Journal of Counseling Psychology
2015, Vol. 62, No. 4, 553–567

0022-0167/15/$12.00 http://dx.doi.org/10.1037/cou0000105

A Meta-Analysis of Confidence and Judgment Accuracy in Clinical Decision Making

Deborah J. Miller, Elliot S. Spengler, and Paul M. Spengler
Ball State University

The overconfidence bias occurs when clinicians overestimate the accuracy of their clinical judgments. This bias is thought to be robust leading to an almost universal recommendation by clinical judgment scholars for clinicians to temper their confidence in clinical decision making. An extension of the Meta-Analysis of Clinical Judgment (Spengler et al., 2009) project, the authors synthesized over 40 years of research from 36 studies, from 1970 to 2011, in which the confidence ratings of 1,485 clinicians were assessed in relation to the accuracy of their judgments about mental health (e.g., diagnostic decision making, violence risk assessment, prediction of treatment failure) or psychological issues (e.g., personality assessment). Using a random effects model a small but statistically significant effect ($r = .15$; CI = .06, .24) was found showing that confidence is better calibrated with accuracy than previously assumed. Approximately 50% of the total variance between studies was due to heterogeneity and not to chance. Mixed effects and meta-regression moderator analyses revealed that confidence is calibrated with accuracy least when there are repeated judgments, and more when there are higher base rate problems, when decisions are made with written materials, and for earlier published studies. Sensitivity analyses indicate a bias toward publishing smaller sample studies with smaller or negative confidence-accuracy effects. Implications for clinical judgment research and for counseling psychology training and practice are discussed.

Keywords: assessment, overconfidence bias, clinical judgment, meta-analysis

> Given that judgment is fallible, our beliefs about our own judgmental ability are likely to be seriously in error.
>
> —Hillel J. Einhorn

Counseling psychologists are responsible for making optimal judgments about client problems, course of treatment, and future behavior among several other decisions (Spengler, Strohmer, Dixon, & Shivy, 1995). They may also be rewarded for displaying a high degree of confidence in their decision-making accuracy as confidence may lend them increased credibility with clients and colleagues. Price and Stone (2004) called this phenomenon the *confidence heuristic*, whereby people prefer decision-makers "who make extreme confidence judgments" (p. 39). Across many applications extreme confidence has been associated with higher credibility, whether it is children's perceptions of adults (Tenney, Small, Kondrad, Jaswal, & Spellman, 2011), eyewitness' accounts of a crime (Semmler, Brewer, & Douglass, 2012), or jurors' perceptions of psychologists as expert witnesses (Cramer, DeCoster, Harris, Fletcher, & Brodsky, 2011). But does confidence equate with accuracy in the realm of psychological assessment?

Conventional wisdom might lead one to believe that counseling and other psychologists who display more confidence in their judgments are also more accurate; however, behavior decision-making theory and research appears to provide support for an *overconfidence bias*, which occurs when individuals report higher confidence in their judgments than is warranted by their actual accuracy (Fischhoff, Slovic, & Lichtenstein, 1977; Kahneman & Tversky, 1973; Koriat, Lichtenstein, & Fischhoff, 1980; Nisbett & Ross, 1980). Overconfidence studies typically assess the relation between judgment confidence and judgment accuracy. For example, in one such study (Carlin & Hewitt, 1990) psychologists made judgments about random versus true responding from psychological test data. Confidence ratings were high whereas accuracy was close to chance resulting in a correlation between confidence and accuracy of −.22. By contrast some studies show a positive association between confidence and judgment accuracy. Leli and Filskov (1984) provided neuropsychologists with test data from actual patients whose diagnoses had been verified through medical tests, surgery, or brain autopsy and were to describe if the patient was unimpaired, or as having diffuse or right-side–left-side lateralized impairment. The correlation between confidence and hit rates (accuracy) was .28. Leli and Filskov's study reflects confidence that is moderately calibrated with accuracy, whereas Carlin and Hewitt's study reflects the commonly hypothesized overconfidence effect resulting in a negative correlation.

Before addressing further the role of confidence in clinical judgment it may be helpful to briefly describe how clinical judg-

This article was published Online First August 17, 2015.

Deborah J. Miller, Elliot S. Spengler, and Paul M. Spengler, Department of Counseling Psychology and Guidance Services, Ball State University.

This study was funded in part by Ball State University internal grants to Deborah J. Miller, Elliot S. Spengler, and Paul M. Spengler. The authors worked equally on this project. Portions of this study were presented in August 2014 at the annual meeting of the American Psychological Association in Washington, DC.

Correspondence concerning this article should be addressed to Deborah J. Miller, Department of Counseling Psychology and Guidance Services, Ball State University, Muncie, IN 47306. E-mail: debmillerphd@gmail.com

ment researchers establish judgment accuracy. Many studies of neuropsychologists' decision making compare their judgments with the results of brain autopsies or other solid evidence as a clear criterion for measuring judgment accuracy. Other measures for establishing decision-making accuracy are also used. Desmarais, Nicholls, Read, and Brink (2010) assessed forensic mental health professionals' confidence related to their predictions of future violence. Follow-up data a year later based on review of legal and psychiatric records determined a 44% recidivism rate. The assessors were overall confident in their predictions although confidence was statistically unrelated to predictive accuracy. Additional methods for establishing a criterion for judgment accuracy include direct behavioral observations, standardized interviews, and objective measures. For example, Haderlie (2011) found that therapists were generally overconfident in their ability to predict client progress from session-to-session when compared with clients' responses to an objective measure (Outcome Questionnaire-45; Lambert et al., 1996). Other studies may not directly measure accuracy but their findings imply disproportionate confidence by clinicians in their intuitive powers (Lilienfeld, Lynn, & Lohr, 2015). A recent sample of 129 clinicians in practice found that the average clinician believed he or she performed at the 80th percentile in terms of client outcomes and a full 25% self-assessed as being at the 90th percentile (Walfish, McAlister, O'Donnell, & Lambert, 2012). Not a single clinician self-rated as being below the 50th percentile. Clearly such flawed self-assessment reflects overestimation of one's professional abilities.

Overconfidence is usually thought to occur with more difficult judgment tasks, whereas underconfidence is another form of flawed self-assessment thought to occur with easier tasks (Lichtenstein & Fischhoff, 1977). Confidence in clinical decision making has also been hypothesized as a positive quality (Glidewell & Livert, 1992). In actuality there are four different possible relations (see Figure 1). Researchers are not always clear in their description of which of the four quadrants they are testing. Boyle (2000) for example hypothesized a negative correlation, which could represent quadrants c (underconfidence) and b (overconfidence). Out of 36 mental health clinical judgment studies, we found the majority of authors conceptualized the relation between confidence and accuracy as prone to overconfidence. Garb (1986) in his review referred to the appropriateness of confidence in relation to accuracy. Because problems with appropriate calibration of confidence ratings can be due to overconfidence or to underconfidence, we use the more inclusive term of *confidence bias*.

The confidence bias has been found to generalize to several types of judgments routinely made by counseling psychologists. For instance, in studies of confirmation bias, or the tendency to seek confirming rather than disconfirming evidence of a clinical hypothesis, judges who were more confident were also more likely to recall confirmatory evidence that supported their hypothesis (Koriat et al., 1980). In an accumulated body of research, counseling psychologists and counselors have been found to be prone to overconfidence associated with confirmatory hypothesis testing strategies (e.g., Martin, 2001; Owen, 2008; Strohmer, Shivy & Chiodo, 1990). Nickerson (1998) noted, "Perhaps the confirmation bias should be thought of as the tendency to seek evidence that increases one's confidence in a hypothesis whether it should or not" (p. 186). Kahneman and Tversky (1973) discussed the "illusion of validity" as occurring when "people are prone to experience much confidence even in their highly fallible judgments" (p. 249), especially when given consistent or extremely consistent information.

Clinical judgment scholars frequently discuss how confidence circumvents the use of a scientific process of assessment (Spengler et al., 1995). Kahneman (2011), in his book *Thinking, Fast and Slow*, noted that confidence is highly associated with intuitive thought processes and less with scientific methods of assessment; he aptly observed "Sustaining doubt is harder work than sliding into certainty" (p. 114). Several examples exist where clinicians inaccurately believe they can form more accurate predictions using clinical (intuitive) compared with statistical (scientific) methods of prediction (see Spengler, 2013), sometimes leading to dramatic reductions in accuracy to no better than chance for such important issues as prediction of violence and recidivism (Harris, Rice, Quinsey, & Cormier, 2015). Misplaced confidence in clinical decision making may be a significant threat to the scientist-practitioner model of training, supervision and practice resulting in

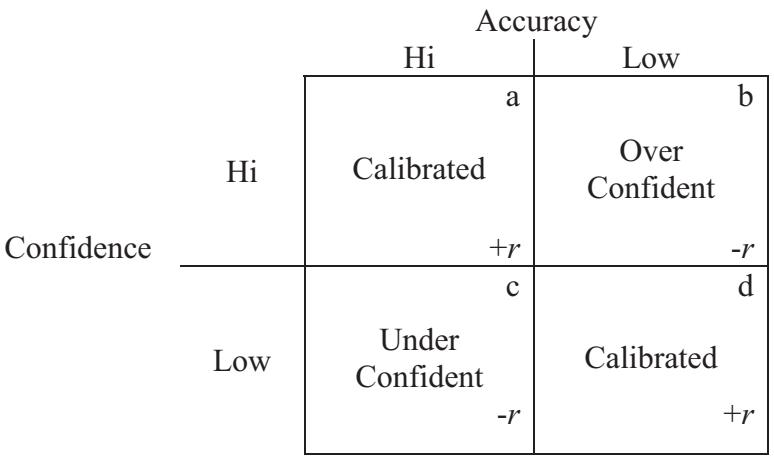

Figure 1. Possible relations between judgment confidence and decision-making accuracy.

practitioners adhering to what Lilienfeld et al. (2015) have termed *pseudoscientific* assessment and treatment methods.

