, which used the "think-aloud" methodology to indirectly elicit participants' clinical judgments and confidence assessments. All previously documented studies used a question-answering format to elicit confidence assessments from research participants. The principal advantage of the more naturalistic and ecologically valid method of

measuring judgment confidence used in this study is that it circumvents a number of potential biasing effects stemming from the question-answering paradigm. An interesting finding arising from this original approach to the problem domain is that experienced clinicians expressed degrees of psychodiagnostic confidence that did not significantly differ from those expressed by clinicians-in-training. This finding diverges from results of earlier studies and can be explained in terms of the advantageous aspects of the think-aloud procedure.

Acknowledgments

I would like to acknowledge the important contributions of a number of people to this project. First and foremost, I want to acknowledge the help provided by my thesis supervisor, Dr. Frank Dumont. Beyond his substantive contributions to this research project, which are innumerable, perhaps the most enduring legacy of his contribution to this project is the indelible mark he has made on my professional development and, in particular, on my abilities to think critically and to write clearly. I also wish to extend my thanks and appreciation to the other members of my advisory committee, Dr. Anastassios Stalikas, Dr. Socrates Rapagna, and Dr. William Talley, for their guidance in this endeavour and their valuable input that enhanced the quality of this project. I would also like to acknowledge the contribution of Dr. Robert Bracewell to this research project and, in particular, his expertise in "think-aloud" protocols.

Finally, I wish to dedicate this dissertation to my wife, Nathalie, and my son, Jonathan. Their unconditional love is a wellspring of joy in my life that no words can adequately express. Thank you for being there.

This dissertation was supported by a doctoral fellowship from the Social Sciences and Humanities Research Council of Canada.

CHAPTER I: INTRODUCTION

Introduction to the Problem Domain

To the extent that mental health professionals are required to make clinical decisions about clients on the basis of limited data and quickly formulated hypotheses, they are vulnerable to a wide array of inferential errors and biases that have been documented in the social attribution literature in the last few decades (for reviews see: Ross, 1977; Ross & Nisbett, 1991; Turk & Salovey, 1988). A large body of scientific evidence compiled in this time has demonstrated that people are limited in the amount of information they can process from among all the data bombarding their senses at any moment. Consequently, people "go beyond the information given" in order to make meaning of and ultimately control events in the world around them (Kelly, 1955; Ross, 1977, Turk, Salovey, & Prentice, 1988).

Research in this domain has focused on uncovering the kinds of heuristics that distort information processing and give rise to biased inferences. The term heuristic, deriving from *heuriskein*, a Greek word meaning "to find", refers in the present context to a cognitive strategy used to reduce the complexity and difficulty of judgment tasks (Tversky & Kahneman, 1974). Heuristics can be very useful decisional aids, especially when the amount of information and the number of variables available to the decision maker reach overwhelming proportions.

However, a voluminous empirical literature has demonstrated that reliance on such cognitive shortcuts can at times result in biased and erroneous inferences. Inferences are the end product of cognitive processes (which implicate heuristics) that transform information that is gathered and which is operative in memory (e.g., Nisbett & Ross, 1980; Kahneman, Slovic, & Tversky, 1982;). They include (but are not limited to) judgments, hypotheses, predictions, estimates, intuitions, and hunches.

A robust finding that has emerged from this line of inquiry is the "overconfidence effect." Research on the overconfidence effect, a phenomenon evidenced across a broad range of judgment and prediction tasks, has consistently demonstrated that people tend to express degrees of subjective confidence in their inferences that significantly exceed the accuracy of those inferences (Lichtenstein, Fischhoff, & Phillips, 1982; Dunning, Griffin, Milojkovic, & Ross, 1990; Vallone, Griffin, Lin, & Ross, 1990). There is a growing recognition that professional psychologists, constrained by limits of the same human information processing system as "lay psychologists", are equally susceptible to inferential errors and biases (Dumont, 1993; Cline, 1985).

Several studies have investigated overconfidence among psychological practitioners. Participants evidenced overconfidence in three studies (Faust, Hart, Guilmette, & Arkes, 1988; Moxley, 1973; Oskamp, 1965) and underconfidence in

one (Wedding, 1983). A larger literature has addressed the relationship between confidence and the validity of clinical judgments. Findings have been mixed, providing only limited support for the notion that experienced clinicians make more appropriate confidence ratings than less experienced or untrained controls (Garb, 1989). However, even when clinicians provide more appropriate confidence ratings relative to another group, they can still be overconfident (cf. Levenberg, 1975).

Confidence and Probability

Clinical decision making is a highly probabilistic enterprise, especially in comparison to the more deterministic systems found among the "hard" sciences (Dumont, 1991). Clinicians regularly work with complex data sets that are usually incomplete and only partially understood. Moreover, there is no agreement among mental health professionals as to what variables are critical for clinical problem solving. Finally, many procedures for identifying and treating psychological problems do not have solid empirical foundations and are often tenuously based on one's clinical experiences--experiences that are shaped significantly by the familial and cultural lore derived from one's pre-clinical experiences (Mahoney, 1991). As such, professional psychology remains largely a fuzzy problem domain, and clinical judgment can be characterized as *decision making under uncertainty*. In the daily work of clinicians, this means that decisions are made in part as

function of the degrees of confidence associated with various diagnostic and treatment alternatives being considered in individual cases (Meehl, 1957). Specifying and delineating the role of variables affecting confidence assessments in clinical practice should serve to improve judgment processes in professional psychology.

Although the concepts of confidence and probability are closely related, they are not identical. Probability can be conceptualized as being an objective phenomenon, and events can be seen as possessing a probability of occurring. Confidence, on the other hand, is a subjective phenomenon and concerns making a judgment about a judgment (Smith & Dumont, 1997). In studies of judgment confidence, individuals are typically instructed (a) to make a judgment about some entity and (b) to assess the accuracy of that judgment in probabilistic terms. The latter judgment (termed a confidence assessment or rating) involves indicating *the degree of one's belief* in the initial judgment (Lichtenstein et al., 1982). This conceptualization finds support in linguistic theory, where verbal confidence expressions are referred to as "modal adjectives" (Lyons, 1977; Reyna, 1981). Research in this domain has found that modal adjectives are used in ordinary speech to "qualify the truth of a statement" and are represented psychologically as continuous values on a unidimensional scale (Reyna, 1981, p. 643).

Although both probability and confidence can be expressed in linguistic terms, their respective numerical representations differ. While probability, in numerical form, can vary between 0 and 1, confidence can only vary between 0.5 and 1, where 0.5 represents complete uncertainty and 1 represents complete certainty. By way of example, consider a coin toss. One may, using an array of data gathering and interpretative strategies, determine that the probability that a coin will turn up heads on any one toss is 0.5 or 50%. This probability is, for all intents and purposes, indisputable. On the other hand, an individual, if asked to state his degree of confidence that a particular coin toss will turn up heads, has many more options available to him. Although he would be wise to say 50 %, this individual, noting that the five previous tosses turned up tails, could likewise say that he is 90% confident that the next toss will turn up heads. Given this conceptualization of confidence, it is clear that producing a confidence assessment implicates a complex network of cognitive functions, including data search and evaluation strategies, and at the end of this process judgment confidence may be experienced as an intuition or a "feeling" of uncertainty.

What is overconfidence? The construct overconfidence has been operationalized in various ways, but the most common measure used by researchers is termed "calibration" (see Lichtenstein et al., 1982). This method involves making observations of the correctness of judgments, on the one hand,

and the correctness of confidence levels assigned to those judgments, on the other. Overconfidence occurs when the mean of confidence ratings assigned to a series of judgments exceeds some objective and comparable measure of their accuracy. Other measurements of overconfidence, including "resolution" (i.e., the degree to which correct responses are assigned higher confidence ratings than incorrect responses; see Sharp, Cutler, & Penrod, 1988) and "confidence intervals" (i.e., specifying an interval having a specific probability of containing an unknown quantity; see Plous, 1995) are essentially variations of calibration inasmuch as they all reflect a quantitative relationship between confidence and accuracy.

While these measures are suitable for gauging the appropriateness of confidence in tasks involving judgments of facts, they are less applicable to ill-structured diagnostic tasks in which it is difficult, sometimes even impossible, to delineate a single correct diagnosis. Moreover, in applied fields, it is extremely important to practitioners to know whether or not variability in their judgments (and the levels of confidence expressed therein) entails clinically meaningful consequences. Another method of measuring overconfidence that accounts for these issues (and a method that has received considerably less attention than calibration) involves the complementary notions of an "action threshold" and "consequential variation" (see Baumann, Deber, & Thompson, 1991). In this formulation, using an action threshold delimits the definition of overconfidence to

those circumstances in which varying degrees of confidence give rise to different courses of action. The concept of consequential variation further delimits the definition to situations in which confidence ratings falling on different sides of an action threshold differentially affect clinical outcomes. In this framework, poor calibration is relevant only when it is associated with divergent intervention strategies that differentially influence treatment outcomes. By this definition, only clinicians whose inferences eventuate in these anomalies are deemed to be overconfident. While this method of operationalizing overconfidence in psychological assessment has much to recommend it, it has not yet been used in this area of empirical inquiry.

Implications of overconfidence. Over the years, various authors have discussed the problem of overconfidence and its implications for clinical judgment (cf. Arkes, 1981; Meehl, 1957). It has been noted that overconfidence seems to obstruct natural learning processes. Overconfident individuals tend not to evaluate their decisions and, therefore, are less likely to learn from their experiences (Faust & Ziskin, 1988). This could lead mental health professionals to perseverate in clinical practices that are invalid and potentially harmful to clients. Overconfidence also may obscure clinicians' ability to realistically assess their competence in making certain decisions or undertaking particular tasks. Arkes, Dawes, and Christensen (1986, as cited in Baumann et al., 1991, p. 167) noted:

"One of the dangers of overconfidence is that one feels that no assistance is needed. If one assumes that his or her judgment is quite good, decision aids would be entirely superfluous."

The clinical sequelae of overconfidence are potentially far-reaching, since it is recognized that clinicians' diagnostic confidence affects their treatment decisions (Garb, 1986). Mental health professionals, like all people faced with difficult decisions, are likely to commit resources on the basis of highly confident assessments (Dunning et al., 1990). The issue of how clinicians commit resources has become more salient since the introduction of managed care, which requires psychologists to justify how they allocate costly health care services. Additionally, clinicians, when they are sure of their judgments, are less likely to undertake "insurance" measures to attenuate negative outcomes in the event of diagnostic error (Dunning et al., 1990). In the most extreme cases, this can entail serious risk to clients' well-being, such as when an acutely suicidal patient or a potentially dangerous patient is misdiagnosed.

Statement of the Problem

The research findings briefly discussed above suggest that the appropriateness of clinicians' diagnostic confidence is affected by a multitude of variables that operate differently across clinical contexts. This is consistent with evidence suggesting that reasoning processes, which implicate heuristics, biases,

and inferential errors, are context-bound and "cannot be adequately described in terms of content-independent formal rules" (Kahneman & Tversky, 1982, p. 130). For example, two variables that mediate the correlation between judgment confidence and validity are (a) length of professional experience and (b) validity of the clinical data on which the judgments are based (Garb, 1986). Both have been shown to vary directly with the appropriateness of diagnostic confidence. Other studies have shown how task characteristics (i.e., level of difficulty, quantity of information and its level of redundancy) affect levels of diagnostic confidence (e.g., Heller, Saltzstein, & Caspe, 1992; Lichtenstein et al., 1982; Oskamp, 1965). Additionally, Dunning et al. (1990) suggested (but did not demonstrate) that inferential biases arising from reliance on heuristics are a source of overconfidence.

A review of the relevant literatures shed light on some of the problems with research that has been conducted in this domain. One problem is conceptual in nature. Earlier studies reported in the clinical judgment literature have addressed the question of whether or not professional psychologists tend to be overconfident in their clinical inferences. Reflecting a shift from description to explanation in this research domain, researchers are now asking the following question: In what contexts does a particular clinician operating with a particular clinical database articulate inferences with unwarranted confidence? Consistent

with this recent trend, this study seeks to identify the variables that impact upon psychodiagnostic confidence.

A second problem bears on an important assumption underpinning research in this problem domain, that is, that confidence in large measure determines post-decision actions of decision makers. In the clinical domain, this means that diagnostic confidence is presumed to influence decisions such as whether or not an individual who is judged to be a poor candidate for treatment will be offered psychotherapy, whether or not hospitalization will be sought for a client assessed to be suicidal, and what specific treatment plan will be implemented given a particular diagnosis. While there are theoretical reasons and some indirect evidence that lend support to this assumption, it has not yet been subjected directly to empirical scrutiny.

A third problem with previous research in this domain is methodological in character. In all studies cited in the foregoing review, psychodiagnostic confidence has been measured exclusively by having participants provide numerical ratings of their degrees of confidence in particular clinical hypotheses. On both empirical and theoretical grounds, this is not an optimal approach. Although it has been argued that decision makers should use precise (i.e., numerical) rather than vague (i.e., lexically descriptive) representations of subjective probabilities in order to optimize decision making (Beyth-Marom,

1982; Bryant & Norman, 1980), several compelling arguments have been made against this recommendation. First, lay people and professionals alike prefer to communicate subjective probabilities verbally rather than numerically because they are more natural, easier to use, and consistent with underlying uncertainty (Merz, Druzdzel, & Mazur, 1991; Wallsten, 1990). Although those receiving this communication tend to prefer numerical probability statements to verbal ones, research has demonstrated that people often misunderstand the events to which the numerical probabilities refer and the statistical meaning of such probabilities (Brun & Teigen, 1988; Murphy, Lichtenstein, Fischhoff, & Winkler, 1980). Second, using verbal probability terms may facilitate thinking about uncertainty. Zimmer (1983) suggested that people, being more familiar with the rules of language than probability, may handle linguistic information better than numerical information. Finally, having to make verbally qualitative probability judgments seems to encourage decision makers to exploit qualitative data they have at hand (Fox, Barber, & Bardhan, 1980; Zimmer, 1984, 1986).

A fourth problem with the method of directly asking clinicians to provide confidence ratings is that it inadvertently draws attention to the variables under study and subsequently alters participants' responses to experimental protocols (Kahneman & Tversky, 1982). Clearly, this raises questions about the quality of data reported in previous studies of psychodiagnostic confidence. Finally,

Kahneman and Tversky (1982) raised broader questions about the question-answering paradigm that predominates the human decision making (and clinical judgment) literature. They suggested that it is inappropriate to assume that inferences generated within an "experimental conversation" simulate perfectly inferences that occur in response to daily interactions with the environment. For example, in contrast to what typically occurs in experimental settings, inferences rarely arise in response to explicit questions in an individual's daily life. Additionally, they noted that study participants apply the "cooperativeness principle" to experimental conversations (cf. Clark & Clark, 1977) and assume that the experimenter is trying to be "informative, truthful, relevant, and clear" (p. 132). On this basis, nothing in an experimenter's questions is presumed to be neutral or irrelevant.

Purpose of the Research

The principal goal of this research project was to investigate the relationship between psychodiagnostic confidence and inferential heuristics. Specifically, this study sought to measure the contribution of four inferential biases evidenced by mental health practitioners when aggregating and interpreting a client casefile to degrees of diagnostic confidence (expressed in linguistic terms) assigned to judgments in the context of a psychological assessment task. The four inferential biases that were examined in this investigation are dispositionalism,

confirmationism, data-search truncation, and narrow problem construal. There is evidence (which is reviewed in the next chapter) suggesting that these biases influence confidence assessments and contribute to overconfidence effects. To date, though, no studies documented in the literature have directly examined this research problem, nor have they examined the role of these heuristics in judgment processes in professional psychology.

For the purposes of this investigation, the four inferential biases are defined as follows:

1. Dispositionalism: This refers to clinicians' tendency to situate problems primarily within clients rather than in the circumstances in which they live (Dumont, 1993; cf. the Fundamental Attribution Error, Ross, 1977).

2. Confirmationism: This refers to clinicians' tendency to perseverate on initial diagnostic impressions throughout an assessment, even in the face of disconfirming clinical evidence (Gauron & Dickinson, 1969; Meehl, 1960; cf. anchoring effect, Tversky & Kahneman, 1974; cf. primacy effect, Asch, 1946; Ross & Nisbett, 1980).

3. Data-search truncation: This bias bears on clinicians' tendency to truncate information-gathering procedures when compiling a clinical database (cf. single-cause etiologies, Dumont, 1993; Nisbett & Ross, 1980).

4. Narrow problem construal: This refers to clinicians' tendency to narrowly construe client problems by formulating and testing fewer than optimal diagnostic hypotheses (cf. situational construal, Griffin, Dunning, & Ross, 1990).

A secondary objective of this study was to improve upon methodological shortcomings of earlier studies by using a data-gathering technique not yet applied to this problem domain, namely the "think-aloud" protocol methodology used in discourse analysis (Ericsson & Simon, 1993). In this procedure, subjects are instructed to verbalize all of their thoughts as they clinically interpret a client file. Their responses are tape-recorded for later transcription and detailed analysis. There are at least three benefits to using this novel approach to measure diagnostic confidence over the more common approach of directly asking subjects to indicate their level of confidence on a Lykert-type scale. First, it allows participants to express their confidence in words rather than numbers, which is the natural and preferred approach among professionals (Merz et al., 1991; Wallsten, 1990). Second, it is a less obtrusive method of gathering data, as it does not draw direct attention to the variables being investigated. Finally, although the think-aloud procedure is still an analog methodology, it roughly approximates some common professional activities of clinicians, especially informal case consultation with colleagues and more formal case conferences. Together, these advantages lend support to the ecological validity of the methodology, thus increasing the potential

meaningfulness of the data and the generalizability of the findings to clinical practice.

Significance of the Research

The vital role that diagnostic assessment (and the confidence expressed therein) play in the delivery of psychological services to the public and the implications it has for the society at large underscore the importance of this research. In a clinical context, accurate diagnosis expressed with appropriate confidence makes possible an effective treatment process (Garb, 1986; McReynolds, 1989). In some legal cases, diagnostic assessments are instrumental in determining the degree of responsibility attributed to an individual in the commission of a criminal act. Finally, diagnostic assessments and the "labels" that inevitably accompany them can detrimentally influence society's perceptions of a person's status, condition, and worth (McReynolds, 1989) and can entail serious personal consequences, such as limitations on career opportunities or advancement.

The ultimate goal of this research, as with the other studies in this domain, is to enhance the integrity of clinical judgment in professional psychology. Confidence is an integral aspect of clinical judgment, and when clinical inferences are articulated with unwarranted degrees of confidence, the validity of those inferences is proportionately diminished. Confidence assessments, to be valid,

must accurately reflect the foundations of the judgments they qualify. In professional psychology, these foundations necessarily include the accuracy of the information on which the judgments are based, the efficacy of the methods and instruments used to collect clinical information, and, more broadly, the canonical knowledge of the discipline. In short, valid confidence assessments embody the realistic limits of our ability to know. Psychodiagnostic confidence that is artificially inflated by inferential biases arising from uncritical use of heuristics threatens the validity of clinical judgment. Understanding what variables contribute to confidence and overconfidence is a necessary antecedent to developing corrective procedures and training modules to sensitize clinical trainees and professional psychologists alike to the pitfalls inherent to clinical judgment.

CHAPTER II: LITERATURE REVIEW

The Overconfidence Effect

Some generalizations. Research on judgment confidence has consistently demonstrated that the subjective certainty that people express in their judgments usually exceeds the overall accuracy of those judgments (Lichtenstein et al.,1982; Dunning et al., 1990; Vallone et al., 1990). The so-called "overconfidence effect" is evidenced across a broad range of judgment tasks. Several generalizations about the overconfidence effect have emerged from the research reported in the literature. One is that difficult tasks yield the highest levels of overconfidence and, conversely, easier tasks systematically yield less overconfidence (Lichtenstein et al., 1982). In fact, for easy tasks in which subjects achieve high accuracy rates, underconfidence rather than overconfidence often results. In an effort to explain this well documented pattern, Block and Harper (1991, experiment 1) had two groups estimate the upper and lower limits of either familiar quantities or unfamiliar quantities such that there would be an equal likelihood that this range would contain the true quantity as not contain it (called a 50% confidence interval). Participants estimating unfamiliar quantities produced wider intervals than those estimating familiar quantities. This suggests that people seem to be aware of their limited knowledge when engaged in difficult estimation tasks (as

evidenced by the wider confidence intervals) but do not increase the intervals enough to compensate for these limitations.

Another robust finding is that high levels of judgment confidence are usually associated with high levels of overconfidence (Fischhoff, Slovic, & Lichtenstein, 1977; Dunning et al., 1990). Although highly confident predictions are associated with higher accuracy rates, confidence tends to increase at a rate that disproportionately exceeds the rate at which accuracy increases. Consequently, highly confident predictions tend to yield the highest levels of overconfidence.

Finally, accumulating evidence suggests that overconfidence is more likely to be extreme when decisions are made spontaneously and with relatively little reflection (Sniezak, Paese, & Switzer, 1990). A number of studies have demonstrated that overconfidence decreases as the amount of cognitive processing involved in a particular decision-making task increases (Block & Harper, 1991; Paese & Sniezak, 1991; Sniezak et al., 1990; Zakay, 1985). The effect of diminished overconfidence on these tasks is attributable to both decreased levels of expressed confidence and increased decision accuracy. An implication of this finding is that overconfidence can be reduced by having decision makers consider evidence bearing on *multiple* solutions for any particular decision task. Moreover, such a strategy may help to improve judgment accuracy.

