COGNITIVE DE-BIASING AND THE ASSESSMENT OF PEDIATRIC BIPOLAR DISORDER

Melissa M. Jenkins

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology (Clinical Psychology).

Chapel Hill
2012

Approved by:

Eric A. Youngstrom, PhD

Deborah J. Jones, Ph.D.

Enrique W. Neblett, Jr., PhD

Jen Kogos Youngstrom, PhD

Melanie C. Green, Ph.D.

ABSTRACT

MELISSA M. JENKINS: Cognitive De-Biasing and the Assessment of Pediatric Bipolar Disorder
(Under the direction of Eric A. Youngstrom, PhD)

Decades of research have demonstrated that relying solely on clinical judgment leaves one prone to a host of cognitive errors that compromise optimal decision-making. Clinical judgment appears to be particularly vulnerable to faulty heuristics and biases when assessing for pediatric bipolar disorder, as evidenced by staggering rates of misdiagnosis and overdiagnosis. Despite abundant evidence documenting the problems associated with clinical judgment, little research to date has explored the effectiveness of targeted interventions, or cognitive de-biasing strategies, for improving clinical judgment in mental health practice. The present project developed an intervention aimed at reducing cognitive-based error in the assessment of pediatric bipolar disorder. The study design was a randomized controlled trial and participants were mental health professionals ($N = 79$) with experience treating pediatric populations. The treatment and control groups responded to similar case vignettes and primary outcome measures were clinicians' diagnoses and treatment decisions. Unlike participants in the control group, participants in the treatment group received the cognitive de-biasing intervention and evidenced greater overall judgment accuracy, $p < .0005$. Participants in the treatment group also committed significantly fewer decision-making errors, $p < .0005$. Study findings can significantly advance the mental health field as

improving clinical judgment, especially in cases of complicated diagnoses such as pediatric bipolar, are scientific and clinical priorities.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

Cognitive De-Biasing and the Assessment of Pediatric Bipolar Disorder

INTRODUCTION

Overview of Proposed Study

Pediatric bipolar disorder (PBD) has been in the spotlight of both clinical and research communities as well as the media and popular press (Kluger & Song, 2002; Papolos & Papolos, 2006). Diagnoses of PBD have risen at alarming rates, with some estimates showing as much as a 40-fold increase in diagnoses over the last decade (Blader & Carlson, 2007; Moreno, et al., 2007). Research suggests that PBD is frequently misdiagnosed (low accuracy, including low sensitivity), *and* overdiagnosed (low diagnostic specificity) (Hirschfeld, Lewis, & Vornik, 2003; Kessler, Rubinow, Holmes, Abelson, & Zhao, 1997; Lish, Dime-Meenan, Whybrow, Price, & Hirschfeld, 1994). Clinicians miss true cases of bipolar and diagnose PBD in many cases where the youth does not actually have the disorder.

Clinicians frequently rely on clinical judgment for making important decisions, yet clinical judgment is vulnerable to cognitive errors such as faulty heuristics (i.e., cognitive shortcuts) and biases that often result in suboptimal diagnostic decisions (Aegisdottir, et al., 2006; Croskerry, 2002; Galanter & Patel, 2005). For example, retrospection, hindsight bias, and availability can hinder judgment accuracy (cf. Hastie & Dawes, 2001), and psychologists inadvertently commit biased hypothesis testing (Pfeiffer, Whelan, & Martin, 2000; Strohmer, Shivy, & Chiodo, 1990) and invoke stereotypes about race, ethnicity, and homosexuality (Casas, Brady, & Ponterotto, 1983; Ridley, Li, & Hill, 1998; Wampold, Casas, & Atkinson, 1981; Wisch & Mahalik, 1999). Focusing on PBD, inadequacies in the decision-making

process occur when clinicians use clinical judgment alone to assess bipolar symptoms, with significant variation in clinicians' interpretation of identical case information [1].

The context of decision-making is an important consideration, especially given that heuristics can also serve as useful templates by expediting processing and synthesizing information (i.e., not always lead to inaccurate judgments). Evidence suggests that humans are more prone to commit faulty "cognitive heuristics" when the condition of interest is rare (Davidow & Levinson, 1993), and judgmental capabilities do not necessarily improve with additional experience (Brehmer, 1980, 1998; Faust, 1986; Faust, 1994; Faust, et al., 1988; Faust & Ziskin, 1988; Garb, 1989; Spengler, et al., 2009; Wedding, 1991; Wiggins, 1973; Ziskin, 1995). There is even evidence suggesting that more experienced clinicians may be at greater risk for biased judgment than novice clinicians (Strohmer & Leierer, 2000). Given the frequent misdiagnosis and overdiagnosis of PBD as well as it being a relatively rare condition, the risk for cognitive error seems to be especially high in clinicians' assessment of bipolar symptomotology. Unfortunately, evidence does not suggest that additional clinical experience will necessarily increase judgment accuracy. In the case of bipolar disorder, correct diagnosis is important for many reasons, including prescribing the most effective pharmacological treatment.

Improving clinical decision-making is a national health care priority (Mass, 2003). Despite a rich literature on the problems associated with clinical judgment, there is a dearth of information about strategies for improvement. Specifically, mechanisms for cognitively de-biasing clinicians' judgment, such as education and training in decision-making, are not readily available. Research suggests that actuarial assessment methods can lead to more accurate and efficient diagnostic decisions for pediatric bipolar illness (Jenkins, et al., 2011);

however, even these methods necessitate clinical judgment to maximize their clinical utility, or for them to be implemented in the first place (Youngstrom, Freeman, & Jenkins, 2009).

Better understanding how to reduce clinicians' use of faulty heuristics and biases can improve clinical judgment and increase the number of accurate diagnoses for difficult, high-stakes conditions such as PBD. In contrast to current research efforts in mental health, the medical decision-making literature has catalogued over thirty failed cognitive dispositions to respond (CDRs) (e.g., heuristics, biases, sanctions, fallacies, and errors) and has recommended strategies for reducing cognitive-based error (Croskerry, 2002, 2003; Graber, Gordon, & Franklin, 2002). Substantial clinical utility can be gained in mental health from unpacking the cognitive errors associated with clinical judgment and developing targeted interventions to rectify them. This objective is consistent with recommendations of Spengler et al. (2009), who explain that additional clinical experience may reinforce existing strategies while increased education experience, particularly training in decision-making, may enhance the quality of clinicians' strategies.

Training clinicians in decision-making has considerable *appeal*, even among those who are pessimistic about the role of experience in professional competence (e.g., Swets, Dawes, & Monahan, 2000). Notably, although cognitive de-biasing strategies are not at present widely taught in training programs, and the mental health literature is still relatively young in this domain, the decision-making and cognitive psychology literatures offer several promising methods that can translate to psychiatric conditions, such as PBD, and can inform intervention (Croskerry, 2003). In fact, some of these strategies have already demonstrated improved judgment accuracy for psychiatric diagnoses, including PBD (Jenkins, et al., 2011).

While clinical judgment is especially vulnerable to cognitive error in assessing cases of bipolar disorder, poor clinical judgment affects other diagnostic and treatment decisions as well (Galanter & Patel, 2005; McClellan, Werry, & Ham, 1993; Reimherr & McClellan, 2004). In general, humans typically do not use purely rational or normative approaches in making decisions (Lau & Coiera, 2007; Tversky & Kahneman, 1974), and this is evident and problematic in a variety of professional settings. For example, exorbitant fees result from unnecessary medical tests and procedures (Kraemer, 1992). The validity of clinical judgment and the amount of clinical experience are unrelated (Lueger, 2002). Clinical judgment can derail accurate decision-making (Dawes, Faust, & Meehl, 1989; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Meehl, 1954), and this derailment does not necessarily self-correct with additional clinical experience alone. Taken together, developing new, concrete strategies to help clinicians make more rational decisions and to avoid cognitive pitfalls is crucial- for PBD and other patient populations.

The overarching goal of the present project included developing and testing the effects of an intervention for improving clinical judgment. Specifically, adapting cognitive de-biasing strategies primarily from the medical decision-making literature, the present study investigated the impact of a new intervention on reducing cognitive-based errors that contribute to the misdiagnosis and overdiagnosis of PBD. Findings facilitate an increased understanding of clinicians' cognitive vulnerabilities as well as model a new approach for improving clinical decision-making.

Clinical Judgment and the Potential for Cognitive Error

Clinical judgment or prediction is judgment relying on informal or intuitive processes to combine or synthesize client data (Aegisdottir, et al., 2006). According to Redelmeier and

colleagues (2001), clinical judgment is the exercise of reasoning under uncertainty; and, often involves a combination of scientific theory, personal experience, patient perspectives, and other insights that missing data, conflicting information, and limited time can obscure (Redelmeier, 2005). With respect to assessment and diagnosis, clinical judgment largely refers to the decision-making task of interpretation. For example, clinicians may predict diagnoses by synthesizing data from clinical interviews, self-report measures, and family history of mental illness. Accuracy tends to be low when people use clinical judgment alone to perform this task (Elstein & Schwartz, 2002). In fact, mechanical or statistical prediction methods frequently perform as well, and even outperform in many cases, clinical judgment methods (Aegisdottir, et al., 2006; Grove, et al., 2000). Why does clinical judgment often lead to suboptimal decisions, especially when it comes to PBD?

PBD *is* an extremely challenging diagnosis (Hirschfeld, Lewis, & Vornik, 2003; Kessler, 1999; Lish, Dime-Meenan, Whybrow, Price, & Hirschfeld, 1994; Marchand, Wirth, & Simon, 2006; Youngstrom, Birmaher, & Findling, 2008; Youngstrom, Findling, Youngstrom, & Calabrese, 2005); however, is it possible that the way clinicians think makes diagnosing PBD *even harder*? Given the abundant associations between clinical judgment and cognitive errors (Elstein & Schwartz, 2002; Klein, 2005; Meehl, 1954), it is likely that clinicians' cognitive processes further complicate the clinical assessment of PBD.

Although the current literature indicates a general problem--a high frequency of diagnostic error--we have little knowledge of specific heuristics that are implicated in misdiagnosis and how to treat them. Examining specific heuristics in cases of PBD may identify especially problematic mismatches between typical habits of thought and

characteristics of the condition. Finding ways to promote changes of habit likely includes both greater awareness and the adoption of cognitive de-biasing strategies.

Pediatric Bipolar Disorder – A Public Health Crisis

The rise in clinical diagnoses of PBD represents a major public health concern. Bipolar is the 6th leading cause of disability in adults (Murray, Lopez, & Mundial, 1996), and it is associated with a 10 to 20 times increase in suicide risk compared to the general US population (Bostwick & Pankratz, 2000; Brodersen, Licht, Vestergaard, Olesen, & Mortensen, 2000; Guze & Robins, 1970; Harris & Barraclough, 1997; Sharma & Markar, 1994). Bipolar is also associated with substantial economic burden and medical conditions, including cardiovascular, endocrine, and alcohol use problems as well as significantly higher rates of diabetes, Hepatitis C, lower back pain, and pulmonary conditions (Dunner, 2003; Kupfer, 2005; Murray et al., 1996; Stang et al., 2006).

Serious consequences are associated with the misdiagnosis of PBD (Dunner, 2003). Untreated cases may follow a progressive and deteriorating course of illness (Geller, Tillman, Craney, & Bolhofner, 2004), and there is some evidence that wrong medication, such as antidepressants or stimulants, can possibly worsen outcome (cf. Joseph, Youngstrom, & Soares, 2009). Overdiagnosing or prematurely starting pharmacological treatment for bipolar is also dangerous because medications used to treat the illness can carry serious side effects (Wilens, et al., 2003). Suicidality is another concern with these medications (Olfson, Marcus, & Shaffer, 2006; Tondo, Isacsson, & Baldessarini, 2003). Diagnostic accuracy seems imperative before starting treatment (Weller, Danielyan, & Weller, 2004); however, heterogeneity of bipolar and the role of developmental influences create complex clinical presentations (Leibenluft, Charney, Towbin, Bhangoo, & Pine, 2003).

Research indicates an average delay of more than ten years between first episode and diagnosis of bipolar disorder (Hirschfeld, et al., 2003; Lish et al., 2004). Another recent study found that five or more years elapsed from the onset of symptoms until a PBD diagnosis was made for more than half of youths in a community mental health setting (Marchand, et al., 2006). Moreover, despite evidence that clinical judgment is prone to a host of errors (Dawes, 1996), practitioners are slow to change their approach (Dawes, et al., 1989). Given that misdiagnosis and treatment delays can have harmful consequences, research is needed to better understand clinicians' cognitive vulnerabilities (i.e., *where* mistakes occur) and the acceptability of reliable and efficient assessment strategies among practitioners (i.e., *how* to improve practice with science).

Cognitive Errors in Assessing Bipolar Presentations – Unpacking Misdiagnosis

Although PBD is a challenging diagnosis due to the complicated nature of the illness, the decision-making literature indicates that a number of other factors may also be at play (Galanter & Patel, 2005). In 1974, Tversky and Kahneman suggested that contrary to the normative model that people make decisions rationally and use all information available, people often use "heuristics" or cognitive shortcuts. Although heuristics can provide necessary and often useful templates for clinical decision-making, they can also lead to biases and cognitive errors.

There have been several investigations of heuristics in decision-making (Eddy, 1982; Lueger & Petzel, 1979b; McNeil, Pauker, Sox, & Tversky, 1982; O'Donohue & Szymanski, 1994; Richards & Wierzbicki, 1990; Sandifer, Hordern, & Green, 1970; Tversky & Kahneman, 1974), a few involving mental health professionals (Dailey, 1952; Garb, 1994; Gauron & Dickinson, 1966; Meehl, 1960; Nisbett & Ross, 1980; Sandifer, et al., 1970;

Waddington & Morley, 2000; Wierzbicki, 1993). Abundant associations have been shown between clinical judgment and cognitive errors (Elstein & Schwartz, 2002; Galanter & Patel, 2005). Clinicians, like many people, resist modifying or discarding ideas despite further evidence to the contrary (Dailey, 1952; Dawes, et al., 1989; Friedlander & Stockman, 1983; Herbert, Nelson, & Herbert, 1988; Hirsch & Stone, 1983; Meehl, 1960; Meehl, 1954; Rosenhan & Heiner, 2008; Snyder & Swan, 1978).

Croskerry (2002) identifies over *thirty* failed heuristics, biases, and cognitive distortions. A number of these cognitive shortcuts are highly relevant in the assessment of PBD; see Table 1 for some of these failed heuristics and their consequent clinical implications. The succeeding paragraphs will provide a framework for how particular heuristics highlighted in Croskerry's (2002) review map on to clinicians' decision-making around cases of pediatric bipolar illness and the related challenges of assessing bipolar disorder.

Challenge 1 – The Cyclical Nature of the Illness

Diagnostic decisions and reliability of clinical impressions are challenged by the cyclical nature of bipolar illness (Bowring & Kovacs, 1992). Case presentations appear very differently depending on whether a client is experiencing florid mania, severe depression, mixed or a euthymic state (Youngstrom, 2005). Consequently, clinicians often fail to detect hypo/mania, and they neglect the possibility of PBD in their diagnoses, case conceptualizations, and treatment plans. This clinical reality supports the likelihood of the heuristic *base-rate neglect*--failing to adequately take into account the prevalence of a particular disease. The frequency and seriousness of *base-rate neglect* is highlighted by epidemiological studies suggesting that bipolar is more common than researchers and

clinicians previously thought (Hasin, Goodwin, Stinson, & Grant, 2005; Merikangas, et al., 2007; Merikangas & Pato, 2009), especially in clinical settings (Biederman, et al., 1996; Blader & Carlson, 2007; Geller, et al., 2002; Youngstrom, Youngstrom, & Starr, 2005). Further, a higher proportion of people with depression will actually have bipolar disorder, particularly child populations (Angst, Sellaro, Stassen, & Gamma, 2005). One likely clinical scenario of base rate neglect includes clinicians failing to assess for hypo/manic symptoms in youth with depression. The seriousness of base-rate neglect is also highlighted by recent FDA requirements: product information sheets for all antidepressants must include a recommendation for bipolar screening prior to prescription (see Warning section) (U.S. Food & Drug Administration, 2004; Akiskal & Benazzi, 2005).

Challenge 2 – Symptom Overlap

Symptom overlap makes it difficult to parse apart symptoms of bipolar from other likely suspects in pediatric populations, namely attention-deficit hyperactivity disorder (ADHD), unipolar depression, and oppositional defiant disorder (ODD). Irritability lacks diagnostic specificity and can be used to support a large number of diagnoses) (Youngstrom, Findling, et al., 2005). For example, irritability can be a symptom of bipolar, ADHD, unipolar depression, and ODD. Semi-structured diagnostic interviews, such as the Kiddie Schedule for Affective Disorders and Schizophrenia (KSADS) (Kaufman, et al., 1997), which are used in research, can help clinicians parse apart symptoms; however, interviews like the KSADS are often impractical in clinical settings due to issues surrounding training, burden, and insurance reimbursement. In part, because real-world clinical settings have these types of barriers, thorough assessments of PBD are often not conducted. Consequently, clinicians are also vulnerable to *diagnosis momentum*: the tendency for a particular initial or

provisional diagnosis to become established without adequate evidence. Research lends support to the possibility that clinicians overestimate the likelihood of bipolar without adequate evidence (Dubicka, Carlson, Vail, & Harrington, 2008; Jenkins, et al., 2011). They may be framing this judgment quickly on the basis of a few features, and then fail to check whether a sufficient number of criteria are met. Similarly, once the bipolar hypothesis gains momentum, disconfirming evidence may be discounted or ignored.

Challenge 3 – High Comorbidity

Evidence suggests that bipolar disorder is associated with ADHD, anxiety disorders, ODD, and conduct disorder (CD) (Biederman, et al., 1997; Biederman, et al., 2004; Geller, et al., 2000; Lewinsohn, Seeley, Buckley, & Klein, 2002; Moreno, et al., 2007; Youngstrom, Findling, et al., 2005). High comorbidity may also be problematic in clinicians' assessment of PBD because some clinicians may not see a "typical" uncomplicated bipolar presentation (Biederman, Klein, Pine, & Klein, 1998; Youngstrom, Findling, et al., 2005). For example, individuals with bipolar have high rates of ADHD (Galanter & Leibenluft, 2008; Kessler, et al., 2006; Lewinsohn, Klein, & Seeley, 1995; Singh, DelBello, Kowatch, & Strakowski, 2006). Table 2 provides an overview of the common symptoms between these two disorders while also highlighting important distinctions in symptom manifestation. As a result of overlapping symptomatology and higher comorbidity rates between bipolar disorder and ADHD, clinicians seem especially prone to commit *search satisficing*-- when one calls off a search once something is found (i.e., finding something may be satisfactory but not finding everything is suboptimal (see Croskerry, 2002)). In the case of PBD and ADHD, clinicians may "call off the search" when they diagnose one of these disorders, causing them to miss the other disorder for youths who actually meet DSM criteria for both conditions. Search

satisficing also probably contributes to the overdiagnosis of unipolar depression at the expense of bipolar disorder (Angst, 2004). Patients are more likely to seek help for symptoms of depression, and once clinicians decide that depression is present, they may call off the diagnostic search before checking for a history of hypomania or mania.

Challenge 4 – Misdiagnosis of Minority Populations

Adult research on bipolar indicates that minority populations are frequently misdiagnosed (Kirov & Murray, 1999; Lawson, Hepler, Holladay, & Cuffel, 1994). Specifically, adult African American populations are at high risk of being misdiagnosed with schizophrenia (Davis & Jones, 1973; Flaskerud & Hu, 1992; Gross, Herbert, Knatterud, & Donner, 1969; Lawson, et al., 1994; Marquez, Taintor, & Schwartz, 1985; Mukherjee, Shukla, Woodle, Rosen, & Olarte, 1983; Neighbors, Jackson, Campbell, & Williams, 1989; Strakowski, Flaum, et al., 1996; Strakowski, et al., 1995; Strakowski, McElroy, Keck, & West, 1996; Strakowski, Shelton, & Kolbrener, 1993), and less likely than Caucasian patients to receive bipolar diagnoses even when all subjects present with psychotic mania (Strakowski, et al., 2003; Strakowski, McElroy, et al., 1996). There is also evidence suggesting that African American pediatric populations are more likely to receive diagnoses of conduct disorder or on the schizophrenia spectrum (DelBello, Lopez-Larson, Soutullo, & Strakowski, 2001; Fabrega, Ulrich, & Mezzich, 1993; Kilgus, Pumariega, & Cuffe, 1995). Currently, no research indicates that the risk of developing bipolar disorder varies by race/ethnicity (American Psychiatric Association, 2001; Van Meter, Moreira, & Youngstrom, 2011). Nevertheless, clinicians may be prone to misdiagnose or inaccurately judged the likelihood of PBD in African American clients (Jenkins, et al., 2011). Clearly, more research

is needed to examine the influence of race/ethnicity bias in the clinical assessment of bipolar populations as well as methods for reducing this bias.

Challenge 5 – Exaggerated Certainty

Studies examining clinicians' ability to make appropriate confidence ratings, or rather their ability to specify when their predications are correct, have generated mixed results. Although there is substantial evidence suggesting that clinicians lack the ability to make appropriate confidence ratings (Goldberg, 1968; Mischel, 1996) and that judges tend to be overconfident (Kahneman & Tversky, 1973; Nisbett & Ross, 1980; Slovic, Fischhoff, & Lichtenstein, 1977), there is also research suggesting that this may not be the case, or that other factors such as clinical experience moderate the relationship between accuracy and confidence (see Garb, 1986, for a review). Given the evidence that PBD may be overdiagnosed frequently in the U.S. (Soutullo, et al., 2005), o*verconfidence*--thinking one knows more than one does, often without having gathered sufficient information--is a serious concern that may impede accurate diagnostic decisions (e.g., Arkes 1981). In particular, overconfident judges may (a) fail to search for information that might cause them to revise their original predictions (Einhorn, 1980; Einhorn & Hogarth, 1978) and (b) disregard evidence contradicting their current judgments (Friedlander & Phillips, 1984; Koriat, Lichtenstein, & Fischhoff, 1980). More research is needed to better understand the role of clinician confidence in of the assessment of PBD.

Cognitive De-Biasing Strategies and Mental Health

Despite substantial evidence suggesting that clinical judgment is prone to a host of errors and is inferior in many cases to statistical prediction methods (Aegisdottir, et al., 2006; Grove, et al., 2000; Spengler, et al., 2009), less research has examined methods for

addressing the cognitive pitfalls that contribute to the inadequacies associated with clinical judgment in mental health. The reduction of diagnostic error in medicine, however, has gained substantial attention in recent years. This may not be surprising given that diagnostic errors have a proportionately higher morbidity compared to other types of medical errors (Brennan, et al., 1991; Thomas, et al., 2000; Wilson, et al., 1995), and delayed or missed diagnoses account for approximately half of all litigation brought against physicians (Ohio Hospital Association & St. Paul Minnesota Insurance Company, 1998). In particular, cognitive de-biasing strategies have taken the forefront in recent medical research for reducing cognitive error.

Cognitive psychology experts recommend similar strategies to those used by experts in medical decision-making (e.g., Croskerry, 2003) for improving clinical judgment, such as *consider the alternatives, think Bayesian*, and *decrease reliance on memory* (Arkes, 1981). Each of these strategies will be discussed in more detail below. In general, there seems to be consensus across these fields, especially regarding the importance of focused professional training (Arkes, 1981; Croskerry, 2003). Arkes (1991) also shares that simply telling individuals about a particular bias and then telling them not to do it is an "absolutely worthless" technique (p. 326). Research supports the ineffectiveness of this approach (Fischoff, 1977; Kurtz & Garfield, 1978; Wood, 1978). Thus, in order to change clinician behavior a more concentrated approach seems warranted.

Although cognitive de-biasing strategies are relatively innovative for mental health practice in the sense that they are not formally organized or packaged as they are presently in medicine (see Croskerry 2003), the concept of educating and training mental health professionals in decision-making is not new and has been recommended by many experts

(e.g., Arkes, 1991; Fischhoff, 1982). For example, based on findings from a recent meta-analysis on clinical judgment, study authors recommend that training incorporate statistical methods to decrease judgment biases and errors (Aegisdottir, et al., 2006). This is consistent with many others' recommendations as well, including calls for increased education in regression formulas, cutoff scores, and hit rates (Anastasi & Urbina, 1982; Crocker & Algina, 1986). Further, there is evidence that training in probability theory and education on common heuristics increases judgment accuracy (Arkes, 1981, 1991a; Jenkins, et al., 2011; Nisbett & Ross, 1980; Spengler & Strohmer, 2001).