The relation between clinicians' confidence and mental health clinical judgment accuracy has been studied for over 60 years. From the outset, researchers uncovered evidence that the most confident judges were oftentimes among the least accurate (Goldberg, 1959; Holsopple & Phelan, 1954; Oskamp, 1965). In Oskamp's (1965) seminal study, psychologists were asked to make clinical judgments about a fictional case with subsequent assessments made as more information was provided in four separate stages. Psychologists reported their burgeoning confidence at each stage of additional information. Confidence and accuracy were poorly aligned from the beginning, but as judges received more case information their confidence increased (a shift from 33% to 53% confidence) whereas accuracy levels remained poor (from stage one 26% to stage four 28% accuracy).[1] Oskamp concluded, ". . . So-called clinical validation, based on the personal feelings of confidence of a clinician, is not adequate evidence for the validity (accuracy) of clinical judgment in diagnosing or predicting human behavior" (p. 265).

Results from subsequent studies on the confidence bias led to the general belief by scholars that counseling and other psychologists are unable to make appropriate confidence ratings (Arkes, 1981; Smith & Dumont, 2002). Garb (1986) reviewed 18 studies on the relation of mental health professionals' confidence to the accuracy of their judgments. Though he did not use statistical methods to aggregate the data, he concluded there was actually mixed empirical support for the confidence bias. Garb commented on a possible relation between accurate confidence ratings and two factors: when the stimulus materials used had high validity and when the judges were more experienced. Oskamp (1965) also concluded that more experienced clinicians appropriately moderated their confidence by reporting greater tentativeness in their assessment of accuracy. By contrast, Smith and Dumont (2002) asserted that overconfidence is more ubiquitous because clinical judgment is fallible as an underlying cause for overconfidence. Dunning, Heath, and Suls (2004) concluded in a comprehensive review that laypersons and professionals, across a wide variety of occupations including health care providers, researchers, educators, lawyers and public policy decision-makers, are inherently prone to overestimating the accuracy of their judgments with few specific exceptions.

To assess the magnitude and variability of the confidence bias by mental health professionals we sought to synthesize clinical judgment research in which confidence and mental health-related judgment accuracy have been studied. This study, which is an extension of the larger Meta-Analysis of Clinical Judgment (MACJ) project (Spengler, White, Ægisdóttir, Maugherman, Anderson, et al., 2009), used meta-analytic techniques to synthesize findings about clinician confidence and mental health clinical judgment accuracy. Meta-analysis is a statistical technique whereby data are pooled from individual studies and an analysis of these analyses is conducted, thus the term *meta*-analysis. Meta-analyses on the confidence bias outside the field of mental health suggest that confidence produces weak to modest correlations with accuracy: correlations from various meta-analyses (a) for eyewitness accounts range from .08 to .20s (Sporer, Penrod, Read & Cutler, 1995), (b) for expert law enforcement and forensic professionals' ability to detect deception equal .05 (Aamodt & Custer, 2006), and (c) for athletes' assessment of their performance equal .24 (Woodman & Hardy, 2003). This possible poor confidence-accuracy calibration reminds us of Garrison Keillor's (2014) reference to the mythical town of Lake Wobegon, "Where the women are strong, the men are good looking, and the children are all above average" (p. xxviii). In short, there may be a pervasive human tendency to overestimate one's personal capabilities (see Dunning et al., 2004) and this tendency may be no different for counseling and other psychologists' assessment of their clinical decision-making skills.

Although the confidence bias has been extensively researched in various realms of human decision making, and is thought to be a core impingement upon accurate mental health decision making, there has been no empirical synthesis of this phenomenon in the mental health field. In response to this absence, and in light of the considerable emphasis by clinical judgment scholars on the confidence bias, we sought to synthesize this area of mental health judgment research and to test relevant moderators suggested in the literature. Moderators are essentially interaction effects assessed in meta-analytic studies. Consistent with other clinical judgment meta-analyses (Spengler & Pilipis, 2015) we sought to assess two classes of moderators: (a) conceptual/theoretical and (b) methodological. The conceptual moderators are rationally selected on the basis of our reading of the confidence-accuracy research and behavior decision-making theory (Kahneman, Slovic & Tversky, 1982; Nisbett & Ross, 1980). The methodological moderators are ones commonly assessed in meta-analyses (e.g., study quality; Cooper, Hedges & Valentine, 2009). As this is one of only a few existing meta-analyses of clinical judgment research (see Ægisdóttir et al., 2006; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Spengler et al., 2009; Spengler & Pilipis, 2015; White et al., 1995) we believe it is relevant to test an array of potential moderators. We also sought to implement a sensitivity analysis strategy (Greenhouse & Iyengar, 2009) by examining such factors as the impact of the quality of the data on the overall analysis (zero effects, outliers included or excluded) and commonly considered threats to validity in meta-analyses (publication bias, file drawer analysis, funnel plot). A sensitivity analysis in short determines if the findings are robust to design and analysis decisions used in the meta-analysis.

Several factors have been studied which may affect the relation between confidence and mental health decision-making accuracy, including whether concurrent or predictive judgments are made. Because of the well-known decline in predictive validity compared with concurrent validity (Cronbach & Meehl, 1955) there might be a higher confidence-accuracy relation for judges making concurrent decisions. Researchers have examined the accuracy of unassisted clinical judgment compared with mechanical methods, such as statistical and actuarial formulas, with mechanical prediction being slightly superior to unassisted clinical judgment (d = .12; Ægisdóttir et al., 2006; Grove et al., 2000). Clinicians provided access to mechanical judgment aids, such as statistical and actuarial formulas, test score cut-offs, and other statistical assistance of their choice may have better calibrated confidence in their accuracy, presuming they would use these tools (Meehl, 1957). Schol-

[1] In Oskamp's (1965) study 20% accuracy represented accuracy at a chance level.

ars have found that confidence increases with the amount of information a judge is given but that accuracy generally does not (Oskamp, 1965; Ryback, 1967). The perceived incremental validity of additional information has also been found to influence judges' confidence in their decisions (Koriat et al., 1980). In this regard, we assessed the impact of when judges were given freedom to engage in assessments using their preferred methods with the availability of additional information.

A key variable proposed by Spengler and Pilipis (2015) is that improvement in decision making is contingent upon receiving feedback. We sought to assess both the impact of training and of feedback on the confidence bias. One might assume that training in a task in close proximity to the moment when the judgment is made would calibrate the confidence-accuracy relation (see Ericsson, Krampe, & Tesch-Römer, 1993). Most confidence-accuracy studies involve repeated judgments made by clinicians. Based on behavior decision-making theory (Nisbett & Ross, 1980) and Oskamp's (1965) original findings, of increasing confidence without increases in accuracy, we anticipated that repeated judgments would fatigue judges and negatively affect the confidence-accuracy relation. Nisbett and Ross (1980) proposed that knowledge of base rates could enhance decision-making accuracy and appropriate calibration of confidence. Even without base rate information more frequent occurring behaviors should be easier to assess and associated with better confidence calibrations. Similarly, confidence may be better calibrated when criteria are highly valid (Cronbach & Meehl, 1955).

One of the most important potential moderating variables may be clinician experience. Garb (1986) speculated in his narrative review that more experienced clinicians acquire better confidence calibration. By contrast, others have suggested that more experienced clinicians become more confident but actually no more accurate (Faust & Faust, 2012). Still other researchers have reported more experience with a specific judgment task actually fortifies overconfidence for developmental levels ranging from preschoolers (Lipko, Dunlosky, & Merriman, 2009) to experienced bankers (Lambert, Bessière, & N′Goala, 2012). Garb (1986) also proposed that the validity of the stimulus material would moderate the confidence-accuracy effect. Lastly, publication bias may occur in the most competitive journals, such as those published by the American Psychological Association (APA), because reviewers prefer studies that report statistically significant results (Rothstein, Sutton, & Borenstein, 2005).

In accordance with behavior decision making theory, where clinicians are thought to be prone to unconsciously invoke judgment heuristics or cognitive shortcuts, and overconfidence findings across other domains of decision making (Dunning et al., 2004), we hypothesized for the overall effect that confidence would be poorly aligned with accuracy for a variety of judgment tasks. If confidence is poorly aligned this would result in a negative effect or a null effect where the 95% CI crosses zero. The considered importance of the confidence bias is highlighted by recent APA (2015) ethical guidelines for clinical supervision; supervisors are presumed to "overestimate their competence and grow in confidence about their abilities, even though it is not necessarily matched by corresponding increases in ability" (p. 39) and therefore cautioned to collect structured supervisee feedback. For counseling and other psychologists, miscalibrated confidence may lead to hasty conclusions, rigid treatment plans, and failure to correct errors, resulting in potentially negative consequences for clients (Smith & Dumont, 1997, 2002). Our hope is that the results of this meta-analysis inform research and education on clinical judgment and enhance practitioners' assessment of their accuracy resulting in better client assessments.