Confidence and clinical judgment. Few documented studies addressed overconfidence directly within the field of professional psychology. Oskamp (1965) had clinicians, graduate students, and undergraduates read a case, answer questions bearing on the target individual's personality, and indicate the level of confidence in their decisions. All three groups manifested overconfidence in their judgments at each of four prediction times. Moxley (1973) had graduate psychology students and clinicians predict the duration of counselling at four different times for each of several cases. Successive judgments were based on incrementally increasing amounts of information. Participants were overconfident in their judgments at each of the four prediction times. Wedding (1983) had clinicians classify patients into one of five diagnostic categories (i.e., schizophrenia, left-hemisphere impairment, right-hemisphere impairment, diffuse brain damage, or normal) using the Halstead-Reitan. In contrast to previous findings, a majority of participants (8 of 14) were found to be underconfident in their judgments. Faust et al. (1988) presented neuropsychologists with pairs of casefiles comprised of diagnostic test results (e.g., Halstead-Reitan, Wechsler Adult Intelligence Scale–Revised, Wechsler Memory Scale, etc.). For each of the two cases evaluated, clinicians indicated whether the patient was malingering or suffering from neurological impairment. Accuracy on this task was well below chance at a 13% true-detection rate. However, 97% of the clinicians in the sample

were moderately confident or more than moderately confident that their judgments were accurate, and 70% were highly or very highly confident.

A larger body of research has evaluated the appropriateness of clinicians' confidence in their diagnostic judgments. In general, this research, which has been correlational in design, has investigated the effects of moderating variables on the relationship between judgment confidence and judgment accuracy. The effect of one moderating variable, namely length of experience, has been probed in a number of studies. In Goldberg's (1959) widely cited study, participants (experienced clinicians, clinical trainees, and secretaries) differentiated "brain-damaged" patients from "psychiatric" patients on the basis of Bender-Gestalt protocols and provided confidence ratings. There were no significant differences in diagnostic accuracy across groups, but there were significant differences regarding their diagnostic confidence. Results revealed an inverse relationship between level of clinical experience and mean confidence rating, with lay participants, on the whole, expressing more confidence in their judgments than experienced psychologists. Oskamp (1962) asked experienced clinicians and students to judge 200 MMPI profiles as indicative of either psychiatric or medical dysfunction and assign a confidence rating to each of their judgments. He reported similar results: the experienced clinicians expressed significantly less confidence in their judgments than the inexperienced judges. Clinicians' tendency toward

lower confidence ratings (but not greater accuracy) resulted in significantly better calibration of confidence levels with judgment accuracy compared to the inexperienced judges. Levenberg (1975) had judges (clinicians, interns, secretaries, and one expert) discriminate Kinetic Family Drawing protocols obtained from children in psychiatric treatment from those belonging to normal children. Results were consistent with earlier findings: the groups did not differ significantly with respect to accuracy. (Of note, the expert performed the worst on this task.) However, with judgment confidence taken into consideration, level of experience correlated positively with the appropriateness of participants' clinical judgments.

Friedlander and Phillips (1984) reported data that conflicted with previous findings. They had undergraduates assess a casefile on two dimensions, severity and prognosis, using the Global Assessment Scale (GAS) and Axis V of the DSM III, respectively. Post hoc analyses indicated that while the students' assessments did not differ significantly from those of the experienced clinicians (who were administered the same procedure in Friedlander and Stockman [1983]), their confidence ratings were significantly lower than clinicians' ratings. Additionally, the clinicians evidenced an anchoring bias, whereas the students did not. The authors speculated that the more confident clinicians had more elaborate clinical prototypes than the students. According to their hypothesis, because new data

were more easily accommodated by the clinicians than inexperienced students, it was less likely the clinicians ever considered that their initial impressions might be in error. The students, on the other hand, might have had more difficulty accommodating new clinical data leading them to make more frequent adjustments to their problem formulations. The authors suggested that these factors possibly undermined their diagnostic confidence.

On the basis of this hypothesis, Richards and Wierzbicki (1990) expected to find an anchoring effect that would be more pronounced among participants who expressed greater degrees of confidence. The predicted interaction between confidence and anchoring did not materialize. A critical methodological difference between this study and Friedlander and Phillips (1984), however, was that the former researchers compared confidence levels within-subjects across their sample of undergraduates students whereas Friedlander and Phillips contrasted the group means of experienced clinicians and undergraduates. These results suggest that the anchoring bias in clinical judgment arises more from features inherent to clinical experience, such as sophistication of clinical prototypes, as Friedlander and Phillips proposed, rather than just simply from higher confidence levels.

Garb (1986; 1989), who reviewed the research bearing on confidence in clinical assessment, concluded that there is only weak evidence supporting the

notion that experienced clinicians tend to make more appropriate confidence ratings than inexperienced judges. He stated that the relative scarcity of studies that bear on this issue along with the fact that this research has been overwhelmingly correlational in design weakens the conclusions that can be drawn from the data. One problem with this type of design is that clinicians who express more appropriate degrees of confidence in their inferences can still be overconfident (cf. Levenberg, 1975).

Some researchers have studied the effect of increasing amounts of information on the relationship between judgment accuracy and confidence levels. Oskamp (1965) found that increasing the amount of information available to judges served to significantly increase confidence over the four stages of the casefile (33, 39, 46, and 53% respectively) in the absence of a corresponding increase in the accuracy of those judgments (which leveled off around 27%). This led the author to conclude that "a psychologist's increasing feelings of confidence as he works through a case are *not* a sure sign of increasing accuracy for his conclusions" (p. 265). Richards and Wierzbicki (1990) had undergraduates assess the level of pathology in four clinical cases, each of which was divided into five paragraphs that were sequentially presented. Results revealed that subjects' confidence in their judgments escalated significantly over the five successive ratings, thus confirming Oskamp's earlier finding.

Another important variable mediating the relationship between judgment accuracy and confidence is the validity of the clinical data on which judgments are based (Garb, 1986). Generally, the more valid the clinical data, the stronger the correlation between validity and confidence. For example, while Moxley (1973) also demonstrated that incremental increases in the data provided to subjects leads to corresponding increases in confidence ratings, she found that confidence ratings were increasingly appropriate (in relation to judgment accuracy) at each of four informational levels. The dissimilarity between this and other studies reporting divergent results may be accounted for by the quality of the information provided to participants. In short, it appears upon examination that the information supplied to judges at successive judgment times in Moxley's study was more informative and clinically valid than the information provided incrementally in other studies (cf. Garb, 1984; 1986). Similarly, Heller, Saltzstein, and Caspe (1992) demonstrated that the informational value of the data can affect confidence ratings. In their study, medical residents estimated the probability that their medical and non-medical inferences were accurate based on information presented in list form. In the experimental condition, the list was longer than in the control condition because redundant information had been added. Virtually all participants expressed greater confidence in judgments based on the list with

redundant information than on the list with no redundant information for both medical and non-medical problems.

There is evidence that environmental variables can alter decision making processes and lead to overconfidence. Schaeffer (1989) had groups of participants respond to a questionnaire that was designed to detect the extent to which they relied on cognitive heuristics to answer questions. Participants provided confidence ratings along with their responses. One experimental group completed the task while being exposed to an unpredictable and uncontrollable stressor (a loud noise) while another completed the task following exposure to the stressor. Participants completing the task following exposure displayed (a) significantly more reliance on cognitive heuristics and (b) significantly greater confidence in their decisions than those completing it during exposure. This suggests that the effects of environmental stress on decision making appear following the cessation of a stressor and not during exposure to it. Schaeffer proposed that stress acts on decision-making processes by constricting data gathering patterns and diminishing problem solving abilities. These have serious implications for clinicians and their patients, both of whom often make critical decisions in the aftermath of personal crises.

Anchoring Errors and Confirmatory Bias

There is abundant evidence that clinicians formulate hypotheses about a client's problems quickly and on the basis of very limited data (cf. Dumont, 1993). Asch's (1946) *primacy* and *recency* effects, which were identified in his extensive social psychological research, are early evidence of this tendency. These two effects refer to the undue influence of information that is presented early (in the case of primacy effects) or presented late (in the case of recency effects) on social judgments. Nisbett and Ross (1980) concluded that "several decades of psychological research has shown that primacy effects are overwhelmingly more probable" (p. 172).

Tversky and Kahneman (1974), working within an information processing framework, delineated several heuristics that people use to compensate for the limitations of their information-processing capabilities and to streamline inferential processes. When using the *anchoring and adjustment* heuristic to estimate uncertain quantities, an individual starts with a specific value and then adjusts it to yield a final answer. Typically, the adjustment is insufficient, and final judgments are biased toward initial values. These authors proposed that the widely documented overconfidence effect "is attributable, in part at least, to anchoring" (p. 1130).

Block and Harper (1991) reported data supporting the anchoring-and-adjustment heuristic and its role in producing this effect. However, they observed that anchoring does not invariably lead to overconfidence and suggested that the anchoring process is more complex than first thought. In a series of experiments, these researchers had participants estimate uncertain quantities under various anchoring conditions. Those who had to explicitly generate and state a point estimate (the anchor) before providing a specified confidence interval (the final answer) displayed significantly less overconfidence than participants in all other anchoring conditions, presumably because it sensitized them to the difficulty of the task. Block and Harper proposed that overconfidence may result from the unrealistic assessment of one's estimation ability rather than the anchoring process *per se*.

Within the domain of professional psychology, the tendency of clinicians to become anchored in their early impressions of clients and to inadequately adjust these impressions as new information becomes available has been widely documented. Meehl (1960) demonstrated that clinicians diagnose a client's problems within the first few sessions and that these formulations remain largely unchanged after 24 sessions. Similarly, Gauron and Dickinson (1969) showed that clinicians' diagnoses, formulated within moments of seeing clients, resist significant alteration thereafter. Mahoney (1976) observed that: "The scientist is

not a paragon of reason. In fact, he may often be expediently illogical and prejudicially confirmatory" (p. 161).

In a more recent study of clinical judgment, Friedlander and Stockman (1983) hypothesized that clinicians who are presented pathognomonic information early in a casefile would display an anchoring effect, assessing the client as more pathological than clinicians who encountered the same information later in the casefile. Results revealed an anchoring effect in clinicians' assessments of a moderately disturbed client but not of a severely disturbed client. In a follow-up study, Friedlander and Phillips (1984) attempted to mitigate the anchoring error by warning experimental subjects about it and instructing them how to avoid this pitfall. Their subjects, undergraduate psychology students, were presented with the casefile of the moderately disturbed client, which had elicited a robust anchoring effect in Friedlander et al. (1983). However, there was no evidence of anchoring errors for either the experimental or control group in this study. In Richards and Wierzbicki's (1990) investigation of anchoring in clinical judgment, undergraduate psychology students assessed four casefiles and displayed a strong anchoring effect for two of the four cases and a moderate and modest effect for the other two cases, respectively.

In an analogue study, Strohmer, Shivy, and Chiodo (1990) examined the way in which counsellors collect and remember information about a client. They

found that their participants selected and remembered more confirmatory than disconfirmatory information, even when the information they read about a client contained disproportionately more disconfirmatory information. Additionally, participants displayed higher levels of confidence about the accuracy of a clinical hypothesis (which had been provided by the experimenter) as the amount of confirmatory information they recalled increased.

The pattern of findings on confirmationism in earlier research is not completely uniform, and divergent findings have been reported. For instance, Strohmer and Chiodo (1984) presented data suggesting that (a) the confirmatory bias is not as pervasive among counsellors as alleged and (b) some of the confirmatory effects reported in the literature (e.g., Snyder, 1981) are, actually, methodological artifacts. They had counsellors (novice and experienced) develop questioning strategies to test a particular clinical hypothesis. In the experimental conditions, the hypothesis was (a) generated by the participants in order to enhance their personal investment in the hypotheses to be tested or (b) consistent with participants' self-schemas (e.g., an extravert testing whether a client was extraverted). Results revealed no evidence of a penchant for confirmatory testing strategies, and, in fact, counsellors demonstrated a strong preference for unbiased questioning strategies in all conditions. However, these data, which bear on information gathering procedures used in the earliest stages of counselling, do not

rule out the possibility that a confirmatory bias is implicated in other clinical processes (e.g., interpretation of data) or at later stages in counselling.

In a study of anchoring and confidence in clinical judgment (Lee, Barak, Uhlemann, & Patsula, 1995), two groups of clinical trainees interpreted a client's casefile and provided confidence assessments with their judgments. In order to induce an anchoring effect, the experimenter presented participants with two different segments of client information prior to beginning the assessment task. Results revealed no significant anchoring effect arising from presentation of pre-interview information about the client, nor was there any significant difference between the two groups in their levels of psychodiagnostic confidence. The data did reveal, however, a significant tendency toward confirmatory memory on the part of participants, but only after the passage of time (in this case, several weeks).

Dispositionalism

Another inferential bias that people evidence as a result of their efforts to reduce demands on their limited information processing capacities is referred to as "dispositionalism" (Dunning et al., 1990) or the "Fundamental Attribution Error" (Ross, 1977; Ross & Nisbett, 1990). This phenomenon occurs when observers overestimate the role of broad personality traits in motivating an actor's behaviour and simultaneously underestimate the impact of situational factors. It is an

abundant source of inferential errors. This tendency is manifested in people's willingness to make judgments about actors' personalities based on particular behaviours, while inadequately adjusting for situational cues and constraints faced by the actors (Ross, 1987). Experiments designed to probe the Fundamental Attribution Error have convincingly demonstrated that subjects are apt to draw dispositional inferences about actors in a broad range of circumstances whose even when the latter's behaviour is motivated less by personality traits than by situational factors.

Observers tend to make dispositional inferences about actors when their behaviour is or appears to be exceptional, that is, when the behaviour is contrary to known base rates or to what observers assume to be normative responses (Paese & Kinnaly, 1991; Ross, 1987). Furthermore, studies have demonstrated that these inferences are expressed with higher levels of overconfidence than when an actor's behaviour is consistent with the known base rates or simply the observer's behaviour (Dunning et al., 1990; Paese & Kinnaly, 1991; Vallone et al., 1990). Presumably because judgments are made against base rates, the accuracy of these judgments is usually significantly worse than judgments that are consistent with base rates.

There is evidence that mental health professionals might be more susceptible to this dispositional bias than non-professionals (Dumont & Lecomte,

1987). Batson (1975) tested Jones and Nisbett's (1971) finding, that observers are more likely to make dispositional attributions about actors than the actors about themselves, using a clinical assessment task with trained and untrained helpers (or observers, in terms of Jones and Nisbett's framework). Participants listened to recorded interviews in which the clients (the actors in this context) complained of circumstantial problems they were facing. Participants were instructed to indicate the locus of the problem (that is, in the client or in the environment) and make a treatment referral to an institution that would provide suitable help. In spite of clients' assertions to the contrary, trained helpers overwhelmingly perceived the client as the problem and subsequently made treatment referrals that were directed toward changing the client rather than the environment.

Paese and Kinnaly (1991) sought to determine (a) whether or not assigning social roles and providing opportunities to verbally interact served to increase people's tendency to rely on individuating (i.e., dispositional) cues; and (b) how these factors impacted on judgment accuracy and confidence. They had participants in the experimental group assume the role of a "job interviewer", which, like the role of counsellor or psychologist, entails an implicit expectation to use individuating information to make professional judgments. Their task was to predict responses of "interviewees" (whom half of the interviewers had the opportunity to interview) on a work-values inventory. Results revealed that both

role assignment and verbal interaction increased participants' tendency to rely on individuating cues and subsequently make interpersonal judgments contrary to base rates (which were inferred from interviewers' own responses to the inventory). Consistent with previous findings, this tendency led participants (in the role condition only) to make less accurate judgments about targets and display more overconfidence in their judgments compared to controls.

The Problem of Situational Construal

Ross (1987) proposed that, of the many possible determinants of the overconfidence people often evidence, one of the most powerful arises from people's failure to understand the role of the construal process when predicting others' behaviour. In order to understand, predict, and control events in their social environment, people are required to discern the minute details of a situation and anticipate their impact on actors. Errors at any point in this perceptual process will give rise to erroneous predictions expressed with unwarranted confidence. Ross, Greene, and House's (1977) account of the false consensus effect illustrates this point. They suggested that observers in their study resolved situational ambiguity by filling in missing details using highly idiosyncratic strategies and then selecting a response they deemed appropriate based on their formulations. From the observers' perspective, anyone choosing a different course of action

would be doing so in spite of the situational demands (as construed by the observer) and, thus, would be manifesting his or her personality traits.

Griffin et al. (1990) demonstrated that people do not make adequate inferential allowance for their uncertainty about relevant details of construals of situations facing actors. Moreover, "to the extent that people naturally and habitually treat their situational construals as if they are error-free representations of reality, their predictions and assessments are bound to be overconfident" (p. 1138). Their data also supported the notion that people typically generate only one construal of an ambiguous situation and then make inferences as if that construal were perfectly correct. The clinical parallel of this phenomenon has been termed "single-cause etiologies", by which clinicians seek an explanation of a client's problem(s), and finding one, go no further (Dumont, 1993). Dumont (1993) argued that such truncated formulations of client difficulties militate against holistic conceptualizations of client problems and could have negative implications for the treatment rendered.

Research Questions

This study investigated the effect of four inferential biases on experienced clinicians' and clinical-trainees' confidence in their problem formulations. The four inferential biases targeted in this study included the following: Clinicians' tendencies (a) to attribute behaviour disproportionately to dispositional factors

and neglect current environmental factors, (b) to confirm initial hypotheses, (c) to truncate data searches, and (d) to narrowly construe client problem formulations. It was hypothesized that psychodiagnostic confidence would vary according to participants' reliance on inferential heuristics (that resulted in biased judgments) to aggregate and interpret a client casefile. Furthermore, it was expected that clinical trainees would evidence higher degrees of aggregated confidence in their problem formulations than the experienced clinicians.

These questions were investigated using archival data in the form of clinician-generated "think-aloud" protocols (see Ericsson & Simon, 1993 for a description and critique of this technique). The conditions under which these data were gathered are delineated in detail below (see Chapter III).

Hypotheses

This study tested the hypothesis that levels of psychodiagnostic confidence will positively vary with participants' manifest reliance on the following inferential biases:

1. confirmationism, that is, a tendency to confirm their initial hypotheses;
2. dispositionalism, that is, the tendency to explain the client's problem with dispositional, as opposed to contextual, inferences;
3. data-search truncation, as evidenced by the range of informational categories used in formulating diagnostic inferences;

4. narrow problem-construal, as evidence by the number of initial diagnostic inferences posited.

The second part of this study examined the relationship between clinical experience and psychodiagnostic confidence using the think-aloud technique as opposed to the question-answering format, which was the data-gathering method used in all previous studies of judgment confidence. Although previous research had identified an indirect relationship between length of clinical experience and psychodiagnostic confidence, the novelty of the think-aloud method in this research domain precluded assuming that a similar group difference would materialize in this study. Therefore, the second research hypothesis for this study was posed as an exploratory question in the following form: Would clinical-trainees differ significantly from experienced clinicians in their aggregate confidence assessments in a clinical assessment task?

CHAPTER III: METHOD

Design

Using a quasi-experimental analogue design (Cooke & Campbell, 1979), this study examined the effects of four independent variables, narrow problem formulation, truncated data search, confirmationism, and dispositionalism, on one dependent variable, psychodiagnostic confidence.

Participants

Thirty-six English-speaking psychological practitioners from the Montreal area were recruited for participation in this study. Half of the participants had trained or were training in clinical psychology and the other half had trained or were training in counselling psychology. Additionally, exactly half of each disciplinary group was experienced professionals with at least five years of full-time clinical experience; the other half was enrolled in either the doctoral program (Ph.D.) in clinical psychology or the master's program (M.Ed.) in counselling psychology at McGill University. All students who participated in this study had completed at least one full year of study and one practicum course as prescribed by their respective programs.

Candidates for participation were selected randomly (using a table of random numbers) from lists provided by (a) the university departments in which they were enrolled for their graduate studies (in the case of students) and (b) the

professional orders in which they were inscribed (in the case of professionals). Those candidates selected through this procedure were contacted by the experimenter by telephone and asked if they wished to participate in a study investigating how clinicians process information about clients. All candidates were informed that they would be monetarily remunerated for their participation upon completion of the tasks. Professionals were paid $60 and students were paid $15.

Stimulus Materials

A dormant casefile (see Appendix C) of a middle-aged man in an outpatient clinic of a local hospital, diagnosed in the mid-1980s as suffering from passive-aggressive personality disorder (301.84 in *DSM-III-R*, but with elements of borderline personality disorder), was edited into 63 segments, each containing one to three sentences. The segmented casefile was presented in a serial fashion on 16 type-written pages to participants individually. The segments, of which there were four to a page, were widely separated on the page to minimize the influence of the succeeding segment on the interpretation of each unit of material. The file was redacted according to the following standards:

1. It contained approximately equal amounts of information about (a) historically remote events and conditions (i.e., dispositional information) and (b) current events and conditions (i.e., contextual information) in the client's life.