Table 3 provides an integrative list of proposed cognitive de-biasing strategies from the decision-making and cognitive psychology literatures along with definitions and examples of each strategy. A few of these strategies will be discussed in the following paragraphs with an emphasis on their generalizability to clinicians' assessment of PBD.

Develop Insight and Awareness

As discussed, the literature identifies common challenges associated with the assessment of PBD that likely account for the difficulty assessing bipolar and the frequency of misdiagnosis and overdiagnosis (Bowring & Kovacs, 1992; Youngstrom, Findling, et al., 2005). Some of these challenges seem to map on to specific cognitive errors, including include base-rate neglect, diagnosis momentum, search satisficing, overconfidence bias, and race/ethnicity bias. Based on recommendations provided in the decision-making and cognitive psychology literatures, giving detailed descriptions and characterizations of these specific faulty heuristics and biases along with clinical examples of each demonstrating potential adverse effects on decision-making can be effective cognitive de-biasing strategies (Croskerry, 2003).

For example, clinicians might learn that base-rate neglect refers to failing to adequately take into account the prevalence of a particular disease. Then, they might be educated about the base rate of bipolar disorder based on the most recent findings from epidemiological research (e.g., Van Meter et al., 2011). Finally, two clinical scenarios might be presented in which a clinician fails to assess for past symptoms of hypo/mania in a client with depression (i.e., an example of committing the cognitive error) whereas in the second scenario, the clinician inquires about past episodes of unusually energetic or irritable mood (i.e., an example of not committing base-rate neglect). Although there is some debate about the effectiveness of merely educating clinicians and increasing their awareness about faulty heuristics and biases (Croskerry, 2003; Gordon & Franklin, 2003), experts agree that increasing insight is an important ingredient of cognitive de-biasing (Fischhoff, 1982; Redelmeier, 2005).

Consider Alternatives

Given both the high rates of comorbidity and the overlapping symptomatology of bipolar disorder and other psychiatric conditions, *considering the alternatives* or routinely contemplating differential diagnoses seems highly relevant. For example, if a youth presents with irritability and high energy, these symptoms might be a sign of a mood disorder or ADHD. Training clinicians to ask themselves, "What else might this be?" can reduce cognitive errors such as search satisficing and diagnosis momentum. Further, always making sure to consider the most common (i.e., high base rate) options in addition to the diagnoses that a clinician is currently considering can prevent base-rate neglect. This strategy may help reduce the rate of "missed" diagnoses as well as lead to more accurate diagnostic decisions by decreasing the likelihood of misdiagnosis.

Decrease Reliance on Memory

Research indicates that for a number of reasons, relying on one's memory can threaten judgment accuracy (Arkes & Harkness, 1980; Lueger & Petzel, 1979a; Ward & Jenkins, 1965). For example, Arkes & Harkness (1980) found that individuals recalled symptoms consistent with a given diagnosis even when these symptoms were not present. Mnemonics, algorithms, and clinical practice guidelines (to name a few) have been suggested as mechanisms for decreasing reliance on memory and increasing judgment accuracy. A few mnemonics have been identified in the literature on PBD (Kowatch, et al., 2005; Quinn & Fristad, 2004; Youngstrom, et al., 2009). For example, "GRAPES" is a mnemonic for more specific symptoms of bipolar including: grandiosity; racing thoughts; increased goal directed activity; pressured speech; elated, euphoric mood; and, decreased need for sleep. Additionally, "FIND" is a mnemonic for assessing the frequency, intensity, number, and duration of bipolar symptoms. Training clinicians how to use these mnemonics can assist with symptom recognition as well as clinical interviewing (especially in time-limited situations when a gold standard assessment such as the KSADS is not feasible).

Specific Training on Bayesian Thinking

Bayesian probability theory is a highly endorsed strategy for overcoming flaws and biases in thinking (Aegisdottir, et al., 2006; Arkes, 1981; Croskerry, 2003; Guyatt & Rennie, 2002; Straus, Glasziou, Richardson & Haynes, 2005; Youngstrom, et al., 2009). Training clinicians in Bayesian thinking or actuarial decision-making might consist of education on diagnostic likelihood ratios (DLRs) and how to correctly synthesize risk information. In the case of PBD, this process might entail learning the base rate of bipolar at one's clinical

setting and learning how to combine this information with family history and/or test score information. This method will be discussed in more detail below.

Addressing the Issue of Inaccurate Judgments

Although there seem to be advantages of training clinicians in decision-making strategies (Berven, 1985; Berven & Scofield, 1980; Falvey & Hebert, 1992; Kurpius, Benjamin, & Morran, 1985; Spengler & Strohmer, 2001), more research is needed to identify effective ways to teach students and professionals how to avoid cognitive pitfalls (Regehr & Norman, 1996). Graber (2002) argues that diagnostic error may never be eradicated, but that it may be significantly reduced by specific training aimed at improving cognition and cognitive awareness. Overall, clinical decision-making could benefit from more reliable and valid clinical judgment; however, strategies or interventions for minimizing cognitive errors in mental health, particularly around diagnostic decisions, remain largely unexplored-especially strategies that have been tested under rigorous experimental conditions.

What We Do Know: Making Bayesian Thinking Easier for Clinicians

One de-biasing strategy recommended for helping professionals reduce cognitive error is specific training in fundamental rules of probability and basic Bayesian probability theory (Croskerry, 2003). At present, many common assessment methods in mental health are not empirically supported (as reviewed in Fletcher, Francis, Morris, & Lyon, 2005; Neisworth & Bagnato, 2004); and those widely used techniques with evidence also have substantial shortcomings. Inter-rater reliability among clinicians conducting unstructured interviews has been shown to be inadequate (Piacentini, et al., 1993), and there is only moderate agreement between diagnoses based on standardized research interviews and those based on clinical chart reviews (Ezpeleta, et al., 1997; Lewczyk, Garland, Hurlburt, Gearity,

& Hough, 2003; Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009; Vitiello, Malone, Buschle, Delaney, & Behar, 1990; Vitiello & Stoff, 1997). To address shortcomings of current assessment methods, researchers have begun testing the effectiveness of actuarial approaches (Jenkins, et al., 2012) that are consistent with recommendations by Croskerry (2003) and others to provide clinicians specific training in Bayesian reasoning (Arkes, 1981; Youngstrom, et al., 2009).

In contrast to traditional clinical assessment, Bayesian reasoning or actuarial approaches employ mathematical and statistical methods to assess risk of illness. Actuarial decision-making has most commonly been used in the medical community to assess the likelihood of illness such as cancer; however, two recent meta-analyses lend support to statistical prediction methodologies in mental health (Aegisdottir, et al., 2006; Grove, et al., 2000). A recent study indicates that actuarial approaches can be used in psychological assessment to improve assessment accuracy and increase agreement surrounding diagnostic decisions for suspected bipolar disorder (Jenkins, et al., 2011).

The nomogram is a simple, practical method for combining information about risk with "likelihood ratios" associated with test results or other clinical findings (Jaeschke, Guyatt, & Sackett, 1994). Clinicians can work directly with probabilities without requiring any mathematical computation (Youngstrom & Duax, 2005). Specifically, the nomogram operates as a type of slide rule (see Appendix A). The nomogram approach correctly combines three pieces of information (i.e., base rate, familial risk, and test score) into *consistent* (less spread in opinion), *unbiased* (neither systematically over- or under-estimating risk), and *efficient* (using a parsimonious amount of information to arrive at the estimate) estimates (Youngstrom et al., 2009; Youngstrom & Duax, 2005). This estimate, a

Bayesian posterior probability, can be used to determine the likelihood that a youth has PBD, and to guide next steps in assessment and treatment (Youngstrom, et al., 2009).

An alternative to the probability nomogram is to use actuarial tables. Tables are familiar, can be user-friendly, and arguably require less cognitive resources compared to the nomogram. However, tables would need to be recalculated and replaced to accommodate each scientific advance such as new base rate estimates or updates about the diagnostic performance of measures. Mental health professionals could potentially access actuarial tables to determine patient risk immediately based on their clinical setting and a patient's test score and/or family history of bipolar illness. Table 4 provides an example of an actuarial table that mental health professionals might use. Overall, training clinicians in actuarial decision-making with convenient tools such as the nomogram or actuarial tables is not only consistent with recommendations in the de-biasing literature but is also gaining recognition in the mental health field as an evidenced-based approach (Chorpita, Yim, & Tracey, 2002; Jenkins, et al., 2011; Youngstrom, et al., 2009).

Evidence-Based Practice and Heuristics

Research suggests that people with severe mental illness seldom receive effective services in community mental health centers (Lehman & Steinwachs, 1998; U.S. Department of Health, 1999; Wang, Demler, & Kessler, 2002; West, et al., 2005). To address this issue, evidence-based practice in psychology (EBPP) promotes effective psychological practice and enhancing public health via empirically supported principles of assessment (American Psychological, 2002; Sackett, Rosenberg, Muir Gray, Haynes, & Richardson, 1996). At present, EBPP remains largely theoretical and hortatory. Several mental health initiatives acknowledge that more effective approaches could be implemented in care (Ganju, 2006;

Institute of Medicine, 2001), yet the research-practice gap persists. Any number of reasons may explain the practice-research gap. For instance, clinicians can feel overwhelmed by the literature and elect not to use it for clinical issues (Haynes, Glasziou, & Straus, 2000). This is problematic: in one study, consulting the literature changed the medicine, diagnostic test, or prognostic information given to the patient in 47% of cases (Crowley, et al., 2003).

The role of clinical judgment in EBPP has been a contentious issue between researchers and clinicians. For instance, a common critique of EBPP by clinicians involves a belief that evidence-based treatments restrict or deemphasize the importance or role of a clinician and his/her judgment in patient care (Addis & Krasnow, 2000; Addis, Wade, & Hatgis, 1999). The *Report of the 2005 Presidential Task Force on Evidence-Based Practice* argues that although cognitive errors can negatively impact clinical decision-making, clinical judgment is also responsible for complex decisions that can lead to well-conceptualized EBPP (Levant, 2005). Thus, it appears that improving clinical judgment may also help facilitate EBPP. Moreover, it is possible that training clinicians in decision-making (i.e., to improve clinical judgment) in tandem with training on or dissemination of evidence-based treatments may help validate the importance of clinical judgment, defusing clinician resistance and leading to more positive attitudes toward EBPP.

One way to prevent furthering the practice-research gap is to involve providers in earlier stages of research (i.e., prior to dissemination). For example, in contrast to the more common and sequential approach of first developing protocols and then examining how they translate into practice, researchers can be proactive and concurrently request feedback from clinicians while developing new interventions. Involving providers in all stages of EBPP research is a model that other prominent researchers are already adopting (Garland,

Plemmons, & Koontz, 2006). Getting clinicians' feedback early in the process by providing them the opportunity to share their concerns and voice their preferences seems pivotal to successful implementation. Learning from clinicians on the front line about potential barriers that might otherwise go undetected can also allow researchers the opportunity to tailor strategies to maximize utility.

Significance and Broader Impact

Among many consequences, research shows that inaccurate decisions can lead to treatment failure and premature dropout (Epperson, Bushway, & Warman, 1983; Lutz, et al., 2006). Evidence also suggests that accurate diagnostic assessment is important for providing effective interventions (Meyers, et al., 1998). In the case of bipolar disorder, delays in diagnosis carry serious consequences for patients, their caregivers, and society, including high rates of suicide and increased economic burden (Dunner, 2003; Stang, et al., 2006). Correct diagnosis is pivotal to starting appropriate medication and treatment (e.g., mood stabilizers versus antidepressant medication, which may worsen outcome; cf. Joseph et. al, 2009).

Redelmeier (2005) recommends that instead of attempting to completely eliminate cognitive shortcuts (which often serve clinicians well), increasing awareness of common errors could improve patient care. Awareness alone may not be enough. Aegisdottir & colleagues (2006) recommend developing and testing more user-friendly heuristics for use in clinical settings to further aid judgment accuracy and to save time. Increasing the accuracy of clinical judgment by targeting specific deficits and equipping clinicians with effective decision aids (e.g., mnemonics, the nomogram), could greatly improve clinical decision-making and enhance the quality of clinical practice. PBD represents an opportunity to make

considerable public health contributions due to the combination of rapidly increasing diagnosis combined with the currently low accuracy of diagnostic decisions.

Overall, the literature on error in psychiatry is relatively small (Grasso, Rothschild, Genest, & Bates, 2003) with only a few studies having investigated the incidence, nature, predictors, and prevention of errors that may occur in mental health settings (Nath & Marcus, 2006). Yet, implications of failed cognitive shortcuts have dire consequences for patient care. Increasing knowledge about interventions to increase awareness and reduce faulty heuristics and biases can accomplish multiple goals, including: (a) improved education and training around assessing bipolar and potentially other disorders; (b) development of effective cognitive de-biasing strategies; and, (c) advancement of the extant literature on clinical decision-making for mental health (Croskerry, 2002). The study aims focused on these goals.

Study and Hypotheses

Can training in cognitive de-biasing strategies increase judgment accuracy in the assessment of bipolar disorder? The proposed study explored the following primary aim: To test the efficacy of a new intervention designed to improve clinical judgment in the assessment of pediatric bipolar disorder by educating participants in common cognitive pitfalls and training them in recommended de-biasing strategies. In addition to examining the influence of this intervention on decreasing the likelihood of several decision-making errors such as base-rate neglect, search satisficing, and diagnosis momentum, we also investigated the relationship between clinician confidence and judgment accuracy as well as the practice implications of decision-making errors and participants' attitudes toward the intervention and vignette study.

Hypotheses

1.0 Overall diagnostic accuracy. Participants' judgments in the treatment group will show higher decision accuracy than participants' judgments in the control condition. Overall diagnostic accuracy will be a composite measure of participants' judgments about diagnosis across case vignettes as well as the probability of disorder for one vignette. Participants' diagnostic judgments in both groups will be coded as inaccurate, somewhat accurate, or accurate based on expert consensus. Risk of disorder will also be coded similarly as inaccurate, somewhat accurate, or accurate compared to the Bayesian posterior probability. Somewhat accurate risk estimates will fall within a range of ± 5% of the Bayesian estimates (Sedlmeier & Gigerenzer, 2001) whereas accurate estimates will be the exact estimate. After all of these decisions are coded for accuracy, they will be compared between the treatment and control groups.

1.1 Base Rate Neglect. Participants in the treatment group are less likely to make decisions consistent with base rate neglect, as demonstrated by their inclusion of bipolar disorder as a probable diagnosis. We predict that without cognitive de-biasing training, clinicians will not routinely assess for hypomania or mania. This pattern of decision-making would be consistent with findings in the literature that suggest delays in diagnosis (Hirschfeld, et al., 2003; Marchand, et al., 2006), as well as the possibility of underdiagnosis of bipolar disorder (Angst, 2007).

1.2 Search Satisficing. Participants in the control condition will be more likely to engage in search satisficing than participants in the treatment condition, resulting in suboptimal diagnostic decisions. Given the high rates of comorbidity of bipolar disorder and other diagnoses--such as ADHD, which shares symptoms of bipolar--we anticipate that

clinicians may stop after diagnosing one disorder when the youth may actually meet criteria for multiple psychiatric conditions. Educating clinicians about this tendency and training them in how to conduct a more comprehensive assessment (i.e., similar to the proposed intervention) will result in a decreased likelihood of search satisficing. Participants in the treatment condition will be more likely to make sensitive diagnostic decisions, accounting for symptoms of both ADHD and PBD in their probable diagnoses.

1.3 Diagnosis momentum. Participants in the treatment group will generate more accurate estimates of the probability of bipolar disorder than participants in the control group. Typical clinical decision-making becomes self-reinforcing, with information that contradicts the initial impression de-emphasized, and corroborating information counting more. Participants in the treatment condition may correct for this tendency. In contrast, participants in the control condition will be more likely to endorse higher probabilities of PBD without adequate supporting evidence.

1.4 Race/ethnicity bias. Participants' diagnoses in the treatment group will not be affected by youths' race/ethnicity status. Given the tendency for clinicians to misdiagnose African-American youths with conduct disorder or schizophrenia, clinicians' rates of bipolar diagnoses in the control condition will vary by youths' race/ethnicity status, all other case information held constant. We hypothesize that educating clinicians that Caucasian and African American children are equally likely to have bipolar disorder (i.e., to date no research shows variation in diagnostic rates of bipolar across racial and ethnic groups (American Psychiatric Association, 2001; Grant, et al., 2005; Jonas, Brody, Roper, & Narrow, 2003; Kessler, et al., 2005), will lead to more accurate diagnoses. Therefore, we

anticipate that participants' diagnoses in the treatment group will be significantly less influenced by a youth's race/ethnicity.

1.5 Confidence and Judgment Validity. There are mixed findings regarding the relationship between judgment accuracy and clinical confidence (see literature review). To further explore this relationship, we hypothesize that for participants in the present study, the discrepancy between accuracy and confidence will be significantly different depending on group status. Specifically, participants in the treatment group who learn de-biasing strategies are expected to demonstrate better agreement between accuracy and confidence. Participants in the control condition are more likely to show *overconfidence bias* as demonstrated by discrepancies between accuracy and confidence in which the accuracy of their judgments is less than their confidence ratings.

1.6 Decision-Making Errors and Practice Implications. The effect or clinical implications of *base rate neglect, search satisficing, diagnosis momentum, overconfidence bias*, and *race/ethnicity bias* will also be examined. We predict that the presence of a faulty heuristic/bias (as demonstrated by inaccurate diagnoses or only somewhat accurate diagnoses) will affect clinicians' decisions regarding next clinical action, suggesting that heuristics change treatment as well as assessment formulations.

1.7 Clinician Experience of the Vignette Exercise. The cognitive de-biasing intervention will help participants in the treatment group feel more prepared for the case vignette exercise. As a result, they will more likely to rate their experience of reading case vignettes and responding to questions as more favorable than participants who do not receive the intervention.

1.8 Clinician Attitudes toward the Intervention. Clinicians who receive the intervention will find it helpful and informative.

In addition, we will explore whether clinicians' perceived ability to perform various mathematical tasks and preference for the use of numerical versus prose information relates to (a) the accuracy of their decisions on specific vignettes, or (b) their attitudes toward the intervention.

METHOD

Overview

In the current study, a randomized controlled trial (RCT) tested the efficacy of a new intervention designed to improve decision accuracy. Study participants were randomly assigned to one of two conditions, the treatment or control group. The randomization process is discussed in more detail below. Participants in both groups received information about mood disorders via a 5 minute pre-recorded presentation, and instructed to read several case vignettes and respond to a series of questions after each vignette regarding judgments about probable diagnoses, next clinical actions, and/or decision confidence. Participants in the treatment group also received a training (20 minutes) on decision-making errors *and* cognitive de-biasing strategies for assessing PBD (hereafter referred to as the cognitive de-biasing intervention) prior to responding to the case vignette exercise.

Participants

Participants were 79 mental health providers with 32 participants in the treatment group and 47 in the control group. Study recruitment consisted of providing study fliers and making study announcements via listservs (e.g., North Carolina Psychological Association) as well as electronic mail (e.g., contacting training directors, chief psychologists, trainees). Table 7 provides information about participant demographics including participants' age, race/ethnicity, and current geographic location. Table 8 provides information about participants' professional backgrounds including professional title, licensure status, and years of clinical experience. Table 9 contains information regarding participants' current caseload

including the ages and diagnostic groups of participants' primary patient populations. The treatment and control groups evidenced no significant differences, $p > .05$, or significant associations, $p > .05$, on any demographic, professional, or patient variables.

Procedure

Study administration was Web-based using Qualtrics, a provider of online survey software solutions, and an automated Powerpoint presentation with narration. Clinicians accessed the study through a Web address. When individuals entered the website, they were thanked for their interest in the study and were requested to answer two questions to determine study eligibility. In order to be eligible to participate, clinicians needed to be: (a) licensed or currently supervised by a licensed mental health professional; and, (b) have experience treating child and adolescent patient populations for mental health issues. If participants respond "no" to either of these screening questions, they were thanked for their interest and informed that they were not eligible to participate, thus ending the survey.

Clinicians who responded "yes" to both screening questions were informed of their eligibility to participate and they were directed to the study informed consent page. Individuals who chose not to participate either closed the browser window or clicked a box indicating that they did not want to participate in which case they were thanked for their interest in the study and exited from the survey. Individuals who reviewed the consent form and wanted to participate clicked a box to indicate agreement to participate in the present study.

After providing informed consent, we asked participants to complete a demographic and background questionnaire (see Appendix B for a copy of the questionnaire). On this questionnaire and throughout the rest of the study, participants were not forced to answer any

question that they intentionally wanted to leave unanswered. Instead, Qualtrics has a feature that *requested* a response for any unanswered items (to help prevent participants from unintentionally skipping items).

Following the demographic and background questionnaire, we programmed Qualtrics to randomly present participants with one of two possible URLs. Participants clicked on this link, taking them to a new window in which they either watched a brief presentation on mood disorders (i.e., control condition) or a brief presentation on mood disorders + the cognitive de-biasing intervention (i.e., treatment condition). At the end of both presentations, participants returned to Qualtrics and continued on to the case vignette exercises.

We informed participants that they would be presented with four case vignettes and asked to read each vignette and respond to the questions that followed. We used only four vignettes to reduce the burden on participants and to maximize response rate. Qualtrics was programmed to present case vignettes to participants in a random order. More information about study vignettes is provided below. After completing the last vignette and corresponding questions, we asked all participants to rate their experience of participating in the vignette exercise (see Appendix C for a copy of the questionnaire). Participants in the treatment group were also asked to complete a brief questionnaire about their attitudes toward the cognitive de-biasing intervention (see Appendix D for a copy of the questionnaire). After completing their respective attitude questionnaires, all participants completed the Subject Numeracy Scale (Fagerlin, et al., 2007). This scale is described below; see Appendix G for a copy of the questionnaire items.

After completing their participation in the study, we provided participants a link in Qualtrics to a separate Qualtrics database where participants provided their email addresses

and selected their preferred gift card. They could choose from Target, iTunes, or Amazon. We used two Qualtrics databases to ensure that participants' responses were kept separate from their email addresses. As noted, participants earned a $20 gift card for their time and effort. We also offered to provide all participants in the control condition a copy of the cognitive de-biasing intervention. Participants were thanked for their participation in this one-time online study.

Randomization process. Two variables were randomized using Qualtrics software including the treatment condition and the race/ethnicity of vignette characters. As described earlier, we programmed Qualtrics to randomly assign participants to either the control (i.e., presentation on introduction to mood disorders only) or treatment condition (i.e., presentation on introduction to mood disorders + cognitive de-biasing intervention) by randomly presenting them with different website links. Qualtrics also randomized different versions of vignettes so that the different vignette characters were African American half of the time and Caucasian half of the time, all other case information identical. Also, the case vignettes were presented in random order to participants as well as the questions on the multiple choice test for the *overconfidence* vignette (see below).

Study Materials

Intervention. Table 5 provides an overview of the didactic materials included in the cognitive de-biasing intervention that was administered to participants in the treatment condition. The intervention consisted of two main parts: The first part provided education about common cognitive pitfalls; and, the second part provided training and tools to avoid cognitive pitfalls when assessing for PBD.

Case vignettes. We crafted vignettes to examine specific decision-making errors. Study procedures varied slightly by case vignette in order to uniquely test for the specific decision-making error of interest. See Table 6 for an overview of the case vignettes and related procedures. For the vignettes testing for base-rate neglect and search satisficing, participants were instructed to read the case vignette and then: (i) make a determination of probable diagnosis; and, (ii) recommend next clinical action from a list of five options (e.g., more assessment; if yes, specify), psychotherapy or refer for psychotherapy (if yes, what type), medication or refer for medication (if yes, what type), no offer of treatment, or "other" (if yes, specify what "other" treatment do you recommend). Participants could select a maximum of two initial treatment methods (see methods from Currin et al., 2007). Note: Vignette instructions and questions for the vignettes testing for diagnosis momentum and overconfidence included additional questions; we present these additional instructions below with the actual vignette description. Brief descriptions and content of all case vignettes are also provided below. See Appendix E for a copy of each vignette.

Base rate neglect vignette. This vignette described a youth with current symptoms characteristic of a major depressive episode.