Method

We combined published and unpublished studies from our search of 1997 to 2011 with confidence-accuracy studies archived in a database from the comprehensive MACJ (Spengler et al., 2009). The MACJ project exhaustively screened and coded over 30,000 studies from 1970 to 1996 related to mental health clinical decision making, and from that time period likely identified every possible study on confidence and clinical judgment accuracy. This database contained 46 confidence-accuracy studies where confidence was treated as a predictor and judgments made by mental health professionals were assessed. An additional 83 clinical judgment studies in the archive treated confidence solely as a criterion (e.g., Hirsch & Stone, 1983) and were not analyzed in the present study. To extend the MACJ database we applied the same strenuous search process to identify confidence-accuracy research from 1997 to 2011. Standardized MACJ definitions of key variables (e.g., judgment accuracy) and moderators (e.g., criterion validity) were used paired with moderator codes specifically identified as having potential relevance to confidence bias research.[2]

Coders and Training

The three authors were the coders for this study. Training procedures used in the MACJ project were used to train coders for study characteristics and statistics. The authors reached agreement on definitions by coding practice articles during six 2-hr training sessions. Practice coding occurred on less challenging, and then more challenging, studies until 90% agreement was reached across each step of coding.

Study Search and Inclusion

The first and second authors searched for published and unpublished research on the confidence-accuracy effect from 1997 to 2011. The authors reviewed each identified study to reach consensus on inclusion or exclusion. The same MACJ inclusion criteria and 207 search terms were used to search electronic databases, namely PsycINFO, ERIC, Dissertation Abstracts, MEDLINE, and Social Science Index, crossed with variations on the terms confidence, overconfidence, underconfidence, certainty and calibration. We also used forward-search (studies citing) and backward-search strategies (references) for the 46 MACJ confidence-accuracy clinical judgment studies and for studies found from 1997 to 2011. This strategy identified 106 mental health clinical judgment studies from 1970 to 2011 that assessed confidence as a predictor in relation to mental health clinical decision making. To be included, each study had to assess some type of mental health judgment (e.g., diagnosis, violence risk assessment, prediction of treatment) and establish what constituted an accurate judgment. Studies were excluded when *all* of the participants were undergraduates or

[2] Training manuals and methods materials are available upon request.

nonmental health professionals (e.g., physicians, nurses, law enforcement). To be accepted, a study had to include professionals or graduate students in mental health fields, such as counseling (mental health, school, rehabilitation, and pastoral), psychiatry, psychiatric nursing, social work or psychology. Included studies had to provide statistics necessary to calculate an effect size. If a study did not report statistics the authors were contacted (e.g., Regehr, Bogo, Shlonsky, & LeBlanc, 2010). From this pool of 106 studies a total of 34 met inclusion criteria. These 34 studies produced 36 effects because two publications reported two experiments with independent samples (Carroll, Rosenberg, & Funke, 1988; Greenfield & Haaga, 2011).

Coding Procedures

The authors independently coded each study using standardized instructions and code sheets for study characteristics, a priori planned moderator variables, and statistics for effect sizes. Discrepancies were resolved by team consensus. The moderator (categorical) codes for each study are provided in Figure 2. Study statistics were reported most commonly in the form of correlation coefficients (Pearson, Kendall, point-biserial), and in some instances p values or chi-square distributions. Cohen's kappa for coding the study characteristics, moderators and statistics ranged from .69 to .95 reflecting substantial to near perfect agreement (Landis & Koch, 1977).

Independent Measure: Confidence

Confidence was most frequently assessed using a single-item Likert-type scale. For example, Boyle (2000) used a 0 to 10 point scale where 10 reflected that judges were 100% confident in their accuracy. Cantor, Smith, French, and Mezzich (1980) used a seven-point scale, ranging from 1 (*the patient fit poorly in the category*) to 7 (*the patient fit optimally into the category*). The most common measures for confidence were 5-, 6-, 7-, and 10-point Likert-type scales. Other researchers used percentage ratings reflecting confidence from a low value up to 100% certainty (e.g., Haderlie, 2011). Most researchers maintained continuous measures although others used categorical splits (e.g., Douglas & Ogloff, 2003). Although there are limitations to using a single-item measure for confidence, the vast majority of studies assessed hit rates through repeated clinical judgments of confidence paired with judgment accuracy. The number of confidence-accuracy judgments per participant ranged from 2 to 166 with a mean of 37.84 (SD = 48.47; excluding 1368 judgments reported by Twaites, 1974).

Dependent Measure: Judgment Accuracy

Judgment accuracy reflected the accuracy of clinicians' judgments related to various mental health constructs, such as diagnosis, classification, prediction of behavior, and assessment of client progress. The way researchers established accuracy varied from study to study with some authors maintaining high levels of accuracy standards (e.g., verification of neuropsychological diagnosis by brain autopsy; Boyle, 2000) or verifiable behavior observations to establish judgment accuracy (e.g., measurable blood alcohol levels; Carroll, Rosenberg, & Funke, 1988). Other studies used less concrete verification of accuracy through expert a priori validation of a vignette, empirical findings, or accepted practice standards (e.g., confirmation of accurate schizophrenia diagnosis through multiple sources; Walker & Lewine, 1990). We evaluated judgment accuracy, however operationalized, as a form of criterion validity (high or low) based on our assessment of the rigor in establishing accuracy. Highly valid and objective standards were rated as having high criterion validity. Less objective methods, such as Kalichman and Craig's (1991) use of a logically constructed clinical vignette portraying sexual or physical abuse, were coded as low criterion validity. This was because logic or consensus rather than an objective measureable criterion was used. In this regard, criterion validity (high, low) for measures of accurate judgment was also treated as a moderator.

Definitions of Moderators

For categorical moderator analyses we established a priori that three studies must have examined a moderator sublevel to test its effects (see Table 1). Clinician experience (high, medium, low, mixed) was coded as high for postgraduate level mental health professionals including psychologists, psychiatrists, and mental health counselors; medium for doctoral level trainees including psychology interns; low for master's level trainees; and mixed for participants across these levels of experience. Judgment timeframe (concurrent, predictive) reflected whether judges formed concurrent classifications or predictions of future behavior. Whether or not clinicians were given access to mechanical prediction aids (yes, no) such as statistical and actuarial formulas was coded. We were interested in whether or not judges were given freedom to engage in assessments using their preferred or personally selected methods (yes, no) or were constrained by research methods. Studies varied in the amount of training provided to the judges before or during the judgment task and were coded accordingly (no training, before judgment task, before and during). The type of decisions judges made was coded, including diagnostic, brain injury, violence risk, abuse risk, substance abuse, malingering and prediction of therapeutic change. A general judgment category of "other" was formed that included mixed types (e.g., scoring confidence on WAIS-III, Hopwood & Richard, 2005; Ryan & Schnakenberg-Ott, 2003; race-based recall accuracy for content of psychotherapy sessions, Pedley, 1994). As a form of publication bias (Rothstein, Sutton & Borenstein, 2005), we assessed for differences between publication outlets (American Psychological Association [APA], non-APA, dissertation). Three continuous conceptual moderators were also coded: (a) number of judgments made, (b) base rate (occurrence) of the target behavior, and (c) age of the study. We initially hoped to code other moderators discussed in the confidence bias literature but they were not sufficiently studied (e.g., validity of stimulus material, Oskamp, 1965; feedback on accuracy, Ericsson et al., 1993; informing judges about base rates, Nisbett & Ross, 1980).

In the context of a sensitivity analysis Greenhouse and Iyengar (2009) suggest that methodological moderators also be assessed. If study quality, for example, is a significant moderator it may be relevant to examine only studies with higher quality as a means of estimating the confidence-accuracy effect.

Study name	Correlation	Lower limit	Upper limit	Z-Value	p-Value	Total	Categorical Variables
Boyle (2000)	0.680	0.234	0.890	2.750	0.006	14	11118321111
Cantor et al. (1980)	0.048	-0.636	0.690	0.118	0.906	9	12118122124
Carlin & Hewitt (1990)	-0.220	-0.801	0.574	-0.500	0.617	8	11123221211
Carroll et al. (1988) - 1	0.098	-0.260	0.432	0.529	0.596	32	12214221312
Carroll et al. (1988) - 2	0.000	-0.443	0.443	0.000	1.000	20	12214221312
Cooper & Werner (1990)	0.430	-0.002	0.727	1.951	0.051	21	21117223112
Desmarais et al. (2010)	-0.069	-0.468	0.353	-0.309	0.757	23	21127233512
Douglas & Ogloff (2003)	0.210	-0.941	0.974	0.213	0.831	4	21127123122
Ekman & O'Sullivan (1991)	0.020	-0.221	0.259	0.160	0.873	67	12213122311
Elkovitch et al. (2008)	0.220	-0.940	0.975	0.224	0.823	4	21127223212
Fero (1975)	0.188	-0.596	0.788	0.425	0.671	8	21234321112
Greenfield & Haaga (2011) - 1	0.120	-0.951	0.969	0.121	0.904	4	22225222113
Greenfield & Haaga (2011) - 2	-0.070	-0.966	0.955	-0.070	0.944	4	22225222113
Haderlie (2011)	-0.420	-0.830	0.285	-1.184	0.236	10	22215321513
Hopwood & Richard (2005)	-0.300	-0.627	0.118	-1.418	0.156	24	11226221321
Kalichman & Craig (1991)	0.440	0.348	0.523	8.500	0.000	327	12211121121
Kalichman et al. (1989)	0.381	0.276	0.477	6.642	0.000	277	12211121121
Kendell (1973)	0.100	-0.347	0.510	0.426	0.670	21	12218212321
Lefkowitz (1973)	0.208	-0.129	0.502	1.213	0.225	36	11238322423
Leli & Filskov (1981)	0.135	-0.387	0.591	0.490	0.624	16	11232222211
Leli & Filskov (1984)	0.280	-0.424	0.773	0.761	0.447	10	11232222211
Levenberg (1975)	0.086	-0.428	0.558	0.311	0.756	16	12218222211
McNeil et al. (1998)	0.210	-0.013	0.413	1.846	0.065	78	22117133412
Moxley (1973)	0.235	-0.392	0.713	0.718	0.472	12	21215222422
Nadler et al. (1994)	0.000	-0.224	0.224	0.000	1.000	77	11212111223
Pedley (1994)	0.530	0.295	0.704	4.046	0.000	50	32216322311
Rabinowitz et al. (1999)	0.030	-0.530	0.572	0.095	0.924	13	22117223512
Rodriguez (2002)	0.090	-0.213	0.377	0.578	0.563	44	12211221121
Ruscio & Stern (2005)	0.125	-0.159	0.390	0.861	0.389	50	12216221125
Ryan et al. (2003)	-0.100	-0.530	0.371	-0.401	0.688	19	11216212223
Stemple (1985)	0.004	-0.308	0.315	0.023	0.982	40	11218321213
Twaites (1974)	-0.020	-0.496	0.465	-0.075	0.940	17	21216323212
Walker & Lewine (1990)	0.000	-0.454	0.454	0.000	1.000	19	22218223311
Walters et al. (1988)	0.000	-0.254	0.254	0.000	1.000	60	12213121221
Wittemann et al. (2007)	0.092	-0.222	0.389	0.569	0.570	41	12218232121
Young (1970)	0.240	-0.459	0.755	0.648	0.517	10	11238321212
Overall random effect *r*	0.150	0.059	0.238	3.225	0.001	1485	