2. The client data it contained were drawn from a broad range of informational categories permitting participants to assess the client's psychological status and functioning in several different and important areas. Aronoff (1997), who also studied clinical judgment processes using the think-aloud methodology, classified each segment of a clinical casefile into one of 20 informational categories. Using a slightly modified version of this "kind-of-inference", a post-hoc analysis revealed that the casefile used in this study included the following 15 different informational categories of client data (the total number of segments drawn from each category appears in parentheses): family of origin (22), family of procreation (3), education (5), career and work (14), sport and leisure (3), social relationships (11), romantic and sexual relationships (11), clinical presentation (2), presenting symptoms or illness (19), medical conditions (2), psychiatric conditions (1), non-symptomatic affect (5), living environment (2), demographic data (1), culture and ethnicity (2).

Procedure

The experimenter met participants individually in their own offices or in university facilities provided for clinical interviews. On each occasion, the experimenter first read the instructions aloud and then presented the participant with three practice tasks (see Appendix A for detailed instructions). The last practice task was a short casefile consisting of five segments of client data, which

participants were asked to read and interpret aloud. While assessing the practice case, participants were frequently and explicitly encouraged to verbalize their thoughts as they occurred. They were then presented with the experimental casefile, which they were asked to read aloud and interpret segment by segment. The assessment task concluded with three questions that requested participants to summarize their understanding of the client's problems and their causes and to indicate the information in the casefile on which they based these judgments. Finally, participants completed a questionnaire in writing that sought information about their educational and professional experiences and posed several questions related to the casefile. Each experimental session was audio-taped for later transcription.

Coding the Independent Variables

The verbal protocols collected through the think-aloud methodology were previously segmented and coded on all but one of the coding dimensions (i.e., psychodiagnostic confidence) used in the present study (see *Coding Dimensions* below). The segmentation and coding of verbal protocols were completed by two graduate students from the Department of Educational and Counselling Psychology (the author of this study being one of them) over a period of about 8 months between 1991 and 1992. This work was completed within the context of another study of inferential processes in clinical judgment.

Segmenting the protocols involved dividing the verbal output of participants, which had been transcribed from the audio-tapes, into units that contained a single idea or thought. The length of segments varied from one word to a phrase or complete sentence containing numerous words. The following are two consecutive segments (separated by "/") articulated by a participant that illustrate the degree of variation typically seen in the think-aloud protocols: "Well it's an important symptom that could indicate a variety of things, ranging all the way from an organic difficulty / to social phobia."

Both raters coded all of the 36 protocols individually. After completing the coding of each of the 36 protocols, the two raters met to compare their work and to discuss and resolve any discrepant codes. This was done in the interests of having interrater agreement at 100% so that a single segment had only one code. This double-coding technique was used in order to circumvent occasional oversights on the part of one or the other rater in this demanding task that required fine-grained analyses and judgments. The large majority of these discrepancies were resolved quickly and easily to the satisfaction of both raters. On the rare occasions that the raters were unable to reach agreement on a segment code even after extensive discussion, the raters consulted the director of the research project for a tie-breaking vote.

Coding Dimensions

Every segment in the verbal protocols was coded on four different dimensions. Three of the dimensions consisted of two coding categories, and one dimension consisted of three categories. The coding dimensions along with the categories with which each is comprised are defined as follows:

Inferential dimension: An *inference* is defined as a statement that not only transforms the text of the case history but goes beyond what is explicitly contained in that text or is necessarily implied by it. This coding category includes, but is not limited to, judgments, hypotheses, diagnoses, hunches, guesses that relate directly to the experimental casefile. A *non-inference* is a segment that simply repeats or paraphrases what is stated in the casefile. All statements made by participants were classified as being either an inference or a non-inference. All statements coded as non-inferences were subjected to no further analysis.

Problem formulations: The first category on this dimension, *diagnostic*, includes all inferences that indicate the presence of clinically relevant problems or disorders, either psychological or organic in nature (e.g., "This guy is looking more and more depressed."). These judgments refer to client behaviours that are (a) inappropriate or excessive under given circumstances or (b) of such intensity that they significantly interfered with some aspect of the client's functioning. The

second category on the problem-formulation dimension includes all *etiologic* inferences, which are inferences that make reference to factors, either temporally recent or remote, that are causally related to the client's problems (e.g., "The client endured punishing abuse at the hands of his father while growing up."). Only diagnostic inferences received further consideration in this study; etiologic inferences were given no further attention.

Attributions: The first category on this dimension includes all *dispositional* inferences, which are attributions that explain behaviour primarily by reference to long-standing intrapsychic dynamics or personality traits (e.g., "It appears that this man has lived with social anxiety all of his life."). By making dispositional attributions, clinicians situate problems within the client and view them as being rooted primarily, but not exclusively, in historically remote events. On the other hand, *contextual* inferences are attributions that explain behaviour by reference to contemporary conditions or events, in other words, the context in which the client presently lives (e.g., "He is very distressed following the deaths of all these people that were close to him."). As such, client problems are viewed as being precipitated by external conditions. All inferences were assigned to one of the two categories on the attribution dimension.

Hypothesis testing: The first of three coding categories on this dimension, called *initial*, includes clinical inferences stated by participants for the first time.

Inferences in the second category, called *confirmatory*, include (a) all repetitions of initial inferences and (b) all inferences that are directly linked to any initial inference and serve to lend support to the latter. The third category, referred to as *disconfirmatory*, includes (a) inferences that explicitly disconfirm an initial inference and (b) all repetitions of disconfirmatory inferences.

As mentioned above, not all segments in the think-aloud protocols were included in this investigation. The basic data set for this study consisted only of inferences that were classified as diagnostic; all other segments in the verbal protocols were ignored. The majority of protocol segments that were excluded were non-inferences, the largest proportion of which were requests from participants for specific information about the client described in the casefile (e.g., "I'd like to know whether or not he ever lost consciousness as a result of beatings from his father."). The remainder of inferences excluded from this investigation were all etiological inferences about the client's problems. This was done in order to eliminate "noise" in the data and facilitate clearer interpretations of the results. It seems safe to assume that there is greater agreement among mental health professionals on the kinds of psychological syndromes and disorders that exist than on the causes of such dysfunctions. Moreover, etiology more than diagnosis appears to vary as a function of theoretical orientation. The preeminence of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-IV*; American

Psychiatric Association, 1994) among nosological systems, at least in North America, is likely most responsible for this phenomenon. The *DSM-IV* is a descriptive system that explicitly attempts to stay neutral with respect to the causes of mental disorders (p. xviii). In doing so, it cuts across the boundaries of theoretical orientation in order to appeal to the widest possible range of users. On this basis, the author decided not to include etiological inferences in the present study, as it would risk introducing theoretical orientation as a confounding variable that would hinder clear interpretations of the data.

Reliability of the Coding Scheme

The two individuals who undertook the coding of verbal protocols received extensive training prior to commencing the coding task. Briefly, in the first part of the training, the coders were oriented to the coding criteria through reading materials and group discussions under the tutelage of the investigators directing the research project. Throughout the training, the coders completed numerous coding exercises on selections of transcribed sessions from the verbal protocols. Their work was reviewed and discussed in detail by the coders themselves and the investigators during the regular weekly meetings of the research team. The actual coding of protocols did not occur until the coders demonstrated their readiness for the task as evidenced by acceptable levels of interrater agreement. Unfortunately, data bearing on the interrater agreement for

this coding task are not available to the author and, therefore, cannot be reported here. However, two closely related investigations into clinicians' inferential processing of casefile material using the think-aloud methodology classified data on the same dimensions that were used in the present study (Aronoff, 1997; Goodin Waxman, 1991). The levels of interrater agreement for the coding done in these investigations provide support for the reliability of the coding scheme used in the present study.

In both of these investigations, interrater agreement was assessed on an item-by-item basis; that is, the codes for all dimensions for a single segment were examined simultaneously, and disagreement on any one dimension yielded a negative result for that particular item. Goodin Waxman (1991) had pairs of trained coders, following a period of extensive training, classify all inferences contained in a set of clinician-generated think-aloud protocols on several dimensions, including the attribution and hypothesis-testing dimensions used in the present study. She reported that on average 95% of all inferences in the think-aloud protocols were identically classified within rating pairs. In another study using the same two dimensions (plus a third dimension not relevant to the present study), a similarly high degree of consistency was found in the classification of inferences in clinician-generated think-aloud protocols (Aronoff, 1997). In this study, a pair of coders, following a period of training, identically classified 87.5%

of protocol segments on the three dimensions. In both studies, a strong practice effect was evident in the coding task. Initial measures of interrater agreement fell around 60%, and these numbers rose steadily through the training phase to a satisfactory level before the actual coding commenced.

Coding the Dependent Variable

For the purposes of the present study, the same 36 verbal protocols were re-coded by two different graduate students from the Department of Educational and Counselling Psychology on the psychodiagnostic-confidence dimension. Unlike the previous coding for the independent variables, protocols were not double-coded on the confidence dimension. Instead, ratings from a single rater were the unit of study in this investigation. This decision was made as success in this task depended more on raters following the explicit criteria developed by the investigator than on their ability to make discriminating judgments about protocol segments. In the earlier coding of the independent variables, the double-coding technique served to compensate for the fallibility of the coders' judgment. Because of the relative simplicity of rating confidence in the verbal protocols, this technique was deemed unnecessary and, therefore, dropped because of its considerable cost, both in terms of time and money. It was subsequently observed by the author that considerably less time was required to train the raters for this task (i.e., to reach acceptable levels of interrater agreement) compared to the time

required to train the assistants who coded the independent variables. This finding was taken as support for the decision not to double-code protocols on the confidence dimension.

Training of raters. Training of raters took place over two months in 1996. During this time, the raters met with the author on a total of six occasions. Half-way through the training process, one of the two raters withdrew from the task and was subsequently replaced by another graduate student from the same department. Both the author and the rater who remained participated in the training of the replacement rater.

Throughout the training phase, the author and the research assistants met regularly to discuss all matters relevant to the task. Selections of segments from verbal protocols were presented to the assistants, whose ratings were solicited and discussed. At the end of each meeting, the raters were assigned homework (rating a selection of protocol segments) that they were instructed to complete individually and bring with them to the next meeting. Succeeding meetings began by reviewing homework and identifying and resolving problems that the raters encountered during the task.

The rating procedure. The rating task involved reading all segments containing diagnostic inferences in each of the think-aloud protocols (which had been identified by the author with a highlighter) and determining the level of

confidence expressed in these inferences. To do this, the raters were instructed (a) to select the word or words in a segment that expressed a participant's confidence and then (b) to assign the verbal confidence expression a numerical rating between .50 (complete uncertainty) and .99 (complete certainty) inclusive. In some instances, a single inference was qualified by more than one verbal confidence expression. In such cases, all confidence expressions were rated, and a single rating for the segment was derived by averaging the ratings for the various expressions in the segment.

To assign a numerical confidence rating to a protocol segment, the rater simply selected the numerical equivalent for the particular verbal confidence expression(s) identified in the protocol segment from an extensive list of numerical equivalents for such expressions that had been provided to the raters (see Appendix G for the complete list of numerical equivalents). This list was generated from published studies that have mapped verbal probabilistic expressions (numbering several hundred in total) over the [0,1] probability interval (e.g., Budescu & Wallsten, 1985; Lichtenstein & Newman, 1967; Reagan, Mosteller, & Youtz, 1989; etc.). Many of the expressions in the list were researched in more than one study. In instances where multiple numerical equivalents existed for a single expression, the mean probability for that particular expression was derived by averaging the numerical equivalents reported in the

various studies weighted by their respective sample sizes. (See Appendix D for a more detailed review of the research literature on linguistic expressions of probability.)

The foregoing procedure applied only to clinical inferences that were explicitly qualified by verbal confidence expressions. However, inferences expressed with complete uncertainty (e.g., "Is this man depressed?", "I don't know whether this is an organic problem.", "He may or may not be a social phobic.") or, conversely, with complete certainty (e.g., "This man is depressed.", "Here's his paranoia surfacing again.") are unqualified by confidence expressions. Consequently, these segments were respectively rated (a) with the lowest rating on the scale (.50) when they expressed complete uncertainty or (b) with the highest rating on the scale (.99) when they expressed complete certainty.

Validation of the Rating Procedure

As this was, to the author's knowledge, the first attempt to develop and use such a procedure to rate linguistic confidence expressions in think-aloud protocols, it was necessary to investigate the validity of the procedure. An important question bearing directly on its validity is whether or not there is adequate consistency, both within and between individuals, in the way that verbal probabilistic expressions are interpreted and used. This was investigated in a validation study that is described below (see *A validation study*).

Several features of the sample of clinicians who participated in this study likely served to enhance the degree of consistency with which they used confidence expressions in their clinical assessments. Since the decision-making context influences the consistency with which verbal probability expressions are interpreted, several contextual variables in this study were held constant in the present study (cf. Brun & Teigen, 1988; Merz et al., 1991; Wallsten, 1990; Zimmer, 1984). In the first place, the people who participated were either professional psychologists or training to be so. It was expected that using a more homogeneous sample (in this case, clinicians who shared relatively similar educational and work histories) would promote consistency in the manner in which they used verbal confidence expressions in the assessment task (Nakao and Axelrod, 1983). Secondly, participants in this study were asked to clinically interpret the same clinical database, a single casefile that was presented to all participants.

A validation study. To further verify the validity of the rating procedure, three empirical issues were addressed in a study of how verbal probabilities are interpreted by a sample of psychologist-trainees. As research mapping probability words over the [0, 1] probability interval had not yet been conducted within the field of professional psychology, the generalizability to this field of findings from earlier studies was not yet established. The first objective of this validation study

was to determine whether or not earlier findings could be generalized to this field and used in the development of the procedure for rating psychodiagnostic confidence. This was accomplished by comparing group (i.e., mean) ratings from this study to comparable data reported in the literature on verbal probabilities.

One assumption underpinning the rating procedure is that there is adequate consistency in the intended meanings of psychologists' linguistic probabilities. Although, as indicated above, this assumption has some empirical support, it had yet to be verified within the field of professional psychology. Thus, the second objective of the validation study was to test this assumption by examining the degrees of consistency, both within and between individuals, with which verbal probabilities are interpreted.

Third, a perusal of the 36 verbal protocols gleaned from the think-aloud assessment task revealed that several probabilistic expressions had not been researched in earlier mapping studies. The third objective of this study was to derive numerical equivalents for these as yet unstudied expressions.

In brief, this validation study (see Appendix E for a detailed description of the methodology) employed a sample of advanced graduate students in counselling and clinical psychology. They were presented 15 verbal expressions, each embedded in a different diagnostic statement, and they were asked to indicate (on a 9-point scale) what numerical probability corresponded to each

expression. This procedure was repeated with the same expressions embedded in different statements. In the final part of the questionnaire, the participants were asked to rank the expressions from lowest to highest probability. The 9-point scale was used instead of the more common 1-99 percentage scale because it is more compatible with the actual capacity of working memory, which imposes considerable limitations on the number of information bits people can work with at any time (cf. "The magic number seven, plus or minus two", Miller, 1956). Furthermore, research has shown that using such simpler, coarser scales has a negligible impact on the precision and reliability of the resulting ratings (Brun & Teigen, 1988).

Measures

For the purposes of this study, the five variables (four independent variables and one dependent variable) were operationalized in the following ways:

1. *Narrow problem formulation*: Each participant's score on this variable was operationalized as the total number of different initial diagnostic inferences contained in the think-aloud protocol.

2. *Truncated data search*: This variable was operationalized as the total number of different categories of client information used by each participant in generating all of the initial diagnostic inferences contained in the verbal protocol.

3. *Confirmationism*: In order to measure each subject's tendency to confirm previously stated judgments while controlling for the total of inferences posited, a ratio of the total number of confirmatory inferences to the total number of diagnostic inferences overall was calculated for each subject and constituted the score on this variable.

4. *Dispositionalism*. The total number of dispositional diagnostic inferences posited in the course of the think-aloud assessment task was obtained for each participant. A ratio of the total number of dispositional inferences to total number of diagnostic inferences was calculated for each subject and constituted the score on this variable.

5. *Diagnostic confidence*. The mean of all confidence ratings made within each protocol was calculated, and this value represented each participant's diagnostic confidence score.

Data Analyses

Reliability and validity of the rating procedure. In order to measure the reliability of the rating procedure, the percentage of identically categorized responses between the two raters was determined at four points during the entire task: Two assessments of interrater reliability were made during training (early and late) and two during the actual rating of protocols (early and late). One problem with using percentage of identically categorized responses as an index of

interrater agreement is that it does not take into account the effect of chance on ratings. While statistical techniques have been derived to correct for this limitation (e.g., Cohen's kappa; Cohen, 1960), the nature of the rating procedure in this study renders such techniques inappropriate choices. Using an index of interrater agreement, such as Cohen's *k*, assumes that raters are presented with a discrete number of coding categories. This number is then used in determining the percentage of agreement expected by chance alone, which is included in the calculation of the coefficient. However, in the present study, the actual number of coding categories available to the raters is, for all intents and purposes, indeterminate. This is because the number of possible categories depends on, among other things, the size of participants' vocabulary of verbal confidence expression and the number of words in a particular protocol segment, both of which vary across participants and segments.

In order to assess the validity of the rating procedure, various descriptive statistics were compiled in order to show the levels of consistency in the numerical interpretations of verbal probability expressions. Indices of within-participant consistency were generated for each expression by correlating participants' in-context ratings at time 1 with their ratings of the same expression (presented in a different psychodiagnostic statement) at time 2. A Pearson correlation between these two vectors (i.e., ratings at time 1 and ratings at time 2)

was calculated for each expression. Indices of between-subject consistency were derived by correlating 15 rankings provided by each participant in the ranking task with each set of 15 ranks provided by each of the other 22 participants. A second matrix of intersubject correlations was generated by correlating each participant's set of 15 mean ratings (the mean ratings were derived by averaging the time 1 and time 2 ratings for each expression) with each of the other 22 mean-rating sets. This resulted in two matrices, each containing 253 intersubject correlations.

Finally, to test consistency with which probabilistic expressions are interpreted across different groups of experimental participants, the grand-mean rating (collapsed across rating times [1 and 2] and participants) for each of the expressions was calculated. These data were compared to mean ratings of probabilistic expressions documented in the literature that were researched in the present study. In total, seven previously researched expressions common to this study (i.e., consistent with, could, maybe, perhaps, possible, unlikely, usually, very probable) were examined for interpretative consistency.

Testing the experimental hypotheses. Hypothesis 1 (see *Hypotheses* above) was tested using multiple regression analysis. In general, multiple regression assesses the separate and combined effects of the independent variables on the dependent variable. In the present study, four independent variables, breadth of

problem formulation, scope of data search, dispositionalism, and confirmationism, were regressed onto one dependent variable, psychodiagnostic confidence, in order to determine the best-fitting linear model that accounted for the relationship between the independent and dependent variables. The beta-weights generated in this analysis were then tested in order to determine (a) whether or not a significant proportion of variance in the dependent variable was explained by the independent variables and (b) to what extent each independent variable contributed to the effect resulting from the multiple regression.

Hypothesis 2 was tested using ANOVA. In this procedure, the experience factor, with two levels (novice and expert), was entered as the independent variable and psychodiagnostic confidence as the dependent variable. The observed variance in the psychodiagnostic confidence scores attributable to the experience factor was then tested for statistical significance.

CHAPTER IV: RESULTS

This chapter reviews pertinent findings for the present study and is divided into three main sections. The first section provides some general comments and observations about how the two studies comprising the present research project unfolded. In particular, comments on the validation study bear on the questionnaires returned by participants and comments on the experimental study address the quality of interviews with participants in the think-aloud task. In the second section, results bearing on the first objective of this research project, to develop a procedure to rate psychodiagnostic confidence in think-aloud protocols, are reviewed. Specifically, the reliability and validity of this rating procedure will be discussed and pertinent statistical data presented. Finally, the third section, using statistical and tabular data, addresses the findings of the experimental study, namely the relationship between the four inferential heuristics clinicians used in aggregating and interpreting a client database, on the one hand, and the confidence expressed in their diagnostic inferences, on the other.

General Observations

The validation study. Two batches of 30 questionnaires (i.e., 60 in total) were distributed to candidates through their university mailboxes. From these 60, 36 questionnaires were returned completed. Thirteen of these were ultimately rejected because they were improperly completed. A perusal of the 13 rejected

questionnaires suggested that the same problem afflicted all of them. Specifically, the participants who completed these questionnaires appeared to interpret the verbal probability expressions as confidence expressions. The evidence of this apparent confusion was that similar ratings were given to two probabilistic expressions, "certainly" and "very improbable", that, in fact, lie at opposite ends of the probability interval. Interpreted as confidence expression, however, these two expressions would be situated at the same end of the confidence interval, as they reflect approximately the same level of confidence. As noted previously, the full probability interval ranges from 0 to 1, whereas the full confidence interval ranges only from 0.5 to 1.

Based on these observations, the following criterion was used to eliminate questionnaires with improper responses: If a participant in the third part of the questionnaire ranked the expression "very improbable" in the upper half of the probability interval (i.e., as corresponding to a probability higher than 0.5), then the questionnaire was rejected. Such a rating was considered as evidence of confusion between the concepts of probability and confidence. It should be noted that this was a conservative means of eliminating faulty questionnaires: A few questionnaires evidenced this confusion in one or other of the two in-context rating tasks, but not on the ranking task. Questionnaires falling into this category were retained.