Search satisficing vignette. This vignette describes a youth with bipolar and attention-deficit hyperactivity disorder (ADHD) symptoms.

Diagnosis momentum vignette. This vignette portrays a youth with bipolar symptoms, a moderately highly elevated test score on widely used assessment instrument, and a family history of bipolar. The vignette resembles the case vignette used in Jenkins et al. (2011).

For this vignette, we asked participants to: (i) make a determination of probable diagnosis; (ii) recommend next clinical action from a list of five options (e.g., more assessment; if yes, specify), psychotherapy or refer for psychotherapy (if yes, what type?), medication or refer for medication (if yes, what type), no offer of treatment, or "other" (if yes, what "other" treatment do you recommend?); and, (iii) report the probability of a bipolar diagnosis (from 0 to 100) on the basis of the available case information.

Overconfidence. This vignette focused on a youth, Lynda, with classic symptoms of mania including grandiosity, hypersexuality as an example of disinhibited and risky behavior, psychomotor agitation, and distractibility – meeting criteria for at least four "B Criterion" symptoms in addition to the episodic disturbance of mood. Consistent with the episodic nature of mood disorders, her symptoms are described as intermittent. Lynda met duration criteria for a manic episode (e.g., 1 week or longer), and experienced impairment as a result of her symptoms.

This vignette was similar to those in Dubicka et al. (2008) (we received permission from study authors to use this vignette). For information regarding the assembly of the case vignette, see Dubicka et al. (2008). Similar to other vignettes, we asked participants to: (i) make a determination of probable diagnosis; and, (ii) recommend next clinical action from a list of five options (e.g., more assessment; if yes, specify), psychotherapy or refer for psychotherapy (if yes, what type), medication or refer for medication (if yes, what type), no offer of treatment, or "other" (if yes, what "other" treatment do you recommend). And, different from other vignettes, we also gave participants additional instructions, similar to procedures used in Oskamp (1965) and Adams (1957). Specifically, participants were asked

to make *confidence judgments*. This process is described in more detail in the following paragraphs.

In order to have a basis for determining the accuracy of the judgments, we constructed a multiple-choice case-study test. We asked participants to respond to 5 multiple choice questions about the vignette character's presenting problem, symptoms, and diagnosis. See Appendix F for a copy of the questions. Per recommendations by Oskamp (1965), items were constructed only where there was fairly objective criterion information presented in the case. The incorrect alternatives for each item were constructed in such a way as to be clearly wrong, based on the published case material, but to be otherwise convincing and "seductive" alternatives. None of the items had their answers contained in the summarized case material; instead, participants were expected to follow the usual procedure in clinical judgment (McArthur, 1954) by forming a diagnostic picture of the youth from the material presented and then imputing her symptoms from their case formulation of the youth.

For each question of the multiple-choice case-study test, we asked participants to provide a confidence judgment. In other words, we asked them to rate their confidence in the correctness of their response from 0 to 100%. Confidence judgments were made using Adams (1957) confidence scale. According to this scale, confidence is defined in terms of expected percentage of correct decisions. Thus, of those responses made with confidence *p*, about *p*% should be correct. This scale has the advantage of facilitating a direct comparison between the level of accuracy and the level of confidence. For example, if a participant responded to 30% of the items correct and had an average confidence level of 85%, he/she demonstrated *overconfidence*.

Measures

Independent variables.

Demographic and background information. We gathered the following demographic and background information from all study participants (see procedures above): age; gender; race/ethnicity; years of clinical experience; clinical setting; theoretical orientation(s); typical client caseloads; and, training in evidence-based practices. See Appendix B for complete information collected.

Treatment condition. Participants in the treatment condition received the cognitive de-biasing intervention (as described previously) whereas participants in the control condition did not. Participants in both groups received the same brief introduction to mood disorders.

Race/ethnicity. Vignette characters were experimentally manipulated so that the youths in each of the vignettes were African American half of the time and Caucasian the other half of the time.

Dependent variables.

Decision accuracy. Several methods for defining judgment accuracy exist in the literature (see Spengler, 2009). The present study employed dependent measures of high criterion validity. Examples of highly valid criteria include extensive a priori validation of written clinical vignettes (e.g., Spengler, Strohmer, et al., 1990) as well as post hoc manipulation checks of clinical vignettes (e.g., Goldsmith & Schloss, 1986). To help assess the accuracy of participants' decision-making, dependent measures included: (a) diagnostic accuracy (per vignette), (b) decision-making error status (per vignette and overall), (c) risk estimate accuracy (for the diagnosis momentum vignette only), (d) overall diagnostic

accuracy (i.e., a composite measure of (a) and (b) above); and, (e) confidence judgment accuracy (i.e., for the overconfidence vignette only).

Diagnostic accuracy. We asked participants to report the youth's probable diagnoses after reading each vignette. These diagnoses were rated on a 3-point Likert scale of (1 = inaccurate diagnosis, 2 = somewhat accurate, and 3 = accurate diagnosis) using criteria established by the research team (which includes expert diagnosticians). In addition to selecting the most probable diagnosis, participants were able to indicate other diagnoses that they were considering.

Decision-making error status. Responses that earned "inaccurate" and "somewhat accurate" scores on *diagnostic accuracy* (see above) were classified as committing a decision-making error; "accurate" responses indicated no decision-making error.

Risk estimate accuracy. For the vignette testing diagnosis momentum, participants' estimates were compared to an objective, actuarial estimate--the Bayesian posterior probability of the likelihood of bipolar disorder. The Bayesian approach is described in detail in several peer-reviewed manuscripts (see Jenkins et al., 2011; Youngstrom & Duax, 2005; Youngstrom, Freeman, & Jenkins, 2009). For this vignette, the true risk estimate was 27% (±5%) based on: clinic setting (e.g., the youth was seen in outpatient clinic, so the starting base rate of PBD is 6%); family history information (e.g., the youth had a second degree relative with bipolar disorder increasing the odds of a bipolar diagnosis by a factor of 2.5), and a test score of 35 on the PGBI (which is associated with a DLR of 2.3) (Youngstrom, et al., 2004). Participant responses were rated on a 3-point Likert scale of (1 = inaccurate diagnosis, 2 = somewhat accurate, and 3 = accurate diagnosis). A 3-point response was 27%, a 2-point response fell within ±5% of the true estimate (from 22 to 32) (Sedlmeier &

Gigerenzer, 2001), and a 1-point response fell outside of a ±5% range (less than 22 or greater than 32).

Overall diagnostic accuracy. Overall diagnostic accuracy was a composite measure of diagnostic accuracy and risk estimate accuracy across all of the case vignettes. This variable serves as a primary outcome measure.

Confidence judgments. The confidence judgments were made using a scale devised by Adams (1957) that defines confidence in terms of expected percentage of correct decisions. In addition to providing a clearly understood objective meaning for confidence, this scale had the advantage of allowing a direct comparison between the level of accuracy and the level of confidence (Oskamp, 1965). The discrepancy between participants' percent accurate and percent confident (i.e., the difference score) was the dependent variable that we examined across treatment and control groups.

Subjective Numeracy Scale. The Subjective Numeracy Scale (SNS) is an 8-item self-report measure that assessed respondents' perceived ability to perform various mathematical tasks as well as their preference for numerical versus prose information [2]. It required no mathematical computations and had no correct answers. Four questions asked respondents about their numerical ability in different settings, and four questions inquired about respondents' preferences for the presentation of numerical and probabilistic information. The SNS is both reliable and highly correlated with the Lipkus, Samsa & Rimer (2001) numeracy measure (Fagerlin et al., 2007). See Appendix G for a copy of the measure.

Clinician attitudes toward the vignette exercise. All participants completed a brief attitude questionnaire about their experience completing the case vignette exercise. See Appendix C for a copy of the questionnaire items. This measure described different aspects

of the vignette exercise (e.g., how challenging it was to make diagnostic decisions), and required participants to make ratings on a 6-point scale: (1) strong agree; (2) agree; (3) somewhat agree; (4) somewhat disagree; (5) disagree; and, (6) strong disagree.

Clinician attitudes toward the intervention. Participants in the treatment group completed a brief questionnaire (4 items) about their attitudes toward the cognitive de-biasing intervention. See Appendix D for a copy of the questionnaire items. This measure described different aspects of the intervention (e.g., how helpful it was in making diagnostic decisions), and required participants to make ratings on a 6-point scale: (1) strong agree; (2) agree; (3) somewhat agree; (4) somewhat disagree; (5) disagree; and, (6) strong disagree.

RESULTS

Power Analyses

Power is the probability that the analysis will reject the null hypothesis when indeed it should be rejected. There was adequate power to detect effects for all primary analyses. A sensitivity analysis was run for the Logistic regression, Polytomous Universal Model (PLUM) regression, Multiple Linear regression, Analysis of Covariance (ANCOVA), *t*-test, and Chi-Square tests.

Logistic regression modeled the likelihood of decision-making errors for each vignette. Specifically, one goal of the proposed study was to test the null hypothesis that the event rate is identical in the treatment and control groups; or, equivalently, the odds ratio is 1.0, the log odds ratio (beta) is 0.0, and the relative risk of making a diagnostic error is 1.0. Power was computed to reject the null under the following alternate hypothesis. Based on the sample size, power was adequate (80% or higher) for relative risks of 2.8 or larger. For example, for the control group, the rate of diagnostic errors might be 0.50 or higher, and for the training group the error rate would be 0.18 or lower. The study included a total of 79 participants, assigned as follows: 60% to the control and 40% to the training group. The criterion for significance (alpha) was set at 0.05. The test is 2-tailed, which means that an effect in either direction would be interpreted. For this distribution, effect size (event rates of 0.50, 0.18), sample size (79), and alpha (0.05, 2-tailed), power was 80%. This means that 80% of studies would be expected to yield a significant effect, rejecting the null hypothesis

that the event rates are identical (Borenstein, Hedges, Rothstein, Cohen, & Schoenfeld, 2010).

PLUM regression modeled diagnostic accuracy for each vignette. Current power analysis programs do not include PLUM regression in their power calculations. The sensitivity analysis for the PLUM regression is an approximation based on the logistic regression: 80% power to reject the null hypothesis given effect size, sample size (79), and alpha (0.05, 2-tailed). Notably, using logistic regression as an approximation is an underestimation of power (Hsieh, Bloch, & Larsen, 1998).

While logistic regression and PLUM regression tested participants' performance on the individual vignettes, Multiple Linear regression (MLR) tested participants' overall performance both in terms of the total number of decision-making errors committed and the sum accuracy of diagnostic decisions across all vignettes. For the MLR, we had 80% power to detect effect sizes of f^2=.10 or larger (Cohen, 1988) for the given sample size (N = 79) and alpha = .05, 2-tailed) (Faul, Erdfelder, Buchner, & Lang, 2009). Cohen described effect sizes of f^2 ~.02 as "small," .15 as "medium," and .35 as "large."

ANCOVA compared differences in Bayesian estimates, clinician confidence, and attitudes toward the vignette exercise across the treatment and control groups. The power sensitivity analysis indicated that we had 80% power to detect a medium effect size (f = .32) (Cohen, 1988) for the given sample size (N = 78) and alpha (.05, 2-tailed). Cohen suggested benchmarks of f ~.10 as "small," .25 as "medium," and .40 as "large."

Next, *t*-tests compared the treatment and control group on continuous scores. Based on the number of participants, we had 80% power to detect effect sizes of d=.65 or larger

(Cohen, 1988) for the given sample size (N = 78) and alpha = .05, 2-tailed), corresponding to somewhere between a medium (d ~.5) and large (d ~.8) effect.

Lastly, chi-square tests examined the association between decision-making error and practice implications. Results from this sensitivity analysis (based on the chi-square test with the highest number of degrees of freedom which was 5) indicated that we had 80% power to detect a critical χ^2 = 12.59 for the given sample size (n = 78) and alpha (.05, 2-tailed).

Quantitative Preliminary Analyses

Prior to quantitative analyses, data were screened to ensure quality and to check standard statistical assumptions. Given that the amount of missing data was small (<2%), missing data were excluded listwise. This approach provides less bias than pairwise deletion and is adequately suited for small amounts of missing data (Allison, 2002). There was only missing data for one participant. Therefore, we did not need to create dummy codes to indicate the presence or absence of data and run correlations between missing data indicators and variables to examine for evidence of patterns or associations with relevant variables. We performed statistical analyses with standard software packages—IBM Statistical Package for the Social Sciences (IBM SPSS Statistics) Version 19.0.

Descriptive statistics, including means, standard deviations and frequencies, quantified all major study variables. We checked the distributions of variables for normality and obtained plots of the data (e.g., box-plot, stem-and-leaf plot) to determine the necessity of data transformation and/or removal of extreme cases. Also, we examined bivariate correlations among the variables.

Overall Diagnostic Accuracy

We modeled overall diagnostic accuracy using two different multiple linear

regression (MLR) models. In both models, we predicted that condition status (IV) would lead to enhanced decision-making across all four vignettes accounting for participants' years of clinical experience. Tables 10 and 11 provide an overview of participants' performance on the vignettes by condition status and across groups. Note that for Table 10 high scores indicate more decision-making errors but for Table 11, high scores indicate greater accuracy.

In the first model, we used MLR to test the hypothesis that participants in the treatment condition would commit fewer decision-making errors than participants in the control condition. See Table 10 for an overview of the total number of errors committed per group. Participants ranged from 0 to 4 (i.e., high scores are bad, reflecting more errors). Results indicate a significant regression equation ($F(2,75) = 10.37$, $p < .0005$) with an R^2 of .22 using treatment condition and years of experience as predictors of errors. Assignment to the training condition predicted significantly lower numbers of errors, $p < .0005$; however, years of clinical experience did not, $p > .05$.

In the second model, MLR tested the hypothesis that participants who received the cognitive de-biasing intervention would make more accurate diagnostic decisions. See Table 11 for an overview of the summary of diagnostic accuracy ratings; participant scores ranged from 5 to 12 (i.e., high scores are good). Results indicate a significant regression equation ($F(2,75) = 10.86$, $p < .0005$) with an R^2 of .23, and assignment to the training condition made a significant unique contribution even after controlling for years of experience. Years of clinical experience did not make a significant unique contribution, $p > .05$. As proposed in the design, we included years of clinical experience in both regression models as a regression covariate, but it did not contribute a significant portion of the variance in either model, $p > .05$.

Race/Ethnicity Bias: Are Clinicians' Diagnostic Decisions Vulnerable to Race/Ethnicity Bias?

For each of the four vignettes, we performed chi-square tests to assess for any associations between vignette characters' race/ethnicity (IV) and participants' diagnostic accuracy. Results from chi-square tests indicated no significant associations or trends between vignette characters' race/ethnicity and judgment accuracy. Out of the four vignettes, the Lynda vignette (e.g., classic mania symptoms) produced the largest chi-square, χ^2 (2) = 3.94, p = .14. Notably, even the largest chi-square did not approach significance, and the pattern of observed versus expected frequencies showed higher rates of accurate decisions when Lynda was presented as African-American – in the opposite direction of what would have been predicted in a bias hypothesis. Because no associations were found between race/ethnicity of vignette characters and diagnostic accuracy, we did not include vignette race as a covariate in the regression models presented.

Performance on Individual Vignettes

We ran separate analyses for the base-rate neglect, search satisficing, diagnosis momentum, and overconfidence vignettes. Logistic and PLUM regression tested study predictions. Logistic regression tested if group status (IV) predicted decision-making errors (present/absent). Diagnostic accuracy served as an operational definition for decision-making errors, with "accurate" diagnoses indicating no decision-making error and "somewhat" and "inaccurate" diagnoses indicating decision-making errors. PLUM regression then tested if group status predicted the level of accuracy (e.g., accurate, somewhat accurate, inaccurate). Results for the individual vignettes are provided below.

Base Rate Neglect: Does the Cognitive De-Biasing Intervention Increase Diagnostic Sensitivity to Mania/Hypomania?

Logistic regression tested if participants in the treatment condition were less likely to commit base rate neglect, as demonstrated by their inclusion of bipolar disorder as a probable diagnosis. Decision-making error was a dichotomous variable, rated '0' if no error was committed and '1' if the participants committed an error (i.e., in this case, *base-rate neglect* by not accounting for the possibility of mania or hypomania in a depressed phase of illness). Condition status (treatment or control) (IV) predicted decision-making error (absent or present) (DV). In the treatment group, 41% committed a decision-making error, versus 76% in the control group, χ^2 (1) = 10.10, $p < .005$.

Given the significant results of the logistic regression, the SPSS Ordinal Regression procedure, or PLUM regression--an extension of the general linear model to ordinal categorical data--tested levels of judgment accuracy (inaccurate, somewhat accurate, accurate) and included clinician demographic variables (e.g., years of clinical experience and professional title) as regression covariates. We included participants' professional title as a dummy code, with non-psychologists coded as "0" and psychologists coded as "1". In the treatment group, 56% were psychologists, versus 51% in the control group, χ^2 (1) = .21, p = .65.

PLUM regression tested the hypothesis that participants in the treatment group make more accurate diagnostic judgments (DV) than participants in the control condition, as evidenced by scores of 3 (e.g., most accurate) than scores of 1 and 2 (e.g., inaccurate and somewhat accurate, respectively). Responses that included depression and inquired about hypo/mania were coded as accurate versus responses that included depression but missed the

possibility of mania/hypomania which were coded as somewhat accurate. Inaccurate responses missed depression and/or consideration of mania symptoms.

The overall chi-square for the PLUM regression was significant, χ^2 (3) = 15.96, p < .005, with a Nagelkerke R^2 of .23. From the observed significance levels, treatment group status was significantly related to diagnostic accuracy in the expected direction, B = -1.64 (1), p < .005. As predicted, years of clinical experience did not make a significant unique contribution, p > .05. Notably, professional title (i.e., psychologist: yes/no) approached significance, p = .051, suggesting a trend in psychologists' diagnostic decisions as being more accurate than non-psychologists' decisions.

Search Satisficing: **Are Clinicians More Sensitive to the Possibility of Comorbid Conditions as a Result of the Cognitive De-Biasing Intervention?**

Similar model building tested hypotheses related to search satisficing. Specifically, logistic regression tested if participants in the treatment condition were less likely to commit search satisficing as evidenced by their consideration of ADHD and mood symptoms. Again, decision-making error was a dichotomous variable, rated '0' if no error was committed and '1' if the participants committed an error. Condition status (treatment or control) (IV) predicted decision-making error (absent or present) (DV). In the treatment group, 41% committed a decision-making error, versus 68% in the control group. Logistic regression revealed a significant relation between condition status and decision-making error, χ^2 (1) = 5.87, p < .05.

Given the significant results of the logistic regression, PLUM regression tested levels of judgment accuracy (inaccurate, somewhat accurate, accurate) and clinician demographic variables (e.g., years of clinical experience and professional title) as regression covariates.

Responses that included consideration of ADHD and a bipolar disorder were coded as accurate versus responses that included only ADHD or a bipolar disorder which were coded as somewhat accurate. Inaccurate responses missed ADHD and bipolar disorder. Consistent with previous analyses, we included participants' professional title as a dummy code (i.e., psychologist, yes/no). PLUM regression tested the hypothesis that participants in the treatment group make more accurate diagnostic judgments (DV) than participants in the control condition, as evidenced by scores of 3 (e.g., most accurate) than scores of 1 and 2 (e.g., inaccurate and somewhat accurate, respectively).

The overall chi-square for the PLUM regression was not significant, χ^2 (3) = 6.07, p = .11, with a Nagelkerke R^2 of .09. From the observed significance levels, group status was significantly related to diagnostic accuracy, B = -1.02 (1), $p < .05$. As predicted, years of clinical experience did not make a significant unique contribution, $p > .05$, nor did professional title (i.e., psychologist: yes/no), $p > .05$.

Diagnosis Momentum: Are Participants in the Control Condition More Likely to Endorse Significantly Higher Probabilities of Bipolar Disorder?

Multiple linear regression tested the hypothesis that participants' probabilities of a bipolar diagnosis in the treatment group were significantly different than probabilities generated by participants in the control condition, accounting for covariates such as years of clinical experience, professional title (i.e., psychology: yes/no), subjective numeracy (i.e., average score on the subjective numeracy scale), and previous exposure to Bayesian approaches. Previous participation in one of Dr. Eric Youngstrom's continuing education (CE) seminars served as a proxy for exposure to Bayesian approaches. Notably, only 6 participants had attended one of Dr Youngstrom's CEs (9% of the control group and 6% of the treatment group).

Results from MLR indicate that the regression equation was not significant ($F(5,71) = 1.27$, $p > .05$) with an R^2 of .08. Condition status made a significant unique contribution, $p < .05$, with the treatment condition's estimates ($M = 37.16$, $SD = 17.68$) being significantly less than the control conditions' ($M = 46.47$, $SD = 18.91$). None of the other predictors were significant, $p > .05$. A regression model that only included treatment status without the other (inert) covariates would have been significant, suggesting that the omnibus model was affected by including so many nonsignificant covariates while working with an only moderate sized sample. Table 12 shows the regression coefficients, *t* values, and *p* values for group status and the regression covariates.

For the diagnosis momentum vignette, the true risk estimate was 27% (±5%) based on: clinic setting (e.g., the youth is seen in outpatient clinic so the starting base rate of PBD is 6%); family history information (e.g., the youth has a second degree relative with bipolar disorder), and a test score in the moderately high range on the PGBI (which is associated with a DLR of 2.22) (Youngstrom, 2005). A one sample *t*-test supports the prediction that participants' estimates of the probability of a bipolar diagnosis in the control group ($M = 46.47$, $SD = 18.91$) are significantly higher than the "true" Bayesian estimate, which in this case was 27%. Notably, control group participants' estimates are also significantly higher when compared to a constant of 32%--the upper limit of +5% from the precise estimate of 27% (Sedlmeier & Gigerenzer, 2001), suggesting that the control condition was at increased risk of diagnosis momentum.

A one sample *t*-test comparing participants' estimates of the probability of a bipolar diagnosis in the treatment group ($M = 37.16$, $SD = 17.68$) to the "true" Bayesian estimate (again, the constant, 27%) was significantly higher as well, $p < .05$. However, when we

compared the treatment group's average estimate to 32% (the upper limit), it was no longer significant, $p > .05$.

Similar to the coding of the other vignettes, participants' risk estimates were coded for decision-making errors and accuracy. Estimates of 27 and estimates lying within ± 5 percentage points around the Bayesian estimate (22 to 32) (Sedlmeier & Gigerenzer, 2001) were coded as evidencing no decision-making error. Estimates falling outside of ± 5 percentage points around the Bayesian estimate (<22 and >32) (Sedlmeier & Gigerenzer, 2001) were coded as evidencing a decision-making error. In the treatment group, 44% committed a decision-making error, versus 85% in the control group. Logistic regression tested if condition status predicated decision-making error (i.e., in this case, estimates falling within +/- 5% of the Bayesian estimate). Results indicate that the cognitive de-biasing intervention lead to less cognitive error in synthesizing the case information, X^2 (1) = 17.26, $p < .0005$.

Notably, we examined outlier diagnostics including the studentized deleted residuals, Mahalanobis' Distance, and Cook's Distance to verify that the data met regression assumptions. Results of the outlier diagnostics revealed two cases where the values predicted were substantially discrepant from the actual scores. In examining these two values, we found that removing them from our data would increase our R^2 and p values but would ultimately result in no substantive changes. For this reason, these values were included in our analyses.

Classic Mania: Does the Cognitive De-Biasing Intervention Help Clinicians Better Understand Manic Symptoms in Pediatric Populations?

Similar to the base-rate neglect and search satisficing vignettes, logistic regression tested if participants in the treatment condition were less likely to commit decision-making

errors when assessing Lynda's manic symptoms. Again, decision-making error was a dichotomous variable, rated '0' if no error was committed and '1' if the participants committed an error. Condition status (treatment or control) (IV) predicted decision-making error (absent or present) (DV).

In the treatment group, 75% committed a decision-making error, versus 83% in the control group. Logistic regression revealed no significant relation between condition status and decision-making error, χ^2 (1) = 0.74, $p > .05$.

We ran PLUM regression to examine if participants' decision-making between groups differed when we applied levels of diagnostic accuracy (accurate, somewhat accurate, and inaccurate). Responses that included bipolar I disorder were coded as accurate versus responses that included a bipolar spectrum disorder which were coded as somewhat accurate. Inaccurate responses missed bipolar altogether. Results indicate the group status was not a predictor of diagnostic accuracy for this vignette, $B = -.32$ (1), $p > .05$.