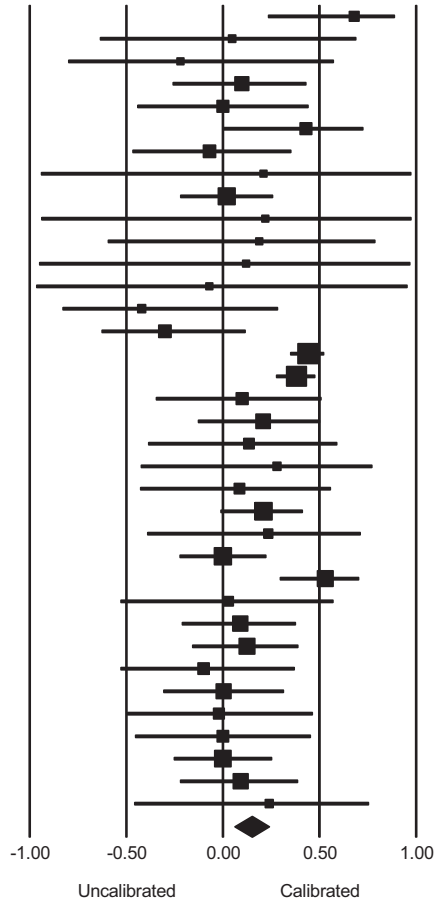

Figure 2. Random effects (*r*) between confidence and accuracy with forest plot and study moderator characteristics. Positive *r* effects indicate confidence is calibrated with accuracy. For each study *r* effects are represented by squares proportionately sized to inverse-variance weights used in calculating the overall effect. Lines represent the 95% confidence intervals (CIs) for each study. The diamond reflects the mean weighted overall confidence-accuracy effect; 95% CI is reflected by the width of the diamond. Categorical moderator variables and codes: Judgment timeframe (1 = concurrent, 2 = predictive, 3 = retrospective); mechanical prediction aids (1 = yes, 2 = no); freedom in assessment (1 = yes, 2 = no); training provided (1 = no, 2 = before judgment task, 3 = before and during); judgment outcome (1 = abuse risk, 2 = brain injury, 3 = malingering, 4 = substance abuse, 5 = therapeutic change, 6 = other, 7 = violence risk, 8 = diagnosis); publication source (1 = non-American Psychological Association [APA], 2 = dissertation, 3 = APA); study quality (1 = acceptable, 2 = good, 3 = excellent); research design (1 = experimental, 2 = quasi-experimental, 3 = correlational); stimulus materials (1 = written case material, 2 = test protocol, 3 = videotape, 4 = more than one source, 5 = live observation); criterion validity (1 = high, 2 = low); standard for accuracy (1 = a priori validation, 2 = observed behavior, 3 = test score cut-offs, 4 = clinical record of diagnosis, 5 = other).

We also assessed research design and validity issues related to the stimulus material and criterion validity used to establish judgment accuracy. We coded study quality as acceptable, good, or excellent using a global rating of methods and analyses based on Shadish, Cook, and Campbell's (2002) threats to validity. Cooper (1998) provides a good discussion of the advantages and disadvantages of such a global rating; compared with multidimensional ratings global ratings tend to have better interjudge agreement and equivalent heuristic value. The studies analyzed for this meta-analysis utilized experimental, quasi-experimental, and correlational research designs and were coded as such. The type of stimulus provided to judges was coded. Some judges were presented with direct client stimulus (through videotape or live observation). Other judges were given indirect client stimulus (written case material or test protocols). Each study was assigned a rating of high or low

Table 1
Categorical Models for Overall Confidence-Accuracy Effects

Moderators and levels	Between-class effect Q_B	p	k	Mean weighted effect size r	95% CI Lower	95% CI Upper
Conceptual moderators						
Clinician experience	6.51	.09				
High			15	.25	.14	.36
Medium			6	.08	−.15	.29
Low			13	.08	−.17	.32
Mixed			11	.04	−.08	.16
Judgment timeframe	.11	.74				
Concurrent			22	.14	.04	.24
Predictive			13	.10	−.09	.29
Mechanical prediction aids	1.10	.30				
Yes			17	.10	−.05	.24
No			19	.19	.09	.29
Freedom in assessment	0.41	.52				
Yes			9	.21	−.01	.42
No			27	.14	.03	.23
Training provided	3.88	.14				
No training			24	.17	.08	.27
Before judgment task			7	−.16	−.45	.17
Before and during			5	.21	−.10	.47
Judgment outcome	16.93	<.05				
Abuse risk			3	.37	.26	.48
Brain injury			3	.05	−.18	.28
Malingering			3	−.002	−.20	.20
Substance abuse			3	.07	−.21	.35
Therapeutic change			4	−.05	−.48	.40
Other			5	.16	−.02	.32
Violence risk			6	.18	−.01	.37
Diagnosis			9	.14	−.02	.29
Publication source	3.50	.17				
Non-APA journal			20	.05	−.05	.16
Dissertation			8	.23	−.01	.45
APA			8	.20	.05	.35
Method moderators						
Study quality	1.57	.46				
Acceptable			3	.00	−.27	.27
Good			30	.18	.08	.27
Excellent			3	.11	.08	.27
Research design	.08	.96				
Experimental			15	.15	.02	.27
Quasi-experimental			13	.17	.00	.32
Correlational			8	.13	−.09	.34
Stimulus material	22.93	<.001				
Written case material			12	.34	.26	.42
Test protocol			11	.02	−.12	.15
Videotape			7	.12	−.02	.26
More than one source			3	.21	.02	.39
Live observation			3	−.12	−.42	.21
Criterion validity	.01	.98				
High			22	.15	.02	.27
Low			14	.15	.03	.27
Standard for accuracy	3.13	.21				
A priori validation			15	.20	.06	.33
Observed behavior			11	.16	.02	.30
Test score cut-off			7	.01	−.14	.16

Note. Positive effects indicate confidence is calibrated with accuracy. Only categories with three or more effects were included in moderator analyses. Experience effects were calculated using subgroup analyses. CI = confidence interval; APA = American Psychological Association.

criterion validity based on the rigor with which judgment accuracy was established. Lastly, studies varied in the methodology used to establish judgment accuracy. Some studies used a priori methods to establish accuracy (e.g., review of stimulus materials by a panel of experts, expert construction of materials). In other studies, the target behaviors were directly observed or an objective test cut-off score was used and were coded accordingly.

Calculation of Correlation Effects

Study effects were reported differently across manuscripts but mostly in the form of correlation coefficients. When an overall *r* was reported we used that as the effect size. When provided other metrics (e.g., *p* value) these were entered into the meta-analysis software program *Comprehensive Meta Analysis 2.0* (CMA; Borenstein, Hedges, Higgins, & Rothstein, 2005) and converted to *r*. In all instances CMA transformed correlations to the Fisher's *z* scale for analysis, using Hedges and Olkin's (1985) *r*-to-*z* transformation, and then the analyses were converted back to correlations for ease of interpretation. Converting from *r* to Fisher's *z* for meta-analytic calculations is generally considered necessary to correct for skewed population distributions as *r* departs from zero (Rosenthal, 1984; for further discussion, see Hafdahl, 2010). For some studies we had to calculate an effect size; for example, Cantor et al. (1980) provided a table with accurate or inaccurate judgments paired with 7-point Likert-type confidence ratings and we calculated an overall point-biserial correlation. When an overall effect was not provided, we combined effect sizes within studies by first converting correlations to Fisher's *z*, averaging, and then converting back to an overall correlation (Corey, Dunlap, & Burke, 1998). Four studies reported statistically nonsignificant results, with data unavailable from the authors, and these were coded as zero effects (Carroll et al., 1988, Study 2; Nadler et al., 1994; Walker & Lewine, 1990; Walters et al., 1988). Only one moderator (experience) could be assessed using subgroup analyses, otherwise for the overall analysis and the moderator analyses one effect size per study was used. In calculating the overall effect size study effects were weighted by their inverse variance related to sample size (Borenstein, Hedges, Higgins, & Rothstein, 2005).