Because this error came to the author's attention after receiving some completed questionnaires following the initial distribution and before the second distribution, slight modifications in the instructions to participants were made in the second batch of 30 questionnaires. These included (a) changing the word "confidence", which appeared once in the first paragraph of the instructions, to "probability" and (b) adding a second example of a diagnostic statement containing a verbal probabilistic expression (see Appendix H). These changes reduced, but did not eliminate, the apparent confusion between confidence and probability, since a smaller proportion of questionnaires was rejected in the second batch.

The experimental study. Without exception, the 36 participants in the experimental study were cooperative with the experimenters throughout the experimental task, completing it from beginning to end as instructed. Overall, nearly all participants completed all think-aloud tasks, both practice and experimental, without any difficulty and without prompting from the experimenters. When problems "thinking aloud" did emerge during the various practice tasks, they were identified promptly and corrected with feedback. Practice tasks were repeated when, in the experimenter's judgment, it seemed necessary. Most participants finished reading and interpreting the experimental casefile in 45-60 minutes. There were only a few exceptions to this pattern.

Among them were the shortest interview, which lasted about 25 minutes, and the longest interview, which lasted about 90 minutes.

During the experimental task, participants were discouraged from lapsing into extended silences. In the rare instances that this occurred, the experimenter prompted the participant by asking, "What are you thinking about?" or something similar. Participants were similarly discouraged from not responding verbally to the casefile material that they were reading aloud. Participants who did not respond to two consecutive segments in the casefile were prompted with a question such as, "Do you have any thoughts about what you are reading?" Such prompts were required infrequently, and never more than once or twice with an individual participant. Occasionally, a participant would pose a question directly related to the casefile. In all such instances, the experimenters responded by stating that they could not answer such questions. Only questions of a procedural nature received direct replies.

A Procedure to Rate Psychodiagnostic Confidence

Interrater reliability. Table 1 shows the percent agreement between the two raters on item-by-item ratings over time for ratings of psychodiagnostic confidence in the think-aloud protocols. At each time interrater reliability was assessed, a different think aloud protocol was rated. Overall, these data show that raters were quite consistent in their ratings throughout the rating task.

Furthermore, these levels of interrater reliability are consistent with the levels of reliability attained in coding the independent variables (Aronoff, 1997; Goodin Waxman, 1991). As in these previous studies, raters became more consistent in their ratings with practice.

Table 1: Interrater Reliability over Time

Time Assessed	% Agreement
During training (early)	62%
End of training	75%
During rating task (early)	80%
During rating task (late)	84%

Validity of the rating procedure. Fundamental to the efficacy of this rating procedure is its validity: Does this procedure measure what it purports to measure? A validation study examined two important questions that speak to the validity of this rating technique (see Appendix E for a detailed description of the methodology). The task presented to participants in this study was constructed such that each participant (N=23) gave two ratings of each of the 15 probabilistic

expression, thus allowing various indices of interpretative consistency to be calculated. In the first section of the questionnaire, each of 15 expressions that participants were instructed to rate was presented in a psychodiagnostic statement. The same expressions were presented in different psychodiagnostic statements in the second section, and participants were instructed to rate them in the same manner as those presented in the first section. For example, the expression "perhaps" was presented in the following two statements: "Perhaps she is denying the true impact that this event had on her" and "Perhaps her mother's death evoked feelings of guilt in her."

To determine the degrees of consistency within-subjects for each of the 15 different confidence expressions, numerical ratings of identical confidence expressions given at times 1 and 2 were compared. Pearson correlations were calculated for all 15 expressions and are reported in Table 2. Correlations ranged from a low of 0.033 ("possible") to a high of 0.896 ("consistent with"), and the mean of the within-subject correlations for all 15 expressions is 0.606. Within-subject consistency was also examined by comparing the rank order of the numerical ratings to the ranks assigned to confidence expression in the third task using a Spearman correlation. The mean of the two numerical ratings given in the first and second rating task constituted the score in the rating vector. This analysis

yielded more modest correlations. The mean of these correlations is 0.284 and the range extends from -0.264 to 0.599. Table 2 lists these correlations by expression.

To determine the degree of between-subject consistency in numerical interpretations of the 15 linguistic probabilities, correlations between the mean of the ratings given at times 1 and 2 in the in-context rating task and the ranks were calculated for each subject pair. Inter-subject correlations for these two data sets are reported in Tables 3 and 4 respectively. The average of the 253 between-subject correlations for the ratings given in-context is 0.705 and for the ranking task is 0.773. These data provide evidence that there is reasonable, but not perfect, consistency both within and among participants in their interpretations of verbal probability expressions. These findings are consistent with previous findings bearing on the consistency of the interpretations of such expressions (Budescu & Wallsten, 1985).

Table 2: Within-subject Correlations by Expression

Expression	*Rating 1* Rating 2*	*Mean Rating * Rank*
very improbable	0.623	0.490
unlikely	0.690	0.490
can't rule out	0.719	0.599
wonder	0.697	-0.264
could be	0.279	-0.004
perhaps	0.392	0.193
possible	0.033	0.298
maybe	0.702	0.295
seems	0.819	0.461
I think	0.855	0.558
tend to believe	0.282	-0.069
consistent with	0.896	0.347
usually	0.728	0.188
very probable	0.778	0.283
certainly	0.591	0.398

Table 3: Intersubject Correlations for Ratings (In-context)

AA	B	BB	CC	D	DD	E	EE	FF	GG	I	J	K	O	Q	R	S	T	V	W	X	Y
0.359																					
0.518	0.795																				
0.837	0.244	0.508																			
0.744	0.350	0.425	0.835																		
0.867	0.428	0.625	0.926	0.844																	
0.885	0.356	0.453	0.876	0.914	0.897																
0.895	0.424	0.557	0.881	0.860	0.869	0.953															
0.788	0.262	0.412	0.857	0.938	0.884	0.942	0.863														
0.822	0.193	0.412	0.892	0.886	0.850	0.926	0.850	0.925													
0.744	0.555	0.701	0.702	0.681	0.810	0.721	0.666	0.684	0.770												
0.388	0.624	0.565	0.198	0.218	0.360	0.203	0.261	0.096	0.098	0.579											
0.534	0.317	0.434	0.362	0.330	0.547	0.364	0.308	0.309	0.385	0.729	0.617										
0.846	0.373	0.589	0.845	0.855	0.936	0.897	0.833	0.915	0.917	0.852	0.244	0.562									
0.820	0.495	0.706	0.743	0.702	0.860	0.748	0.764	0.772	0.737	0.839	0.395	0.560	0.903								
0.751	0.476	0.589	0.815	0.907	0.854	0.907	0.875	0.910	0.891	0.808	0.337	0.345	0.886	0.788							
0.692	0.248	0.450	0.727	0.670	0.702	0.694	0.761	0.691	0.712	0.703	0.336	0.399	0.726	0.801	0.792						
0.841	0.273	0.544	0.889	0.811	0.902	0.909	0.879	0.905	0.897	0.736	0.083	0.408	0.922	0.842	0.817	0.717					
0.889	0.437	0.639	0.868	0.854	0.936	0.902	0.881	0.910	0.879	0.813	0.324	0.424	0.952	0.927	0.898	0.746	0.908				
0.880	0.459	0.658	0.848	0.830	0.906	0.900	0.893	0.900	0.876	0.832	0.292	0.479	0.943	0.941	0.890	0.809	0.948	0.960			
0.822	0.191	0.472	0.910	0.862	0.867	0.927	0.894	0.924	0.970	0.697	0.024	0.306	0.920	0.762	0.877	0.719	0.936	0.897	0.893		
0.861	0.237	0.495	0.936	0.860	0.902	0.920	0.899	0.914	0.938	0.687	0.061	0.284	0.914	0.794	0.843	0.682	0.934	0.931	0.891	0.973	
0.889	0.318	0.548	0.846	0.828	0.874	0.931	0.922	0.886	0.877	0.744	0.215	0.382	0.897	0.838	0.849	0.742	0.946	0.924	0.927	0.909	0.925

Table 4: Intersubject Correlations for Rankings

	AA	B	BB	CC	D	DD	E	EE	FF	GG	I	J	K	O	Q	R	S	T	V	W	X	Y
)	0.468																					
B	0.704	0.682																				
C	0.789	0.529	0.746																			
)	0.521	0.786	0.693	0.775																		
D	0.889	0.611	0.864	0.893	0.696																	
	0.918	0.607	0.825	0.839	0.682	0.975																
E	0.925	0.629	0.857	0.857	0.675	0.979	0.979															
F	0.796	0.668	0.825	0.825	0.746	0.907	0.875	0.918														
G	0.846	0.571	0.836	0.854	0.561	0.929	0.914	0.929	0.793													
	0.696	0.764	0.757	0.811	0.636	0.814	0.746	0.814	0.775	0.875												
	0.250	0.600	0.389	0.336	0.411	0.389	0.339	0.421	0.486	0.339	0.582											
;	0.489	0.282	0.339	0.521	0.171	0.575	0.529	0.575	0.514	0.646	0.657	0.514										
)	0.786	0.718	0.900	0.889	0.768	0.950	0.914	0.950	0.921	0.925	0.889	0.511	0.557									
)	0.757	0.786	0.875	0.825	0.814	0.889	0.868	0.921	0.904	0.846	0.861	0.554	0.539	0.957								
	0.732	0.761	0.911	0.843	0.786	0.918	0.875	0.921	0.886	0.882	0.882	0.539	0.525	0.982	0.975							
	0.618	0.625	0.775	0.904	0.925	0.818	0.768	0.764	0.807	0.689	0.671	0.329	0.236	0.850	0.807	0.836						
	0.854	0.546	0.696	0.846	0.686	0.811	0.793	0.871	0.832	0.757	0.761	0.407	0.411	0.836	0.857	0.821	0.746					
	0.832	0.693	0.882	0.864	0.739	0.961	0.932	0.971	0.939	0.907	0.854	0.482	0.532	0.986	0.950	0.964	0.825	0.864				
	0.829	0.646	0.821	0.893	0.721	0.954	0.914	0.950	0.911	0.907	0.861	0.436	0.546	0.968	0.911	0.932	0.829	0.861	0.986			
	0.729	0.832	0.857	0.836	0.796	0.896	0.871	0.907	0.904	0.882	0.911	0.589	0.589	0.971	0.961	0.961	0.807	0.786	0.950	0.925		
	0.868	0.646	0.889	0.861	0.682	0.975	0.950	0.982	0.907	0.932	0.854	0.464	0.571	0.971	0.932	0.961	0.786	0.857	0.975	0.950	0.918	
	0.800	0.621	0.782	0.854	0.636	0.907	0.850	0.904	0.861	0.889	0.896	0.461	0.525	0.936	0.857	0.904	0.761	0.854	0.946	0.968	0.879	0.932

In order to assess the degree of between-group consistency in the interpretation of verbal probabilities, descriptive statistics for expressions from this study were compared to those reported in other studies that investigated a number of identical expressions. Table 5 lists these expressions, along with their means. A perusal of this table shows that for most expressions there is considerable consistency in the manner that different groups of subjects interpret probability expressions. Two noteworthy exceptions to this are the expressions "unlikely" and "possible" for which there are, respectively, a 23 and 20 percentage-point spread between the highest and lowest sample means. Based on these data, it was concluded that enough consistency exists among the various sample means of numerical equivalents for verbal probability expressions to support their use in developing the rating procedure in the present study.

Table 5: Between-group Comparison of Mean Ratings (in percentages)

	STUDY								
EXPRESSION	L & N, 67 (N=188)	B & N, 80 (N=16)	B & W, 85 (N=32)	K, B, M, & Y, 86 [N in (...)]	H, 77 (N=60)	B & T, 88 [N in (...)]	S, L, T, S, B, & T, 91 (N=25)		S, 98 (N=23)
consistent with		66							67
could					51				50
maybe					50				47
perhaps						43(64)			44
possible	37	43	38	27 (155)	36	50(64)			46
unlikely	18	20	20	14 (150)			11	29	31
usually	77		78		74		80	84	70
very probable	87					86 (16)			82

Legend

L & N: Lichtenstein and Newman (1967); sample of male employees of System Development Corporation
B & N: Bryant and Norman (1980); sample of physicians
B & W: Budescu and Wallsten (1985); sample of faculty and graduate students in university psychology department
K, B, M, & Y: Kong, Barnett, Mosteller, and Youtz (1986); sample of physicians, medical students
H, 77: Hartsough (1977); sample of introductory psychology students
B & T, 88: Brun and Teigen (1988); sample of advanced psychology students
S, L, T, S, B, &T, 91: Sutherlalnd, Lockwood, Tritchler, Sem, Brooks, and Till (1991); sample of cancer patients
S, 98: Smith (1998; this study); sample of advanced graduate students in counselling and clinical psychology

The Contribution of Four Biases to Psychodiagnostic Confidence

In the experimental study, four independent variables, breadth of problem formulation, scope of data search, dispositionalism, and confirmationism, were regressed onto one dependent variable, psychodiagnostic confidence, using the Multiple Regression program of *SYSTAT for Windows, Version 5* (1992). Multiple regression allows one to examine the separate and collective contributions of the independent variables to the variation in scores on the dependent variable. In general, the purpose of this analysis is to select a line that passes through a set of data points such that the average square error (i.e., the distance between the observed scores and the regression line) is minimized. In the present study, it was hypothesized that a statistically significant proportion of the observed variance in participants' scores on the dependent variable, psychodiagnostic confidence, would be explained by the four independent variables, that is, the inferential heuristics used by participants in interpreting the client casefile.

This multiple regression analysis yielded the following best-fitting linear model to explain the relationship between the four independent variables and the dependent variable:

$$Y = -0.001X_1 + -0.002X_2 + 0.081X_3 + 0.237X_4 + 0.58,$$

where X_1–X_4 are, respectively, the independent variables, breadth of problem formulation, scope of data search, confirmationism, and dispositionalism.

The multiple R derived in the present analysis was 0.481. The multiple R statistic represents the correlation between the predicted scores for the dependent variable generated from the regression equation and the obtained scores on the dependent variable. The square of this statistic, the squared multiple R, represents the proportion of variance in the dependent variable explained by the independent variables and in this case equals 0.232. ANOVA was used to test whether or not the variance in the dependent variable explained by the independent (i.e., the squared multiple R) was statistically significant. As is shown in Table 6, this analysis yielded an effect ($F [4, 31]= 2.337$, $p=0.077$) whose corresponding alpha level falls just short of the conventionally accepted alpha cutoff of $p=0.05$.

Table 6: Analysis of Variance: Regression

Source	df	Mean-square	F	p
Regression	4	0.008	2.337	0.077
Residual	31	0.003		

Examining the size of the coefficients permits one to infer the relative contribution of each variable to the variance in the dependent variable. The coefficients for the four independent variables are listed in Table 7. The weight of the dispositionalism variable relative to the weights of the other three independent

variables suggests that the former accounts for considerably more variance in psychodiagnostic confidence scores than the other three independent variables. Table 7 also shows the results of *t*-tests that were conducted on the regression coefficients. Three of the variables, breadth of problem formulation, scope of data search, and confirmationism failed to reach statistical significance. whereas the variable dispositionalism did reach statistical significance (t=2.975, p=0.006).

Table 7: Regression Output for Independent Variables

Variable	Coefficient	*t*	p
Narrow problem formulation	-0.001	-0.518	0.608
Truncated data search	-0.002	-0.360	0.721
Confirmationism	0.081	0.761	0.452
Dispositionalism	0.237	2.975	0.006

Stepwise regression was used in order to determine exactly how much variance each of the independent variables explained in the dependent variable. In the *SYSTAT for Windows* (1992) stepwise (forward) regression program, the variables are entered into the regression equation individually beginning with the variable that accounts for the largest portion of variance in the dependent variable. The sequence continues by entering variables one-by-one into the equation that

account for the largest increase in the squared multiple R, but only if this value equals or is less than a criterion alpha level, which in this case was 0.15. The sequence ends when there are no independent variables that add significantly to the explained variance (i.e., no variables have corresponding alpha values of 0.15 or less).

In the present analysis, the variable dispositionalism was entered at step 1. Its corresponding multiple R was 0.451 and squared multiple R was 0.203. An F-test performed on the latter value reached statistical significance (F=8.684, p=0.006). The stepwise sequence ended at this point, as no other independent variables met the entry criterion. Comparing the results of this analysis to the previous multiple regression and, in particular, the squared multiple Rs revealed that of the 23% of variance in the dependent variable explained by the four independent variables, 20 % is explained by dispositionalism alone. The remaining three heuristics account for only 3% of the variance in confidence scores.

Psychodiagnostic Confidence and Experience

The analysis of variance program of *SYSTAT for Windows* (1992) was used in order to examine the effect of level of clinical experience on the degrees of confidence participants expressed in their psychodiagnostic formulations in the think-aloud assessment task. The single factor included in this analysis consisted of two levels, experienced and novice. Experienced clinicians had a minimum of

five years of full-time clinical experience and the novices were advanced graduate students in clinical and counselling psychology. As in the previous analysis, the mean of all confidence ratings made within each protocol constituted the score on the dependent variable. It had been hypothesized that experienced clinicians would have lower overall confidence than the clinical trainees. Table 8 lists the means for the two groups comprising the experience factor.

Table 8: Group Means for Confidence Scores

Group	Mean
Clinical trainees (N=18)	0.749
Experienced clinicians (N=18)	0.754

The results of the analysis of variance, shown in Table 9, reveal no statistically significant difference between the two groups in their mean confidence ratings ($F [1, 34]=0.061$, $p=0.806$), therefore, the null hypothesis cannot be rejected.

Table 9: ANOVA: The Effect of Clinical Experience on Confidence Scores

Factor	df	Mean-square	F	p
Experience	1	<0.0005	0.061	0.806
Error	34	0.004		

CHAPTER V: DISCUSSION

The present study represents a distinct segment of a growing body of research that is attempting to explain the cognitive processes by which mental health professionals gather information in clinical interviews and use this information to generate hypotheses about their clients. Numerous studies in various disciplines, including social cognition, human decision making, medicine, and clinical and counselling psychology, have demonstrated that the processes underlying human judgment are often flawed and lead to biased, even erroneous, inferences in many instances. This dissertation focuses on one aspect of clinical judgment in professional psychology, namely the confidence with which clinicians express their diagnostic judgments. In particular, the following questions were addressed in a study of clinical judgment in the context of a psychological assessment task:

1. Does reliance on cognitive shortcuts (which implicate inferential heuristics and often eventuate in biased judgments) for aggregating and interpreting a clinical database lead mental health practitioners to express high degrees of confidence in their inferences?

2. Are there differences between experienced and novice clinicians with respect to the overall confidence with which they state their problem formulations?

The data for this study were drawn from 36 think-aloud protocols collected in the context of an earlier study of inferential processes in clinical assessment. The experimental procedure involved presenting a sample of clinicians, both experienced and novice, with a client casefile that they were instructed to read and clinically interpret aloud. The four inferential biases served as the independent variables for the investigation bearing on the first research question above. These variables were operationalized using criteria developed in a number of earlier investigations in related problem domains (Aronoff, 1997; Dumont, 1996; Goodin Waxman, 1991). For the second question, the experience factor with two levels was the independent variable. In both of these investigations, psychodiagnostic confidence was the dependent variable.

In order to operationalize psychodiagnostic confidence, it was necessary to develop a procedure to quantify linguistic expressions of confidence that clinicians used in the course of their verbal analyses of the casefile. A separate study was conducted to ascertain the validity of this procedure. In it, a sample of advanced trainees in counselling and clinical psychology completed a questionnaire that asked them to specify in numeric terms their understanding of 15 linguistic probability expressions. Evaluations of interpretative consistency were made at several levels, including within individuals, between individuals, and between groups. High levels of interpretative consistency would lend support

to the validity of the procedure to rate verbal confidence expressions that occurred in the verbal protocols.

The findings of this study as they relate to each of the objectives described above are reviewed in this chapter. The implications of these findings are also discussed, along with limitations of the study and directions for future research. Before addressing these issues, results of the validation study and their implications for the procedure to rate psychodiagnostic confidence are first reviewed.

The Validation Study

The procedure to rate linguistic confidence expressions in verbal protocols is founded on the assumption that people are generally consistent in the way they interpret and use linguistic expressions of probability. Furthermore, this consistency is high enough to permit the use of norms (derived from aggregate group data) to numerically rate single expressions used by individual study participants. In other words, the intended meaning of linguistic probabilities used by individuals can be approximated within reasonable limits by norms.

Intra-individual consistency. Examination of the ratings assigned to the verbal probabilistic expressions embedded within psychodiagnostic statements revealed a high level of interpretative consistency within individual participants. Ten of the 15 expressions studied yielded correlations exceeding 0.62; an eleventh expression approached this cut-off, falling at 0.59. Correlations for four

expressions fell below 0.4. These four expressions with weak intra-individual correlations fall within the middle of the [0,1] probability interval, which is consistent with earlier findings in this domain (Reagan, Mosteller, & Youtz, 1989). It can be explained in part as an artifact of the boundedness of the full probability interval. Specifically, the structural limitations of this interval (i.e., it starts at 0 and ends at 1) permit broader numeric interpretations of verbal probabilities falling within the mid-range of the interval as opposed to those expressions falling at the extremes of the interval. Additionally, many words falling in the middle of the probability range represent very fuzzy concepts with very vague meanings. Graphically, these expressions are represented by wide membership functions (e.g., Reagan et al., 1989; see Appendix D for an explanation of this concept). Data for the words "perhaps" and "possible" in this validation study reflect this effect. These expressions yielded lower than average intra-individual correlations.