Confidence and Judgment Validity

In addition to examining participants' diagnostic decision-making for classic manic symptoms (prior analyses, above), we investigated participants' confidence and judgment validity. Specifically, regression tested whether the discrepancy scores (DV) between participants' accuracy and confidence judgments are significantly different depending on treatment group status. Notably, this regression model included discrepancies that reflected over and under confident decision-making. Results from the regression show a positive trend, in that group status approaches significance in predicting discrepancies, B=-14.28, $p = .053$, with participants in the treatment condition showing smaller discrepancies between the accuracy of their judgments and their confidence in those judgments.

We ran a different set of analyses to specifically examine the study prediction that participants in the treatment condition would demonstrate less *overconfidence* (i.e., smaller discrepancies between accuracy and confidence in their decision-making) than the control group. In this set of analyses, we excluded 22 cases of *underconfidence*. ANCOVA tested the effect of condition status on participant overconfidence (57 cases), covarying out years of clinical experience and professional title (psychology: yes/no). Results from the one-way between-subjects ANCOVA indicated that years of clinical experience was significantly related to overconfidence ($F(1,53) = 4.60$, $p < .05$) as was professional title ($F(1,53) = 4.82$, $p < .05$).

Participants with more clinical experience evidenced significantly less *overconfidence*. Non-psychologists ($n = 31$) were significantly more *overconfident* ($M = 31.90$, $SD = 21.08$) than psychologists ($n = 26$; $M = 20.77$, $SD = 20.47$). The main effect for treatment condition was significant, ($F(1,53 = 4.36$, $p < .05$), with the treatment condition's discrepancies between accuracy and overconfidence ($M = 18.62$, $SD = 17.45$) significantly less than the control conditions' ($M = 32.17$, $SD = 22.40$). On average, participants in the treatment condition displayed significantly less overconfidence than participants in the control condition.

Decision-Making Errors and Practice Implications

Chi-squared analyses tested study predictions regarding practice implications of faulty heuristics and biases. Specifically, we hypothesized that somewhat accurate and inaccurate diagnoses would be associated with significantly different assessment and treatment choices. To examine the clinical implications of cognitive errors, for each heuristic/bias investigated in the case vignettes, chi-squared tests evaluated the relationship

between the particular heuristic/bias and participant recommended clinical action (e.g., heuristic [present/absent] x clinical action [more assessment/therapy/medication /no offer of treatment/other]). We used the accuracy of participants' diagnoses (including participants in both conditions) as the operational definition for the presence of decision errors. Consistent with other coding, inaccurate or somewhat inaccurate diagnostic judgments were coded as "positive" for the occurrence of a decision-making error (i.e., decision-making error present); accurate diagnostic decisions were "negative" for decision-making errors. Chi-squared tests for each vignette are provided below.

Chi-square tests for the search satisficing (χ^2 (4) = 6.21, $p > .05$) and base rate neglect (χ^2 (6) = 7.27, $p > .05$) showed no significant associations between decision-making error and next clinical action, considered globally. However, for the base-rate neglect vignette, we tested the association between decision-making error and selection of antidepressant medication (or referral for antidepressant medication) as the next clinical action and we found a significant association, χ^2 (1) = 4.91, $p < .05$. The pattern of observed versus expected frequencies showed that when people failed to consider hypomania/mania history, they were significantly more likely to recommend consideration of an antidepressant.

For the diagnosis momentum vignette, which included participants providing likelihood estimates of the probability of bipolar disorder, chi-squared revealed a trend but no significant association between decision-making error and next clinical action recommendation, χ^2 (4) = 8.57, $p = .07$.

Finally, for the vignette depicting a youth with classic mania symptoms, chi-square revealed a positive association between decision-making error and next clinical action, χ^2 (3) = 18.60, $p < .0005$. The pattern of observed versus expected frequencies showed that when

participants detected bipolar disorder they were more likely to recommend medication and assessment.

Clinician Experience of the Vignette Exercise

The majority of participants reported that the case vignettes used in the study resembled clients that they have seen in their respective practice settings. Specifically, nineteen percent strongly agreed, 37% agreed, and 40% somewhat agreed to this statement. Three percent somewhat disagreed and one percent disagreed.

ANCOVA tested the prediction that the cognitive de-biasing intervention would have a positive effect on participant experience of the case vignette portion of the study. In particular, we anticipated that participants in the treatment group would demonstrate significantly more favorable attitudes toward the case vignette exercise than participants in the control group. Using ANCOVA, we tested whether group status had an effect on participant attitudes after adjusting for covariates such as years of clinical experience and participants' professional title (e.g., psychology: yes/no).

Two main dimensions of participants' attitudes toward the vignette exercise were evaluated, including: (a) how challenging participants found the vignette exercise; and, (b) participants' overall confidence in their diagnostic and treatment decisions. On average, participants in both groups agreed that it was challenging to make diagnostic decisions for vignette characters. Psychologists ($M = 3.67$, $SD = 1.41$) were slightly yet significantly more confident than non-psychologists ($M = 3.06$, $SD = .89$).

Results from the one-way between-subjects ANCOVA indicated that years of clinical experience was not significantly related to how challenging participants found the vignette exercise $F(1,73) = .01$, $p = .94$); however, participant title was significant, $F(1,73) = 4.66$, $p =$

.03, with psychologists (M = 2.02 SD = .84) finding the vignette exercise significantly less challenging than non-psychologists (M = 2.44 SD = .93). Treatment status was not significant, ($F(1,73) = .04$, $p = .84$ with the treatment condition's rating (M = 2.22, SD = .83) comparable to the control conditions' (M = 2.22, SD = .96).

We found a similar pattern of findings with respect to participants' overall confidence in their diagnostic and treatment decisions for the case vignettes. Specifically, ANCOVA indicated that years of clinical experience was not significantly related to participants' self-reported confidence in their diagnostic decisions $F(1,73) = 1.04$, $p > .05$); however, participant title was significant, $F(1,73) = 8.45$, $p < .05$, with psychologists reporting higher confidence (M = 3.67, SD = .89) than non-psychologists (M = 3.06, SD = 1.14). Intervention status was not significant, ($F(1,76) = .34$, $p > .05$) with the treatment condition's rating (M = 3.50, SD = 1.05) comparable to the control conditions' (M = 3.30, SD = 1.09).

Finally, in terms of participants' self-reported confidence in their treatment decisions, ANCOVA indicated no significant results. Specifically, years of clinical experience was not significant ($F(1,73) = 2.82$, $p > .05$) nor was psychologist versus other professional status--$F(1,73) = 1.44$, $p > .05$. Also, the main effect for condition status was not significant, $F(1,73) = 1.76$, $p > .05$. On average, participants in both groups experienced similar levels of confidence in their treatment decisions.

Clinician Attitudes toward the Intervention

Descriptive statistics supported the hypothesis that clinicians who received the intervention would find it helpful and informative. Out of thirty-two participants in the treatment condition, 22% strongly agreed to *learning something new*, 56% agreed, and 22% somewhat agreed. Sixteen percent strongly agreed to *finding the tutorial helpful to for*

making diagnostic decisions in the case vignettes, with 50% agreeing and 34% somewhat agreeing.

When asked if they would *recommend the intervention to a colleague*, 25% strongly agreed, 47% agreed, and 28% somewhat agreed. Finally, twenty-two percent strongly agreed that *information in the tutorial was easy to understand*, with 59% agreeing and 16% somewhat agreeing. Notably, the majority of participants in the treatment condition (91%) reported using material from the cognitive de-biasing intervention when responding to the case vignettes.

Additionally, we ran post hoc analyses to examine the potential influence of participants' preference for numerical versus prose information on their attitude toward the cognitive de-biasing intervention. Specifically, given the Bayesian component of the de-biasing intervention, we predicted that higher ratings on the Subjective Numeracy Scale (IV) would lead to more positive ratings of the cognitive de-biasing intervention (DV). Results from the simple linear regression indicate no significant relationship between participants' scores on the SNS and participants' attitudes toward the vignette--$F(1,30) = 0.00$, $p > .05$, with an R^2 of 0.00. This suggests that the de-biasing methods may not be limited to those of a more quantitative persuasion.

DISCUSSION

The overarching goal of the present study was to test the efficacy of a new cognitive de-biasing intervention aimed at improving clinicians' assessment of PBD and common comorbid conditions. In short, we sought to better understand *where* mistakes occur and *how* to improve practice with science. Recent research documents the clinical utility of evidence-based tools for decreasing the overdiagnosis of PBD and helping clinicians detect true cases of PBD (Jenkins, et al., 2011; Youngstrom, et al., 2009); however, we are unaware of any existing research that has (a) tested these strategies using experimental methods; (b) with multiple case vignettes; (c) through a decision-making lens; and, (d) as part of a broader yet targeted intervention package (i.e., participants learned about cognitive vulnerabilities specific to PBD and de-biasing tools for assessment; see Table 5 for a comprehensive overview of intervention components). Main hypotheses and findings are discussed in more detail below.

Clinicians have demonstrated substantial difficulty accurately and efficiently diagnosing PBD (Kessler, 1999; Marchand, Wirth, & Simon, 2006; Youngstrom, Birmaher, & Findling, 2008; Youngstrom, Findling, Youngstrom, & Calabrese, 2005) as evidenced by frequent delays in diagnosis, misdiagnosis, and overdiagnosis (Hirschfeld, Lewis, & Vornik, 2003; Kessler, et al., 1997; Lish et al., 1994). Unfortunately, misdiagnosis of PBD and associated treatment delays carry serious consequences (e.g., see Dunner, 2003; Geller, Tillman, Craney, & Bolhofner, 2004; cf. Joseph, Youngstrom, & Soares, 2009). From a decision-making perspective, the complex presentation of bipolar illness, coupled with its

relatively low prevalence and high stakes status, increases the likelihood of cognitive errors (Croskerry, 2003; Davidow & Levinson, 1993). Also, there is a tendency for mental health professionals to rely solely on clinical judgment in making important diagnostic decisions. This approach is problematic given the decades of research showing that clinical judgment can compromise optimal decision-making (Aegisdottir, et al., 2006; Croskerry, 2002; Elstein & Schwartz, 2002; Klein, 2005; Meehl, 1954).

At present, there is relatively little information about decision-making and the clinical utility of cognitive de-biasing strategies in mental health. Fortunately, related disciplines including cognitive science and medical decision-making have more well-established decision-making literatures about cognitive error and de-biasing interventions (Croskerry, 2003). These literatures can help inform new approaches for improving diagnostic and treatment decisions for challenging psychiatric disorders such as pediatric bipolar disorder.

Overall Diagnostic Accuracy

As predicted, participants' judgments in the treatment group evidenced higher overall decision accuracy than participants' judgments in the control condition. We measured overall decision accuracy in two ways: total decision-making errors and diagnostic accuracy ratings across all four vignettes. In both cases, the treatment group performed significantly better than the control group. In particular, participants who received the cognitive de-biasing intervention committed fewer decision-making errors and generated more accurate diagnoses for multiple vignettes with various case presentations.

Our investigation of overall diagnostic accuracy examined the influence of participants' amount of clinical experience. Notably, participants had a wide range of clinical experience (e.g., early stage career professionals to professionals with 20+ years), and

participants had comparable amounts of clinical experience between the treatment and control groups. As predicted, amount of clinical experience consistently had no effect on participants' diagnostic accuracy, including their overall accuracy and their accuracy on individual vignettes. This non-significant result is noteworthy in light of the literature suggesting that additional clinical experience does not necessarily enhance clinical judgment, and instead can reinforce faulty cognitive habits (Brehmer, 1980, 1998; Faust, 1986; Faust, 1994; Faust, et al., 1988; Faust & Ziskin, 1988; Garb, 1989; Spengler, et al., 2009; Wedding, 1991; Wiggins, 1973; Ziskin, 1995). One may be surprised by the finding that clinicians with decades of more experience performed similar to early stage professionals. It is important to point out that diagnosing bipolar disorder in youth populations is a relatively new practice, and that more seasoned professionals were likely not trained in the assessment or treatment of PBD in their respective training programs. Given the significant treatment effect, it appears that clinicians can benefit from cognitive de-biasing strategies regardless of how long they have been practicing in the field.

Race/Ethnicity Bias

Contrary to study hypotheses, participants' diagnoses were not affected by the race/ethnicity of the vignette character. Specifically, clinicians' rates of bipolar diagnoses did not vary as a function of the youth being African American or Caucasian, with all other case information held equal by design. It is worth noting that in addition to describing the youth's race/ethnicity in the case vignettes, we included a picture of the fictitious youth to make race/ethnicity more salient to study participants (see Appendix E). Our non-significant race/ethnicity findings are intriguing given the tendency for clinicians in the community to misdiagnose African-American youths with conduct disorder or schizophrenia (DelBello, et

al., 2001; Fabrega, Ulrich, & Mezzich, 1993; Kilgus, Pumariega, & Cuffe, 1995) when there is no empirical support for variation in diagnostic rates of bipolar across racial and ethnic groups (American Psychiatric Association, 2001; Grant, et al., 2005; Jonas, et al., 2003; Kessler, et al., 2005; Van Meter, et al., 2011).

Although contrary to our study predictions, this finding is consistent with findings from recent ethnographic studies (Carpenter-Song, 2009). Specifically, qualitative methodologies have revealed potential differences in how minority populations describe mood symptoms, possibly emphasizing behavioral issues versus emotional problems (Carpenter-Song, 2009). In turn, clinicians may mistakenly focus on externalizing symptoms and fail to detect mood symptoms. This possibility raises the question of whether another heuristic may be at play that better accounts for clinicians' cognitive error than race/ethnicity bias. The present study's vignette methodology may not have been as sensitive to this type of race/ethnicity bias (or other heuristic at play), particularly inasmuch as participants could not ask follow-up questions that might have amplified the effects of faulty heuristics; thus, future research may want to consider alternative methods for testing race/ethnicity bias as well as how to best operationalize this cognitive error.

In general, race/ethnicity bias in clinical decision-making continues to be a significant concern, especially for African Americans for whom recent evidence indicates a rise in mental health disparities (Cook et al., 2007). Researchers advocate for additional mixed-methods research to elucidate the role, forms and effects of bias (Carpenter-Song, Whitley, Lawson, Quimby, & Drake, 2011). Although experimental methods have certain benefits, it is possible that we currently do not have a good enough understanding of race/ethnicity bias in the role of assessing childhood mood disorders to test for this bias using experimental

designs. Alternatively, we were not powered to detect small effect sizes for analyses examining race/ethnicity bias, in which case experimental methods may be well suited for testing race/ethnicity but require increased power. Taken together, it will likely be advantageous to consider alternative methods (e.g., qualitative interviews) as follow up studies to the present project as well as increasing power in experimental studies to detect smaller effect sizes.

Performance on Individual Vignettes

Although each vignette tested a different decision-making error, we took a similar approach to examining the treatment effect across all of the vignettes. Notably, the diagnosis momentum vignette involved some additional analyses: It was the only vignette that required participants to estimate the risk of a bipolar diagnosis using percentages. For each vignette, we also investigated similar covariates, including years of clinical experience and professional title (e.g., psychology or non-psychology professional). Consistently across all four vignettes, both years of clinical experience and professional title made no significant impact in participants' decision-making performance. Findings from each vignette as well as the broader implications are discussed below.

Base Rate Neglect. Unlike participants in the control condition, who received a brief introduction to PBD, participants in the treatment group were more likely to assess for hypomania or mania and/or consider bipolar disorder as a probable diagnosis in the presence of depressive symptoms. The intervention appears to help safeguard clinicians' decision-making, particularly from cognitive errors associated with the cyclical nature of bipolar illness (e.g., florid mania, severe depression, mixed or a euthymic state)--a major challenge associated with assessing bipolar disorder (Youngstrom, 2005). This finding is especially

meaningful because going into this vignette, all clinician participants were primed for bipolar as a function of the brief overview of PBD provided before the case vignette exercise. However, only the treatment group demonstrated diagnostic sensitivity to hypo/mania when a case presented with depression and no current symptoms of elevated mood. This finding suggests that the observed significant effect was likely the result of the intervention and not priming or recency effects from the PBD overview. Minimizing base-rate neglect (as the cognitive de-biasing intervention accomplished) has important practice implications including decreasing long delays in diagnosis (Hirschfeld, et al., 2003; Marchand, et al., 2006) as well as rates of underdiagnosis (Angst, 2007). Moreover, roughly one third of mood disorder diagnoses may be on the bipolar spectrum, but the majority of individuals seek treatment when feeling depressed. Thus, failure to probe for hypomania or past mania contributes heavily to misclassification.

Search Satisficing. Per recommendations by Croskerry (2003), the intervention provided clinicians *specific training* on potential overlap between ADHD and bipolar symptoms as well as mechanisms for *considering alternatives* (e.g., checklists). These specific cognitive de-biasing strategies seem effective for helping clinicians tease apart symptoms and correctly identify comorbid conditions. Consistent with study predictions, clinicians who received the cognitive de-biasing intervention were significantly less likely to commit search satisficing by accounting for symptoms of both ADHD and PBD in their diagnostic decisions (and not "calling off the search" when they diagnosed one of the disorders). Given that bipolar disorder and ADHD co-occur more frequently than by chance (Singh, DelBello, Kowatch, & Strakowski, 2006) and that comorbid bipolar disorder and ADHD has been associated with poorer global functioning, greater symptoms severity, and

more additional comorbidity than for either single disorder (Arnold, et al., 2011), our observed improvement in clinicians' decision-making has substantial value for real-world practice.

Recently in the medical decision-making literature, cognitive forcing strategies such as checklists have gained particular attention, especially for diagnostic purposes (Ely, Graber, & Croskerry, 2011). Checklists function to decrease clinicians' reliance on intuition and memory in clinical problem-solving, and they are especially attractive tool for complex diagnoses (Ely, Graber, & Croskerry, 2011). Overall, given the high comorbidity rates of psychiatric disorders in pediatric populations (Angold, Costello, & Erkanli, 1999), including bipolar disorder as well as other disorders (Kowatch, Youngstrom, Danielyan, & Findling, 2005), and the preliminary support for the effectiveness of checklists found in the present study, more research focused on cognitive forcing strategies and user-friendly checklists seems warranted.

Diagnosis Momentum. In addition to some evidence of staggering rates of overdiagnosis (Blader & Carlson, 2007; Moreno, et al., 2007), research has shown a tendency for clinicians to overestimate the likelihood of PBD compared to the supporting evidence when presented with case information (Jenkins, et al., 2011). Clinicians in the control condition of the present study demonstrated a similar susceptibility to diagnosis momentum. Clinicians' risk estimates in the control group were significantly higher than both the true risk estimate (based on Bayesian reasoning) and the treatment group's risk estimates. Similar to the other vignettes, clinicians' years of experience and professional title (i.e., psychology or non-psychology professional) did not play a significant role in diagnostic accuracy. For this vignette, we also examined clinicians' previous exposure to Bayesian approaches and

their self-ratings on the Subjective Numeracy Scale. Our finding that neither one of these factors significantly predicted risk estimate accuracy, supports generalizability of the intervention. Specifically, our findings provide preliminary support that clinicians do not need prior knowledge of Bayesian reasoning or to endorse a preference for or ability in numeracy (including working with probabilities) to use actuarial assessment methods effectively.

Further, we examined whether the cognitive de-biasing intervention predicted fewer decision-making errors. Findings indicate that the intervention resulted in less cognitive error in synthesizing the case information. Consistent with previous research, teaching clinicians actuarial methods helped decrease overdiagnosis of PBD *and* generated more accurate decision-making (Jenkins, et al., 2011). Of significance, this was our first attempt to teach actuarial methods using the Web. The significant treatment effect observed in the present study supports future research examining the effectiveness of online training of Bayesian approaches.

Detecting Mania. Contrary to study hypotheses, training condition status did not influence diagnostic accuracy of clinicians' assessment of more classic mania symptoms. Our definition of *accurate* diagnoses was relatively stringent. Only responses that included bipolar I disorder were coded as *accurate*; responses that included a bipolar spectrum disorder were coded as *somewhat accurate*. *Inaccurate* responses missed bipolar altogether. A more lenient approach could have assigned bipolar spectrum diagnoses full credit versus the partial credit participants received by our guidelines. Although we considered this alternative classification, several factors emerged during our review of the case vignette that prompted us to go forward with the more stringent guidelines. The details in the vignette are

completely sufficient to satisfy all DSM-IV criteria for a manic episode: duration criteria (e.g., 1 week or longer); severity (e.g., impairment occurred); intermittent presentation of symptoms (i.e., consistent with episodic nature of mood disorders); and, number of symptoms (e.g., grandiosity, hypersexuality as an example of disinhibited and risky behavior, psychomotor agitation, and distractibility – meeting criteria for at least four "B Criterion" symptoms in addition to the episodic disturbance of mood) (4th ed. [DSM-IV]; American Psychiatric Association, 1994). Also, we conducted a post hoc analysis to examine the potential treatment effect if we assigned full credit for bipolar spectrum diagnoses. Again, we observed no treatment effect.

Overall, findings on this vignette likely reflect the extreme difficulty a clinician encounters when diagnosing manic symptoms in youth in the real-world (Carlson, 1988). Clinicians' low diagnostic sensitivity to mania in the study vignette suggests: (1) an area of the cognitive de-biasing intervention to strengthen; and, (2) a potential weakness of the vignette methodology.

First, the intervention may benefit from incorporating role plays with individuals evidencing manic-like symptoms, or to find ethical and respectful mechanisms for showing real-life depictions of symptoms from individuals with bipolar illness. Given the complicated nature of this diagnosis, more intensive training such as role-plays may be necessary to bring about significant improvement in interviewing and cognitive interpretation and diagnostic decision-making. Second, as an artifact of the case vignette methodology, participants could not ask follow-up questions that might have helped confirm suspicions of bipolar I disorder and helped clinicians be more confident in making this high stakes diagnosis. However, this possibility should be considered within the context of a large body of literature that

unstructured clinical interviewing rarely addresses a consistent body of content, and agreement between clinical diagnoses and semi-structured interviews is modest at best [3]. Thus, allowing more follow-up questions without also providing cognitive decision-supports or de-biasing strategies could as easily make judgments worse.

Confidence and Judgment Validity

Based on the vignette depicting a youth with more classic manic symptoms, clinicians responded to a series of multiple choice questions presented in a randomized order following their diagnosis and treatment decisions for the vignette character. For each answer clinicians provided to the five multiple choice questions, they also provided a confidence estimate indicating their confidence in their response to that particular question. This process established a direct comparison between the level of accuracy and the level of confidence (see study procedures for a review; Oskamp, 1965).

We anticipated that clinicians in the treatment group, who learned de-biasing strategies, would have a more accurate sense of confidence, operationally defined here as smaller discrepancies between the accuracy of their judgments and their confidence in those judgments. Results confirmed the hypothesis, both when we examined combined discrepancies (i.e., instances of under and over confidence) as well as when we focused solely on overconfidence (i.e., discrepancies between accuracy and confidence in which the accuracy of their judgments was less than their confidence ratings). Specifically, we found a positive trend ($p = .053$) when looking at under and over confidence, and we found significant differences between groups for overconfidence despite the reduced sample size used for the more focal hypothesis.

Overconfidence Bias. As hypothesized, clinicians who received the intervention showed significantly less *overconfidence bias* than clinicians in the control condition. In this set of analyses, we also examined the influence of years of experience and professional title (psychologist: yes/no), and we found that both were significantly related to overconfidence. Notably, psychologists and clinicians with more experience demonstrated significantly less *overconfidence bias.* Given the exploratory nature of this secondary aim, we are curious if additional research testing similar effects will generate comparable findings.

The significant treatment effect for *overconfidence bias* is especially intriguing given the insignificant treatment effect we found for diagnostic accuracy and decision-making error for the same vignette character. Although the dependent variables are different (i.e., overconfidence is a different construct than diagnostic accuracy and decision-making error), the content that clinicians are using to make their decisions is similar. Specifically, the multiple choice questions required participants to identify classic manic symptoms (e.g., grandiosity, hypersexualism, psychomotor agitation), the vignette character's probable disorder (e.g., bipolar I disorder), and the likely cause of her change in functioning (e.g., mood disturbance). See Appendix F for a copy of the multiple choice questions and response options. In light of the significant treatment effect when clinicians worked with more concrete diagnostic decisions (versus the more open ended vignette and list of 26 possible disorders from which to choose), it appears that the intervention was helpful for accurately assessing mania. In particular, clinicians in the treatment group were significantly more accurate on multiple choice questions, and they demonstrated a better understanding of their understanding or misunderstanding of the characters' behaviors and problems (i.e., the smaller discrepancies and less *overconfidence*). Clearly, we need more research to better

understand the role of clinician confidence as well as how to best assess discrepancies between accuracy and confidence including the impact of clinicians' amount of experience and their professional title.