Results

Overall Analysis

The overall analysis was conducted using a random effects model based on the assumption there would be differences in effect sizes because of the various applications and ways in which studies were conducted. A random effects model is most appropriate in such an instance when the goal is to generalize to a number of scenarios (Borenstein et al., 2009). Across the 36 studies and multiple judgments made by 1,485 participants, the random effects weighted average effect size ($r = .15$) indicating a small relation between confidence and accuracy. The 95% confidence interval (CI [.059, .238]) does not include zero meaning this is a statistically significant effect ($p = .001$). The studies, types of prediction tasks, and aggregated confidence-accuracy correlations are presented along with a forest plot in Figure 2. The forest plot is a method used widely in the medical sciences and provides an overview of meta-analyses in a readily understandable visual mode (e.g., see Cochrane Library, 2015). Effect sizes ranged from −.42 indicating poor calibration between confidence and accuracy to +.68 indicating good calibration. A positive skew was found with 69% of the studies reporting a positive relation between confidence and accuracy. The effect size estimate demonstrated variability as evidenced by a statistically significant homogeneity index, $Q(35) = 69.52$, $p < .001$. The I-squared index was $I^2 = 49.66$, which means that approximately 50% of the total variance between studies is due to heterogeneity and not to chance; this is considered a moderate level of heterogeneity (Higgins, Thompson, Deeks, & Altman, 2003).

Sensitivity Analyses

Greenhouse and Iyengar (2009) recommend several steps subsumed under the concept of sensitivity analyses designed to examine the robustness of meta-analytic findings related to methodological and analysis decisions. To this end, we assessed the impact of including or excluding outliers and studies where we assigned zero effects in the absence of available data. To assess for outliers we utilized the outlier-labeling rule suggested by Tukey (1977) and refined by Hoaglin and Iglewicz (1987). With a lower critical value ($r = -0.479$) and an upper critical value ($r = .696$) no outliers were present, indicating no need to assess the overall effect with removal of outlier studies. Zero effects were entered for four studies when authors said the confidence-accuracy analyses were statistically nonsignificant and we were unable to obtain their data. Removal of these studies from the analyses results in a random effects ($r = .178$; CI [.083, .270]), $Q(32) = 57.70$, $p = .002$, and $I^2 = 46.27$, reflecting relatively similar findings compared with the more conservative overall correlation with the zero effects. We consider the more conservative correlation with the zero effects to represent the overall confidence-accuracy effect.

Categorical Moderator Analyses

Testing moderators is an essential part of an overall sensitivity analysis, particularly when 50% of the variance is unaccounted for due to heterogeneity between studies (Higgins et al., 2003). The overall effect size was further analyzed as a function of mixed-effects moderator analyses using a procedure analogous to analysis of variance (ANOVA; Hedges & Olkin, 1985). The between-class effect (Q_B) is conceptually equivalent to the main effect in ANOVA with each variable having more than one level. Mixed-effect models are thought to provide a more stringent test of moderators than fixed effects models by reducing the possibility of Type I errors (Overton, 1998). We treated each study as the unit of analysis meaning that an average of effects occurred when there were multiple judgments, or in some instances effect sizes reported for subgroups within a study were averaged into one aggregate study effect, weighted by the number of participants (Cooper, 1998). Studies either systematically varied clinician experience, or we were able to code studies for levels of experience; therefore but only for this moderator we were able to conduct subgroup analyses. As seen in Table 1, the type of stimulus used and the type of judgment made were statistically significant moderators of the confidence-accuracy effect. To assist in interpretation, moderators

are statistically significant when the 95% CI does not cross zero, and differences between levels of moderators are statistically significant when there is no overlap in their respective 95% CIs (Borman & Grigg, 2009). Stimulus material was found to moderate the relation between confidence and accuracy, $Q(4) = 22.93$, $p < .001$. Written case material was found to produce the largest effect size ($r = .34$), with more naturalistic stimulus materials such as live observation ($r = -.12$) and videotape ($r = .12$) producing effect sizes whose confidence intervals crossed zero, indicating no true relation. A difference was found based on the type of judgment, $Q(7) = 16.93$, $p < .05$. Abuse risk had the largest effect size ($r = .37$) and was the only effect with 95% CI not crossing zero, but is not statistically significant from several of the other judgments (because of overlap of 95% CIs).

Regression Moderator Analyses

Three moderator variables, number of judgments made, base rate of the target behavior, and age of study, were continuous and therefore analyzed using mixed effects (maximum likelihood) meta-regression (Borenstein et al., 2009).[3] For number of judgments three studies were not included in this moderator analysis due to omission of information about the number of judgments (Haderlie, 2011; Hopwood & Richard, 2005; Pedley, 1994). We also excluded Twaites (1974) who had participants make 1,368 judgments as we considered this to be an outlier. Consistent with our assumption as the number of judgments increased the relation between confidence and accuracy worsened, $Q(30) = 4.16$, $p < .05$. Second, as the base rate of the target behavior increased so does the positive relation between confidence accuracy, $Q(19) = 10.82$, $p = .001$. This finding is qualified by only 21 of the 36 studies reported base rates. Finally, we had no specific hypothesis for the age of the study but did find that older studies had larger positive confidence-accuracy effects, $Q(34) = 3.68$, $p = .05$.

Publication Bias

As part of a sensitivity analysis strategy we also assessed for publication bias. Publication bias occurs when editors and reviewers are more inclined to publish statistically significant as opposed to statistically nonsignificant findings (Rothstein et al., 2005). A fail-safe analysis (Rosenthal, 1979) resulted in a two-tailed Z value for the observed studies of 5.04 ($p < .00001$). To obtain a statistically nonsignificant Z value, under the cutoff of 1.96, 203 zero-effect studies would be needed which seems unlikely. The fail-safe analysis has been criticized because it assumes missing studies would have zero effects and does not take into consideration study weights (Becker, 2005). To provide a more comprehensive assessment we visually inspected a funnel plot with confidence-accuracy effect sizes on the horizontal axis and a measure of study sample size, or precision (one standard error) in this case, on the vertical axis (see Figure 3). Larger studies cluster toward the top of the funnel, usually closer to the mean effect size, whereas smaller studies are distributed at the bottom of the funnel. The bottom of the funnel shows asymmetry, with smaller studies toward the left, reflecting possible publication bias to publish small or even negative effects. Egger et al. (1997) recommend using the inverse of the standard error (precision) to predict the confidence-accuracy effect, in this case represented by Fisher's r to z, which results in a two-tailed, $t(34) = 3.46$, $p = .001$, supporting the statistical significance of this asymmetry. These analyses suggest there may be a bias of publishing small sample confidence-accuracy studies with small or negative effects, reflecting poorly calibrated confidence accuracy.

Discussion

> There are three things extremely hard: steel, a diamond, and to know one's self.—Benjamin Franklin, *Poor Richard's Improved Almanack* (1750)

Contrary to prevailing assumptions our analysis of over 40 years of research on the appropriateness of confidence uncovered that clinicians' confidence has a small but statistically significant relation with psychological assessment judgment accuracy ($r = .15$). This is a statistically significant effect because the 95% CI does not cross zero, but it is also an unreliable effect due to heterogeneity in variance between studies. This positive effect reflects a small improvement in accuracy associated with greater confidence. It could also reflect better calibration of confidence with accuracy where low confidence is associated with lower accuracy. A small effect is not without impact on current ways of thinking about the confidence bias in clinical decision making. Scholars consistently stress lowering confidence in one's judgments as a debiasing strategy to increase the likelihood of accurate judgments (Arkes, 1981; Borum et al., 1993; Ridley et al., 1998; Smith & Agate, 2004; Smith & Dumont, 1997; Spengler et al., 1995). In accordance with behavior decision-making theory (Kahneman et al., 1982; Nisbett & Ross, 1980), laypersons and professionals alike are theorized to be overconfident in the accuracy of their judgments due to unchecked use of judgmental heuristics and other cognitive shortcuts. On the basis of these assertions, we expected confidence to be uncalibrated with accuracy, which is not supported by this meta-analysis.

Do the findings suggest that counseling and other psychologists should trust their feelings of confidence in their clinical decision making? On the one hand an r of this magnitude in psychotherapy process-outcome studies (e.g., Norcross & Lambert, 2011) is considered a beneficial effect. Alternatively an r of .15 reflects that confidence accounts for 2% of variance in judgment accuracy ($r^2 = .0225$), which by any standard seems inconsequential. If counseling and other psychologists do in reality have the ability to appropriately gauge the accuracy of their own judgments, one would expect the aggregated effect size to be much larger. In practical terms, this means that while counseling and other psychologists may think they are very accurate (and indeed, in most studies clinicians reported that they were quite confident in their judgments), the feeling of confidence should continue to be viewed as marginally indicative of decision-making accuracy. In addition to awareness of the overconfidence bias, counseling psychologists should be aware of the closely aligned confidence heuristic (Price & Stone, 2004), or the environmental press to express high levels of confidence and the internal tendency to be influenced by highly confident decision-makers (e.g., Cramer et al., 2011; Semmler et al., 2012; Tenney et al., 2011). Apparently, human tendency is to

[3] Graphs of continuous moderators regressed on Fisher's Z scores are available upon request.