To test the consistency between repeated rankings for each expression. ranks were inferred from the mean of the two ratings for each expression given by participants in the in-context portion of the questionnaire and correlated with the explicit ranks for each expression provided in the final part of the questionnaire. Compared to the data for the ratings given in-context, there were considerably lower correlations yielded from this analysis. The majority of expressions (11 of 15) had intra-individual correlations below 0.4, and no correlation exceeded 0.6.

The finding of greater consistency in numeric ratings of verbal probabilities compared to rankings is divergent with earlier findings. Typically, people display greater variability in the numeric values they assign probabilistic expressions than in the rankings they give expressions (Budescu & Wallsten, 1985; Kong, Barnett, Mosteller, & Youtz, 1986). However, several aspects of the task may have led to a reversal of this more common finding.

In the first place, at least one contextual variable in the questionnaire likely influenced the data in this respect. Specifically, it appears that asking the participants to assign numeric ratings to probabilistic expressions within a particular context increases the consistency of their ratings. In the validation study, the repeated numeric ratings for each of the 15 expressions were given within highly similar task environments; that is, all expressions were embedded in psychodiagnostic statements. Correlations of the rank data, however, were derived from data generated in two different tasks: (a) the rating task in which expressions were embedded within psychodiagnostic statements, and (b) the ranking task in which participants were instructed to rank the 15 expressions, which were presented in a randomly ordered list without a specific psychodiagnostic context. This notable difference in the task contexts may have engendered the higher correlations between the two rating sets compared to the ranking scores.

An interesting implication of these data is that it appears that various aspects of the context in which verbal probabilistic expressions occur can

systematically affect how people interpret these expressions. Therefore, when a task is contextualized, such as in psychodiagnostic statements in the present study, the consistency with which individuals interpret probabilistic expressions is enhanced. This conclusion finds support in other studies of the effects of context on numeric interpretations of verbal probabilities (cf. Brun & Teigen, 1988; Nakao & Axelrod, 1983; Zhu, 1992).

A second factor that may account in part for the difference between the results for the rating and ranking of probability expressions involves the scales on which scores were generated in this study. By virtue of ranking the 15 expressions in ascending order, a 15-point scale with ordinal properties was created. Numeric probability ratings though were given on a 9-point probability scale. In general, a scale with more intervals, such as the 15-point ordinal scale in this study, permits a broader range of scores and, all things being equal, such a scale would be more likely to produce lower intra-individual correlations.

Inter-individual consistency. To assess the level of consistency between individuals, two sets of correlations, one derived from the means of the repeated rating scores and the other derived from the ranking scores, were calculated for each subject pair, resulting in 253 intersubject correlations in each of the two sets. In both sets, correlations spanned nearly the entire range between 0 and 1 but were more numerous in the high range, as reflected in the strength of the mean

correlation for each set (0.71 and 0.77 for the rating and ranking sets, respectively).

These data reveal that different people tend to assign similar, but not identical, meanings (in numeric terms) to linguistic expressions of probability. The difference between the mean correlations for the rating and ranking sets suggests that there is higher agreement among individuals on relative ordering of expressions than on ratings with absolute numeric values. The finding is consistent with results of earlier studies that showed that the ranks of probabilistic expressions tend to be more stable than the numeric values (on ratio scales) they are assigned in rating tasks (Budescu & Wallsten, 1985; Kong et al., 1986). It contrasts, however, with the weaker intra-individual correlations for the ranking task discussed above. It is noteworthy that the means for the intersubject correlations are derived from many more data points than the mean of the intra-individual correlations (253 versus 15, respectively). Thus, the considerably stronger mean correlations for the intersubject data may be explained in part as a regression-to-the-mean effect. Again, the contextual uniformity of the task from which the inter-individual correlations were derived likely enhanced the consistency of scores. This conclusion is based on the fact that the correlations reflect comparisons on the same task rather than two different tasks, as was the case with the intra-individual ranking data.

Inter-group consistency. One of the more consistent findings in this domain of research is that when aggregate group ratings are compared with one another, there is relatively little variation in scores. The results of the present study appear to be no exception to this trend. Aggregate group ratings of 8 probabilistic expressions researched in this study were compared to aggregate ratings documented in earlier published studies. For the majority of these expressions (5 out of 8), there was no meaningful variation in these ratings; specifically, variation was confined to 1 percentage point for three of the expressions, 3 percentage points for one expression, and 5 percentage points for one expression (see Table 5). One expression showed modest variation in a 14-percentage-point spread, and two expressions showed moderate variation in their percentage-point spreads of 20 and 23. Overall, these data for aggregate group ratings display considerably greater consistency than data bearing on either intra-individual consistency or inter-individual consistency. This finding replicates results of other studies that have looked at variations in group-mean ratings of probabilistic expressions across studies (Reagan et al., 1989; Robertson, 1983).

Looking at these data across studies also permits an assessment of the effect of the passage of time on numeric interpretations of probabilistic expressions. According to the data reported in Table 5, there seems to be no discernible effect of the passage of time on interpretations of these expressions, at least within a range of a few decades. Consequently, data from studies that

researched probabilistic expressions in earlier decades were used to develop the procedure to rate verbally expressed confidence in think-aloud protocols.

Conclusion. The purpose of the validation study was to assess the degrees of consistency in clinical trainees' interpretations of verbal probability expressions at three levels, within individuals, between individuals, and between groups. The findings of the validation study at each of these three levels can be summarized as follows:

1. When individuals make repeated ratings of a single expression, they tend to give the same or a similar ratings for that expression over time. This consistency is enhanced by contextualizing the task, as this tends to foster greater uniformity within individuals in their numeric interpretations of expressions. Interpretative consistency also varies depending where expressions fall along the probability interval: Consistency tends to diminish when very vague expressions that fall around the middle of the probability interval are rated (e.g., "perhaps" or "possible").

2. There is reasonably high, but not perfect, agreement between individuals when they interpret linguistic probability expressions. Moreover, agreement increases when people are asked to indicate the relative location of these expressions on an ordinal scale as opposed to generating absolute numeric values on a ratio scale like a 100-point percentage scale or even a coarser 9-point

scale as was used in this study. Again, it appears that contextualizing a task serves to enhance interpretative consistency.

3. Aggregate ratings derived from groups reveal high consistency in mean ratings of verbal probability expressions. Moreover, meanings of probabilistic expressions do not vary over periods of time limited to a few decades.

Together, findings of this validation study, which for the first time examined how people belonging to the field of professional psychology interpret linguistic probability expressions, lend support to conclusions reached by other researchers, namely (a) that there is reasonable interpretative consistency of expressions and (b) that the goal of codifying these expressions is both realistic and attainable (Kong et al., 1986; Reagan et al., 1989). While these data lend empirical support to the validity of the procedure to rate verbally expressed confidence in think-aloud protocols, they obviously do not provide unequivocal support for the assumptions underpinning the procedure. Several caveats seemed to be indicated by the data, and these will now be discussed. The limitations and shortcomings of the procedure are considered later in this chapter (see *Delimitations and Limitations* below).

Caveats. Confidence is in essence a multifaceted psychological state. The process of generating confidence assessments involves the interaction of cognitive and affective phenomena operating at both conscious and unconscious levels in

the minds of individual judges. Because of the sheer complexity of these phenomena, relatively little is known about precisely how confidence assessments are generated. Additionally, several researchers have demonstrated that linguistic expressions of probability communicate more than just subjective confidence. For example, the severity of consequences associated with a decision outcome and non-commitment to a choice or judgment are other communicative aspects of probabilistic language (Merz et al., 1991; Teigen, 1988). These other aspects of judgment confidence, which are poorly understood and inadequately specified, serve to complicate efforts to operationalize the construct with precision.

It is noteworthy that other researchers who presented data showing interpretative consistencies of a similar magnitude to those reported in this study have suggested that the use of verbal probabilistic expressions is unacceptable in a professional context (in this instance, a medical one) (Nakao & Axelrod, 1983; Robertson, 1983). One reason that may account, at least in part, for these apparently discrepant perspectives on the same or similar findings is that one's tolerance for the fuzziness conveyed by verbal probabilities is not independent of the fuzziness inherent in one's clinical or research discipline (see fuzzy-set theory, e.g., Azevdo, Lajoie, & Fleiszer, 1996; Derry & Hawkes, 1993). Verbal probabilities may not serve well those disciplines that are capable of defining problems with relatively high degrees of precision using highly advanced technologies. Other disciplines, professional psychology among them, operate

from knowledge bases and with technologies that limit them to defining problems using fuzzy descriptors (Smith & Dumont, 1997). In counselling and clinical psychology, the data sets clinicians use to construct clinical portraits are ambiguous, based as they are on clients' flawed inferences and their idiosyncratic construals of reality. Moreover, psychological theories and models consist in part of intrinsically fuzzy sets, which are constituted by numerous continuous variables to which it is difficult to assign anything but rough numeric values. In these disciplines, using verbal probabilities to qualify problem definitions appears to be appropriate and, therefore, acceptable to their practitioners.

Inferential Biases and Psychodiagnostic Confidence

Recent findings in several studies suggest that levels of confidence and overconfidence inversely vary as a function of the amount of mental effort applied to judgment tasks (Block & Harper, 1991; Sniezek et al., 1990). Inferential heuristics presumably underlie this effect, as they are used to reduce the difficulty and complexity of judgment tasks. It was therefore hypothesized that clinicians who manifested greater reliance on heuristics to aggregate and interpret a client database would evidence higher levels of psychodiagnostic confidence.

Four inferential biases (i.e., dispositionalism, confirmationism, truncation of data search, and narrow problem construal) were studied in order to determine their separate and combined effects on the dependent variable, psychodiagnostic confidence. Multiple regression analysis revealed that the proportion of variance

in the dependent variable explained by the four independent variables achieved a marginal level of statistical significance (F=2.337, p=0.077). Stepwise regression, however, revealed a clearer picture of the individual and collective contributions of the four heuristics to scores on psychodiagnostic confidence. It showed that one bias, namely dispositionalism, accounted for 20% of the total 23% of variance in psychodiagnostic confidence scores that was explained by the four inferential biases considered together. The regression coefficient for dispositionalism was statistically significant (F=8.684, p=0.006), whereas the regression coefficients for other three heuristics fell far short of statistical significance. These results do not support the broad hypothesis tested in this study that an indirect relationship exists between the amount of mental effort applied to a psychological assessment task, as revealed by clinicians' reliance on the four inferential biases, and psychodiagnostic confidence.

There are two mutually exclusive explanations for these particular results. The first is that the research hypothesis is essentially untenable, and the present data provide conclusive evidence to this effect. The second possibility is that the present study was a less than completely adequate test of the hypothesis due to problems in the design and/or execution of this study. In order to explore each of these possibilities and arrive at a judgment about which of the two explanations is most plausible, data bearing on each inferential bias along with its relationship with psychodiagnostic confidence are reviewed below.

Dispositionalism. The present study revealed a strong, direct relationship between dispositionalism, which is the tendency of clinicians to attribute clients' problems to aspects of their personalities rather than features of the environments in which they live, and the overall degrees of confidence expressed in diagnostic judgments. In practical terms, this means that the more clinicians favour dispositional explanations of clients' problems, the more confidence they tend to place in the accuracy of these explanations. This finding is consistent with number of studies in the area of social attribution have examined the relationship between these two variables (Dunning et al., 1990; Paese & McKinnaly, 1991; Vallone et al., 1990).

Most previous studies investigating the relationship between these variables have employed undergraduate students who made various kinds of social judgments. This research revealed that predictions people make about others tend to be highly confident and highly overconfident when the predictions are based on inferred aspects of personality. The results of the present study represent an interesting extension of these findings to professional psychologists as a group, since it appears that a similar conclusion can be made about the clinicians in this study, both experienced professionals and advanced clinical trainees alike. Specifically, clinicians displayed more confidence in their judgments when they favoured dispositional over situational explanations of the client's problems.

Although this was by no means an unexpected finding, it is still interesting to consider why professional psychologists continue to place such high degrees of confidence in dispositional judgments when numerous studies have demonstrated the tenuousness of this kind of inferencing (e.g., Mischel, 1968; Nisbett & Ross, 1980; Ross, 1977). Their persistence in doing so may be in part a function of the professional role that psychologists fulfill. Specifically, making judgments about people and, in particular, aspects of their personalities based on limited and selective data sets is what psychologists train many years to do and what they are usually expected to do to earn a living. We should not be surprised to see them fulfilling this role. Conversely, we might expect sociologists to resort more often than psychologists to social constructs than psychological ones to explain a person's behaviour. (In passing, this could be an interesting way of assessing the impact of education on clinical problem solving.)

These findings also suggest that psychologists could benefit from educational interventions designed to sensitize them to the pitfalls inherent in the work they do. To be beneficial, interventions would have to target specific shortcomings of clinical judgment. Although much research remains to be done to specify the precise nature of these pitfalls, we are beginning to understand how judgment overconfidence arises. In particular, it has been shown that a source of overconfidence (among numerous possible sources, some of which are discussed later) is the implicit personality theories that people use to generate predictions

and judgments about others (Dunning et al., 1990; Vallone et al., 1990). It appears, based on the findings of this study, that psychologists, like lay judges in earlier studies, place too much stock in the personality theories that inform their clinical judgments. The accuracy of clinical judgments is additionally diminished at times by the questionable validity of some of these theories, many of which probably pre-date clinicians' formal education in such matters. For example, research on illusory correlations suggest that implicit theories that have no scientific foundation are as tenacious and pervasive among professional psychologists as among lay individuals (Chapman & Chapman, 1967, 1969). Similarly, the continued use of projective devices, such as the human figure drawing (see Machover, 1949), to derive personality judgments based on features of a drawing despite overwhelming empirical evidence contraindicating such practices demonstrates that implicit (and flawed) personality theories that constitute part of the lore of our culture are not easily eliminated (Smith & Dumont, 1995).

Confirmationism. This investigation revealed no significant relationship between confirmationism, which is the tendency to use a positive or confirmatory hypothesis-testing strategy when interpreting a clinical history, and degrees of psychodiagnostic confidence. Although data from earlier studies pointed to a possible relationship between these two variables, the overall pattern of findings in the literature has been somewhat inconsistent. For example, Block and Harper

(1991) found that anchoring does not lead to overconfidence in all cases and suggested that other variables mediate the relationship between these two variables. Specifically, they proposed that judgment overconfidence arises from overestimation of one's judgment abilities rather than the anchoring heuristic per se.

Looking at the literature from a broader perspective, it is not yet clear whether or not an anchoring or confirmatory bias is a pervasive problem among clinicians. Earlier research on this inferential bias showed clear evidence of anchoring in prediction tasks of considerably less complexity than clinical judgment(e.g., numerical estimation; see Tversky & Kahneman, 1974). Evidence of confirmationism and anchoring in clinical judgment is actually mixed. In a recent study of clinical judgment, researchers found no evidence of anchoring among clinical trainees who were presented pre-interview information about a client prior to assessing the client's casefile, although participants displayed a tendency to remember information that confirmed rather than disconfirmed their clinical judgments following a delay of several weeks (Lee et al., 1995). Aronoff (1997) had professional clinicians interpret different versions of a client casefile and solicited concurrent verbal reports as well as final judgments about the client. His results revealed an anchoring effect, but only among clinicians who were initially presented with healthy information about the client. Moreover, this effect appeared only in their final clinical judgments, but not in their think-aloud

protocols. On the other hand, the think-aloud protocols revealed an anchoring effect among clinicians who were initially presented with pathonomonic information, but the effect was absent in their final clinical judgments. Such a variety of findings in this problem domain indicates that there likely are intervening variables that moderate the effect of the anchoring and confirmatory biases on clinical judgment and the cognitive processes involved in generating them. The precise nature of these variables remains to be explored and specified by future research.

A methodological limitation of the present study makes the interpretation of the results difficult and attenuates the test of the original research hypothesis. This limitation bears on the operational definition of the confirmatory bias that was used in this study. Specifically, all repetitions of any initial diagnostic judgment were coded individually as confirmatory inferences and interpreted as instances of a confirmatory hypothesis-testing strategy. Although this operational definition reveals what hypotheses are retained (and, conversely, which are rejected) as clinicians proceed through the casefile, it may be an insensitive measure of the construct for other reasons, as the following example highlights: A clinician makes 50 initial inferences and 50 confirmatory inferences in the course of his assessment. Each initial inference is confirmed once, resulting in 50 different confirmatory inferences. Another clinician also makes 50 initial inferences and 50 confirmatory inferences, however, in her case the same initial

inference is confirmed 50 times. Although the scores on this variable in the present study would be the same for each clinician, to say that each relied to the same extent on a confirmatory hypothesis-testing strategy would not be an accurate representation of the data. Admittedly, this an extreme example that did not actually occur. Nevertheless, data in the think-aloud protocols do reveal that some clinicians evidenced more diagnostic variety among their confirmatory inferences than others. It would be interesting and informative if future research in this area explored this aspect of the confirmatory bias in think-aloud protocols using a more sensitive measure of the construct.

Narrow problem construal. Previous research in social judgment has indicated that a source of overconfidence is the tendency of observers to narrowly construe the situations facing actors (about whom the observers are making inferences), which leads the observers to underestimate or even overlook the influence of situational variables on actors' behavior (Griffin et al., 1990). In the present study, it was hypothesized that an analogous phenomenon would occur among mental health practitioners making clinical inferences: That is, clinicians who construed the client's problems narrowly by making fewer initial diagnostic inferences would display higher degrees of psychodiagnostic confidence relative to those who made more initial inferences. This hypothesis was based on the presumption that construing the client's problem narrowly reflected an underestimation or oversight of important aspects of the client's clinical history,

which would subsequently lead clinicians to express higher confidence in their inferences. In other words, clinicians conceptualizing the case in simpler terms would have less to be uncertain about. However, the findings of this study failed to confirm this hypothesis. The data revealed no relationship between the number of initial inferences posited and the overall degrees of psychodiagnostic confidence.

While this finding diverges with results of previous research, an important aspect of the data gathered in this study might account for this inconsistency. Specifically, Griffin et al. (1990) found that observers typically generate only *one* construal of the situation facing an actor, uncritically filling in informational blanks and overlooking influential aspects of an actor's circumstances while making social judgments. In the present study, the clinicians without exception generated more than one initial inference about the client's problems. In fact, the lowest number of initial inferences generated by a participant in this study was 15, and the mean for all participants was 39. When taken together, the results of these studies raise the possibility that beyond a certain threshold for the number of construals judges generate (perhaps 1 is the threshold), the effect of having a narrower construal of an individual's situation on degrees of judgment confidence ceases to exist.

One variable possibly responsible for attenuating this effect is the extensive training in differential diagnosis that psychologists normally receive in

the course of their clinical studies and training. In essence, the process of arriving at an accurate diagnosis of a client's problems normally requires that the clinician generate a list of all plausible diagnostic options and then evaluate each in light of the clinical evidence at hand. It seems reasonable to speculate that professional training in clinical assessment and differential diagnostics may offer some protection against the pitfall of narrowly formulating clients' problems. One piece of anecdotal evidence in support of this hypothesis was offered by a participant who was a trainee in clinical psychology. She spontaneously remarked following the think-aloud procedure that a component of her training involved developing as plausibly broad a diagnostic formulation as possible by generating many different hypotheses about the client being assessed. Dumont (1993) stated that clinicians may be susceptible to narrowly construing their clients' problems in an analogous manner to the way lay people making social judgments narrowly construe the situations facing actors (Griffin et al., 1990; Nisbett & Ross, 1980, p. 127). However, results of the present study indicate that this may not be true, although it is up to future research to investigate the validity of this hypothesis.

Truncated data search. Another aspect of social judgment examined in this study is the tendency to truncate the search for relevant data about individuals and the situations in which they find themselves. Research has shown that observers typically have less than optimal amounts of information at hand when they make judgments about others (Griffin et al., 1990; Ross & Nisbett, 1991, p.

136). People go about filling informational lacunae using idiosyncratic strategies that often lead them down a path of inferential error (Dumont, 1993; Nisbett & Ross, 1980). Extrapolated to the clinical domain, it was hypothesized that this same shortcoming would manifest itself among the sample of clinician-participants as truncated searches for relevant data about the client being assessed.

The present study looked into a possible link between the scope of casefile data clinicians integrated into their problem formulations and the degrees of confidence they expressed in those formulations. The results, however, revealed no significant relationship between these variables. As this variable is conceptually related to the variable discussed in the preceding section (i.e., narrow problem construal), it is possible that the same explanation accounts for these results: That is, participants in this study may have sampled a range of clinical data that was sufficiently large enough to surpass a critical threshold and show no effect on their confidence assessments. On this point, it is noteworthy that out of a maximum of 15 information categories available to participants in the casefile, the mean number of categories used by all 36 participants was 9.8. The fewest number of categories used by a participant in generating a client problem formulation was 5.

From a methodological perspective, the particular way the think-aloud technique was used in this study obscures somewhat the meaning of the data and prevents one from drawing a firm conclusion about these results. A firm

conclusion would rest on the assumption that the scope of casefile data participants used in generating their problem formulations was *explicitly* reflected in their verbal reports. However, it seems imprudent to make this assumption, as all participants in this study were given the entire casefile to examine and read aloud. It is possible that casefile segments to which they did not explicitly respond were, nonetheless, implicitly integrated into their evolving problem formulations. On this point, Ericsson and Simon (1993, p. xxxv) acknowledge that concurrent verbal reports, while a rich source of information about thinking processes, are inevitably incomplete; not all mental events occurring during the completion of a task are reflected in the corresponding verbal reports.