Decision-Making Errors and Practice Implications

A secondary goal of the present study involved testing the potential influence of decision accuracy on clinician's recommended treatment approaches. Accurate diagnosis of PBD is especially important before starting treatment (Weller, Danielyan, & Weller, 2004; Kowatch, et al., 2005). We examined the clinical implications of the primary cognitive errors of interest including *base rate neglect, search satisficing, diagnosis momentum, and overconfidence bias.* Specifically, we predicted that committing one of these faulty heuristics would affect clinicians' treatment formulations. Consistent with our previous coding scheme, *inaccurate* and *somewhat accurate* decisions were coded as positive for committing an error while *accurate* decisions were coded as error free. For each of the primary diagnoses clinicians assigned to the vignette characters, they were asked to recommend a next clinical action (i.e., representing their treatment formulation). Their choices for next clinical action consisted of: more assessment, therapy, medication, no treatment necessary, and other. Participants could select up to two of these options for any given diagnosis.

Several findings warrant attention. First, for the vignette character depicting more classic mania symptoms--diagnostically, a particularly challenging vignette for clinicians regardless of condition status--we found a significant association between diagnostic error and next clinical action. Specifically, when participants detected bipolar disorder, they were more likely to recommend medication and assessment. In this case, an accurate diagnosis aligned with recommended treatment guidelines in the literature (Kowatch, et al., 2005).

Second, although *base-rate neglect* was not significantly related with global choices about next clinical action in terms of assessment, psychotherapy, or medication, our analyses revealed a significant association between this decision-making error and type of medication recommendation. In particular, when clinicians overlooked the possibility of hypomania/mania history, they were significantly more likely to recommend antidepressant medication. Given that antidepressants are largely ineffective for treatment of bipolar depression (Nierenberg, et al., 2006) and can possibly worsen outcome (cf. Joseph, Youngstrom, & Soares, 2009), it is worth noting this significant finding. Improved detection of hypomania or mania in people seeking treatment for depression is consistent with FDA recommendations for antidepressant medication use and would decrease the lag between onset of mood problems and recognition of bipolar disorder (Hirschfeld, et al., 2003; Lish, et al., 1994).

Third, for the *diagnosis momentum* vignette, we found a positive trend such that when clinicians did not overestimate the likelihood of bipolar disorder without adequate evidence (as evidenced by a risk estimate in the "acceptable" range compared to the Bayesian estimate), they were more likely to recommend additional assessment. This treatment formulation for this level of risk coincides with current recommendations in the literature (see Youngstrom, et al., 2009).

Finally, it is important to point out that we were underpowered to detect small effect sizes for our investigation of practice implications of cognitive errors. It is possible that with an increased sample size, the positive trend we found for *diagnosis momentum* would become significant. Likewise, it is possible that with more power, we would have detected additional smaller yet significant associations for the other heuristics of interest (e.g., base-

rate neglect). Additional research with higher powered analyses will be advantageous for further examining the relationship between cognitive errors and treatment formulation.

Clinician Attitudes

The Vignette Exercise. Contrary to study predictions, control and treatment group participants provided comparable ratings of their experiences participating in the vignette exercise. Specifically, we examined how challenging clinicians found the vignettes as well as their confidence in their diagnostic and treatment decisions across all of the vignettes. Given that clinicians in the treatment group displayed higher diagnostic accuracy than clinicians in the control group on three out of four of the individual vignettes as well as on overall diagnostic accuracy, it is surprising that we did not observe a treatment effect on these dimensions.

Notably, psychologists found the case vignettes significantly less challenging than the other professionals did, and they also endorsed significantly higher confidence in their diagnostic decisions. Because the emphasis on assessment tends to differentiate psychologists from other mental health professionals, it is possible that psychologists had greater familiarity with this type of exercise and/or assessment in general. It will be interesting and important to further examine differences among mental health professionals' attitudes in future studies.

Related, it is likely that in order to better understand clinician attitudes we need to develop a more comprehensive questionnaire. Due to feasibility constraints (i.e., limited funds to compensate clinicians for their time), we kept our vignette attitudes questionnaire brief. In future studies, it will be important to consider how we can expand our questionnaire

to more fully understand clinicians' experience of case vignettes, or to incorporate interviews and other qualitative methods.

Most participants agreed that vignette characters resembled clients in their respective clinical settings. This is noteworthy because it: (1) reinforces our belief that vignette characters represented real-world cases, simulating actual practice as much as case vignette methodology can; and, (2) it suggests that cognitive de-biasing strategies will generalize to community settings (i.e., the assessment tools are not just effective in a "lab" setting).

The Cognitive De-Biasing Intervention. Overall, clinicians were very positive about the cognitive de-biasing intervention. When presented with six possible response options ranging from strongly disagree to strongly agree, participants in the treatment group unanimously provided favorable responses (e.g., *somewhat agree*, *agree*, or *strongly agree*). In particular, clinicians endorsed *learning something new*, *recommending the intervention to a colleague*, *finding information in the intervention easy to understand*, and *finding the intervention helpful for making diagnostic decisions in the case vignettes*. Perhaps most importantly from an interventionist perspective, the majority of participants in the treatment condition (91%) reported using material from the cognitive de-biasing intervention when responding to the case vignettes.

Historically, the uptake of empirically driven approaches into practice has been slow (Glasziou & Haynes, 2005; Rogers, 2003; Straus, et al., 2011). For example, the literature abounds with evidence-based treatments; however, the adoption of these approaches into applied settings often gets lost in translation (Dawes, Faust, & Meehl, 1989; Mass, 2003; Meehl, 1986). Further, several mental health initiatives recognize that effective approaches exist that are not being widely implemented in care (Ganju, 2006; Institute of Medicine,

2001, 2006; U.S. Department of Health and Human Services, 2001). Given the exploratory nature of the present study, the positive feedback from clinicians in the community and their use of intervention materials in responding to case vignettes is very encouraging. Nevertheless, it will be important to continue eliciting clinician's attitudes and expand the role of clinician feedback as we tailor the intervention and look for ways to improve its clinical utility.

It is also promising that clinicians' preferences for the use of numerical versus prose information, as measured by the Subjective Numeracy Scale (Fagerlin, et al., 2007), did not influence their attitudes toward the intervention. We predicted that preferences for the presentation of numerical and probabilistic information would lead to more positive attitudes toward the intervention. This prediction was not substantiated. Despite a notable range in clinician preferences, we found a near zero relationship between preferences for numerical preference and attitude toward the intervention. Given the Bayesian component of the cognitive de-biasing intervention, this finding is compelling because it suggests that the intervention is user-friendly and may generalize to a wide audience of mental health professionals.

Limitations

The advantages of this project must be weighed against the possible limitations. First, the external validity of studies with case vignettes can be limited because vignettes may not fully reflect complex interpersonal situations and contextual pressures present in real life diagnosis. Vignettes can also fall short in eliciting intense emotion -- which may itself increase judgment distortion. Participants also may process vignette content less carefully than under actual conditions (Abramowitz & Herrera, 1981; Heverly, Fitt, & Newman, 1984;

Kazdin, 1978; Lopez, 1989; Mikton & Grounds, 2007; Stolte, 1994). Nevertheless, case vignettes are still the most favorable method for investigating the present study aims given the exploratory nature of this research (i.e., a new intervention designed to improve clinical judgment in the assessment of PBD). Results suggest that extension to a trial using cognitive de-biasing training in the context of new, live therapy referrals would be a promising next step.

A second limitation is the restricted number of cognitive errors that the project could feasibly investigate (i.e., over *thirty* heuristics have been identified in the literature) (Croskerry, 2002). To account for this limitation, heuristics for the present study were carefully selected. For example, a thorough literature review informed the selection process. Heuristics were chosen that seemed most probable given the specific challenges associated with the clinical assessment of PBD (e.g., symptom overlap, base rate neglect). By focusing on heuristics likely to operate in more cases, it is possible to generate good coverage with a smaller number of targets – a principle often referred to as Pareto's 80:20 Rule, or the "Law of the Vital Few."

Another limitation of the study was the imbalance of study participants in the treatment and control conditions. Future investigations could benefit from an urn randomization design that would accommodate randomization as well as offer an even balance of participants across study conditions (Project MATCH Research Group, 1997). Additionally, we did not conduct any pre/post tests that may have better informed the quality of clinicians' decision-making before and after the intervention, possibly increasing internal validity. This design feature was initially contemplated; however, several practical considerations including feasibility concerns (e.g., participants' time and limited funding for

compensation) made this option unattractive. Notably, the randomization process and the experimental manipulation of key variables (e.g., race/ethnicity) serve a similar purpose. Participants' similar demographic and professional backgrounds across the two conditions support the effectiveness of the randomization.

Related, a larger number of participants in both conditions would increase power and permit additional analyses. For example, a larger sample with more clinicians from different professional backgrounds (e.g., psychiatry, clinical psychology) would allow us to examine professional title as a meaningful covariate in the regression models. Further, a larger sample size could clarify some of the trends we found (e.g., the relationships between cognitive errors and treatment formulations could change from trends to significant associations). Likewise, as previously noted, some of the analyses for our secondary aims were underpowered to detect small effect sizes. Again, a larger sample size may reveal more significant results. Taken together, increasing power could provide yet greater empirical support for the cognitive de-biasing intervention. On balance, the intervention appears to generate a variety of medium to large effects on clinical decision-making.

Further, it is important to highlight the possibility of potential design biases. The first is self-selection bias: Clinicians who participated in the study may reflect a subset of clinicians who are particularly eager to learn new assessment methods and/or interested in childhood mood disorders. One indicator of this possibility is the high number of participants (96%) reporting prior training in evidence-based practice. It is unclear whether this reflects selection bias, versus a frequently observed tendency for practitioners' self-ratings to indicate higher levels of evidence-based practice than would be discernible based on independent ratings (Shapiro, Youngstrom, Youngstrom, & Marcinick, 2012). Self-selection bias may

decrease generalizability of study findings, but it may also have resulted in diminished effect sizes (i.e., participants in the control condition may be more assessment savvy than the average clinician). We advocate for additional research to disentangle these potentially competing tendencies.

The second possible bias may have occurred via priming from the information communicated to participants prior to their responding to the case vignettes. Participants were primed for childhood mood disorders regardless of group assignment. This inevitable event may have increased participants' sensitivity to bipolar disorder in their diagnostic decisions. This sensitivity may have enhanced control condition participants' performance in vignettes such as the search satisficing and base-rate neglect vignettes. Again, the effect of this priming would be to increase the sensitivity of control participants to bipolar features, akin to enhancing the rate of placebo response in a treatment study, thus attenuating treatment effects. The fact that the de-biasing strategies continued to show statistically significant and moderate to large treatment effects attests to the potency and promise of the approach.

Despite these limitations, this project stands to significantly contribute to the mental health field, scientifically and clinically. In addition to training more than thirty clinicians in cognitive de-biasing strategies and making the intervention available to all study participants, participants responded favorably to this new approach to improving assessment of childhood mood disorders. This positive feedback from providers on the front line is encouraging, especially given the innovative nature of the intervention (i.e., adapting strategies from cognitive science and medical literatures to mental health issues) *and* the historically slow uptake of evidence-based practices in the community.

Future Directions

Despite many encouraging findings, substantial research is needed to elucidate the role of cognitive de-biasing in improving mental health decision-making. Several avenues for future research have already been highlighted; however, to summarize, it appears as though the most critical next steps involve tailoring some of the strategies (e.g., more intensive cognitive de-biasing strategies involving role-plays and simulation for bipolar I disorder), and expanding our sample of clinicians. To our knowledge, this project was the first attempt to develop a cognitive de-biasing intervention to address diagnostic error. In light of the significant findings, it seems advantageous to use this exploratory investigation as an opportunity to refine the intervention and study design (e.g., flesh out attitudes questionnaire), and to continue testing its efficacy.

Along similar lines, more research is needed to examine the effectiveness of online teaching resources for mental health practitioners as well as the best mediums for presenting information to audiences (e.g., recorded conference presentation, youtube, etc.). Fortunately, there are several "successful" models for web-based learning courses (e.g., Trauma Focused Cognitive Behavioral Therapy *Web*, Medical University of South Carolina, http://tfcbt.musc.edu/) that have started paving the way for mental health interventions. Important considerations in evaluating the effectiveness of web-based seminars also include clinicians' adoption of strategies into practices--something that was outside the scope of the current study but that would be a meaningful future study.

Finally, strategies provided in the cognitive de-biasing intervention likely generalize to other aspects of patient care, including the assessment of other mental disorders and treatment decisions. Thus, although the current study concentrated on PBD, strategies may

generalize to other aspects of care by increasing clinicians' awareness of common cognitive-based errors and teaching them new ways of thinking. Further examination of generalizability to other challenging psychiatric diagnostic and treatment decisions will be important.

Conclusion

This project used a randomized trial design to test the effects of cognitive de-biasing training on clinical decision-making about childhood mood disorders. Randomization assigned participating mental health professionals in a balanced manner in terms of professional training and experience (i.e., no evidence of randomization failure), and the training intervention produced statistically significant and medium to large effect sizes across most outcome measures. Participants rated the training as well-tolerated, as well as teaching them new content that they applied when working with the clinical vignettes. An important secondary aim of the study was experimentally manipulating the race of the clinical vignettes to test the possibility of clinical bias in the diagnosis of bipolar disorder. There was no significant difference in clinical decisions due to race. Overall, the study represents the beginning of a program of research that investigates how applying a decision-making lens to crucial diagnostic and treatment decisions can enhance clinical practice. Findings support future investigations of cognitive de-biasing interventions to improve the assessment of PBD and common comorbid conditions and, more generally, decision-making in mental health research and practice.

Table 1

Implications of faulty heuristics and biases for bipolar populations

Heuristic/Bias	Synonyms	Examples Specific to Bipolar	Practice Implication(s)
Base rate neglect	Representativeness exclusivity	Not assessing for mania in cases of depression	Underdiagnosis; Misdiagnosis; Delay in diagnosis; Suboptimal medication and treatment planning
Diagnosis Momentum	Diagnostic creep	Placing too much faith in one's early/first impressions	Delay or missed diagnosis; Unwarranted interventions, costly
Over-confidence bias	Premature closure	Thinking one knows more than one does; exaggerated certainty	Delay or missed diagnosis; Unwarranted interventions, costly
Race/ethnicity bias	N/A	Diagnosing African American youth with conduct disorder or schizophrenia instead of bipolar disorder	Misdiagnosis; Delay in diagnosis & treatment; Stigma
Search satisficing	Bounded rationality, keyhole viewing	Stopping assessment after diagnosing ADHD (or bipolar disorder)	Calling off the search once something has been found can lead to significant further findings being missed.

Note. Source adapted from Bornstein & Emler, 2001; Croskerry, 2002; Elstein & Schwartz, 2002; Galanter & Patel, 2005

Table 2

Examples of some potential symptom overlap between mania and ADHD

Mania Presentation	**ADHD Presentation**
Elevated, expansive mood	"Class clown'', attention-seeking, silly behavior in children with ADHD which may be difficult to distinguish from elation*
Irritable mood	Low frustration tolerance, temper tantrums
Increase in goal-directed activity or psychomotor agitation	Runs about, leaves seat, on the go
More talkative than usual; pressure to keep talking	Talks excessively
Distractibility	Often distracted by extraneous stimuli
Excessive involvement in activities with high potential for painful consequences	Impulsivity may lead to accidents and to engagement in potentially dangerous activities without consideration of possible consequences
Decreased need for sleep; sleeping less but no decrease in functioning	Sleep disturbance; often trouble settling at night and rising early
Inflated self esteem or grandiosity	Children with ADHD may not appreciate danger or consequences of their actions; may engage in activities beyond their abilities; may be interpreted as grandiosity*
Flight of ideas/or subjective experience of racing thoughts	Distractibility may result in problems with topic maintenance
These symptoms should represent a change in function; episodic symptoms instead of chronic presentation of illness	Symptoms should be chronic, starting before age 7

Note. *Symptoms not part of DSM criteria but common in populations with ADHD. Table adapted from Dubicka, et al. (2008).

Table 3

Cognitive de-biasing strategies

Strategy	Mechanism/Action
Develop insight/awareness	Provide detailed descriptions and thorough characterizations of known cognitive biases, together with multiple clinical examples illustrating their adverse effects on decision-making and diagnosis and formulation
Consider alternatives	Establish forced consideration of alternative possibilities, e.g., the generation and working through of a differential diagnosis. Encourage routinely asking the question: What else might this be?
Metacognition	Train for a reflective approach to problem solving: stepping back from the immediate problem to examine and reflect on the thinking process
Decrease reliance on memory	Improve the accuracy of judgments through cognitive aids: mnemonics, clinical practice guidelines, algorithms, hand-held computers
Specific training	Identify specific flaws and biases in thinking and provide directed training to overcome them: e.g., instruction in fundamental rules of probability, distinguishing correlation from causation, basic Bayesian probability theory
Simulation	Develop mental rehearsal, ‘‘cognitive walkthrough’’ strategies for specific clinical scenarios to allow cognitive biases to be made and their consequences to be observed; Construct clinical training videos contrasting incorrect (biased) approaches with the correct (de-biased) approach
Cognitive forcing strategies	Develop generic and specific strategies to avoid predictable bias in particular clinical situations
Make task easier	Provide more information about the specific problem to reduce task difficulty and ambiguity; Make available rapid access to concise, clear, well-organized information

Table 4

Actuarial estimates of likelihood of bipolar disorder for an outpatient setting

		Family History			
PGBI Test Score	**PGBI Categorical Description**	*1st degree*	*2nd degree*	*Fuzzy*	*None*
>11	*Low*	3%	2%	1%	1%
11 to 20	*Moderately Low*	13%	7%	6%	3%
21 to 30	*Neutral*	30%	18%	15%	8%
31 to 42	*Moderately High*	42%	27%	23%	13%
43 to 50	*High*	61%	44%	38%	24%
> 51	*Very High*	67%	50%	45%	29%

Note. Fuzzy family history refers to a past diagnosis of uncertain validity, or else a different diagnosis for which bipolar is often mistaken in a minority population (e.g., schizophrenia or conduct disorder) (DelBello, Lopez-Larson, Soutullo, & Strakowski, 2001; Strakowski, McElroy, Keck, & West, 1996). PGBI = Parent General Behavior Inventory. Table adapted from Youngstrom, et al. (2004).

Table 5

Components of the cognitive de-biasing intervention for clinical judgment

Strategy	Mechanism/Action	Adapted to mental health for Assessing PBD
Develop insight/awareness	Provide detailed descriptions and thorough characterizations of known cognitive biases, together with multiple clinical examples illustrating their adverse effects on decision-making and diagnosis and formulation	Education on: base-rate neglect; search satisficing; diagnosis momentum; overconfidence bias; and, race/ethnicity bias
Consider alternatives	Establish forced consideration of alternative possibilities, e.g., the generation and working through of a differential diagnosis.	Symptom checklists; Encourage routinely asking the question: What else might this be?
Metacognition	Train for a reflective approach to problem solving: stepping back from the immediate problem to examine and reflect on the thinking process	Train to safeguard thinking by examining one's own decision-making during and after assessment (before diagnosing a disorder)
Decrease reliance on memory	Improve the accuracy of judgments through cognitive aids: mnemonics, clinical practice guidelines, algorithms, hand-held computers	Mnemonics: GRAPES; FIND
Specific training	Identify specific flaws and biases in thinking and provide directed training to overcome them: e.g., instruction in fundamental rules of probability, distinguishing correlation from causation, basic Bayesian probability theory	Train on actuarial methodologies such as the nomogram, diagnostic likelihood ratios, and actuarial tables
Simulation	Develop mental rehearsal, ''cognitive walkthrough'' strategies for specific clinical scenarios to allow cognitive biases to be made and their consequences to be observed; Construct clinical training videos contrasting incorrect (biased) approaches with the correct (debiased) approach	Opportunities to practice decision-making with case information (e.g., how to apply actuarial methods with client data)

Table 5 cont. *Components of the cognitive de-biasing intervention for clinical judgment*

Strategy	Mechanism/Action	Adapted to mental health for Assessing PBD
Cognitive forcing strategies	Develop generic and specific strategies to avoid predictable bias in particular clinical situations	Provide examples of race/ethnicity bias and search satisficing errors
Make task easier	Provide more information about the specific problem to reduce task difficulty and ambiguity; Make available rapid access to concise, clear, well-organized information	Identify what pediatric bipolar disorder is (and is not); present mood graphs and highlight cognitive vulnerabilities related to heterogeneity of the disorder

Table 6

Overview of case vignettes

	Vignette			
Cognitive error	*Base-rate neglect*	*Search satisficing*	*Diagnosis momentum*	*Overconfidence*
Demographics	14-year-old female	8-year-old male	10-year-old male	11-year-old female
Symptoms	Depression	ADHD; some bipolar	Some bipolar symptoms	Classic mania
Experimental manipulation	Race/ethnicity	Race/ethnicity	Race/ethnicity	Race/ethnicity
Information gathered from participant	Probable diagnosis: Pull down menu listing options which were recoded into "inaccurate", "somewhat accurate", or "accurate"	Probable diagnosis: Pull down menu listing options which were recoded into "inaccurate", "somewhat accurate", or "accurate"	Probable diagnosis: Pull down menu listing options which were recoded into "inaccurate", "somewhat accurate", or "accurate"	Probable diagnosis: Pull down menu listing options which were recoded into "inaccurate", "somewhat accurate", or "accurate"
	Next clinical action: Family of variables including more assessment, therapy, medication, other, or no Tx	Next clinical action: Family of variables including more assessment, therapy, medication, other, or no Tx	Next clinical action: Family of variables including more assessment, therapy, medication, other, or no Tx	Next clinical action: Family of variables including more assessment, therapy, medication, other, or no Tx
	N/A	N/A	Probability of PBD: Participants rate the probability of a bipolar diagnosis from 0 to 100	N/A
	N/A	N/A	N/A	Case-study multiple test: 5 M/C questions, participants rate confidence for each answer
Analyses	Logistic and PLUM Regression to predict DME & Dx accuracy	Logistic and PLUM Regression to predict DME & Dx accuracy	Logistic and PLUM Regression to predict DME& Dx accuracy; ANCOVA; one sample *t*-test	PLUM Regression to predict DME & Dx accuracy; ANCOVA

Note. Race/ethnicity of the vignette character was Caucasian or African American. The order in which vignettes was presented to participants was random. Dx = diagnostic. Tx = treatment. DME = decision-making error. M/C = multiple choice. PLUM = Polytomous Universal. Model. ANCOVA = Analysis of Covariance

Table 7

Participant Demographics

Characteristic	**Control Group (*n* = 47) *n* (%)**	**Treatment Group (*n* = 32) *n* (%)**	**Total (*N* = 79) *N* (%)**
Gender: Female	42 (89)	26 (81)	68 (86)
Age			
< 30	17 (36)	13 (40)	30 (38)
31-45	24 (51)	14 (44)	34 (43)
> 45	6 (13)	5 (16)	11 (14)
Race/ethnicity			
Asian	3 (6)	4 (13)	7 (9)
African American	0 (0)	2 (6)	2 (3)
Caucasian	38 (81)	23 (72)	61 (77)
Latino	3 (6)	2 (6)	5 (6)
Other	2 (4)	1 (3)	3 (4)
Current Geographical Location in US			
Southern	8 (17)	6 (19)	14 (18)
Northeast	15 (32)	14 (44)	29 (37)
Western	15 (32)	10 (31)	25 (32)
Midwest	9 (19)	2 (6)	11 (14)

Note. Tx = treatment. Chi-square tests indicate that the treatment and control groups did not significantly differ on any demographic factors, $p > .05$.