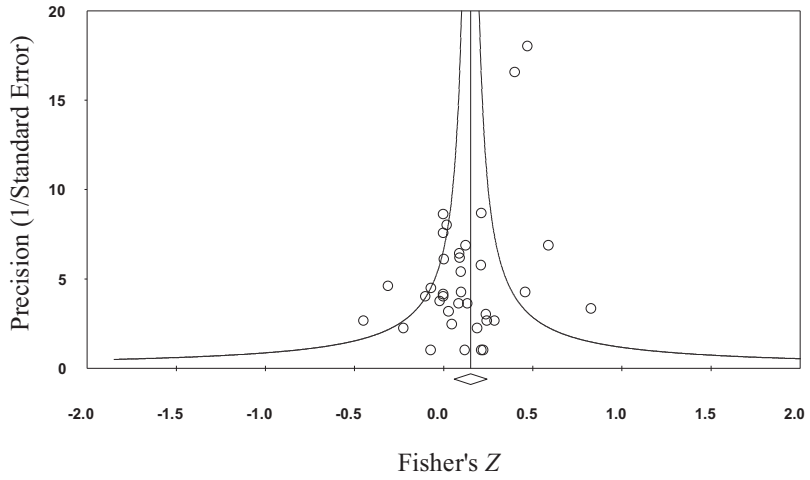

Figure 3. Funnel plot of precision by Fisher's *Z*.

trust more confident decision makers. Based on our meta-analysis, and similarly small effect sizes from confidence-accuracy meta-analyses in other domains (Aamodt & Custer, 2006; Sporer et al., 1995; Woodman & Hardy, 2003), the tendency to trust confident decision-makers, whether self or other, should probably be tempered.

Moderator Analyses

Few statistically significant moderators of the confidence-accuracy effect were found. This finding may in part be an artifact of low statistical power related to a modest sample size of 36 studies with a mean of 41.25 participants per study. With these values and assuming measurement error varying across studies, Sackett, Harris, and Orr's (1986) Monte Carlo computer simulation suggests that statistical power to detect moderator effects (r = .1, .2, and .3) in the present study was, respectively, .104, .316 and .760. Thus, small moderator effects may not be detected (for further discussion, see Hedges & Pigott, 2004). We found that older studies had the largest effect sizes, indicating better calibration of confidence-accuracy judgments, which might explain Garb's (1986) earlier narrative review where he concluded there is only mixed support for the confidence bias. We also found that confidence-accuracy calibration became worse over the course of repeated judgments, consistent with Oskamp (1965). In the absence of judges receiving feedback we assumed repeated decision making would fatigue judges and increase error. A body of research on deliberate practice suggests that feedback about errors (not accuracy), with subsequent opportunity to improve performance, is necessary to achieve optimal performance across a variety of domains (e.g., chess masters, elite athletes, virtuoso musicians, lifetime achievement scientists; Ericsson et al., 1993). We have no reason to believe that the positive correlation between feedback and performance would be different in the realm of mental health clinical decision making, and assume that this would help with calibration of confidence, but were only able to assess the impact of repeated judgments devoid of feedback.

We were unable to assess Garb's (1986) assertion that clinicians would be better able to calibrate their confidence when presented with highly valid stimuli. Of the 13 studies that provided decision-makers with test protocols, only Desmarais et al. (2010) provided readers information about test reliability and validity. Counseling psychologists are trained as scientist-practitioners and would be expected to take into account the validity of a predictor test when weighing confidence in their clinical judgments. The validity of the stimulus materials, in our opinion, is an important variable to consider in future research. We assessed the type of stimulus material and found that less directly experienced written case materials were associated with a larger confidence-accuracy effect (r = .34), whereas what might be considered more directly experienced stimuli (e.g., videotape) produced unreliable effects. Test protocols served as the stimuli for about one third of the studies and also were associated with an unreliable effect. It would seem that psychological tests would have the highest validity of all of the stimuli, but without information on validity for all of the various stimuli Garb's assertion remains untested.

Garb (1986) further proposed that more experienced clinicians would be better calibrated compared with less experienced clinicians. It seems worth noting, given possible low statistical power for moderator tests, that this was partially supported by a statistical trend (p = .09) where experienced clinicians produced a larger and the only reliable effect size (r = .25) of the experience groups. In a comprehensive meta-analysis of experience related to clinical decision-making accuracy, Spengler and Pilipis (2015) reported an overall random effects (d = .12; equivalent r = .06), reflecting some improvement in accuracy with clinical or educational experience. The results of the current meta-analysis suggest that experience and improvement in decision making may also include modest gains in confidence calibrations, but further research is needed to test this assumption.

Given resistance to the use of mechanical prediction techniques, we were not surprised that providing access to these decision aids, thought to be a key method for improving clinical decision making, failed to moderate the confidence-accuracy effect. A robust body of research establishes the slight superiority of statistical prediction over clinical prediction (Spengler, 2013), with a gain in accuracy reflected by a reliable effect (d = .12; equivalent r = .06;

see Ægisdóttir et al., 2006; Grove et al., 2000). Use of statistical reasoning in clinical decision making, however, commonly meets with resistance by practitioners and may only occur under certain compelling conditions perhaps not present in this body of confidence bias research. One such application where statistical prediction has been found to far outweigh clinical prediction is violence risk assessment and the plausible argument has been made that it may be unethical to confidently present clinical (vs. statistical) judgments in this realm of decision making (Harris et al., 2015). Contrary to our expectations training (without feedback) and the amount of information provided had no moderating effect. Finally, higher base rate (percentage occurrence) behaviors were associated with improved confidence-accuracy calibrations.

Given restrictions in how the data are reported the vast majority of moderator analyses were conducted on a between-studies basis in comparison to the more powerful within-studies or subgroups basis. This produces comparisons with lower certainty because the moderator represents an attribute of an entire study. Wood and Eagly (2009) argue that such a strategy reduces statistical conclusion validity because the moderator can be a proxy or a covariate with other study attributes. The alternative is to analyze subgroup comparisons within studies, as we were able to do for experience. This became an intractable problem for other moderators, however, as there was inconsistency between studies in terms of comparisons made not allowing for subgroup analyses. In other areas of research, such as meta-analyses of randomized controlled psychotherapy trials, where research is more systematic than we observed in this body of research, this type of subgroup or within-study moderator analysis is possible.

Research Implications

We found substantive heterogeneity between studies indicating there may be unidentified subclasses of studies that differ in the confidence-accuracy population effect (Matt & Cook, 2009). Future researchers should heed this finding by systematically studying potentially relevant moderators within studies, rather than assessing the mere correlation between confidence and accuracy, to better understand the confidence bias. The magnitude of heterogeneity found means that the included studies have genuinely different results not related to chance, reducing confidence in the findings and producing a lower grading of the evidence (Higgins et al., 2003). We also found publication bias with a disproportionate number of small sample studies producing effect sizes smaller than the overall effect ($r = .15$). Based on publication bias, and the fact that 32 of the 36 studies have 95% CIs that cross zero (see Figure 2), no wonder clinical judgment researchers perceive there is overall support for the confidence bias.

The most prominent shortcoming in extant confidence bias research is the need for stronger theory-driven studies, especially given the weight scholars put on the notion that reducing confidence improves decision-making accuracy. Many studies provided no rationale for collecting confidence measures (e.g., Carlin & Hewitt, 1990; Ryan & Schnackenberg-Ott, 2003). Others included measures of confidence but no analysis of data (Dauber, 1980; Regehr et al., 2010). Dauber (1980), for example, only provided visual plots of their data and no analyses. Several moderators we wished to test could not be tested due to insufficient studies or an absence of research. These include (a) the validity of the stimulus material (Garb, 1986), (b) feedback about accuracy of decision making (Ericsson et al., 1993), (c) whether the judges were informed about the base rate (Nisbett & Ross, 1980), and (d) when predictive accuracy is assessed the length of time of the outcome from the time of judgment. Related to being informed about base rates (not just higher base rates), the third author has served on the Division 17 Society for Counseling Psychology Programming Committee where decisions are made about proposals for presentations at the annual convention of the APA. Committee members are told the base rate of acceptance (e.g., 50%) and instructed to accept no more than 50% of the reviewed proposals unless there is something unusual about that sample. This type of base rate information assists with decision making (Ægisdóttir et al., 2006) and would presumably help calibrate the confidence-accuracy correlation.

Future research should also consider that the relation between confidence and clinical judgment accuracy might be more complex than a simple correlation (Olsson & Juslin, 2003). While nearly all studies included in this meta-analysis examined the confidence-accuracy relation in terms of a linear correlation, there is the possibility that confidence and accuracy have a curvilinear or other polynomial relation. That is, a certain level of confidence aids in accuracy but those benefits may decrease as confidence increases or decreases past certain points. Olsson and Juslin (2003) provided alternative metrics for researchers to use for assessing confidence and accuracy by eyewitnesses and earwitnesses and concluded that the point-biserial correlation coefficient may underestimate a stronger relation between confidence and accuracy. Additionally, we were unable in our review to determine in all instances the direction of correlations reported in studies. A positive correlation might reflect high confidence paired with high accuracy, or low confidence paired low accuracy, obfuscating the relation (see Figure 1).

Practice and Training Implications

We recognize that counseling and other psychologists do not rely solely on the feeling of being confident when making judgments. In fact, they should not as this would lead to mere impressionistic assessments. Although there is a statistical relation between confidence and accuracy, it is small and certainly not enough to allow counseling psychologists to disregard functioning like scientists in their assessments. Decision-making research reveals other factors that have empirical support as aids to forming accurate psychological assessments, such as the use of mechanical prediction techniques (for a comprehensive review, see Spengler, 2013), or the use of objective measures of change in psychotherapy (for a mega-analysis, see Shimokawa et al., 2010), but these approaches are not widely adopted by practitioners. Nonetheless, some increase in confidence does relate to an increase in accuracy. Therefore, it might be helpful for counseling psychologists to learn how to estimate appropriate levels of confidence in their judgments.