This limiting factor could be resolved in a future study examining the relationship between scope of data search and psychodiagnostic confidence by using a form of the think-aloud technique that is slightly different than the one used in this study. Rather than presenting clinicians with the entire casefile to read aloud and interpret, participants could collect client information they considered pertinent by directing questions to an experimenter, who would answer these questions by referring to a detailed client casefile that she or he had at hand. This technique would give a much clearer view of the kind and quantity of information clinicians seek in the course of an assessment. This approach also has the advantage of being even more ecologically valid than the method used in the

present study, as it more closely approximates the true manner by which clinicians gather information about their clients.

Conclusion. While findings of this study do not support the hypothesis that reliance on heuristics to simplify clinical problem solving fosters higher levels of psychodiagnostic confidence, it seems premature to entirely rule out this hypothesis. As alluded to in the preceding discussion, data from this study raise the possibility that there are intervening clinician variables that moderate the relationship between confidence and specific inferential heuristics and biases, such as confirmationism. As previously mentioned, professional training in making clinical diagnoses may be one such variable, but there are likely others, such as theoretical orientation and clinical discipline. There is some preliminary data indicating that these latter variables impact on clinical judgment processes. For example, Goodin Waxman (1991) found a significant difference between psychodynamically-oriented clinicians and behaviourally-oriented clinicians in their attributions of a client's problems. Dumont, Sladeczek, & Martel (1998) found that length of clinical experience moderated the relationship between order of client information and attributions of a client's problems. Future research could investigate whether or not such clinician variables impact on the relationship between psychodiagnostic confidence and inferential heuristics. Additionally, this research could improve on the methodological limitations of this study and arrive at more definitive conclusions about the effects of biases such as truncated data

search and narrow problem formulation on psychodiagnostic confidence. Data bearing on these matters would be informative at this juncture for clinicians and researchers alike, as it would contribute to the development of more comprehensive and, presumably, more externally valid models of human problem solving.

Experience and Psychodiagnostic Confidence

Results of a number of earlier studies suggested that experienced clinicians tend to make more appropriate confidence assessments (i.e., display smaller discrepancies between judgment confidence and judgment accuracy) than inexperienced clinicians or untrained participants (Goldberg, 1959; Oskamp, 1962; Levenberg, 1975; Garb, 1986). In each case, the source of this difference observed across the various levels of experience was not higher accuracy but lower confidence assigned to judgments.

The present study also looked at this question in order to determine whether or not the same result would occur using a different and more ecologically valid technique, namely the think-aloud procedure, to probe clinical judgment processes. A univariate analysis of overall confidence ratings for experienced clinicians versus clinical trainees showed no significant difference ($F[1,34]=0.061$, $p=0.806$). The least-square means of overall confidence for the two groups (75.4 and 74.9 for experienced clinicians and clinical trainees, respectively) revealed a statistically non-significant difference that was in the

direction opposite to what was hypothesized. It appears on the basis of these results that experienced clinicians do not differ from clinical trainees in the overall degrees of confidence assigned to problem formulations when they use a natural and ecologically valid means of expressing psychodiagnostic confidence. This finding raises a number of interesting questions, two of which are now considered. Firstly, why would using the think-aloud technique instead of the common method of simply asking participants to indicate their level of confidence on a numerical scale affect participants' confidence assessments? Secondly, in consideration of the present results along with earlier findings, why would there be an interaction of the type of data-gathering method with level of clinical experience?

The research literature provides several clues about possible answers to the first question. Kahneman and Tversky (1982) discussed the biasing effects of the question-answering paradigm that dominates the social and clinical judgment literatures. The relevance of these effects to the present discussion bears on the information or cues inadvertently given to study participants when they are asked a question intended to measure an experimental variable. The authors state: "It is often difficult to ask a question without giving (useful or misleading) clues regarding the correct answer, and without conveying information about the expected response" (p. 135). Similarly, Fischhoff and Bar-Hillel (1980) demonstrated that study participants are sensitive to irrelevant information

provided in experimental protocols if that is the only variable obviously manipulated across a set of judgment problems. By applying these arguments to the previous studies of psychodiagnostic confidence, one would expect that by simply asking research participants to indicate their level of confidence in their clinical judgments will alter, albeit indiscernibly, subsequent responses. For example, in the brief interlude between hearing or reading a question and making a response, it does not seem difficult to imagine a clinician's internal monologue occurring along the following lines: "I don't want to look foolish by appearing to be too sure of a judgment that may be wrong. And I remember reading some time ago about research showing that professional clinicians are often no more accurate in their judgments than untrained subjects. So I'd probably be better off proceeding cautiously." The point illustrated in this example is that ostensibly innocuous questions can cue thoughts and memories that can bias subsequent responses.

The second question regarding the interaction of data-gathering method and level of experience on confidence assessments leads the present discussion on a more speculative path in order to explain the results of this study. This explanation goes as follows: Experienced professionals generally may be more sensitive than their inexperienced counterparts to the implicit cues that are inadvertently made available to study participants in the traditional question-answering format, the format used in virtually all previous studies of

psychodiagnostic confidence. This sensitivity may develop over time through professional experiences that expose clinicians to pertinent information about relevant clinical and research issues. Conferences, discussions with colleagues, and clinical practice are some of the mediums by which this information is likely transmitted. For instance, recalling the internal monologue described above, it is not inconceivable, as the example indicates, that experienced clinicians would be more sensitive than clinical trainees to the problems and pitfalls inherent in clinical judgment as a function of their broader and more frequent exposure to the literature bearing on this question. Therefore, by asking a direct question about confidence, the experimenter may be inadvertently cueing traces in long-term memory bearing on this literature.

In contrast to the traditional question-answering format, the think-aloud technique circumvents this unintended cueing of previously encoded information, because participants' attention is not oriented to the variable under study (psychodiagnostic confidence in the present context). Consequently, the differential effect of experience on confidence assessments would be expected to disappear, as it did in the present study. Of course, further research will be required to verify this hypothesis.

Delimitations and Limitations

Several factors related to the design, the methodology, and the individuals who participated in this study limit the generalizability of the findings of this

study. A number of limitations of the methodology specific to the independent variables have already been discussed in previous sections of this chapter. In this section, several more general limitations and delimitations will be discussed.

One limitation of this project bears on the nature of the think-aloud methodology. Although this method has been demonstrated to yield valid data in many problem domains (Ericsson & Simon, 1993; Jones, 1989), it remains an analogue methodology and, as such, only approximates the clinical assessment process. Some limitations of this method arise from the manner by which information is provided to study participants (Aronoff, 1997). Rather than gathering clinical data from an emotionally charged and spoken exchange, clinician-participants read information that is considerably less voluminous and is organized and paced differently than typical clinical intake sessions. Additionally, this method provides no opportunities for clinicians to test their hypotheses about the client being assessed. All of these factors could potentially affect the character of participants' clinical judgments.

This findings of this study are delimited (a) to the initial assessment phase of the counselling process and (b) to diagnostic judgments made therein. The results cannot be assumed to apply to clinicians' inferential processes at later phases of the counselling process, as it is conceivable that their problem formulations develop and change as the process unfolds over time and progresses into later therapeutic phases. On this point, it is noteworthy that relatively little is

known about clinicians' inferential processes at later points in the therapeutic process, as most research to date has focused on the assessment phase. Nor can it be assumed that the results apply to inferences that clinicians make about the specific causes of the problems they diagnose. The etiologies of psychological disorders are likely influenced to a greater extent by clinicians' theoretical orientations than are psychodiagnoses, and, as mentioned previously, we cannot presume that this variable does not impact on clinicians' judgment confidence.

One limitation of the procedure to rate linguistic probabilities in the think-aloud protocols is that it uses norms derived from group data to quantify individual responses. Despite the reasonable consistency in the interpretation of such expressions among individuals, this consistency is not perfect. The variability in the interpretation and use of probabilistic expressions that was observed in this study and others, particularly between individuals, makes it reasonable to expect that a certain proportion of individuals' responses in the think-aloud protocols were inaccurately scored on the basis of the norms that were used. In all likelihood, some true scores were under-estimated while others were over-estimated. However, (a) assuming that this error is randomly distributed and (b) because all participants used numerous different expressions, under-estimated and over-estimated confidence scores likely balanced each other. Under these conditions, it seems reasonable to assume that the data derived from the experimental study were protected from this limitation and escaped bias, however,

the data gathered in this study cannot be verified in this respect. In the final analysis, the rating procedure, although imperfect, represents a tradeoff between rough estimates of psychodiagnostic confidence scores, on the one hand, and the disadvantages of soliciting confidence assessments directly from study participants (low ecological validity among them), on the other.

Although an important part of this research project was devoted to verifying the validity of the procedure to rate psychodiagnostic confidence, it cannot be assumed that its validity has been definitively established. Establishing the validity of any data-gathering method is an ongoing process, and, ultimately, validity can only be inferred on the basis of the empirical evidence that accumulates over time. While the validation study completed as part of this project yielded data in support of the procedure, there remain outstanding issues that have yet to be submitted to empirical scrutiny. One such issue involves the extent to which confidence expressions verbalized in the think-aloud protocols represent the degrees of confidence participants actually place in their clinical hypotheses. While this procedure assumes this to be true (and there is no compelling evidence to suggest otherwise), there are at least two arguments that qualify this assumption, suggesting it should not be taken for granted. The first is that studies have shown that verbal confidence expressions have communicative properties that extend beyond simply expressing a probability. These include (but are not limited to) (a) the severity of consequences associated with a judgment

and (b) degree of commitment to a judgment or choice (Merz et al., 1991; Teigen, 1988). It is therefore possible that these other communicative functions confound the relationship between clinicians' actual confidence and the manner in which they choose to express it. The second point arises from some anecdotal data gathered in the study. A small number of participants in this study showed idiosyncratic predilections for using particular expressions in their verbal protocols. In these cases, it is difficult to know whether the expressions represented the clinician's actual confidence assessment or whether they were simply "pet" phrases these participants habitually used in conversation. Future research will be required to verify this assumption.

Several limitations and delimitations bear on the sample of clinicians who chose to participate in this research project. Findings of these investigations are delimited to the population of counselling and clinical psychologists. They are generalizable to this population only to the degree that it is accurately represented by the sample. Although steps were taken to maximize the representativeness of the sample (e.g., in terms of discipline, gender, experience, etc.), it is impossible to control or even account for all variables that potentially impact upon the self-selection of clinicians.

Participants may have been also unduly influenced by concerns about self-presentation and the social desirability of their responses. Abramowitz and Dokecki (1977) suggested that clinician-participants are usually able to detect the

purposes of such studies and, as a result, often respond atypically to study instructions or questions. These influences might have led participants in this study to make more conservative clinical judgments than they typically would in practice. It is possible that these factors might even have potentiated the inferential biases under investigation. For example, participants may have felt obliged to demonstrate mastery in making judgments about a client's personality (since this is one of the distinguishing functions of the profession) and consequently evidenced a dispositional bias. It should be noted, however, that no data from this study, either empirical or anecdotal, lend strong support to this hypothesis. The fact that participants were paid for completing the experimental protocol is another potentially biasing factor that may limit the meaningfulness of the data. For example, participants might feel obliged to produce ideas and inferences about the casefile in order to justify being paid when, under other conditions, they might have delayed their responses until they had more information about the client. The fact that a small number of participants asked the experimenters questions such as, "How am I doing?" or "Am I doing all right?", raises the possibility that these individuals had some concerns about self-presentation that could have influenced their responses.

Implications of Research in Psychodiagnostic Confidence

Options for Training and Practice. Research in several domains has revealed that overconfidence pervades human judgment, especially judgments

produced in decision environments characterized by uncertainty and ambiguity. It is apparent that many variables in differing combinations are involved in generating confidence assessments and contribute to overconfidence effects. The present research project focused on a subset of those variables, namely inferential heuristics. From a broader perspective, enduring solutions to overconfidence can be as complex as understanding (a) the nature of human judgment processes, along with the heuristics that often undermine their efficacy, (b) the belief systems that drive these processes and (c) the characterological profiles of those who perseverate in their biased products.

Thousands of variables and many solutions can present themselves when one is faced with a problem. Cognitive heuristics, which are strategies that reduce the complexity of mental tasks, make such problems manageable. These shortcuts befriend us when the problem is in reality a simple and straightforward one. They betray us when the problem conceals unusual features that only deeper probing can reveal. Accumulating evidence suggests that professional psychologists will be overconfident to the degree that they make uncritical use of cognitive heuristics. In consideration of the results of the present study and previous research in this domain, it appears that heuristics do play a role in generating confidence assessments, although the precise nature of the relationship (or relationships) between psychodiagnostic confidence and heuristics is not completely understood, and more research is certainly needed to probe this issue.

While there is scant research on the topic, personological factors are in all likelihood causally related to overconfidence (cf. Spengler & Strohmer, 1994). Clinicians' confidence in their judgment derives partly from the socio-educational culture in which they were raised and which disposes them to construe the world in a way that makes other visions opaque to them. The simpler one's world view–and the more one has been indoctrinated into it by family, religious affiliations, and political community–the more difficult will it be for that person to entertain hypotheses that are not in accord with it. It cannot be denied that the lengthy professional training of psychologists also shapes the biases that inevitably play in their assessment of clients. Graduates of certain programs display diagnostic propensities that graduates of others do not. We have long known that the theories in which we have been trained and our indoctrination into a particular therapeutic approach, say a cognitive-behavioral as opposed to a psychodynamic one, have a profound influence. (See Nisbett and Ross [1980] for a discussion of how theory-biased data sampling can distort and even thoroughly corrupt the analysis of client problems.)

Solutions for overconfidence. A number of researchers have developed and tested strategies to reduce overconfidence in judgments (e.g., Block & Harper, 1991, Plous, 1995; Sharp, Cutler, & Penrod, 1988; Sniezek, 1991). It is noteworthy that the tasks used in these studies are almost exclusively general knowledge or estimation of unknown quantities tasks, which are only remotely

related to the inferential tasks facing psychologists in their daily work. Despite this obvious shortcoming, they do provide some guidelines for developing educational modules for mitigating psychodiagnostic overconfidence.

Perhaps most fundamental to any method is that students must be brought *experientially* to the realization that their diagnostic hypotheses are precisely that–hypotheses. Feedback on one's judgments, with special emphasis on one's fallibility in these matters, needs to be frequently repeated until such time as tentativeness becomes second nature in arriving at a diagnosis. To achieve this goal, students should be presented with numerous occasions to experience the pitfalls of diagnostic inferencing. Though one failure would seem to be sufficient to undermine one's sense of infallibility in some domains, repeated failure would seem to be necessary in probabilistic domains such as psychotherapy where feedback is spotty and never seems to be conclusive. Presenting trainees with many case histories that are complex and difficult to assess would allow them to experience the enigmatic quality of much human behavior, stumble in their problem-solving, and collectively work through the pitfalls that threaten even the most experienced and expert of therapists. These exercises can be especially instructive when conducted in groups where trainees are exposed to divergent ideas, but only if they remain cognizant of the conditions that detrimentally inflate confidence in group judgments (see Sniezek, 1992 for a review). Although we

have framed these interventions as instructional techniques for clinical trainees, with slight modifications they could also be useful for professional practitioners.

To develop such a training module, one could begin by selecting segments of a case history that present multifaceted aspects of a client's personality. These could be compiled in such a manner that one could easily misjudge the character of the client if one had available the information in one pile but were missing salient information of one sort or another in the other pile (which, of course, often happens in the course of many therapeutic relationships). One group of students would then be given half of the case material and a second group the other half of the same casefile. Important diagnostic events that took place in one setting would be presented to one of the groups; important events that took place in a very different kind of setting would be presented to the other. The behavior of a client in a bar or at a football rally will suggest one profile. The behavior of that same client at his mother's funeral or in a courtroom or in a job interview will suggest another. The great divergence in the assessments of the two groups can then be made evident to the entire class, and they are asked to discuss the reasons why this one client appeared to be such a different kind of person as a function of the data set selected for analysis. The instructor would then lead the class to a maximum likelihood assessment of the integral casefile.

Giving students problems in covariation is another method of reducing the penchant for making snap judgments and then clinging to them tenaciously. In

general, people like to think that they are skilled in judging covariations. The reality is otherwise. "Much research," stated Nisbett and Ross (1980, pp. 90-91) "that has dealt explicitly with people's abilities to recognize and estimate covariation has not been flattering to the layperson's abilities." In all likelihood, the same applies to mental health professionals also. To combat this problem, an instructor could fabricate some illustrative sets of data in which symptoms both present and absent are correlated with disorders both present and absent. These sets could then be presented to students in a simple but classic fourfold table (see Peterson & Beach, 1967). Such a graphic demonstration of the complexity of clinical judgment will lead them to readily grasp the challenge that diagnosticians face in estimating the correlation of two dichotomous variables. Collectively studying a 2 X 2 matrix of symptom and disorder (simply present vs. absent in both cases) readily convinces them, statisticians as they all are, that they must examine all four cells of the matrix. Progressing to more complex examples involving continuous and multifarious variables involving both current situational and personal historical data can truly be mind-boggling.

Another effective method of breeding caution is to ask students to develop a realistic fictional case history that they could use for the purpose of sensitizing others (their classmates, the professoriate, and experienced clinicians) to the hazards of unwarranted confidence in their clinical judgments. Omitting critical information as well as sprinkling the case history with plausible red herrings

generates a difficult diagnostic task. The simple exercise of developing such a scenario, independent of the discussions and case conferencing it can prime, gives students a heightened sense of the pitfalls they must be alert to in their own client-problem formulations.

These are a sample of training strategies that may have a certain psychopedagogic warrant. There are other methods, of course, of reducing the overconfidence that some if not all clinicians manifest in their work. Readers will note that their effectiveness has yet to be scientifically tested, a point that leads us to the next section.

Future Research

The literature on psychodiagnostic confidence and overconfidence suggests a number of important directions for future research in this problem domain. First, accumulating research findings imply that, rather than psychologists having a pervasive tendency to be inappropriately confident, their diagnostic confidence is affected by a multitude of variables operating differently across clinical contexts (cf. Dunning et al., 1990). For example, two variables that mediate the relationship between judgment confidence and validity are (a) length of professional experience and (b) validity of the clinical data on which the judgments are based. Both have been shown to vary directly with the appropriateness of diagnostic confidence (for a review, see Garb, 1986; 1998). Other studies have shown how task characteristics (e.g., level of difficulty,

quantity of information and its level of redundancy) affect levels of diagnostic confidence (e.g., Heller et al., 1992; Lichtenstein et al., 1982; Oskamp, 1965). Therefore, future research should address the question, in what contexts does a particular clinician operating with a particular clinical database articulate inferences with unwarranted confidence?

A subset of this research could further investigate the relationship between psychodiagnostic confidence and cognitive heuristics. This study, along with earlier research on these variables, suggests that heuristics play a primary role in the generation of confidence assessments and in fostering overconfidence. However, such intervening variables likely mediate the relationship between heuristics and confidence. At this juncture, data bearing on these questions would be highly informative to the discipline and useful inasmuch as this knowledge could be used to enhance the validity of the assessments that mental health professionals make.

An important area for future research is to delineate precisely the clinical sequelae of overconfidence for clinicians, clients, and clinical processes. Although research on overconfidence is underpinned by the assumption that the confidence people place in their judgments largely determines how they will subsequently act, this assumption has never been directly tested, and there is presently only very limited, indirect support for it (Sniezek, 1992). This point is of particular relevance to professional psychologists, as it is not difficult to imagine the far-

reaching and potentially detrimental ramifications when confidence arises from idiosyncratic clinician variables unrelated to pertinent client data. The concepts of action threshold and consequential variation described earlier could be useful points of departure for this research (Baumann et al., 1991). The use of action thresholds delimits the definition of overconfidence to those circumstances in which varying degrees of confidence give rise to different courses of action. Consequential variation further delimits the definition to situations in which confidence ratings falling on different sides of an action threshold differentially affect clinical outcomes. In this framework, poor calibration is important only when it is associated with divergent intervention strategies that differentially influence treatment outcomes. By this definition, only clinicians whose inferences eventuate in these anomalies are deemed to be overconfident. These measures are more suited for gauging the appropriateness of confidence in ill-structured diagnostic tasks in which it is difficult, sometimes even impossible, to delineate a single correct diagnosis. Moreover, in applied fields, it is extremely important to practitioners to know whether or not variability in their judgments (and the levels of confidence expressed therein) entails clinically meaningful consequences. While this method of operationalizing overconfidence in psychological assessment has much to recommend it, it has not yet been used in this area of empirical inquiry.

Finally, psychologists would appear to need to develop and evaluate strategies for eliminating unwarranted confidence in clinical judgment. It is clear that the implications of research based on general knowledge and numerical estimation tasks can only be stretched so far for professional psychology. Formalizing such strategies as those described in the previous section and submitting them to empirical scrutiny could prove valuable to a field that presently lacks effective and enduring solutions to psychodiagnostic overconfidence.