Table 8

Participant Characteristics

Characteristic	**Control Group (*n* = 47) *n* (%)**	**Tx Group (*n* = 32) *n* (%)**	**Total (*N* = 79) *N* (%)**
Professional Title			
Counselor	6 (13)	4 (13)	10 (13)
Psychiatry	5 (11)	2 (6)	7 (9)
Psychology	23 (49)	17 (53)	40 (51)
Social Work	11 (23)	8 (25)	19 (24)
Other	2 (4)	1 (3)	3 (4)
Psychology Professional	24 (51)	18 (56)	42 (53)
Unlicensed, supervised professionals	15 (32)	11 (34)	26 (33)
Primary theoretical orientation*			
Cognitive Behavioral	35 (74)	19 (59)	54 (68)
Family Systems	9 (19)	7 (22)	16 (20)
Humanistic	5 (11)	0 (0)	5 (6)
Integrative	4 (9)	11 (34)	15 (19)
Psychoanalytic	4 (9)	1 (3)	5 (6)
Emotion Focused Therapy	3 (6)	4 (13)	7 (9)
Other	1 (2)	2 (6)	3 (4)
Primary clinical setting*			
Community Mental Health	11 (23)	10 (31)	21 (27)
University Setting	5 (11)	5 (16)	10 (13)
Hospital (inpatient and outpatient)	14 (30)	11 (34)	25 (32)
School	12 (26)	2 (6)	14 (18)
Private Practice	2 (4)	3 (9)	5 (6)
Residential	3 (6)	3 (9)	6 (8)
Other	4 (9)	4 (13)	8 (10)
Primary professional activity*			

Table 8 cont. *Participant Characteristics*

Clinical service	42 (89)	27 (84)	69 (87)
Research	6 (13)	3 (9)	9 (11)
Administration	0 (0)	0 (0)	0 (0)
Teaching	1 (2)	1 (3)	2 (3)
Other	1 (2)	2 (6)	3 (4)
Years of experience	9.81 (SD=7.85)	9.78 (SD=6.86)	9.8 (SD=7.24)
Trained in evidence-based practice	44 (94)	32 (100)	76 (96)
Attended Dr. Eric Youngstrom's continuing education seminar	4 (9)	2 (6)	6 (8)

Note. *Percentages may exceed 100% as a result of some participants endorsing more than one primary clinical setting, theoretical orientation, and/or professional activity. Tx = treatment. Chi-square tests indicate that the treatment and control groups did not significantly differ on any demographic factors, $p > .05$; results from a *t*-test indicate that treatment and control groups did not significantly differ in years of clinical experience, $p > .05$.

Table 9

Participants' Primary Client/Patient Population

Patient/Client Demographic	**Control Group (*n* = 47) *n* (%)**	**Tx Group (*n* = 32) *n* (%)**	**Total (*N* = 79) *N* (%)**
Primary Age Group			
Children, ages 5-10	8 (17)	4 (13)	12 (15)
Adolescents, ages 11-17	16 (34)	12 (38)	18 (23)
Pediatric all ages, 5-17	7 (15)	5 (16)	12 (15)
Families	1 (21)	3 (9)	4 (5)
Adults	15 (32)	7 (22)	22 (28)
Couples	1 (2)	1 (3)	2 (3)
Primary Diagnostic Group			
Anxiety Disorder	6 (13)	2 (6)	8 (10)
Mood Disorder	12 (26)	8 (25)	20 (25)
Oppositional Defiant and Conduct Disorders	3 (6)	2 (6)	5 (6)
Attention Deficit Hyperactivity Disorder	13 (28)	9 (28)	22 (28)
Substance Disorders	2 (4)	1 (3)	3 (4)
Schizophrenia	0 (0)	2 (6)	2 (3)
Personality Disorder	2 (4)	1 (3)	3 (4)
Adjustment Disorder	6 (13)	3 (9)	9 (11)
Trauma	5 (11)	9 (28)	14 (18)
Other	6 (13)	3 (9)	9 (11)

Note. Percentages exceed 100% as a result of some participants endorsing more than one primary age group and/or diagnostic category; Tx = treatment. Chi-square tests indicate that the treatment and control groups did not significantly differ on any demographic factors, $p > .05$.

Table 10

Performance on Case Vignettes: Summary of Decision-Making Errors

Total Number of Errors	**Control Group (*n* = 47) *n* (%)**	**Tx Group (*n* = 32) *n* (%)**	**Combined Total (*N* = 79) *N* (%)**
0	0 (0)	3 (9)	3 (4)
1	2 (4)	12 (38)	14 (18)
2	7 (15)	6 (19)	13 (17)
3	20 (44)	4(13)	24 (31)
4	17 (17)	7 (22)	24 (31)

Note. High scores indicate more decision-making errors. An independent-samples t-test comparing the mean scores of the treatment and control groups found a significant difference between the means of the two groups ($t(76) = 4.24$, $p < .0005$). The mean of the treatment group (M =2.00 , SD =1.34) than the mean of the control group (M =3.13 , SD =.83).

Table 11

Performance on Case Vignettes: Summary of Diagnostic Accuracy Ratings

Accuracy Rating	**Control Group (*n* = 47) *n* (%)**	**Tx Group (*n* = 32) *n* (%)**	**Combined Total (*N* = 79) *N* (%)**
4	0 (0)	0 (0)	(0)
5	2 (4)	0(0)	2 (3)
6	9 (20)	3 (9)	12 (15)
7	14 (30)	6 (19)	20 (26)
8	13 (28)	5 (16)	18 (23)
9	6 (13)	1 (3)	7 (9)
10	2 (4)	10 (31)	12 (15)
11	0 (0)	5 (16)	5 (6)
12	0 (0)	2 (6)	2 (3)

Note. High scores indicate more accurate diagnostic decisions. An independent-samples t-test comparing the mean scores of the treatment and control groups found a significant difference between the means of the two groups ($t(76) = -4.36$, $p < .0005$). The mean of the treatment group (M =9.00 , SD =1.83) than the mean of the control group (M =7.39 , SD =1.20).

Table 12

Predictors of Risk Estimates

Model	*B*	*t*	*p*
Group Status	**-8.79**	**-2.03**	**.04**
Years of Clinical Experience	.21	.75	.45
Professional Title	-4.38	-1.01	.31
Subjective Numeracy Scale Score	2.05	-.69	.49
Previous exposure to Bayesian approach	-1.77	-.22	.82

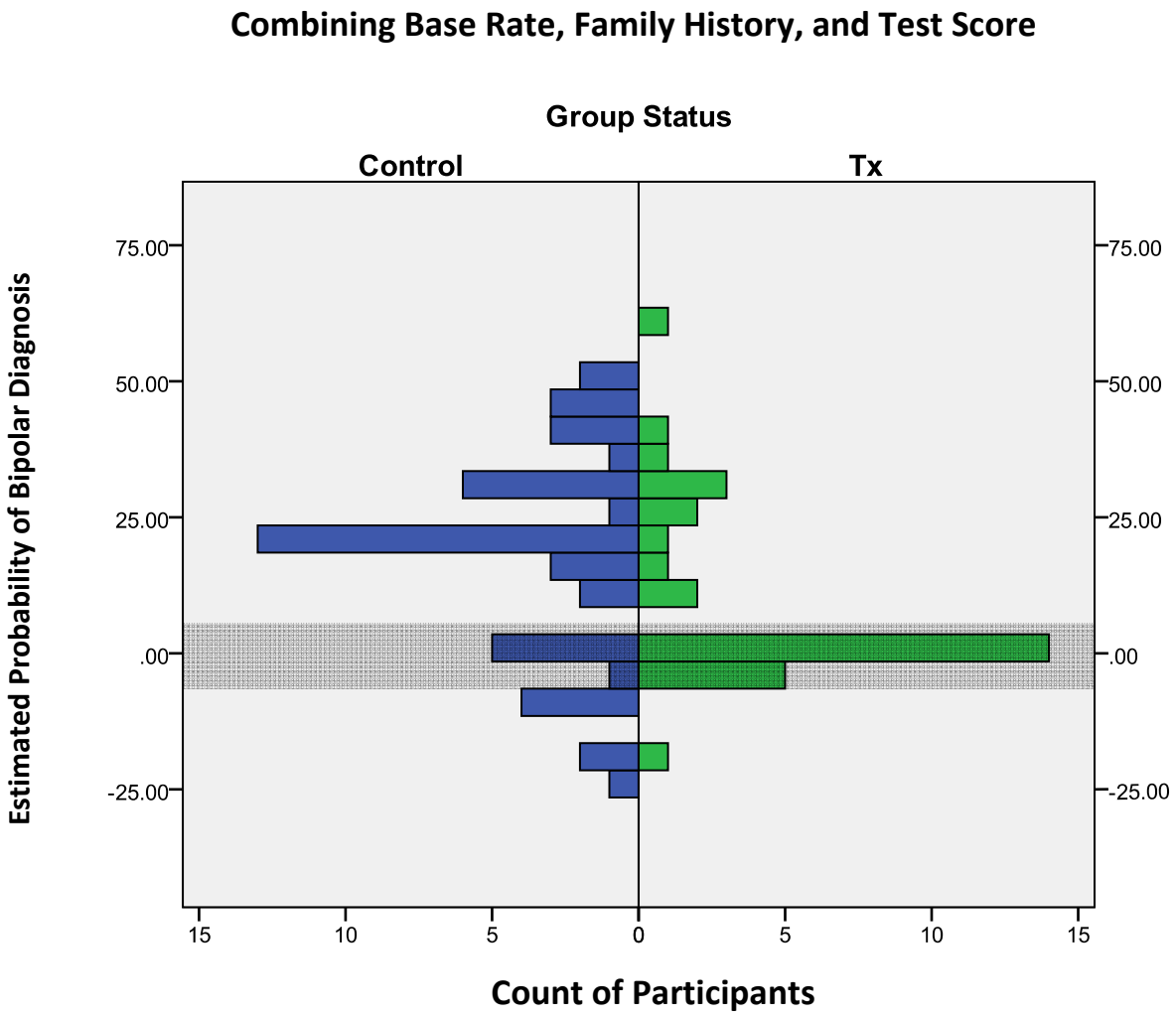
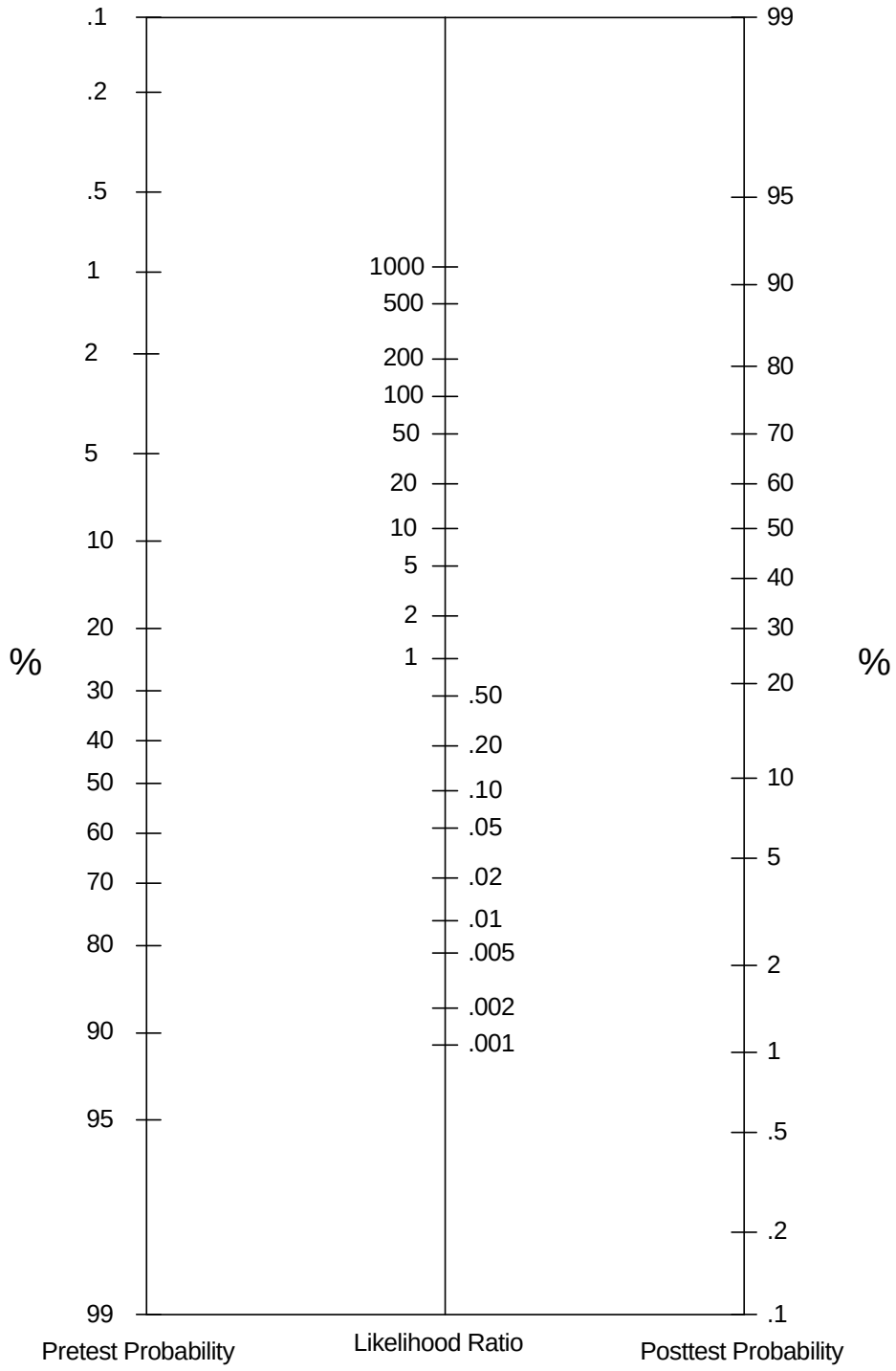


Figure 1. Comparison of Participant Risk Estimates by Control and Treatment Groups. Gray bars indicate the range within +/- 5% of the true risk estimate which was 27%. We adjusted the true risk estimate to 0% (and subtracted 27 from participants' estimates) for the purposes of the regression equation. Tx = treatment.

Appendix A

Nomogram for combining probability with diagnostic likelihood ratios



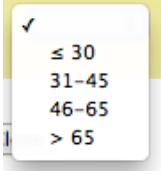
Appendix B

Demographic and background questionnaire

Please indicate your gender.

- ○ Male
- ○ Female
- ○ Transgender

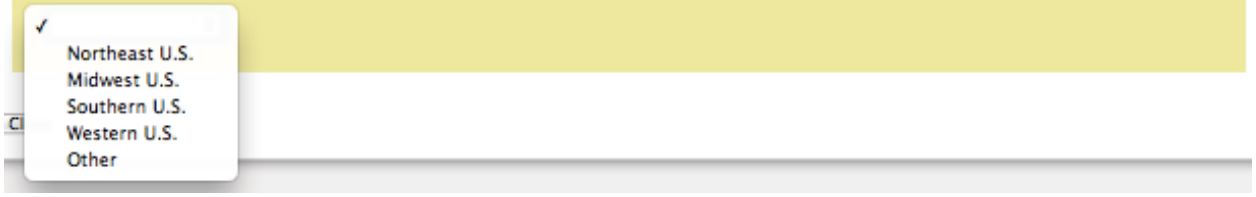
Please select the option that corresponds to your age.



Please indicate your race/ethnicity. If applicable, you may select more than one.

- ☐ Asian
- ☐ African American
- ☐ Caucasian
- ☐ Latino
- ☐ Native American
- ☐ Other, please specify:

I currently practice in the following geographical region:



Please indicate your current professional title, degree, and status by selecting from the options below.

- ○ Psychiatrist
- ○ Licensed MSW
- ○ Unlicensed MSW
- ○ Licensed Clinical Psychologist
- ○ Unlicensed Clinical Psychologist
- ○ Licensed Counseling Psychologist
- ○ Unlicensed Counseling Psychologist
- ○ Licensed Social Worker
- ○ Unlicensed Social Worker
- ○ Licensed School Psychologist
- ○ Unlicensed School Psychologist
- ○ Other Mental Health Professional. Please Specify: [ ]

Please use percentages to indicate the current division of your clinical work in different setting(s).

Community Mental Health	0
University Training Clinic	0
Hospital - Inpatient	0
Hospital - Outpatient	0
Forensic	0
Residential Facility	0
School	0
Other	0
Total	0

Please indicate your current theoretical orientation(s) by using percentages.

Behavioral	0
Cognitive	0
Cognitive Behavioral	0
Emotion Focused	0
Family/Systems	0
Humanistic	0
Integrative	0
Psychodynamic	0
Other	0
Total	0

Please indicate the division of your current case load composition by age by using percentages.

Children, 0 to 10 years old	0
Children, 11 to 18 years old	0
Adults	0
Couples	0
Family/Systems	0
Other	0
Total	0

Please indicate the division of your current professional activities by using percentages.

Clinical Service	0
Research	0
Teaching	0
Administration	0
Other	0
Total	0

Please indicate the division of your case load composition by psychiatric problems by using percentages (which may not add up to 100% due to comorbidity).

Anxiety Disorders	0
Mood Disorders	0
ODD or CD	0
ADHD	0
Substance Abuse	0
Schizophrenia	0
Personality Disorders	0
Adjustment Disorders	0
Trauma	0
Other	0
Total	0

I have experience providing mental health services to the following racial/ethnic groups (select all that apply):

- ☐ Asian
- ☐ African American
- ☐ Caucasian
- ☐ Latino
- ☐ Native American
- ☐ Other, please specify:

I have attended one of Dr. Eric Youngstrom's continuing education seminars on bipolar disorder.

Yes ☐ No ☐

I have received training in one or more evidence-based practices.

Yes ☐ No ☐

Appendix C

Brief questionnaire about participants' experience completing vignettes

For the following statements, select ONE response that best describes your experience responding to the case vignettes.

Vignette characters resemble clients that I have seen in practice.

Strongly Agree	Agree	Somewhat Agree	Somewhat Disagree	Disagree	Strongly Disagree
○	○	○	○	○	○

It was challenging to make diagnostic decisions for the vignette characters.

Strongly Agree	Agree	Somewhat Agree	Somewhat Disagree	Disagree	Strongly Disagree
○	○	○	○	○	○

I am confident in my diagnostic decisions for the majority of the case vignettes.

Strongly Agree	Agree	Somewhat Agree	Somewhat Disagree	Disagree	Strongly Disagree
○	○	○	○	○	○

I am confident in my treatment decisions and recommendations for the case vignettes.

Strongly Agree	Agree	Somewhat Agree	Somewhat Disagree	Disagree	Strongly Disagree
○	○	○	○	○	○

Apendix D

Clinician attitudes toward intervention (for participants in the treatment group only)

For the following statements, select ONE response that best describes your attitude toward the presentation.

I learned something new.

Strongly Agree	Agree	Somewhat Agree	Somewhat Disagree	Disagree	Strongly Disagree
☐	☐	☐	☐	☐	☐

I found the tutorial helpful for making diagnostic decisions in the case vignettes.

Strongly Agree	Agree	Somewhat Agree	Somewhat Disagree	Disagree	Strongly Disagree
☐	☐	☐	☐	☐	☐

I would recommend the tutorial to a colleague.

Strongly Agree	Agree	Somewhat Agree	Somewhat Disagree	Disagree	Strongly Disagree
☐	☐	☐	☐	☐	☐

Information presented in the tutorial was easy to understand.

Strongly Agree	Agree	Somewhat Agree	Somewhat Disagree	Disagree	Strongly Disagree
☐	☐	☐	☐	☐	☐

I used material from the presentation to assist me in answering questions about the vignettes.

☐ Yes

☐ No

Appendix E

Case vignettes

Vignette 1: Joey, a 7-year-old Caucasian male, is in the second grade and has been referred by his school teacher for concentration difficulties and extreme hyperactivity. Joey's teacher reports that he cannot stay in his seat and that he frequently blurts out answers in class. In addition to getting in trouble for his disruptive behavior in class, Joey's school performance is also suffering from his inattention to detail and his problems following instructions. When you meet with Joey's mother, she describes his behavior as if he is "driven by a motor" and "constantly on the go." You also learn from her that Joey has bouts of irritability, increased energy, and sleep disturbance- sometimes sleeping very little and not seeming tired the next day.

Vignette 2: Lizzy is a 14-year-old Caucasian female who has been referred to your clinic by her mother for sadness and frequent crying spells. Over the last month, Lizzy has become less interested in extracurricular activities that she used to really enjoy, including soccer and debate team. She has also experienced decreases in energy and appetite. When you meet with Lizzy, she describes feeling guilty a lot of the time and wanting to sleep more than normal. According to Lizzy and her mother, she has been experiencing these symptoms for about two and a half weeks.

Vignette 3: Michael, a Caucasian male age 10, has been referred to an outpatient community mental health center for extreme motor agitation and aggression. Michael's mother reports that Michael "cycles" frequently; she believes he has bipolar disorder. Michael's biological paternal grandfather was diagnosed with bipolar I disorder and has been treated with lithium for several years. Michael's mother completes the Parent General Behavior Inventory (PGBI), a questionnaire that asks specific questions about children's mood symptoms. Michael earns a score of 35, which is considered a moderately high score, indicating some increased probability of having bipolar disorder.

Vignette 4: Lynda is an 11-year-old Caucasian female in mainstream school. Psychological testing described week long periods of frequent impulsivity, tendencies to discuss topics unrelated to tasks she was completing, intermittent outbursts of anger and anxiety, significantly elevated levels of physical activity, difficulties sitting still, and touching everything. Over the past year, Lynda has also become very angry, irritable, destructive and capricious. She is provocative and can be cruel to pets and small children. She has been sexually inappropriate with peers and family members including "expressing interest in lewd material on the internet, 'Play Girl' magazine, and she has been hugging and kissing peers." She tells her family that she will be attending medical school, or will become a record producer, a professional wrestler or an acrobat. Overall, Lynda's behavior changes are causing her substantial problems at home and at school. Throughout this period, there have been substantial marital difficulties between the parents with resultant family stress and upheaval; however, none of Lynda's siblings have been affected to a marked extent.

Appendix F

Multiple Choice Questions for Mania Vignette

Lynda's difficulty sitting still and touching everything is most likely a sign of:

a) hyperactivity common to ADHD
b) psychomotor agitation
c) oppositional behavior
d) developmentally appropriate behavior
e) attention seeking behavior

I am ___% confident in my response.

Lynda telling her family that she will attend medical school or become a record producer, professional wrestler, or an acrobat, is most likely a sign of:

a) deceitfulness
b) typical aspirations of someone her age
c) irritability
d) grandiosity
e) delusional thinking

I am ___% confident in my response.

Lynda currently meets DSM criteria for:

a) Schizophrenia
b) Conduct Disorder
c) Bipolar I Disorder
d) Bipolar II Disorder
e) Cyclothymic Disorder

I am ___% confident in my response.

Lynda's interest in lewd material on the Internet, "Play Girl" magazine, and her hugging and kissing her peers is likely a sign of:

a) past trauma
b) poor socialization skills
c) age appropriate behavior
d) hypersexuality
e) attention seeking behavior

I am ___% confident in my response.

The change in her functioning is most likely due to:

a) asperger's or autistic traits
b) her parents' marital conflict
c) age appropriate behavior, within normal limits for her age
d) attention problems
e) mood disturbance

I am ___% confident in my response.

Appendix G

Subjective Numeracy Scale (Fagerlin, et al,, 2007)

For each of the following questions, please check the box that best reflects **how good you are at doing the following things**:

1. How good are you at working with fractions?

❑1	❑2	❑3	❑4	❑5	❑6
Not at all good					**Extremely good**

2. How good are you at working with percentages?

❑1	❑2	❑3	❑4	❑5	❑6
Not at all good					**Extremely good**

3. How good are you at calculating a 15% tip?

❑1	❑2	❑3	❑4	❑5	❑6
Not at all good					**Extremely good**

4. How good are you at figuring out how much a shirt will cost if it is 25% off?

❑1	❑2	❑3	❑4	❑5	❑6
Not at all good					**Extremely good**

For each of the following questions, please check the box that **best reflects your answer**:

5. When reading the newspaper, how **helpful** do you find tables and graphs that are parts of a story?

❑1	❑2	❑3	❑4	❑5	❑6
Not at all helpful					**Extremely helpful**

6. When people tell you the chance of something happening, do you prefer that they use **words** ("it rarely happens") or **numbers** ("there's a 1% chance")?

❑1	❑2	❑3	❑4	❑5	❑6
Always Prefer Words					**Always Prefer Numbers**

7. When you hear a weather forecast, do you prefer predictions using **percentages** (e.g., "there will be a 20% chance of rain today") or predictions using only **words** (e.g., "there is a small chance of rain today")?