For instance, using the scientist-practitioner model of psychological assessment (Spengler et al., 1995), counseling psychologists could be taught to engage in a rigorous, structured method of scientific hypothesis testing using empirical evidence and debiasing techniques. In this way, they could learn to rely on empirical strategies for gauging their accuracy instead of relying on gut-level

instincts, such as confidence, which are strongly associated with the use of heuristics and biased judgments (Nisbett & Ross, 1980). Other variables linked to the confidence bias are the use of dispositional (as opposed to situational) explanations of client problems (Smith & Dumont, 2002) and engaging in confirmatory bias when information gathering (Martin, 2001). Combining this information with other known factors that increase accuracy (such as receiving feedback about accuracy; Ericsson et al., 1993) and decreasing reliance on other biased decision-making techniques (e.g., Smith & Agate, 2004), the counselor as scientist-practitioner might better calibrate confidence in her or his ability to make accurate judgments. Appropriate confidence would allow scientist-practitioners to move forward with a clinical decision while consistently acknowledging alternative possibilities, assessing for disconfirming evidence, and adjusting decisions as needed. In short, although our review is limited to a modest number of studies, and mostly to diagnostic and prognostic judgments, our advice for counseling psychologists based on this meta-analysis is "don't be so sure" in your unique clinical decision-making skills. Confidence is apparently not a good proxy for accuracy.

References

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

Aamodt, M. G., & Custer, H. (2006). Who can best catch a liar? A meta-analysis of individual differences in detecting deception. *Forensic Examiner, 15,* 6–11.

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., . . . Rush, J. D. (2006). The Meta-Analysis of Clinical Judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34,* 341–382. http://dx.doi.org/10.1177/0011000005285875

American Psychological Association. (2015). Guidelines for clinical supervision in health service psychology. *American Psychologist, 70,* 33–46. http://dx.doi.org/10.1037/a0038112

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

Becker, B. J. (2005). The failsafeN or file-drawer number. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 111–126). West Sussex, UK: Wiley.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. (2005). *Comprehensive Meta-Analysis Version 2*. Engelwood, NJ: Biostat.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley. http://dx.doi.org/10.1002/9780470743386

Borman, G. D., & Grigg, J. A. (2009). Visual and narrative interpretation. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 497–520). New York, NY: Russell Sage Foundation.

Borum, R., Otto, R., & Golding, S. (1993). Improving clinical judgment and decision making in forensic evaluation. *The Journal of Psychiatry & Law, 21,* 35–76.

*Boyle, P. A. (2000). *Decision-making strategies used by neuropsychologists in the differential diagnosis of dementia* (Unpublished doctoral dissertation). Amherst, MA: University of Massachusetts.

*Cantor, N., Smith, E. E., French, R. S., & Mezzich, J. (1980). Psychiatric diagnosis as prototype categorization. *Journal of Abnormal Psychology, 89,* 181–193. http://dx.doi.org/10.1037/0021-843X.89.2.181

*Carlin, A. S., & Hewitt, P. L. (1990). The discrimination of patient-generated and randomly generated MMPIs. *Journal of Personality Assessment, 54,* 24–29. http://dx.doi.org/10.1207/s15327752jpa5401&2_3

*Carroll, N., Rosenberg, H., & Funke, S. (1988). Recognition of intoxication by alcohol counselors. *Journal of Substance Abuse Treatment, 5,* 239–246. http://dx.doi.org/10.1016/0740-5472(88)90046-3

Cochrane Library. (2015). *The Cochrane Collaboration*. Chichester, UK: Wiley. Retrieved from http://www.cochranelibrary.com

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

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.

*Cooper, R. P., & Werner, P. D. (1990). Predicting violence in newly admitted inmates. *Criminal Justice and Behavior, 17,* 417–431. http://dx.doi.org/10.1177/0093854890017004004

Corey, D. M., Dunlap, W. P., & Burke, M. J. (1998). Averaging correlations: Expected values and bias in combined Pearson *r*s and Fisher's *z* transformations. *The Journal of General Psychology, 125,* 245–261. http://dx.doi.org/10.1080/00221309809595548

Cramer, R. J., DeCoster, J., Harris, P. B., Fletcher, L. M., & Brodsky, S. L. (2011). A confidence-credibility model of expert witness persuasion: Mediating effects and implications for trial consultation. *Consulting Psychology Journal: Practice and Research, 63,* 129–137. http://dx.doi.org/10.1037/a0024591

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302. http://dx.doi.org/10.1037/h0040957

Dauber, L. M. (1980). *An evaluation of the process of clinical judgment based on case history information*. (Unpublished doctoral dissertation). Buffalo, NY: State University of New York at Buffalo.

*Desmarais, S. L., Nicholls, T. L., Read, J. D., & Brink, J. (2010). Confidence and accuracy in assessments in short-term risks presented by forensic psychiatric patients. *Journal of Forensic Psychiatry & Psychology, 21,* 1–22. http://dx.doi.org/10.1080/14789940903183932

*Douglas, K. S., & Ogloff, J. R. (2003). The impact of confidence on the accuracy of structured professional and actuarial violence risk judgments in a sample of forensic psychiatric patients. *Law and Human Behavior, 27,* 573–587. http://dx.doi.org/10.1023/B:LAHU.0000004887.50905.f7

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5,* 69–106. http://dx.doi.org/10.1111/j.1529-1006.2004.00018.x

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315,* 629–634. http://dx.doi.org/10.1136/bmj.315.7109.629

Einhorn, H. J. (1980). Overconfidence in judgment. *New Directions for Methodology of Social and Behavioral Sciences, 4,* 1–15.

*Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist, 46,* 913–920. http://dx.doi.org/10.1037/0003-066X.46.9.913

*Elkovitch, N., Viljoen, J. L., Scalora, M. J., & Ullman, D. (2008). Assessing risk of reoffending in adolescents who have committed a sexual offense: The accuracy of clinical judgments after completion of risk assessment instruments. *Behavioral Sciences & the Law, 26,* 511–528. http://dx.doi.org/10.1002/bsl.832

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100,* 363–406. http://dx.doi.org/10.1037/0033-295X.100.3.363

Faust, D., & Faust, K. A. (2012). Experts' experience and diagnostic and predictive accuracy. In D. Faust (Ed.), *Coping with psychiatric and psychological testimony: Based on the original work by Jay Ziskin* (6th ed., pp. 131–146). New York, NY: Oxford University Press.

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

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 552–564. http://dx.doi.org/10.1037/0096-1523.3.4.552

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

Glidewell, J. C., & Livert, D. E. (1992). Confidence in the practice of clinical psychology. *Professional Psychology, Research and Practice, 23,* 362–368. http://dx.doi.org/10.1037/0735-7028.23.5.362

Goldberg, L. R. (1959). The effectiveness of clinicians' judgments; the diagnosis of organic brain damage from the Bender-Gestalt test. *Journal of Consulting Psychology, 23,* 25–33. http://dx.doi.org/10.1037/h0048736

*Greenfield, M. F., & Haaga, D. A. F. (2011). Accuracy and confidence of training therapists' recognition of sessions before sudden gains. *Journal of Cognitive and Behavioral Psychotherapies, 11,* 157–172.

Greenhouse, J. B., & Iyengar, S. (2009). Sensitivity analysis and diagnostics. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 417–434). New York, NY: Russell Sage Foundation.

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,* 19–30. http://dx.doi.org/10.1037/1040-3590.12.1.19

*Haderlie, M. M. (2011). *Enhancing therapists' clinical judgments of client progress subsequent to objective feedback* (Unpublished doctoral dissertation). Las Vegas, NV: University of Nevada.

Hafdahl, A. R. (2010). Random-effects meta-analysis of correlations: Monte Carlo evaluation of mean estimators. *The British Journal of Mathematical and Statistical Psychology, 63,* 227–254. http://dx.doi.org/10.1348/000711009X431914

Harris, G. T., Rice, M. E., Quinsey, V. L., & Cormier, C. A. (2015). *Violent offenders: Appraising and managing risk* (3rd ed.). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/14572-000

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

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9,* 426–445.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327,* 557–560. http://dx.doi.org/10.1136/bmj.327.7414.557

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327,* 557–560. http://dx.doi.org/10.1136/bmj.327.7414.557

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

Hoaglin, D. C., & Iglewicz, B. (1987). Fine tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association, 82,* 1147–1149. http://dx.doi.org/10.1080/01621459.1987.10478551

Holsopple, J. Q., & Phelan, J. G. (1954). The skills of clinicians in analysis of projective tests. *Journal of Clinical Psychology, 10,* 307–320. http://dx.doi.org/10.1002/1097-4679(195410)10:4<307::AID-JCLP2270100402>3.0.CO;2-3

*Hopwood, C. J., & Richard, D. C. (2005). Graduate student WAIS-III scoring accuracy is a function of full scale IQ and complexity of examiner tasks. *Assessment, 12,* 445–454. http://dx.doi.org/10.1177/1073191105281072

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus & Giroux.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511809477

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

*Kalichman, S. C., & Craig, M. E. (1991). Professional psychologists' decisions to report suspected child abuse: Clinician and situation influences. *Professional Psychology, Research and Practice, 22,* 84–89. http://dx.doi.org/10.1037/0735-7028.22.1.84

*Kalichman, S. C., Craig, M. E., & Follingstad, D. R. (1989). Factors influencing the reporting of father-child sexual abuse: Study of licensed practicing psychologists. *Professional Psychology, Research and Practice, 20,* 84–89. http://dx.doi.org/10.1037/0735-7028.20.2.84

Keillor, G. (2014). *The Keillor reader*. New York, NY: Penguin Group.