Conclusion

The baneful consequences of prematurely locking onto a diagnosis or any other judgment in the domain of uncertain contingencies and unidentified variables and doing this with unshakable confidence need to be examined, and remedies thereto need to be developed. This is especially true in psychotherapy and medicine where the personal costs of errors can be serious if not tragic. The profession is not for those who cannot live with some degree of ambiguity and uncertainty. Neither is promoting a false certitude warranted. In that perspective, it is evident that increasing attention must be paid by clinicians, educators, and students to the numerous threats to the validity of the judgments they are making. It is clear that no single, specific strategy will eliminate overconfidence in clinical judgment. Rather, successful interventions will foster realistic beliefs about how much clinical knowledge one possesses, and like any good therapeutic

intervention, they will have to be repeated frequently throughout training. Ultimately, clinicians working in this fuzzy domain must confront the dilemma posed by excessive reliance on dubious psychotherapeutics, on the one hand, and the inability to definitively formulate problems and appropriate plans of action, on the other.

References

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders*, fourth edition. Washington, DC: American Psychiatric Association.

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

Aronoff, D. N. (1997). *Errors in clinical judgment: The effect of temporal order of client information on anchoring, adjustment, and adjustment mitagation and category of clinical inferences.* Unpublished doctoral dissertation, McGill University, Montreal.

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*, 258-290.

Azevedo, R., Lajoie, S. P., & Fleiszer, D. M. (1996, August). *Complex clinical decision making in an ill-structured task (SAFARI).* Paper presented at the XXVI International Congress of Psychology, Montreal.

Batson, C. D. (1975). Attribution as a mediator of bias in helping. *Journal of Personality and Social Psychology, 32*, 455-466.

Baumann, A. O., Deber, R. B., & Thompson, G. G. (1991). Overconfidence among physicians and nurses: The 'micro-certainty, macro-certainty' phenomenon. *Social Science in Medicine, 32*, 167-174.

Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. *Journal of Forecasting, 1,* 257-269.

Block, R. A., & Harper, D. R. (1991). Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. *Organizational Behavior and Human Decision Processes, 49,* 188-207.

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes, 41,* 390-404.

Bryant, G. D., & Norman, G. R. (1980). Expressions of probability: Words and numbers. *The New England Journal of Medicine, 302(7),* 411.

Budescu, D. V., & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes, 36,* 391-405.

Canadian Psychological Association (1991). *Canadian Code of Ethics for Psychologists*. Author.

Chapman, L. J., & Chapman, J. P. (1967) Genesis of popular but erroneous diagnostic observations. *Journal of Abnormal Psychology, 72,* 193-204.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlations as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74,* 271-280.

Clark, H. H., & Clark, E. V. (1977). *Psychology and language.* New York: Harcourt Brace Jovanovich.

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

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

Cooke, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology, 25*, 1-17.

Derry, S. J., & Hawkes, L. W. (1993). Logical cognitive modelling of problem-solving behavior: An application of fuzzy theory. In S. P. Lajoie & S. J. Derry (Eds.), *Computers as cognitive tools: Technology in education* (pp. 107-140). Hillsdale, NJ: Lawrence Erlbaum.

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

Dumont, F., Sladeczek, E., & Martel, A. (1998). *Individual differences in the generation of clinical inferences: Relations among order of information, attribution, and experience.* Unpublished manuscript.

Dumont, F., & Lecomte, C. (1987). Inferential processes in clinical work: Inquiry into logical errors that affect diagnostic judgment. *Professional Psychology: Research and Practice, 18*, 433-438.

Dunning, D., Griffin, D.W., Milojkovic, J. D., & Ross, L. (1990). The overconfidence effect in social prediction. *Journal of Personality and Social Psychology, 58*, 568-581.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data. Second Ed.* MA: MIT.

Faust, D., & Ziskin, J. (1988). The expert witness in psychology and psychiatry. *Science, 241*, 31-34.

Fischhoff, B., & Bar-Hillel, M. (1980). Focusing techniques as aids to inference. *Decision Research Report, 80-9*, Decision Research, Eugene, Oregon.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 552-564.

Fox, J., Barber, D. C., & Bardhan, K. D. (1980). Alternative to Bayes? A quantitative comparison with rule-based diagnostic inference. *Method of Information in Medicine, 19*, 210-215.

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.

Garb, H. N. (1986). The appropriateness of confidence ratings in clinical judgment. *Journal of Clinical Psychology, 42*, 190-197.

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

Gauron, E. F., & Dickinson, J. K. (1969). The influence of seeing the patient first on diagnostic decision-making in psychiatry. *American Journal of Psychiatry, 126*, 199-205.

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

Goodin-Waxman, T. (1991). *Inferential reasoning during the psychodiagnostic assessment: Attribution, hypothesis-testing strategies, and final inferences as a function of theoretical orientation, level of experience, and temporal order.* Unpublished doctoral dissertation, McGill University, Montreal.

Griffin, D. W., Dunning, D., & Ross, L. (1990). The role of construal processes in overconfident predictions about the self and others. *Journal of Personality and Social Psychology, 59*, 1128-1139.

Heller, R. F., Saltzstein, H. D., & Caspe, W. B. (1992). Heuristics in medical and non-medical decision-making. *The Quarterly Journal of Experimental Psychology, 44A*, 211-235.

Hoch, S. J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *Journal of Personality and Social Psychology, 53*, 221-234.

Jones, E. E., & Nisbett, R. E. (1972). The actor and the observer: Divergent perceptions of the causes of behavior. In E. E. Jones, D. Kanouse, H. Kelly, R. E. Nisbett, S. Vallins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 79-94). Morristown, NJ: General Learning Press.

Jones, J. A. (1989). The verbal protocol: A research technique for nursing. *Journal of Advanced Nursing, 14*, 1062-1070.

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition, 11*,123-141.

Kelly, G. A. (1955). *The psychology of personal constructs, Vol. 1 & 2.* New York: Norton.

Kong, A., Barnett, G. O., Mosteller, F., & Youtz, C. (1986). How medical professionals evaluate expressions of probability. *The New England Journal of Medicine, 315(12)*, 740-744.

Lawlis, G. F., & Lu, E. (1972). Judgment of counseling process: Reliability, agreement, and error. *Psychological Bulletin, 78*, 17-20.

Lee, D. Y., Barak, A., Uhlemann, M. R., & Patsula, P. (1995). Effects of preinterview suggestion on counselor memory, clinical impression, and confidence in judgments. *Journal of Clinical Psychology, 51*, 666-675.

Levenberg, S. B. (1975) Professional training, psychodiagnostic skill, and Kinetic Family Drawings. *Journal of Personality Assessment, 39*,389-393.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306-334). New York: Cambridge University Press.

Lichtenstein, S., & Newman, J. R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Science, 9(10),* 563-564.

Machover, K. (1949). *Personality projection in the drawing of the human figure*. Springfield, IL: Charles C Thomas.

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

Mahoney, M. J. (1991). *Human change processes: The scientific foundations of psychotherapy.* New York: Basic Books

Mapes, R. E. A. (1979). Verbal and numerical estimates of probability in therapeutic contexts. *Social Science in Medicine, 13A,* 277-282.

McReynolds, P. (1989). Diagnosis and clinical assessment: Current status and major issues. *Annual Review of Psychology, 40*, 83-108.

Meehl, P. (1957). *Journal of Counselling Psychology, 4*, 268-253.

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

Merz, J. F., Druzdzel, M. J., & Mazur, D. J. (1991). Verbal expressions of probability in informed consent litigation. *Medical Decision Making, 11*, 253-281.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

Moore, P. G. (1977). The manager's struggle with uncertainty. *Journal of Royal Statistical Society, 140*(2), 129-165.

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

Murphy, A. H., Lichtenstein, S., Fischhoff, B., & Winkler, R. L. (1980). Misinterpretation of precipitation probability forecasts. *Bulletin of the American Meteorological Society, 6*, 695-701.

Nakao, M. A., & Axelrod, S. (1983). Numbers are better than words: Verbal specifications of frequency have no place in medicine. *The American Journal of Medicine, 74*, 1061-1065.

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

Ross, L., & Nisbett, R. E. (1991). The person and the situation: Perspectives of social psychology. New York: McGraw-Hill

Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs: General and Applied, 76*, no. 547.

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

Paese, P. W., & Kinnaly, M. (1991). Effects of role assignment and verbal interaction on accuracy and overconfidence in interpersonal judgment. *Journal of Applied Social Psychology, 21*, 1418-1439.

Paese, P. W., & Sniezak, J. A. (1991). Influences on the appropriateness of confidence in judgment: Practice, effort, information, and decision making. *Organizational Behavior and Human Decision Processes, 48*, 100-130.

Phillips, L. D. (1973). *Bayesian statistics for social scientists.* London: Nelson.

Reagan, R. T., Mosteller, F., & Youtz, C. (1989). Quantitative meanings of verbal probability expressions. *Journal of Applied Psychology, 74*, 433-442.

Reyna, V. F. (1981). The language of possibility and probability: Effects of negation on meaning. *Memory and Cognition, 9*, 642-650.

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

Robertson, W. O. (1983). Quantifying the meanings of words. *Journal of the American Medical Association, 249*, 2631-2632.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (vol. 10, pp. 337-384). New York: Academic Press.

Ross, L. (1987). The problem of construal in social inference and social psychology. In N. Grunberg, R. E. Nisbett, & J. Singer (Eds.), *A distinctive approach to psychological research: The influence of Stanley Schachter* (pp.118-150). Hillsdale, NJ: Erlbaum.

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology, 13*, 279-301.

Ross, L., & Nisbett, R.E. (1991). *The person and the situation: Perspectives of social psychology*. Montreal: McGraw-Hill.

Schaeffer, M. H. (1989). Environmental stress and individual decision-making: Implications for the patient. *Patient Education and Counseling, 13*, 221-235.

Smith, D., & Dumont, F. (1997). Eliminating overconfidence in psychodiagnosis: Strategies for training and practice. *Clinical Psychology: Science and Practice, 4*, 335-345.

Smith, D., & Dumont, F. (1995). A cautionary study: Unwarranted interpretations of the Draw-A-Person Test. *Professional Psychology: Research and Practice, 26*, 298-303

Smithson, M. (1987). *Fuzzy set analysis for behavioral and social sciences.* New York: Springer-Verlag.

Sniezak, J. A., Paese, P. W., & Switzer, F. S.(1990). The effect of choosing on confidence in choice. *Organizational Behavior and Human Decision Processes, 46*, 264-282.

Snyder, M. (1981). Seek, and ye shall find: Testing hypotheses about other people. In E. T. Higgins, C. P. Herman, & M. P. Zanna (Eds.), *Social cognition: The Ontario Symposium.* (Vol.1, pp. 277-303). Hillsdale, NJ: Erlbaum.

Strohmer, D. C., & Chiodo, A. L. (1984). Counselor hypothesis testing strategies: The role of initial impressions and self-schema. *Journal of Counseling Psychology, 31*, 510-519.

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

Tinsley, H. E., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology, 22*, 358-376.

Turk, D. C., & Salovey, P. (1988). Psychotherapy: An information-processing perspective. *Reasoning, inference, and judgment in clinical psychology*. New York: The Free Press.

Turk, D. C., Salovey, P., & Prentice, D. A. (1988). Psychotherapy: An information-processing perspective. In D. C. Turk & P. Salovey (Eds.), *Reasoning, inference, and judgment in clinical psychology* (pp.1-14). New York: The Free Press.

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

Vallone, R., Griffin, D., Lin, S., & Ross, L. (1990). Overconfident prediction of future actions and outcomes by self and others. *Journal of Personality and Social Psychology, 58*, 582-592.

Wallsten, T. S. (1990). The costs and benefits of vague information. In R. M. Hogarth (Ed.) *Insights in decision making: A tribute to Hillel J. Einhorn,* pp.28-43. Chicago: University of Chicago.

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General, 115(4)*, 348-365.

Zakay, D. (1985). Post-decisional confidence and conflict experienced in a choice process. *Acta Psychologica, 58*, 75-80.

Zedah, L. A. (1965). Fuzzy sets. *Information and Control, 8*, 338-353.

Zimmer, A. C. (1983). Verbal versus numerical processing of subjective probabilities. In R. W. Scholz (Ed.) *Decision making under uncertainty.* pp. 159-182. North Holland: Elsevier.

Zimmer, A. C. (1984). A model for the interpretation of verbal predictions. *International Journal of Man-Machine Studies, 20*, 121-134.

Zimmer, A. C. (1986). What uncertainty judgments can tell about underlying subjective probabilities. In L. H. Kanal & J. F. Lemmer (Eds.) *Uncertainty in artificial intelligence.* Amsterdam: Elsevier.

Zhu, S.-H. (1992). Contexts effects in semantic interpretations: A study of probability words. *Dissertation Abstracts International, 53B*, 593.

Zwick, R., & Wallsten, T. S. (1989). Combining stochastic uncertainty and linguistic inexactness: Theory and experimental evaluation of four fuzzy probability models. *International Journal of Man-Machine Studies, 30*, 69-111.

APPENDICES

Appendix A: Instructions to Participants

You are going to be presented with a case file of a person who has sought help from a mental health professional. This case file is not complete, nor is it organized in a traditional assessment format. Your task is to provide a clinical assessment of this client in the same manner you do for other clients you meet in your clinical practice. Please use the assessment style or diagnostic method that you are most comfortable with. We are not, however, necessarily looking for a technical diagnostic label. Rather, we would like you to furnish us with your understanding of this client's problems or disorder.

The information in the case file will be presented in short segments, one at a time, which you are to read aloud. After reading each segment, or, preferably, at any time while reading the segment, we want you to express verbally any thoughts you may have bearing on your assessment. If the information does not enhance your understanding of this client's problems or disorder in any way, please indicate this aloud. Because it is the manner in which you process this information that is of interest to us, it is important that you express ALL YOUR THOUGHTS AT THE MOMENT THEY OCCUR.

Before you begin, we will do a short task in which you can practice reading and thinking aloud. Do you have any questions before you begin?

Appendix B: Consent Form (Main Study)

I am consenting to participate in a research project in which I will be making an assessment of a client who has sought psychological help on the basis of a written case file. I have been informed that the purpose of the project is to understand some of the ways in which clinicians process client information during the assessment procedure. I understand that my responses to the information in the case file will be recorded on a cassette recorder and anonymously subjected to techniques of discourse analysis.

I have been informed that any information obtained in connection with this study that can be identified with me will remain strictly confidential and that any written reports or publications will only include aggregated data. I am aware that I am free at any time or for any reason to discontinue participation in this study.

_____

Signature of Participant

_____ _____

Signature of Researcher Date

Appendix C: Casefile

Presenting Problem: Mr. B., a 42-year old single male, complains of constant hand tremors.

- He has difficulty sleeping and reports having no energy.
- He suffers from a serious case of psoriasis (a skin disorder).
- Mr.B. reports that although he has been troubled by hand tremors for the past 20 years, these tremors have become worse in the past couple of years.
- He notices that his hand trembles most often when he is required to write, particularly in the presence of others.
- He is especially anxious when he goes to the bank and needs to endorse his cheques.
- He notices that some of the tellers look at him strangely. He worries that people will notice his shakiness.
- The client, of Ukrainian extraction, was born in a small farming community in Ontario in 1948.
- The family moved to an upper-lower class suburb in Montreal when he was 8 years old.
- Mother, now aged 67, lived with his father until he died 2 years ago.
- When Mr. B. was a child, his mother worked in a large office equipment firm as a maintenance woman.

- Mother had a nervous breakdown when client was in middle childhood and was hospitalized as a day patient.
- The client's social and psychosexual activities reveal a checkered pattern. He had his first sexual experience when he was 16 years old.
- He then hooked up with an older woman; this relationship lasted until he was 21.
- He reports that he was mean with her, although he dreaded becoming like his father, who had always mistreated his mother.
- He then had no steady girl friends but continued dating steadily, until he met his wife.
- He had a brief homosexual relationship in his mid-20's but reports no sexually related illness.
- Thirteen years ago, Mr. B. became involved with a divorced woman who had a baby daughter. He lived amicably with this woman and his adoptive daughter until 18 months ago when the wife died.
- The daughter then returned to live with her biological father. He was devastated by this double loss.
- Patient had attended group meetings for one-parent families with his common-law wife.
- The client describes his mother as a strong woman whom he admires and still feels close to.

- However, he was distraught by her sale of the family farm soon after the father's death. He was very attached to that farm.
- Mother was diagnosed as having cancer 2 years ago. Her condition has worsened but she seems to be coping well.
- Client lost his uncle 2 and 1/2 years ago.
- The father had worked as a night foreman in his later years. Prior to that he had also worked as a boxing coach.
- Client reports that his father drank heavily, and was an insecure and irresponsible man.
- He recalls that the father had fits of rage when he was drunk, smashing objects about the house. He also beat the mother as well as the patient frequently.
- Physical examination reveals scars on the right buttock due, reports the client, to strappings by his father.
- He is the second of three children. He has two sisters, one 4 years older than he and the other 6 years younger.
- Both sisters are living with their husbands and their children.
- He has a soft spot for his younger sister who was also abused by his father. She has recently moved back to Ontario with her family.
- The mother is also closest to the younger sister.

- Of his early education Mr. B. recalls that the principal of the school struck him on the knuckles with a ruler for failing for answer a question. He was then in the second grade.
- In another elementary school year client recalls failing math. The client reports that he failed grade 6, a year when he was heavily involved in amateur boxing.
- The client the client reports that he completed secondary school and did 2 years of undergraduate work in organization and management in evening courses.
- At this time he was working at a large office equipment firm as a maintenance worker.
- The client has few friends. He finds himself drinking daily, about 4 bottles of beer per day, but he doesn't want to talk about it because it reminds him of his father.
- He gets together with his buddy Joe, once a week to have dinner and share a bottle of wine. The two discuss their problems and despair over their solutions.
- His friend is married but has no children.
- Even as a child, the client kept pretty much to himself. He did not feel like one of the group. The kids at school made fun of him. He preferred to be alone, though secretly he feared others would not accept him.
- He was not active in social sports, but under his father's guidance he began to train as a boxer.

- The client's hand tremors are a hurdle for social involvement.
- He worries that his hands will shake so badly that his coffee or other drinks will spill. He thinks that if this happens he will have to explain this problem to the person he is with.
- He is afraid that they will think he is weird. This prevents him from becoming socially and intimately involved with others.
- He has recently met a woman he would like to get to know but he is afraid that she will find out about his tremors, especially in a restaurant.
- Although he has always been nervous and shaky inside, his problems in writing and holding cups of coffee have begun to seriously trouble him during the past 2 years.
- Although the client has experienced difficulties in the past in falling asleep, he is having even more trouble now.
- He thinks about the nervousness and finds that this prevents him from falling asleep.
- His recent insomnia has begun to incapacitate him, and he feels less energetic than in the past. This chronic fatigue has kept him from doing much outside of his work.
- He still suffers from this fatigue but pushes himself to get going.
- During most of the client's work history he has worked as a maintenance worker.

- He has worked in large institutions but also in a small neighbourhood community centre.
- About 5 years ago he left the community centre after he had become chief superintendent. He quit that job, for unspecified reasons, to return to a large firm.
- He is still employed there although he is now on sick leave.
- His present boss has shown much hostility towards him. He thinks his superiors want to fire him.
- He attributes the hostility to the fact that he is the only person of his ethnic background in the maintenance-worker group.
- The others are all of a different (but homogeneous) ethnic background. They treat him like an outsider; they'd like to get rid of him.
- His work situation has worsened drastically since his wife's death. One superior has officially reported that he has been angry with this employee (that is, the client).
- When he has to fill out reports in front of his boss his hand tremors increase to the point where his writing is not legible.
- His leave of absence is now running out. He fears that the hospital staff who will file the insurance report will not justify extension of his leave.
- The client has recently broken off with a girl friend who thinks that she has picked up genital warts from him.
- He reports having no other women friends outside of his family.

- He relates well with hospital staff. He reports that they, including the psychiatrist, the clinical psychologist, and art therapist who have worked with him, are warm and supportive. He has reciprocated with warmth and cordiality.

Appendix D: Verbal Probabilities: Selective Literature Review

Fuzzy Sets and Subjective Probability

Much of the information that we communicate and receive is vague rather than precise. A statement or word is *precise* if it can be understood in one and only one sense and *vague* if it is not clearly defined or cannot be understood in at least one specified and precise way (Wallsten, 1990). Fuzzy set theory provides a framework for conceptualizing these notions (cf. Zedah, 1965). Within fuzzy set theory, an element may partially belong to a set rather than belong completely or not at all to a set, as defined in classical set theory (Smithson, 1987).

Fuzzy sets have gradations of set membership and, as such, resemble categories of meaning used in natural language. Fuzzy set theory has been used to conceptualize and quantify the meanings of common verbal expressions of probability (e.g., unlikely, possible, probable, etc.) (Wallsten, Budescu, Rapoport, Zwick, & Forsyth, 1986; Zimmer, 1984). The subjective probability of a fuzzy concept, defined as the measure of one's degree of belief in that concept, can be represented formally as a function on the [0, 1] probability interval, as is illustrated in Figure 1 (Zwick & Wallsten, 1989). Functions usually take zero as their minimum value to represent probabilities that are definitely not in the concept represented by the expression and one as their maximum value to represent probabilities that are definitely in the concept. Probabilities with intermediate degrees of membership in the concept are represented by

intermediate values (between zero and one). Within fuzzy set theory, this function is called a *membership function*, which is generally defined as a rule that assigns a number between zero and one to each element within the universe of discourse indicating the degree of membership of that element in a particular set (Wallsten et al., 1986). As applied to linguistic concepts such as verbal probability expressions, a membership function graphically represents the degrees of vagueness inherent to a concept, a characteristic captured by the breadth of the function.