❑1	❑2	❑3	❑4	❑5	❑6
Always Prefer Percentages					**Always Prefer Words**

8. How **often** do you find numerical information to be useful?

❑1	❑2	❑3	❑4	❑5	❑6
Never					**Very Often**

REFERENCES

Abramowitz, S. I., & Herrera, H. R. (1981). On controlling for patient psychopathology in naturalistic studies of sex bias: a methodological demonstration. *Journal of Consulting and Clinical Psychology, 49*(4), 597-603. doi:10.1037/0022-006X.49.4.597

Adams, J. K. (1957). A confidence scale defined in terms of expected percentages. *The American Journal of Psychology, 70*(3), 432-436. doi:10.2307/1419580

Addis, M. E., & Krasnow, A. D. (2000). A national survey of practicing psychologists' attitudes toward psychotherapy treatment manuals. *Journal of Consulting and Clinical Psychology, 68*(2), 331-339. doi:10.1037/0022-006X.68.2.331

Addis, M. E., Wade, W. A., & Hatgis, C. (1999). Barriers to dissemination of evidence-based practices: Addressing practitioners' concerns about manual-based psychotherapies. *Clinical Psychology Science and Practice, 6*(4), 430-441. doi:10.1093/clipsy/6.4.430

Aegisdottir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., et al. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34*(3), 341. doi:10.1177/0011000005285875

Akiskal, H. S., & Benazzi, F. (2005). Psychopathologic correlates of suicidal ideation in major depressive outpatients: is it all due to unrecognized (bipolar) depressive mixed states? *Psychopathology, 38*(5), 273-280. doi:10.1159/000088445

Allison, P. D. (2002). *Missing data*. Thousand Oaks: Sage.

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

American Psychological, A. (2002). Criteria for evaluating treatment guidelines. *American Psychologist, 57*, 1052-1059. doi:10.1037/0003-066X.57.12.1052

Anastasi, A., & Urbina, S. (1982). *Psychological testing*. New York.

Angold, A., Costello, E. J., & Erkanli, A. (1999). Comorbidity. *Journal of Child Psychology and Psychiatry, 40*, 57-87. doi:10.1111/1469-7610.00424

Angst, J. (2004). Bipolar disorder. *European Archives of Psychiatry and Clinical Neuroscience, 254*(2), 59-60. doi:10.1007/s00406-004-0502-5

Angst, J. (2007). The bipolar spectrum. *The British Journal of Psychiatry, 190*(3), 189. doi:10.1192/bjp.bp.106.030957

Angst, J., Sellaro, R., Stassen, H. H., & Gamma, A. (2005). Diagnostic conversion from depression to bipolar disorders: results of a long-term prospective study of hospital admissions. *Journal of Affective Disorders, 84,* 149-157. doi:10.1016/S0165-0327(03)00195-2

Arkes, H. R. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. *Journal of Consulting and Clinical Psychology, 49*(3), 323-330. doi:10.1037/0022-006X.49.3.323

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin, 110(3),* 486-498. doi:10.1037/0033-2909.110.3.486

Arkes, H. R., & Harkness, A. R. (1980). The effect of making a diagnosis on subsequent recognition of symptoms. *Journal of Experimental Psychology, 6*(5), 568-575. doi: 10.1037/0278-7393.6.5.568

Arnold, L. E., Demeter, C., Mount, K., Frazier, T. W., Youngstrom, E. A., Fristad, M., et al. Pediatric bipolar spectrum disorder and ADHD: comparison and comorbidity in the LAMS clinical sample. *Bipolar Disorders, 13*(5-6), 509-521. doi: 10.1111/j.1399-5618.2011.00948.x

Berven, N. L. (1985). Reliability and validity of standardized case management simulations. *Journal of Counseling Psychology, 32*(3), 397-409. doi:10.1037/0022-0167.32.3.397

Berven, N. L., & Scofield, M. E. (1980). Evaluation of clinical problem-solving skills through standardized case-management simulations. *Journal of Counseling Psychology, 27*(2), 199-208. doi:10.1037/0022-0167.27.2.199

Biederman, J., Faraone, S., Mick, E., Wozniak, J., Chen, L., Ouellette, C., et al. (1996). Attention-deficit hyperactivity disorder and juvenile mania: an overlooked comorbidity? *Journal of the American Academy of Child and Adolescent Psychiatry, 35*(8), 997-1008. doi:10.1097/00004583-199608000-00010

Biederman, J., Faraone, S. V., Hatch, M., Mennin, D., Taylor, A., & George, P. (1997). Conduct disorder with and without mania in a referred sample of ADHD children. *Journal of affective disorders, 44,* 177-188. doi:10.1016/S0165-0327(97)00043-8

Biederman, J., Faraone, S. V., Wozniak, J., Mick, E., Kwon, A., & Aleardi, M. (2004). Further evidence of unique developmental phenotypic correlates of pediatric bipolar disorder: findings from a large sample of clinically referred preadolescent children assessed over the last 7 years. *Journal of Affective Disorders, 82,* 45-58. doi:10.1016/j.jad.2004.05.021

Biederman, J., Klein, R. G., Pine, D. S., & Klein, D. F. (1998). Resolved: Mania is mistaken for ADHD in prepubertal children. Journal of the American Academy of Child & Adolescent *Psychiatry, 37*(10), 1091-1093. doi:10.1097/00004583-199810000-00020

Blader, J. C., & Carlson, G. A. (2007). Increased rates of bipolar disorder diagnoses among U.S. child, adolescent, and adult inpatients, 1996-2004. *Biological Psychiatry, 62*(2), 107-114. doi:10.1016/j.biopsych.2006.11.006

Borenstein, M., Hedges, L., Rothstein, H., Cohen, J., & Schoenfeld, D. (2010). Power and precision (Version 4.0). New York.

Bostwick, J. M., & Pankratz, V. S. (2000). Affective disorders and suicide risk: a reexamination.*The American Journal of Psychiatry, 157*(12), 1925-1932. doi:10.1176/appi.ajp.157.12.1925

Bowring, M. A., & Kovacs, M. (1992). Difficulties in diagnosing manic disorders among children and adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry, 31*(4), 611-614. doi:10.1097/00004583-199207000-00006

Brehmer, B. (1980). In one word: Not from experience. *Acta Psychologica, 45*(1-3), 223-241. doi:10.1016/0001-6918(80)90034-7

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

Brennan, T. A., Leape, L. L., Laird, N. M., Lawthers, A. G., Localio, A. R., Barnes, B. A., et al. (1991). Incidence of adverse events and negligence in hospitalized patients: Results of the Harvard Medical Practice Study 1. *The New England Journal of Medicine, 324*(6), 377-384. doi:10.1056/NEJM199102073240604

Brodersen, A., Licht, R. W., Vestergaard, P., Olesen, A. V., & Mortensen, P. B. (2000). Sixteen-year mortality in patients with affective disorder commenced on lithium. *The British Journal of Psychiatry, 176,* 429. doi:10.1192/bjp.176.5.429

Carlson, G. A. (1998). Mania and ADHD: Comorbidity or confusion. *Journal of Affective Disorders, 51*, 177-187. doi: 10.1016/S0165-0327(98)00179-7

Carpenter-Song, E. (2009). Caught in the psychiatric net: Meanings and experiences of ADHD, pediatric bipolar disorder and mental health treatment among a diverse group of families in the United States. *Culture, medicine and psychiatry, 33*(1), 61-85. doi:10.1007/s11013-008-9120-4

Carpenter-Song, E., Whitley, R., Lawson, W., Quimby, E., & Drake, R. E. (2011). Reducing disparities in mental health care: suggestions from The Dartmouth “Howard Collaboration. *Community mental health journal, 47*(1), 1-13. doi: 10.1007/s10597-009-9233-4

Casas, J. M., Brady, S., & Ponterotto, J. G. (1983). Sexual preference biases in counseling: An information processing approach. *Journal of Counseling Psychology, 30*(2), 139-145. doi:10.1037/0022-0167.30.2.139

Chorpita, B. F., Yim, L. M., & Tracey, S. A. (2002). Feasibility of a simplified and dynamic Bayesian system for use in structured diagnostic interviews. *Journal of Psychopathology and Behavioral Assessment, 24*(1), 13-23. doi:10.1023/A:1014097007097

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum.

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

Croskerry, P. (2002). Achieving quality in clinical decision making: cognitive strategies and detection of bias. *Academic Emergency Medicine, 9*(11), 1184-1204. doi:10.1111/j.1553-2712.2002.tb01574.x

Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine, 78*(8), 775. doi:10.1097/00001888-200308000-00003

Crowley, S. D., Owens, T. A., Schardt, C. M., Wardell, S. I., Peterson, J., Garrison, S., et al. (2003). A Web-based compendium of clinical questions and medical evidence to educate internal medicine residents. *Academic Medicine, 78*(3), 270-274. doi:10.1097/00001888-200303000-00007

Currin, L., Schmidt, U., & Waller, G. (2007). Variables that influence diagnosis and treatment of the eating disorders within primary care settings: A Vignette study. *International Journal of Eating Disorders, 40*(3), 257-262. doi:10.1002/eat.20355

Dailey, R. (1952). The effects of premature conclusions upon the acquisition of understanding about a person. *Journal of Psychology, 33*, 113-152. doi: 10.1080/00223980.1952.9712825

Davidow, J., & Levinson, E. M. (1993). Heuristic principles and cognitive bias in decision making: Implications for assessment in school psychology. *Psychology in the Schools, 30*(4), 351-361. doi:10.1002/1520-6807(199310)30:4<351::AID-PITS2310300410>3.0.CO;2-X

Davis, W. E., & Jones, M. H. (1973). Race-related variations in psychiatric diagnosis. *Newsletter for Research in Mental Health & Behavioral Sciences, 15*(4), 31-32.

Dawes, R. M. (1996). *House of cards: Psychology and psychotherapy built on myth.* Free Press.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668-1674. doi:10.1126/science.2648573

DelBello, M. P., Lopez-Larson, M. P., Soutullo, C. A., & Strakowski, S. M. (2001). Effects of race on psychiatric diagnosis of hospitalized adolescents: A retrospective chart review. *Journal of Child and Adolescent Psychopharmacology, 11*(1), 95-103. doi:10.1089/104454601750143528

Dubicka, B., Carlson, G. A., Vail, A., & Harrington, R. (2008). Prepubertal mania: diagnostic differences between US and UK clinicians. *European Child & Adolescent Psychiatry, 17*(3), 153-161. doi:10.1007/s00787-007-0649-5

Dunner, D. L. (2003). Clinical consequences of under-recognized bipolar spectrum disorder. *Bipolar Disorders, 5*(6), 456-463. doi:10.1046/j.1399-5618.2003.00073.x

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. *Judgment under uncertainty: Heuristics and biases*, 249-267.

Einhorn, H. J. (1980). Overconfidence in judgment. *New Directions for Methodology of Social and Behavioral Science, 4*(1), 1–16.

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review, 85*(5), 395-416. doi:10.1037/0033-295X.85.5.395

Elstein, A. S., & Schwartz, A. (2002). Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *British Medical Journal, 324*(7339), 729-732. doi: 10.1136/bmj.324.7339.729

Ely, J. W., Graber, M. L., & Croskerry, P. (2011). Checklists to reduce diagnostic errors. *Academic Medicine, 86*(3), 307. doi: 10.1097/ACM.0b013e31820824cd

Epperson, D. L., Bushway, D. J., & Warman, R. E. (1983). Client self-terminations after one counseling session: Failure of counselors to recognize client's problem definitions. *Journal of Counseling Psychology, 30*, 307-315. doi:10.1037/0022-0167.30.3.307

Ezpeleta, L., de la Osa, N., Domenech, J. M., Navarro, J. B., Losilla, J. M., & Judez, J. (1997). Diagnostic agreement between clinicians and the Diagnostic Interview for Children and Adolescents--DICA-R--in an outpatient sample. *Journal of Child Psychology and Psychiatry, 38*(4), 431-440. doi:10.1111/j.1469-7610.1997.tb01528.x

Fabrega, H., Ulrich, R., & Mezzich, J. E. (1993). Do Caucasian and Black adolescents differ at psychiatric intake? *Journal of the American Academy of Child & Adolescent Psychiatry, 32*(2), 407-413. doi:10.1097/00004583-199303000-00023

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the subjective numeracy scale. *Medical Decision Making, 27*(5), 672-680. doi:10.1177/0272989X07304449

Falvey, J. E., & Hebert, D. J. (1992). Psychometric study of the Clinical Treatment Planning Simulations (CTPS) for assessing clinical judgment. *Journal of Mental Health Counseling, 14*, 490-490.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160. doi: 10.3758/BRM.41.4.1149

Faust, D. (1986). Research on human judgment and its application to clinical practice. *Professional Psychology: Research and Practice, 17*(5), 420-430. doi:10.1037/0735-7028.17.5.420

Faust, D. (1994). *Are there sufficient foundations for mental health experts to testify in court*? No. Controversial issues in mental health, 196-201.

Faust, D., Guilmette, T. J., Hart, K., Arkes, H. R., Fishburne, F. J., & Davey, L. (1988). Neuropsychologists' training, experience, and judgment accuracy. *Archives of Clinical Neuropsychology, 3*(2), 145-163. doi: 10.1016/0887-6177(88)90060-1

Faust, D., & Ziskin, J. (1988). The expert witness in psychology and psychiatry. *Science, 241*(4861), 31-35. doi:10.1126/science.3291114

Fischhoff, B. (1982). Debiasing. *Judgment under uncertainty: Heuristics and biases*, 422–444.

Fischoff, B. (1977). Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance, 3*(2), 349-358. doi:10.1037/0096-1523.3.2.349

Flaskerud, J. H., & Hu, L.-t. (1992). Relationship of ethnicity to psychiatric diagnosis. *Journal of Nervous and Mental Disease, 180*(5), 296-303. doi:10.1097/00005053-199205000-00003

Fletcher, J. M., Francis, D. J., Morris, R. D., & Lyon, G. R. (2005). Evidence-based assessment of learning disabilities in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 34*(3), 506-522. doi:10.1207/s15374424jccp3403_7

Friedlander, M. L., & Phillips, S. D. (1984). Preventing anchoring errors in clinical judgment. *Journal of Consulting and Clinical Psychology*, 52(3), 366-371. doi:10.1037/0022-006X.52.3.366

Friedlander, M. L., & Stockman, S. J. (1983). Anchoring and publicity effects in clinical judgment. *Journal of Clinical Psychology, 39*(4), 637-643. doi:10.1002/1097-4679(198307)39:4<637::AID-JCLP2270390433>3.0.CO;2-Q

Galanter, C. A., & Leibenluft, E. (2008). Frontiers between attention deficit hyperactivity disorder and bipolar disorder. *Child and Adolescent Psychiatric Clinics of North America, 17*(2), 325-346. doi:10.1016/j.chc.2007.11.001

Galanter, C. A., & Patel, V. L. (2005). Medical decision making: a selective review for child psychiatrists and psychologists. *Journal of Child Psychology and Psychiatry, 46*(7), 675-689. doi:10.1111/j.1469-7610.2005.01452.x

Ganju, V. (2006). Mental health quality and accountability: the role of evidence-based practices and performance measurement. *Administration and Policy in Mental Health, 33*(6), 659-665. doi:10.1007/s10488-006-0071-1

Garb, H. N. (1986). The appropriateness of confidence ratings in clinical judgment. *Journal of Clinical Psychology, 42*(1), 190-197. doi:10.1002/1097-4679(198601)42:1<190::AID-JCLP2270420133>3.0.CO;2-6

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin, 105*(3), 387-396. doi:10.1037/0033-2909.105.3.387

Garb, H. N. (1994). Social and clinical judgment: Fraught with error? *American Psychologist, 49*(8), 758-759. doi:10.1037/0003-066X.49.8.758.b

Garland, A. F., Plemmons, D., & Koontz, L. (2006). Research-Practice Partnership in Mental Health: Lessons from Participants. *Administration and Policy in Mental Health and Mental Health Services Research, 33*(5), 517-528. doi:10.1007/s10488-006-0062-2

Gauron, E. F., & Dickinson, J. K. (1966). Diagnostic decision making in psychiatry. I. Information usage. *Archives of General Psychiatry, 14*(3), 225-232. doi: 10.1001/archpsyc.1966.01730090001001

Geller, B., Craney, J. L., Bolhofner, K., Nickelsburg, M. J., Williams, M., & Zimerman, B. (2002). Two-Year Prospective Follow-Up of Children With a Prepubertal and Early Adolescent Bipolar Disorder Phenotype. *American Journal of Psychiatry, 159*, 927-933. doi:10.1176/appi.ajp.159.6.927

Geller, B., Tillman, R., Craney, J. L., & Bolhofner, K. (2004). Four-year prospective outcome and natural history of mania in children with a prepubertal and early adolescent bipolar disorder phenotype. *Archives of General Psychiatry 61*(5), 459-467. doi:10.1001/archpsyc.61.5.459

Geller, B., Zimerman, B., Williams, M., Bolhofner, K., Craney, J. L., DelBello, M. P., et al. (2000). Diagnostic characteristics of 93 cases of prepubertal and early adolescent bipolar disorder phenotype by gender, puberty and comorbid attention deficit hyperactivity disorder. *Journal of Child and Adolescent Psychopharmacology, 10*(3), 157-164. doi:10.1089/10445460050167269

Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgments. *American Psychologist, 23*(7), 483-496. doi:10.1037/h0026206

Goldsmith, L., & Schloss, P. J. (1986). Diagnostic Overshadowing among School Psychologists Working with Hearing-Impaired Learners. *American Annals of the Deaf, 131*(4), 288-293.

Gordon, R., & Franklin, N. (2003). Cognitive underpinnings of diagnostic error. *Academic Medicine, 78*(8), 782. doi:10.1097/00001888-200308000-00005

Graber, M. M. D., Gordon, R. M. A., & Franklin, N. P. (2002). Reducing Diagnostic Errors in Medicine: What's the Goal? *Academic Medicine, 77*(10), 981-992. doi:10.1097/00001888-200210000-00009

Grant, B. F., Stinson, F. S., Hasin, D. S., Dawson, D. A., Chou, S. P., Ruan, W. J., et al. (2005). Prevalence, correlates, and comorbidity of bipolar I disorder and axis I and II disorders: results from the National Epidemiologic Survey on Alcohol and Related Conditions. *The Journal of Clinical Psychiatry, 66*(10), 1205-1215. doi:10.4088/JCP.v66n1001

Grasso, B. C., Rothschild, J. M., Genest, R., & Bates, D. W. (2003). What do we know about medication errors in inpatient psychiatry? *Joint Commission Journal on Quality and Safety, 29*(8), 391-400.

Gross, H. S., Herbert, M. R., Knatterud, G. L., & Donner, L. (1969). The effect of race and sex on the variation of diagnosis and disposition in a psychiatric emergency room. The *Journal of Nervous and Mental Disease, 148*(6), 638-642. doi:10.1097/00005053-196906000-00007

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*(1), 19-30. doi:10.1037/1040-3590.12.1.19

Guyatt, G. H., & Rennie, D. (Eds.). (2002). *Users' guides to the medical literature*. Chicago: AMA Press.

Guze, S. B., & Robins, E. (1970). Suicide and primary affective disorders. *The British Journal of Psychiatry, 117*(539), 437-438. doi: 10.1192/bjp.117.539.437

Harris, E. C., & Barraclough, B. (1997). Suicide as an outcome for mental disorders. A meta-analysis. *The British Journal of Psychiatry, 170*(3), 205. doi:10.1192/bjp.170.3.205

Hasin, D. S., Goodwin, R. D., Stinson, F. S., & Grant, B. F. (2005). Epidemiology of major depressive disorder: results from the National Epidemiologic Survey on Alcoholism and Related Conditions. *Archives of General Psychiatry, 62*(10), 1097. doi:10.1001/archpsyc.62.10.1097

Hastie, R., & Dawes, R. M. (2001). *Rational Decision in an uncertainty world: The psychology of judgment and decision making*. London: Sage Publications.

Haynes, B., Glasziou, P., & Straus, S. (2000). Advances in evidence-based information resources for clinical practice. *ACP Journal Club, 132*(1), A11-14.

Herbert, D. L., Nelson, R. O., & Herbert, J. D. (1988). Effects of psychodiagnostic labels, depression severity, and instructions on assessment. *Professional Psychology: Research and Practice, 19*(5), 496-502. doi:10.1037/0735-7028.19.5.496

Heverly, M. A., Fitt, D. X., & Newman, F. L. (1984). Constructing case vignettes for evaluating clinical judgment: an empirical model. *Evaluation and Program Planning, 7*, 45-55. doi:10.1016/0149-7189(84)90024-7

Hirsch, P. A., & Stone, G. L. (1983). Cognitive strategies and the client conceptualization process. *Journal of Counseling Psychology, 30*(4), 566-572. doi:10.1037/0022-0167.30.4.566

Hirschfeld, R. M., Lewis, L., & Vornik, L. A. (2003). Perceptions and impact of bipolar disorder: how far have we really come? Results of the national depressive and manic-depressive association 2000 survey of individuals with bipolar disorder. *Journal of Clinical Psychiatry, 64*(2), 161-174. doi:10.4088/JCP.v64n0209

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). New York: Wiley. doi:10.1002/0471722146

Hsieh, F. Y., Bloch, D. A., & Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in medicine, 17*(14), 1623-1634. doi: 10.1002/(SICI)1097-0258(19980730)17:14<1623::AID-SIM871>3.0.CO;2-S

Institute of Medicine. (2001). *Crossing the quality chasm: A new health system for the 21st century*. Washington, D.C.: National Academy Press.

Institute of Medicine. (2006). *Crossing the quality chasm: A new health system for the 21st century*. Washington, D.C.: National Academy Press.

Jaeschke, R., Guyatt, G. H., & Sackett, D. L. (1994). Users' guides to the medical literature: III. How to use an article about a diagnostic test: A. Are the results of the study valid? *The Journal of the American Medical Association, 271*(5), 389-391. doi: 10.1001/jama.1994.03510290071040

Jenkins, M.M., Youngstrom, E.A., & Youngstrom, J.K. (2012). Generalizability of evidence-based assessment recommendations for pediatric bipolar disorder. *Psychological Assessment, 24*(2), 269-281. doi: 10.1037/a0025775

Jenkins, M. M., Youngstrom, E. A., Washburn, J. J., & Youngstrom, J. K. (2011). Evidence-based strategies improve assessment of pediatric bipolar disorder by community practitioners. *Professional Psychology: Research and Practice, 42*(2), 121. doi: 10.1037/a0022506

Jonas, B. S., Brody, D., Roper, M., & Narrow, W. E. (2003). Prevalence of mood disorders in a national sample of young American adults. *Social Psychiatry and Psychiatric Epidemiology, 38*(11), 618-624. doi:10.1007/s00127-003-0682-8

Joseph, M. F., Youngstrom, E. A., & Soares, J. C. (2009). Antidepressant-coincident mania in children and adolescents treated with selective serotonin reuptake inhibitors. *Future Neurology, 4*(1): 87–102. doi: 10.2217/14796708.4.1.87

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237-251. doi:10.1037/h0034747

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., et al. (1997). Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime version (K-SADS-PL): Initial reliability and validity data. *Journal of the American Academy of Child & Adolescent Psychiatry, 36*(7), 980-988. doi:10.1097/00004583-199707000-00021

Kazdin, A. E. (1978). Evaluating the generality of findings in analogue therapy research. *Journal of Consulting and Clinical Psychology, 46*(4), 673-686. doi:10.1037/0022-006X.46.4.673

Kessler, R. C. (1999). Comorbidity of unipolar and bipolar depression with other psychiatric disorders in a general population survey. In M. Tohen (Ed.), *Comorbidity in Affective Disorders* (pp. 1-25). New York: Marcel Dekker, Inc.

Kessler, R. C., Adler, L., Barkley, R., Biederman, J., Conners, C. K., Demler, O., et al. (2006). The prevalence and correlates of adult ADHD in the United States: results from the National Comorbidity Survey Replication. *The American Journal of Psychiatry, 163*(4), 716-723. doi:10.1176/appi.ajp.163.4.716

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry, 62*(6), 593. doi:10.1001/archpsyc.62.6.593

Kessler, R. C., Rubinow, D. R., Holmes, C., Abelson, J. M., & Zhao, S. (1997). The epidemiology of DSM-III-R bipolar I disorder in a general population survey. *Psychological Medicine, 27*(5), 1079-1089. doi:10.1017/S0033291797005333

Kilgus, M. D., Pumariega, A. J., & Cuffe, S. P. (1995). Influence of race on diagnosis in adolescent psychiatric inpatients. *Journal of the American Academy of Child & Adolescent Psychiatry, 34*(1), 67-72. doi:10.1097/00004583-199501000-00016

Kirov, G., & Murray, R. M. (1999). Ethnic differences in the presentation of bipolar affective disorder. *European Psychiatry, 14*(4), 199-204. doi:10.1016/S0924-9338(99)80742-1

Klein, J. G. (2005). Five pitfalls in decisions about diagnosis and prescribing. *British Medical Journal, 330*(7494), 781-783. doi: 10.1136/bmj.330.7494.781

Kluger, J., & Song, S. (2002). Young and bipolar. Time(August 19), 39-47, 51.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 107-118. doi:10.1037/0278-7393.6.2.107

Kowatch, R. A., Fristad, M. A., Birmaher, B., Wagner, K. D., Findling, R. L., & Hellander, M. (2005). Treatment guidelines for children and adolescents with bipolar disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 44*(3), 213-235. doi:10.1097/00004583-200503000-00006

Kowatch, R. A., Youngstrom, E. A., Danielyan, A., & Findling, R. L. (2005). Review and meta-analysis of the phenomenology and clinical characteristics of mania in children and adolescents. *Bipolar Disorders, 7*, 483-496. doi: 10.1111/j.1399-5618.2005.00261.x

Kraemer, H. C. (1992). *Evaluating medical tests: Objective and quantitative guidelines.* Newbury Park, CA: Sage Publications.

Kupfer, D. J. (2005). The increasing medical burden in bipolar disorder. *The Journal of the American Medical Association, 293*(20), 2528-2530. doi:10.1001/jama.293.20.2528

Kurpius, D. W. J., Benjamin, D., & Morran, D. K. (1985). Effects of teaching a cognitive strategy on counselor trainee internal dialogue and clinical hypothesis formulation. *Journal of Counseling Psychology, 32*(2), 263-271. doi:10.1037/0022-0167.32.2.263

Kurtz, R. M., & Garfield, S. L. (1978). Illusory correlation: A further exploration of Chapman's paradigm. *Journal of Consulting and Clinical Psychology, 46*(5), 1009-1015. doi:10.1037/0022-006X.46.5.1009

Lambert, L. E., & Wertheimer, M. (1988). Is diagnostic ability related to relevant training and experience? Professional *Psychology: Research and Practice, 19*(1), 50-52. doi:10.1037/0735-7028.19.1.50

Lau, A. Y., & Coiera, E. (2007). How do clinicians search for and access biomedical literature to answer clinical questions? *Studies in Health Technology and Informatics, 129*(Pt 1), 152-156.

Lawson, W. B., Hepler, N., Holladay, J., & Cuffel, B. (1994). Race as a factor in inpatient and outpatient admissions and diagnosis. *Hospital & Community Psychiatry, 45*(1), 72-74.

Lehman, A. F., & Steinwachs, D. M. (1998). Patterns of usual care for schizophrenia: initial results from the Schizophrenia Patient Outcomes Research Team (PORT) Client Survey. *Schizophrenia Bulletin, 24*(1), 11-20; discussion 20-32. doi: 10.1093/oxfordjournals.schbul.a033303

Leibenluft, E., Charney, D. S., Towbin, K. E., Bhangoo, R. K., & Pine, D. S. (2003). Defining clinical phenotypes of juvenile mania. *The American Journal of Psychiatry, 160*, 430-437. doi:10.1176/appi.ajp.160.3.430

Levant, R. F. (2005). Report of the 2005 Presidential Task Force on Evidence Based Practice. Washington: American Psychological Association.

Lewczyk, C. M., Garland, A. F., Hurlburt, M. S., Gearity, J., & Hough, R. L. (2003). Comparing DISC-IV and clinician diagnoses among youths receiving public mental health services. *Journal of the American Academy of Child & Adolescent Psychiatry, 42*(3), 349-356. doi:10.1097/00004583-200303000-00016

Lewinsohn, P. M., Klein, D. N., & Seeley, J. R. (1995). Bipolar disorders in a community sample of older adolescents: Prevalence, phenomenology, comorbidity, and course. *Journal of the American Academy of Child & Adolescent Psychiatry, 34*(4), 454-463. doi:10.1097/00004583-199504000-00012

Lewinsohn, P. M., Seeley, J. R., Buckley, M. E., & Klein, D. N. (2002). Bipolar disorder in adolescence and young adulthood. *Child and Adolescent Psychiatric Clinics of North America, 11*(3), 461-475, vii. doi:10.1016/S1056-4993(02)00005-6

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21*(1), 37-44. doi:10.1177/0272989X0102100105

Lish, J. D., Dime-Meenan, S., Whybrow, P. C., Price, R. A., & Hirschfeld, R. M. (1994). The National Depressive and Manic-depressive Association (DMDA) survey of bipolar members. *Journal of Affective Disorders, 31*(4), 281-294. doi:10.1016/0165-0327(94)90104-X

Lopez, S. R. (1989). Patient variable biases in clinical judgment: conceptual overview and metodological considerations. *Psychological Bulletin, 10*6(2), 184-203. doi:10.1037/0033-2909.106.2.184

Lueger, R. J. (2002). Practice-informed research and research-informed psychotherapy. *Journal of Clinical Psychology, 58*(10), 1265-1276. doi:10.1002/jclp.10110

Lueger, R. J., & Petzel, T. P. (1979). Illusory Correlation in Clinical Judgment: Effects of Amount of Information to Be Processed. *Journal of Consulting and Clinical Psychology, 47*(6), 1120-1121. doi:10.1037/0022-006X.47.6.1120

Lutz, W., Lambert, M. J., Harmon, S. C., Tschitsaz, A., Schurch, E., & Stulz, N. (2006). The probability of treatment success, failure and duration-what can be learned from empirical data to support decision making in clinical practice? *Clinical Psychology & Psychotherapy, 13*(4), 223-232. doi:10.1002/cpp.496

Marchand, W. R., Wirth, L., & Simon, C. (2006). Delayed diagnosis of pediatric bipolar disorder in a community mental health setting. *Journal of Psychiatric Practice, 12*(2), 128-133. doi:10.1097/00131746-200603000-00011

Marquez, C., Taintor, Z., & Schwartz, M. A. (1985). Diagnosis of manic depressive illness in blacks. *Comprehensive Psychiatry, 26*(4), 337-341. doi:10.1016/0010-440X(85)90088-4

Mass, J. (2003). Evidence-based assessment of children with behavioral and emotional disorders. *Report of Emotional and Behavioral Disorders in Youth, 3*, 31-34.

McClellan, J. M., Werry, J. S., & Ham, M. (1993). A follow-up study of early onset psychosis: comparison between outcome diagnoses of schizophrenia, mood disorders, and personality disorders. *Journal of Autism and Development Disorders, 23*(2), 243-262. doi:10.1007/BF01046218

McNeil, B. J., Pauker, S. G., Sox, J. R., & Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *New England Journal of Medicine, 306*(21). doi:10.1056/NEJM198205273062103

Meehl, P. (1960). The cognitive activity of the clinician. *American Psychologist, 15*(1), 19-27. doi:10.1037/h0041744

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press. doi:10.1037/11281-000

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment, 50*(3), 370-375. doi: 10.1207/s15327752jpa5003_6

Merikangas, K. R., Akiskal, H. S., Angst, J., Greenberg, P. E., Hirschfeld, R., Petukhova, M., et al. (2007). Lifetime and 12-month prevalence of bipolar spectrum disorder in the National Comorbidity Survey replication. *Archives of General Psychiatry, 64*(5), 543. doi:10.1001/archpsyc.64.5.543

Merikangas, K. R., & Pato, M. (2009). Recent developments in the epidemiology of bipolar disorder in adults and children: Magnitude, correlates, and future directions. *Clinical Psychology: Science and Practice, 16*(2), 121-133. doi:10.1111/j.1468-2850.2009.01152.x

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

Mikton, C., & Grounds, A. (2007). Cross-cultural clinical judgment bias in personality disorder diagnosis by forensic psychiatrists in the UK: a case-vignette study. *Journal of Personality Disorders, 21*(4), 400-417. doi:10.1521/pedi.2007.21.4.400

Mischel, W. (1996). *Personality and assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Moreno, C., Laje, G., Blanco, C., Jiang, H., Schmidt, A. B., & Olfson, M. (2007). National trends in the outpatient diagnosis and treatment of bipolar disorder in youth. *Archives of General Psychiatry, 64*(9), 1032-1039. doi:10.1001/archpsyc.64.9.1032

Mukherjee, S., Shukla, S., Woodle, J., Rosen, A. M., & Olarte, S. (1983). Misdiagnosis of schizophrenia in bipolar patients: a multiethnic comparison. *The American Journal of Psychiatry, 140*(12), 1571-1574.

Murray, C. J. L., Lopez, A. D., & Mundial, B. (1996). *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*. Harvard University Press Cambridge, MA.

Nath, S. B., & Marcus, S. C. (2006). Medical errors in psychiatry. *Harvard Review of Psychiatry, 14*(4), 204-211. doi:10.1080/10673220600889272

Neighbors, H. W., Jackson, J. S., Campbell, L., & Williams, D. (1989). The influence of racial factors on psychiatric diagnosis: A review and suggestions for research. Community *Mental Health Journal, 25*(4), 301-311. doi:10.1007/BF00755677

Neisworth, J. T., & Bagnato, S. J. (2004). The Mismeasure of Young Children: The Authentic Assessment Alternative. *Infants and Young Children, 17*(3), 198-212. doi:10.1097/00001163-200407000-00002

Nierenberg, A. A., Ostacher, M. J., Calabrese, J. R., Ketter, T. A., Marangell, L. B., Miklowitz, D. J., et al. (2006). Treatment-resistant bipolar depression: a STEP-BD equipoise randomized effectiveness trial of antidepressant augmentation with lamotrigine, inositol, or risperidone. *American Journal of Psychiatry, 163*, 210-216. doi:10.1176/appi.ajp.163.2.210

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

O'Donohue, W., & Szymanski, J. (1994). How to win friends and not influence clients: Popular but problematic ideas that impair treatment decisions. *Behaviour Therapist, 17*, 29-29.

Ohio Hospital Association, & St. Paul (MN) Insurance Company. (1998). Data from the U.S. General Accounting Office Retrieved 07/04/2010

Olfson, M., Marcus, S. C., & Shaffer, D. (2006). Antidepressant drug therapy and suicide in severely depressed children and adults: A case-control study. *Archives of General Psychiatry, 63*(8), 865-872. doi:10.1001/archpsyc.63.8.865

Oskamp, S. (1965). Overconfidence in Case-Study Judgments. *Journal of Consulting Psychology, 29*, 261-265. doi: 10.1037/h0022125

Papolos, D. F., & Papolos, J. (2006). *The bipolar child: the definitive and reassuring guide to childhood's most misunderstood disorder*. New York : Random House, Inc.

Pfeiffer, A. M., Whelan, J. P., & Martin, J. M. (2000). Decision-making bias in psychotherapy: Effects of hypothesis source and accountability. *Journal of Counseling Psychology, 47*(4), 429-436. doi:10.1037/0022-0167.47.4.429

Piacentini, J., Shaffer, D., Fisher, P., Schwab-Stone, M., Davies, M., & Gioia, P. (1993). The Diagnostic Interview Schedule for Children-Revised Version (DISC-R): III. Concurrent criterion validity. *Journal of the American Academy of Child & Adolescent Psychiatry, 32*(3), 658-665. doi:10.1097/00004583-199305000-00025

Project MATCH Research Group. (1997). Matching Alcoholism Treatments to Client Heterogeneity: Project MATCH Posttreatment Drinking Outcomes. *Journal of Studies on Alcohol, 58*, 7-29.

Quinn, C. A., & Fristad, M. A. (2004). Defining and identifying early onset bipolar spectrum disorder. *Current Psychiatry Reports, 6*(2), 101-107. doi:10.1007/s11920-004-0049-1

Redelmeier, D. A. (2005). Improving patient care. The cognitive psychology of missed diagnoses. *Annals of Internal Medicine, 142*(2), 115-120.

Redelmeier, D. A., Ferris, L. E., Tu, J. V., Hux, J. E., & Schull, M. J. (2001). Problems for clinical judgment: introducing cognitive psychology as one more basic science. *Canadian Medical Association Journal, 164*(3), 358.

Regehr, G., & Norman, G. R. (1996). Issues in cognitive psychology: implications for professional education. *Academic Medicine, 71*(9), 988. doi:10.1097/00001888-199609000-00015

Reimherr, J. P., & McClellan, J. M. (2004). Diagnostic challenges in children and adolescents with psychotic disorders. *Journal of Clinical Psychiatry, 65*, 5-11.

Rettew, D. C., Lynch, A. D., Achenbach, T. M., Dumenci, L., & Ivanova, M. Y. (2009). Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *International Journal of Methods in Psychiatric Research, 18*(3), 169-184. doi:10.1002/mpr.289

Richards, M. S., & Wierzbicki, M. (1990). Anchoring errors in clinical-like judgments. *Journal of Clinical Psychology, 46*(3), 358-365. doi:10.1002/1097-4679(199005)46:3<358::AID-JCLP2270460317>3.0.CO;2-7

Ridley, C. R., Li, L. C., & Hill, C. L. (1998). Multicultural Assessment: Reexamination, Reconceptualization, and Practical Application. *The Counseling Psychologist, 26*(6), 827. doi:10.1177/0011000098266001

Rosenhan, D. L., & Heiner, R. (2008). On being sane in insane places: Deviance across cultures. (pp. 235-242). New York, NY: Oxford University Press.

Sackett, D. L., Rosenberg, W., Muir Gray, J. A., Haynes, R. B., & Richardson, W. S. (1996). Evidence-based medicine: what it is and what it isn't. *British Medical Journal, 312*, 71-72. doi: 10.1136/bmj.312.7023.71

Sandifer, M. G., Hordern, A., & Green, L. M. (1970). The psychiatric interview: The impact of the first three minutes. *American Journal of Psychiatry, 126*(7), 968-973.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology General, 130*(3), 380-400. doi:10.1037/0096-3445.130.3.380

Shapiro, J., Youngstrom, J., Youngstrom, E., & Marcinick, H. (2012). Transporting a Manualized Treatment for Children's Disruptive Behavior to a Community Clinic. *Journal of Contemporary Psychotherapy*, 1-11. doi: 10.1007/s10879-012-9206-

Sharma, R., & Markar, H. R. (1994). Mortality in affective disorder. *Journal of Affective Disorders, 31*(2), 91-96. doi:10.1016/0165-0327(94)90112-0

Singh, M. K., DelBello, M. P., Kowatch, R. A., & Strakowski, S. M. (2006). Co-occurrence of bipolar and attention-deficit hyperactivity disorders in children. *Bipolar Disorders, 8*(6), 710-720. doi:10.1111/j.1399-5618.2006.00391.x

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology, 28*, 1-39. doi:10.1146/annurev.ps.28.020177.000245

Snyder, M., & Swan, W. (1978). Hypothesis testing in social interaction. *Journal of Personality and Social Psychology, 36*, 1202-1212. doi:10.1037/0022-3514.36.11.1202

Soutullo, C. A., Chang, K. D., Diez-Suarez, A., Figueroa-Quintana, A., Escamilla-Canales, I., Rapado-Castro, M., et al. (2005). Bipolar disorder in children and adolescents: international perspective on epidemiology and phenomenology. *Bipolar Disorders, 7*(6), 497-506. doi:10.1111/j.1399-5618.2005.00262.x

Spengler, P. M., & Strohmer, D. C. (1994). Clinical judgmental biases: The moderating roles of counselor cognitive complexity and counselor client preferences. *Journal of Counseling Psychology, 41*(1), 8-17. doi:10.1037/0022-0167.41.1.8

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

Spengler, P. M., White, M. J., Aegisdottir, S. A., Maugherman, A. S., Anderson, L. A., Cook, R. S., et al. (2009). The meta-analysis of clinical judgment project: Effects of experience on judgment accuracy. *The Counseling Psychologist, 37*(3), 350-399. doi:10.1177/0011000006295149

Stang, P. E., Frank, C., Kalsekar, A., Yood, M. U., Wells, K., & Burch, S. (2006). The clinical history and costs associated with delayed diagnosis of bipolar disorder. *Medscape General Medicine, 8*(2), 18.

Stolte, J. F. (1994). The context of satisficing in vignette research. *The Journal of Social Psychology, 134*(6), 727-733. doi: 10.1080/00224545.1994.9923007

Strakowski, S. M., Flaum, M., Amador, X., Bracha, H. S., Strakowski, S. M., Pandurangi, A. K., et al. (1996). Racial differences in the diagnosis of psychosis. *Schizophrenia Research, 21*(2), 117-124. doi:10.1016/0920-9964(96)00041-2

Strakowski, S. M., Keck, P. E., Jr., Arnold, L. M., Collins, J., Wilson, R. M., Fleck, D. E., et al. (2003). Ethnicity and diagnosis in patients with affective disorders. *The Journal of Clinical Psychiatry, 64*(7), 747-754. doi:10.4088/JCP.v64n0702

Strakowski, S. M., Lonczak, H. S., Sax, K. W., West, S. A., Crist, A., Mehta, R., et al. (1995). The effects of race on diagnosis and disposition from a psychiatric emergency service. *Journal of Clinical Psychiatry, 56*(3), 101-107.

Strakowski, S. M., McElroy, S. L., Keck, P. E., Jr., & West, S. A. (1996). Racial influence on diagnosis in psychotic mania. *Journal of Affective Disorders, 39*(2), 157-162. doi:10.1016/0165-0327(96)00028-6

Strakowski, S. M., Shelton, R. C., & Kolbrener, M. L. (1993). The effects of race and comorbidity on clinical diagnosis in patients with psychosis. *Journal of Clinical Psychiatry, 54*(3), 96-102.

Straus, S. E., Glasziou, P., Richardson, W. S., & Haynes, R. B. (2011). *Evidence-based medicine: How to practice and teach EBM* (4th ed.). New York: Churchill Livingstone.

Strohmer, D. C., & Leierer, S. J. (2000). Modeling rehabilitation counselor clinical judgment. *Rehabilitation Counseling Bulletin, 44*(1), 3. doi:10.1177/003435520004400102

Strohmer, D. C., Shivy, V. A., & Chiodo, A. L. (1990). Information processing strategies in counselor hypothesis testing: The role of selective memory and expectancy. *Journal of Counseling Psychology, 37*(4), 465-472. doi:10.1037/0022-0167.37.4.465

Swets, J. A., Dawes, R. M., & Monahan, J. L. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*, 1-26. doi:10.1111/1529-1006.001

Thomas, E. J., Studdert, D. M., Burstin, H. R., Orav, E. J., Zeena, T., Williams, E. J., et al. (2000). Incidence and types of adverse events and negligent care in Utah and Colorado. *Medical Care, 38*(3), 261. doi:10.1097/00005650-200003000-00003

Tondo, L., Isacsson, G., & Baldessarini, R. J. (2003). Suicidal behaviour in bipolar disorder: Risk and prevention. *CNS Drugs, 17*(7), 491-511. doi:10.2165/00023210-200317070-00003

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131. doi:10.1126/science.185.4157.1124

U.S. Department of Health, Health State Department. (1999). *A report of the surgeon general*. In P. S. Wang, O. Demler & R. C. Kessler (Eds.). Washington D.C.

U.S. Department of Health, Health State Department. (2001). *A report of the surgeon general*. Rockville, MD.

U.S. Food and Drug Administration, FDA Consumer Health Information (2004). *Understanding antidepressant medications.* Retrieved from http://www.fda.gov/downloads/ForConsumers/ConsumerUpdates/ucm095990.pdf

Van Meter, A. R., Moreira, A. L. R., & Youngstrom, E. A. (2011). Meta-analysis of epidemiologic studies of pediatric bipolar disorder. *The Journal of clinical psychiatry, 72*(9), 1250-1256. doi:10.4088/JCP.10m06290

Vitiello, B., Malone, R., Buschle, P. R., Delaney, M. A., & Behar, D. (1990). Reliability of DSM-III diagnoses of hospitalized children. *Hospital Community Psychiatry, 41*(1), 63-67.

Vitiello, B., & Stoff, D. M. (1997). Subtypes of aggression and their relevance to child psychiatry. *Journal of the American Academy of Child & Adolescent Psychiatry, 36*(3), 307-315. doi:10.1097/00004583-199703000-00008

Waddington, L., & Morley, S. (2000). Availability bias in clinical formulation: the first idea that comes to mind. *British Journal of Medical Psychology, 73* (Pt 1), 117-127. doi:10.1348/000711200160345

Wampold, B. E., Casas, J. M., & Atkinson, D. R. (1981). Ethnic bias in counseling: An information processing approach. *Journal of Counseling Psychology, 28*(6), 498-503. doi:10.1037/0022-0167.28.6.498

Wang, P. S., Demler, O., & Kessler, R. C. (2002). Adequacy of treatment for serious mental illness in the United States. *American Journal of Public Health, 92*(1), 92-98. doi:10.2105/AJPH.92.1.92

Ward, W. C., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology, 19*(3), 231-241.doi:10.1037/h0082908

Wedding, D. (1991). Clinical judgment in forensic neuropsychology: A comment on the risks of claiming more than can be delivered. *Neuropsychology Review, 2*(3), 233-239. doi:10.1007/BF01109046

Weller, E. B., Danielyan, A. K., & Weller, R. A. (2004). Somatic treatment of bipolar disorder in children and adolescents. *Psychiatric Clinics of North America, 27*(1), 155-178. doi:10.1016/S0193-953X(03)00116-3

West, J. C., Wilk, J. E., Olfson, M., Rae, D. S., Marcus, S., Narrow, W. E., et al. (2005). Patterns and quality of treatment for patients with schizophrenia in routine psychiatric practice. *Psychiatric Services, 56*(3), 283-291. doi:10.1176/appi.ps.56.3.283

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

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

Wilens, T. E., Biederman, J., Kwon, A., Chase, R., Greenberg, L., Mick, E., et al. (2003). A systematic chart review of the nature of psychiatric adverse events in children and adolescents treated with selective serotonin reuptake inhibitors. *Journal of Child and Adolescent Psychopharmacology, 13*(2), 143-152. doi:10.1089/104454603322163862

Wilson, R. M., Runciman, W. B., Gibberd, R. W., Harrison, B. T., Newby, L., & Hamilton, J. D. (1995). The quality in Australian health care study. *The Medical Journal of Australia, 163*(458), 71.

Wisch, A. F., & Mahalik, J. R. (1999). Male therapists' clinical bias: Influence of client gender roles and therapist gender role conflict. *Journal of Counseling Psychology, 46*(1), 51-60. doi:10.1037/0022-0167.46.1.51

Wood, G. (1978). The knew-it-all-along effect. *Journal of Experimental Psychology: Human Perception and Performance, 4*(2), 345-353. doi:10.1037/0096-1523.4.2.345

Youngstrom, E. A., Birmaher, B., & Findling, R. L. (2008). Pediatric bipolar disorder: validity, phenomenology, and recommendations for diagnosis. *Bipolar Disorders, 10*, 194-214. doi:10.1111/j.1399-5618.2007.00563.x

Youngstrom, E. A., & Duax, J. (2005). Evidence based assessment of pediatric bipolar disorder, part 1: Base rate and family history. *Journal of the American Academy of Child & Adolescent Psychiatry, 44*, 712-717. doi:10.1097/01.chi.0000162581.87710.bd

Youngstrom, E. A., Findling, R. L., Youngstrom, J. K., & Calabrese, J. R. (2005). Toward an evidence-based assessment of pediatric bipolar disorder. *Journal of Clinical Child and Adolescent Psychology, 34*(3), 433-448. doi:10.1207/s15374424jccp3403_4

Youngstrom, E. A., Freeman, A. J., & Jenkins, M. M. (2009). The assessment of children and adolescents with bipolar disorder. *Child and Adolescent Psychiatric Clinics of North America, 18*(2), 353-390, viii-ix. doi:10.1016/j.chc.2008.12.002

Youngstrom, E. A., Youngstrom, J. K., & Starr, M. (2005). Bipolar Diagnoses in Community Mental Health: Achenbach CBCL Profiles and Patterns of Comorbidity. *Biological Psychiatry, 58*, 569-575. doi:10.1016/j.biopsych.2005.04.004

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