Verbal Probabilities

People's verbal interactions are regularly and frequently laced with expressions of subjective probabilities bearing on an immense range of judgments and predictions (e.g., it might snow today; interest rates are sure to drop; he's unlikely to be acquitted.). This applies as much to communication among experts and professionals as it does to conversations among lay individuals. Because of the importance of probability assessments in professional decision making processes, the rich lexicon of probability expressions in the English language, and the potential for ambiguity and confusion in their use, investigators have undertaken to specify precisely (i.e., numerically) what these words and phrases mean. Ultimately, the goal of this research is to improve human decision making processes by reducing the ambiguity of these expressions.

Intra-individual consistency. Research has demonstrated that the numerical estimates or rank ordering of probability words and phrases that an individual provides are largely consistent over time. Budescu and Wallsten (1985) had subjects compare and rank 19 probability expressions on three different occasions. Their subjects provided stable rankings of these expressions over time. Beyth-Marom (1982) had subjects provide multiple numerical estimates of the same probability expressions in two experiments. In the first, expressions were presented alone and in the second, they were embedded in a meaningful context (i.e., a paragraph). In both parts, subjects' ratings were "highly consistent.» Finally, Wallsten et al. (1986) demonstrated that, in addition to high test-retest reliability of ratings, the shapes of membership functions (i.e., the numerical distribution represented by a probability expression) remain constant over time within-subjects.

Inter-individual consistency. In contrast, studies mapping probabilistic expressions over the probability interval [0, 1] have consistently revealed that there is a considerable variability *between* individuals in the numerical values they assign to specific probability words or phrases (Budescu & Wallsten, 1985). However, there is some systematic variability depending where the expression falls on the probability interval. Expressions that fall near anchor points (0, 0.5, and 1) show the most consistency across subjects compared to expressions falling between these anchor points (Wallsten et al., 1986). Another robust finding is that

there is considerable numerical overlap between probability terms (Budescu & Wallsten, 1985). For example, in one study (Lichtenstein & Newman, 1967), the terms "likely and "unlikely" together covered a range of probabilities from .01 to .99 across subjects and overlapped between .25 and .45. The medians of the two phrases were .16 and .75. The same terms in another study (Bryant & Norman, 1980) covered a range of .00 to .95, and their medians were, respectively, .20 and .75. These results highlight two other generalizations that can be made about verbal expressions of probability. Firstly, in spite of individual differences, measures of central tendency calculated in many studies reveal that numerical estimates of verbal probabilities are quite consistent among groups (Kong, Barnett, Mosteller, & Youtz, 1986; Reagan, Mosteller, & Youtz, 1989), as Table 1 illustrates. Secondly, "verbally symmetric" terms (e.g., likely and unlikely) are not "numerically symmetric" as one might intuitively expect; that is, they tend not be equidistant from the midpoint of the [0, 1] probability interval (Budescu & Wallsten, 1985).

Context effects. In order to explain the variation among numerical estimates for specific probability expressions, researchers have investigated how different contexts in which verbal probabilities are interpreted affect the numerical ratings they are assigned.

Beyth-Marom (1982) had expert political forecasters provide single number equivalents for probability expressions (a) presented in isolation and (b)

embedded within text extracted from political forecasting documents. Her results revealed greater variability across participants when ratings were given in-context rather than in isolation. She suggested that, aside from factors controlled in this study (participant's' field of expertise and the textual context in which expressions were embedded), other contextual variables might have contributed to this result, including (a) the ambiguity with which events in the texts were described and (b) participants' differing personal opinions and values regarding outcomes. Brun and Teigen (1988) reported evidence supporting the latter. Their data revealed that participants' ratings of probability expressions were biased in various ways by personal opinions about the events under consideration, suggesting that less controversial topics elicit more consistency among raters.

Several researchers have studied the effects of context within the domain of medical decision making. Mapes (1979) had physicians interpret probability expressions that were used to describe the side effects of different drugs. Ratings tended to be more extreme (i.e., less probable) when an expression was associated with a drug with severe as opposed to mild side effects, leading Mapes to conclude that the meaning of an expression changes with the context. For example, 20.7% of participants matched the term "rare" with the category "less than 1 per 1000" when considering the relatively mild side effects of an antihistamine compared with 59.4% who matched it with "less than 1 per 1000" when considering the more severe side effects of a beta-blocker. Alternatively,

this difference might reflect divergent perceptions of the rates of each drug's side effects (Kong, Barnett, Mosteller, & Youtz, 1986).

Kong et al. (1986) had medical personnel interpret probability expressions indicating the likelihood that a particular symptom would occur in a patient with a disease that remained unspecified. Ratings were made on three structurally different scales: (a) free choice (0-100 in increments of 1), (b) uniform scale (0-10 in increments of 1), and (c) high or low probability scale (75-95 in increments of 5 and 95-100 in increments of 1 or, conversely, 25-5 in increments of 5 and 5-0 in increments of 1). They found that ratings did not differ significantly among the different groups of participants (physicians, medical students, and other medical professionals), although the scale structure had a significant effect on ratings. Comparing their data to results of five other similar studies involving medical personnel, they noted that the order in the expressions were ultimately ranked varied considerably less than the actual probability values they were assigned, which they deemed as "encouraging for the future prospects of codifying the meaning of such expressions" (p.740).

Merz, Druzdzel, and Mazur (1991) studied how physicians acting as expert witnesses in informed consent litigation used verbal probabilities to characterize the risks of medical procedures. They found that probability expressions typically represented broad numeric interpretations, and their analyses only distinguished opposing extreme categories of probability expressions (i.e.,

"low" and "extremely low" corresponded to significantly lower numeric probabilities than "high" and "very high"). Consistent with Mapes (1979), they also found that a significant proportion of variance in numeric interpretations was explained by the severity of consequences being characterized by the verbal probability. As such, expressions characterizing a severe medical complication (e.g., a "low chance" of death) relative to a less consequential outcome (e.g., a "low chance" of infection) tended to correspond to lower probabilities.

Brun and Teigen (1988) varied knowledge domain and the presentation mode (i.e., written text versus video) to investigate their effects on numeric interpretations of probability expressions. Their results indicated that embedding expressions in a particular context often, but not always, impacted significantly on numeric ratings leading to higher inter-subject variability relative to control conditions in which expressions were presented in isolation. Ratings also varied within-subjects across contexts, in contrast to previous predictions (Budescu & Wallsten, 1985). Among the exceptions, variability was not significantly higher among physicians interpreting expressions within a medical treatment context compared to controls. Moreover, physicians as a group evidenced significantly less variability than a group of lay people interpreting the same expressions in the same context. This is consistent with a previous finding suggesting that individuals with homogeneous backgrounds interpret expressions more

consistently than individuals with more heterogeneous backgrounds (Moore, 1977).

Zhu (1992) developed and tested a model to explain the effect of context on the interpretation of verbal probabilities. His results suggested that, when interpreting probability expressions, people start with a prototypical meaning of a word and then make adjustments given the contextual cues. The cues to which people attend in any context, however, are not always obvious nor the same. Contexts can differentially effect people, leading to either decreased or increased variability across individuals (cf. Brun and Teigen, 1988).

Appendix E: Validation Study

Participants

20 participants will be solicited from the cohort of second-year graduate students in counselling psychology at McGill University to participate in a study of "how psychologists interpret verbal probability expressions commonly used in clinical practice.» All participants in this study will be treated in accordance with the Canadian Psychological Association's (CPA) "Canadian Code of Ethics for Psychologists" (CPA, 1991).

Materials

All 20 candidates will be presented: (a) 2 copies of a consent form and (b) the research protocol.

I. Consent form (see Appendix F). Those students who choose to participate will be asked to complete a consent form. It will inform them of their right to terminate their participation in the study at any time and assure them that their responses will remain confidential and anonymous.

II. Research protocol. The research protocol, which is essentially a paper-and-pencil rating task, will contain 30 different diagnostic statements. The diagnostic statements will be drawn from the clinician-generated think-aloud protocols. These think-aloud protocols, which contain clinicians problem formulations based on an anonymous individual's hospital file, were gathered previously and will be the subject of analyses in the main study in this research project. Each diagnostic statement will contain one probability expression (highlighted with an underline) that participants will be asked to rate. In total, 15 different probability expressions will be presented, and each expression will appear twice. Participants will be asked to assign a rating to each expression on a 7-point scale (cf. "The magic number seven plus or minus two", Miller, 1956) ranging from 1 (lowest probability) to 7 (highest probability). Brun and Teigen (1988) demonstrated that ratings on this scale corresponded closely to ratings of the same terms on a 100-point probability scale and, thus, concluded that the precision and reliability of estimates was not significantly affected.

On the final page of the research protocol, the list of 15 probability expressions will be listed in random order. Below this will be a list of 15

numbered blank lines, which participants will use to rank order the expressions. Lines 1 and 15 will be accompanied by linguistic anchors, "lowest probability" and "highest probability" respectively. In contrast to the first part of this study in which participants assigned expressions to a set number of ordered categories, the ranking task places fewer constraints on participants interpretations and leaves the final number of categories open (cf. Beyth-Marom, 1982).

Item selection. Two principal sources will be used for compiling the list of 15 probability expressions to be presented to participants. The first is the think-aloud protocols, described above, since these constitute instances of professional discourse from a representative sample of (novice and expert) clinicians. The second source is previous studies that have investigated the meanings of verbal probabilities from a related professional domain, namely medicine. Using these two sources, it is expected that the final list of fifteen words will meet two conditions: (a) it will reflect probability expressions currently in use among health professionals, including psychologists and (b) it will cover the full numerical probability range.

Procedure

Participants will be solicited by distributing a memorandum by internal mail. It will invite 20 second-year counselling psychology students to participate in this study and offer a remuneration (of about $5) to participants who complete

the experimental tasks. Participants will be instructed to assemble at a predetermined time and place to complete the experimental tasks.

When the 20 participants are assembled, they first will be asked to read the Statement of Informed Consent (see Appendix F), and if they decide to participate in the study, to sign and date the form. Next, they will be presented with the research protocol. Participants will be instructed to read each statement and indicate how they interpret the underlined probability expression contained in each statement by assigning an appropriate rating on a 7-point scale. Following this, participants will be asked to interpret the meaning of the 15 probability expressions by rank ordering them from lowest to highest probability.

Data Analyses

The data collected in this study will be analyzed in light of determining: (a) to what degree are participants consistent within themselves in their numerical interpretations of probability expressions; (b) to what degree are numerical ratings consistent across all participants; (c) to what degree are group ratings in this study consistent with aggregated data (e.g., medians) reported in previous studies; and (d) what levels of probability are typically assigned to particular probabilistic expressions.

Testing consistency. Three sets of descriptive statistics will be used to determine the degree of intra-individual consistency. First, participants' probability ratings from the in-context task (the first part of the experimental

procedure) made at time 1 will be cross-tabulated with ratings made at time 2. A separate table will be constructed for each expression. Second, the correlation between individual ratings given at time 1 and time 2 will be calculated for each expression. Finally, two separate correlation analyses of aggregated group data for the 15 expressions will be undertaken. The interquartile mean (i.e., the mean of the middle 50% of ratings) will be used in these analyses, thus eliminating extreme and presumably unrepresentative responses (Beyth-Marom, 1983; cf. 50% credible interval, Phillips, 1973). In the first analysis, interquartile mean ratings for the 15 expressions at time 1 will be correlated with interquartile mean ratings at time 2. In the second analysis, data from the in-context rating task will be compared to data from the ranking task, which constitutes the second part of the experimental procedure. This analysis will involve correlating the 15 grand means from the in-context rating task (calculated as the averages of the two interquartile means at times 1 and 2) with the mean ranks assigned to the 15 expressions.

Several statistical procedures will be used in determining the degree of inter-individual consistency among ratings. First, a table will be constructed that includes the mean, standard deviation, interquartile range, and the full range of numeric ratings for each expression resulting from the rating (in-context) task. Similarly, a table of means, standard deviations, and ranges derived from the ranking task will be constructed. The degree of intersubject agreement will be

evident in the amount of dispersion in both the numeric ratings and rankings as represented by these descriptive statistics. Finally, inter-individual agreement will also be assessed by generating a table of intersubject correlations. In this analysis, each subject will be represented by a list of 15 numerical ratings, each of which is the mean of the two ratings given for each probability expression. The mean of the intersubject correlations will be calculated and will represent the overall level of agreement among participants for the 15 verbal probabilities.

Appendix F: Consent Form (Preliminary Study)

Counsellors frequently use expressions of verbal probability (e.g., "improbable", "most likely") in their communications to others when they wish to characterize the likelihood of some event occurring. However, because the meanings of such expressions are by their nature vague rather than precise, they are not always correctly understood. The main objective of this research project is to determine what counsellors typically interpret different probability expressions to mean in numerical terms. Results of this research will help to clarify the intended meaning of probability expressions when they are used by counsellors in their professional work.

If you agree to participate in this project, you will simply be asked to indicate what each of several verbal probability expressions means by furnishing a numerical equivalent of that expression as instructed in the task. It should take you about 15-20 minutes to complete the entire protocol. Your participation in this research project entails no conceivable risks to your personal welfare. The benefits you stand to gain by participating in this project include compensation for your time (a meal voucher worth about $5) and the personal satisfaction of contributing to a scientific endeavor.

Be advised that your responses will be held in strictest confidence, and any public presentations or publications stemming from this study will include only

aggregated data. Also, it is your right to withdraw your consent to participate in this study at any time.

I, _____, having been fully informed about the purpose and methods of this study, the risks and benefits it entails, and my rights to confidentiality and to withdraw this consent at my discretion, agree freely to participate in this study being conducted by:

David Smith, Doctoral Candidate

Department of Educational and Counselling Psychology, McGill University

Signature of Participant:

_____Date:_____

APPENDIX G: CONFIDENCE RATING CHART

EXPRESSION	PROBABILITY: WEIGHTED MEANS	CONFIDENCE RATING
almost always	89.00	89.00
almost certain	78.00	78.00
almost never	4.00	96.00
always	94.00	94.00
apparently	68.00	68.00
atypical	11.00	89.00
barely possible	13.00	87.00
best bet	76.00	76.00
better than even	58.00	58.00
bound to	89.00	89.00
can't rule out	38.00	62.00
cannot be excluded	47.00	53.00
certain	95.00	95.00
certainly	85.00	85.00
chance (for)	47.00	53.00
characteristic	81.00	81.00
characteristically	89.00	89.00
classic	86.00	86.00
common	71.00	71.00
commonly	72.00	72.00
compatible with	65.00	65.00
conceivably	60.00	60.00

consistent with	69.00	69.00
consistently	89.00	89.00
could	51.00	51.00
danger	56.00	56.00
definitely	95.00	95.00
doubtful	20.00	80.00
doubtfully	16.00	84.00
doubtlessly	88.00	88.00
effectively excludes	46.00	54.00
exceptionally	5.00	95.00
expected	75.00	75.00
extremely common	86.00	86.00
faintly possible	13.00	87.00
fair chance	51.00	51.00
fairly likely	66.00	66.00
fairly unlikely	25.00	75.00
frequent	68.00	68.00
frequently	73.00	73.00
generally	72.00	72.00
guess	47.00	53.00
good chance	72.00	72.00
great chances	76.00	76.00
has got to	86.00	86.00
high probability	88.00	88.00
highly improbable	6.00	94.00

highly probable	89.00	89.00
impossible	3.00	97.00
improbable	13.00	87.00
inconclusive	43.00	57.00
infrequent	19.00	81.00
invariably	89.00	89.00
likely	69.00	69.00
likely not	14.00	86.00
low probability	18.00	82.00
majority	70.00	70.00
may	36.00	64.00
maybe	49.00	51.00
might	45.00	55.00
moderate probability	62.00	62.00
moderate risk	54.00	54.00
more often than not	64.00	64.00
most	75.00	75.00
must	87.00	87.00
never	2.00	98.00
no chance	3.00	97.00
normally	79.00	79.00
not certain	38.00	62.00
not inconsistent with	65.00	65.00
not infrequently	48.00	52.00
not likely	13.00	87.00

not much chance	16.00	84.00
not often	15.00	85.00
not possible	3.00	97.00
not probable	13.00	87.00
not quite even	44.00	56.00
not unreasonable	33.00	77.00
not usual	20.00	80.00
not very probable	20.00	80.00
obviously	86.00	86.00
occasionally	21.00	79.00
odds on	74.00	74.00
often	67.00	67.00
on occasion	27.00	73.00
perhaps	43.00	57.00
possible	36.00	64.00
possibly	39.00	61.00
predictable	72.00	72.00
pretty good chance	67.00	67.00
probable	67.00	67.00
probably	72.00	72.00
quite certain	82.00	82.00
quite likely	79.00	79.00
quite unlikely	11.00	89.00
rare	6.00	94.00
rarely	9.00	91.00

rather	58.00	58.00
rather likely	69.00	69.00
rather unlikely	21.00	79.00
scarcely	10.00	90.00
seems	62.00	62.00
seldom	16.00	84.00
should	71.00	71.00
significant chance	58.00	58.00
slight odds against	45.00	55.00
slight odds in favor	55.00	55.00
small chance	16.00	84.00
sometimes	26.00	74.00
somewhat likely	59.00	59.00
somewhat unlikely	31.00	69.00
sounds	68.00	68.00
suggests	58.00	58.00
supports	65.00	65.00
sure	93.00	93.00
tend to believe	63.00	63.00
think	63.00	63.00
tossup	50.00	50.00
typical	77.00	77.00
uncertain	40.00	60.00
undoubtedly	91.00	91.00
unlikely	18.00	82.00

unusual	13.00	87.00
usual	76.00	76.00
usually	76.00	76.00
usually not	18.00	82.00
vast majority	89.00	89.00
very doubtful	8.00	92.00
very improbable	25.00	75.00
very likely	86.00	86.00
very often (happens)	80.00	80.00
very probable	86.00	86.00
very seldom	12.00	88.00
very unlikely	9.00	91.00
very unusual	5.00	95.00
well might	60.00	60.00
wonder	41.00	59.00
would (happen)	65.00	65.00

Appendix H: Instructions to Participants

In the following pages, there are a number of statements that any mental health professional might make about a client in any one of various professional contexts, such as a case conference, an informal consultation with a colleague, or an assessment report. In each of these statements you will find an underlined word or phrase that expresses how much confidence the speaker places in his or her problem formulation. The following is an example of such a statement:

There is a <u>good chance</u> that this man is depressed.

Underneath each statement is a scale with numbered points ranging from 1 to 9, which appears as follows:

---

1 2 3 4 5 6 7 8 9

This scale is simply a coarser version of the full probability scale ranging from 0 (or 0%) to 1 (or 100%). Therefore, the number "1" on this 9-point scale represents probabilities falling at or close to 0 and the number "9" stands for probabilities falling at or close to 1. All other numbers represent values between these two extremes.

Your task is to indicate what probability the underlined expression corresponds to by circling the appropriate number on the 9-point probability scale. In other words, if a counsellor says that there is a "good chance" that someone is depressed, what numerical probability does the expression "good

chance" refer to? If you feel that a particular expression actually falls between two numbers on the scale, circle the number which you believe the expression falls closest to, even if it is only marginally closer. Please circle only one number for each item, and please respond to every item. In the last section, you are asked to rank order a set of expression from lowest to highest probability. Please read the instructions carefully before beginning.

The entire task is divided into three sections. When you are responding to items in each of these sections, please **do not refer back** to previous sections you have completed to help you select any of your responses. Remember that what I am interested in is your interpretations of these probabilistic expressions, so it is impossible for you to respond incorrectly.

Instructions to Participants (Revised)

In the following pages, there are a number of statements that any mental health professional might make about a client in any one of various professional contexts, such as a case conference, an informal consultation with a colleague, or an assessment report. In each of these statements you will find an underlined word or phrase that expresses **the probability that the clinician's judgment is correct**. The following is an example of such a statement:

(a) There is a good chance that this man is depressed.

(b) It is very unlikely that she will benefit from treatment.

Underneath each statement is a scale with numbered points ranging from 1 to 9, which appears as follows:

---

1 2 3 4 5 6 7 8 9

This scale is simply a coarser version of the full probability scale ranging from 0 (or 0%) to 1 (or 100%). Therefore, the number "1" on this 9-point scale represents probabilities falling at or close to 0 and the number "9" stands for probabilities falling at or close to 1. All other numbers represent values between these two extremes.

Your task is to indicate what probability the underlined expression corresponds to by circling the appropriate number on the 9-point probability scale. In other words, if a clinician says that there is a "good chance" that

someone is depressed, what numerical probability does the expression "good chance" refer to? What numerical probability does the expression "very unlikely" refer to? If you feel that a particular expression actually falls between two numbers on the scale, circle the number which you believe the expression falls closest to, even if it is only marginally closer. Please circle only one number for each item, and please respond to every item.

The entire task is divided into three sections. When you are responding to items in each of these sections, please **do not refer back** to previous sections you have completed to help you select any of your responses. Remember that what I am interested in is your interpretations of these probabilistic expressions, so it is impossible for you to respond incorrectly.