*Kendell, R. E. (1973). Psychiatric diagnoses: A study of how they are made. *The British Journal of Psychiatry, 122,* 437–445. http://dx.doi.org/10.1192/bjp.122.4.437

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

Lambert, J., Bessière, V., & N′Goala, G. (2012). Does expertise influence the impact of overconfidence on judgment, valuation and investment decision? *Journal of Economic Psychology, 33,* 1115–1128. http://dx.doi.org/10.1016/j.joep.2012.07.007

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

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174. http://dx.doi.org/10.2307/2529310

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

*Leli, D. A., & Filskov, S. B. (1981). Clinical-actuarial detection and description of brain impairment with the W-B form I. *Journal of Clinical Psychology, 37,* 623–629. http://dx.doi.org/10.1002/1097-4679(198107)37:3<623::AID-JCLP2270370330>3.0.CO;2-V

*Leli, D. A., & Filskov, S. B. (1984). Clinical detection of intellectual deterioration associated with brain damage. *Journal of Clinical Psychology, 40,* 1435–1441. http://dx.doi.org/10.1002/1097-4679(198411)40:6<1435::AID-JCLP2270400629>3.0.CO;2-0

*Levenberg, S. B. (1975). Professional training, psychodiagnostic skill, and kinetic family drawings. *Journal of Personality Assessment, 39,* 389–393. http://dx.doi.org/10.1207/s15327752jpa3904_11

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? The calibration of probability judgments. *Organizational Behavior and Human Performance, 10,* 159–183. http://dx.doi.org/10.1016/0030-5073(77)90001-0

Lilienfeld, S. O., Lynn, S. J., & Lohr, J. M. (2015). *Science and pseudoscience in clinical psychology* (2nd ed.). New York, NY: Guilfrod Press.

Lipko, A. R., Dunlosky, J., & Merriman, W. E. (2009). Persistent overconfidence despite practice: The role of task experience in preschoolers' recall predictions. *Journal of Experimental Child Psychology, 103,* 152–166. http://dx.doi.org/10.1016/j.jecp.2008.10.002

Martin, J. M. (2001). *Confirmation bias in the therapy session: The effects of expertise, external validity, instruction set, confidence and diagnostic accuracy* (Unpublished doctoral dissertation). Memphis, TN: The University of Memphis.

Matt, G. E., & Cook, T. D. (2009). Threats to the validity of generalized

inferences. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 537–560). New York, NY: Russell Sage Foundation.

*McNiel, D. E., Sandberg, D. A., & Binder, R. L. (1998). The relationship between confidence and accuracy in clinical assessment of psychiatric patients' potential for violence. *Law and Human Behavior, 22,* 655–669. http://dx.doi.org/10.1023/A:1025754706716

Meehl, P. E. (1957). When shall we use our heads instead of the formula? *Journal of Counseling Psychology, 4,* 268–273. http://dx.doi.org/10.1037/h0047554

*Moxley, A. W. (1973). Clinical judgment: The effects of statistical information. *Journal of Personality Assessment, 37,* 86–91. http://dx.doi.org/10.1080/00223891.1973.10119834

*Nadler, J. D., Mittenberg, W., DePiano, F. A., & Schneider, B. A. (1994). Effects of patient age on neuropsychological test interpretation. *Professional Psychology, Research and Practice, 3,* 288–295. http://dx.doi.org/10.1037/0735-7028.25.3.288

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2,* 175–220. http://dx.doi.org/10.1037/1089-2680.2.2.175

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

Norcross, J. C., & Lambert, M. J. (2011). Evidenced-based therapy relationships. In J. C. Norcross (Ed.), *Psychotherapy relationships that work: Evidence-based responsiveness* (2nd ed., pp. 3–24). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199737208.003.0001

Olsson, N., & Juslin, P. (2003). Calibration of confidence among eyewitnesses and earwitnesses. In P. Chambres, M. Izaute, & P. Marescaux (Eds.), *Metacognition: Process, function and use* (pp. 203–218). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of Consulting Psychology, 29,* 261–265. http://dx.doi.org/10.1037/h0022125

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3,* 354–379. http://dx.doi.org/10.1037/1082-989X.3.3.354

Owen, J. (2008). The nature of confirmatory strategies in the initial assessment process. *Journal of Mental Health Counseling, 30,* 362–374. http://dx.doi.org/10.17744/mehc.30.4.55281735v473x055

*Pedley, M. (1994). *The influence of client race on therapists' judgment making and memory for session-related material: An investigation of information processing among psychologists in training* (Unpublished doctoral dissertation). Binghamton, NY: State University of New York.

Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making, 17,* 39–57. http://dx.doi.org/10.1002/bdm.460

*Rabinowitz, J., & Garelik-Wyler, R. (1999). Accuracy and confidence in clinical assessment of psychiatric inpatients risk of violence. *International Journal of Law and Psychiatry, 22,* 99–106. http://dx.doi.org/10.1016/S0160-2527(98)00032-6

Regehr, C., Bogo, M., Shlonsky, A., & LeBlanc, V. (2010). Confidence and professional judgment in assessing children's risk of abuse. *Research on Social Work Practice, 20,* 621–628. http://dx.doi.org/10.1177/1049731510368050

Ridley, C. R., Li, L. C., & Hill, C. L. (1998). Multicultural assessment: Reexamination, reconceptualization, and practical application. *The Counseling Psychologist, 26,* 827–910. http://dx.doi.org/10.1177/0011000098266001

*Rodriguez, C. M. (2002). Professionals' attitudes and accuracy on child abuse reporting decisions in New Zealand. *Journal of Interpersonal Violence, 17,* 320–342. http://dx.doi.org/10.1177/0886260502017003006

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86,* 638–641. http://dx.doi.org/10.1037/0033-2909.86.3.638

Rosenthal, R. (1984). Meta analytic procedures for social science. *Applied Social Science Research Methods* (Vol. 6). Beverly Hills, CA: Sage.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. West Sussex, UK: Wiley. http://dx.doi.org/10.1002/0470870168

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

*Ryan, J. J., & Schnakenberg-Ott, S. D. (2003). Scoring reliability on the Wechsler Adult Intelligence Scale (3rd ed.). *Assessment, 10,* 151–159. http://dx.doi.org/10.1177/1073191103010002006

Ryback, D. (1967). Confidence and accuracy as a function of experience in judgment-making in the absence of systematic feedback. *Perceptual and Motor Skills, 24,* 331–334. http://dx.doi.org/10.2466/pms.1967.24.1.331

Sackett, P. R., Harris, M. M., & Orr, J. M. (1986). On seeking moderator variables in the meta-analysis of correlational data: A Monte Carlo investigation of statistical power and resistance to Type I error. *Journal of Applied Psychology, 71,* 302–310. http://dx.doi.org/10.1037/0021-9010.71.2.302

Semmler, C., Brewer, N., & Douglass, A. B. (2012). Jurors believe eyewitnesses. In B. L. Cutler (Ed.), *Convictions of the innocent: Lessons from psychological research* (pp. 185–209). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/13085-009

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton-Mifflin and Company.

Shimokawa, K., Lambert, M. J., & Smart, D. W. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology, 78,* 298–311. http://dx.doi.org/10.1037/a0019247

Smith, J., & Agate, J. (2004). Solutions for overconfidence: Evaluation of an instructional module for counselor trainees. *Counselor Education and Supervision, 44,* 31–43. http://dx.doi.org/10.1002/j.1556-6978.2004.tb01858.x

Smith, J., & Dumont, F. (1997). Eliminating overconfidence in psychodiagnosis: Strategies for training and practice. *Clinical Psychology: Science and Practice, 4,* 335–345. http://dx.doi.org/10.1111/j.1468-2850.1997.tb00125.x

Smith, J., & Dumont, F. (2002). Confidence in psychodiagnosis: What makes us so sure? *Clinical Psychology & Psychotherapy, 9,* 292–298. http://dx.doi.org/10.1002/cpp.336

Spengler, P. M. (2013). Clinical versus mechanical prediction. In J. Graham & J. Naglieri (Eds.), Handbook of psychology (2nd ed.). *Assessment Psychology* (Vol. 10, pp. 26–49). Hoboken, NJ: Wiley.

Spengler, P. M., & Pilipis, L. A. (2015, March 23). A comprehensive meta-reanalysis of the robustness of the experience-accuracy effect in clinical judgment. [Advance online publication]. *Journal of Counseling Psychology*. http://dx.doi.org/10.1037/cou0000065

Spengler, P. M., Strohmer, D. C., Dixon, D. N., & Shivy, V. A. (1995). A scientist-practitioner model of psychological assessment: Implications for training, practice and research. *The Counseling Psychologist, 23,* 506–534. http://dx.doi.org/10.1177/0011000095233009

Spengler, P. M., White, M. J., Ægisdóttir, S., Maugherman, A. S., Anderson, L. A., Cook, R. S., . . . Rush, J. D. (2009). The Meta-Analysis of Clinical Judgment project: Effects of experience on judgment accuracy. *The Counseling Psychologist, 37,* 350–399. http://dx.doi.org/10.1177/0011000006295149

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. L. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy

relation in eyewitness identification studies. *Psychological Bulletin, 118,* 315–327. http://dx.doi.org/10.1037/0033-2909.118.3.315

*Stemple, D. M. (1985). *The clinical judgment process of prediction of behavior problems in hospitalized adolescents using the MMPI and the Rorschach* (Unpublished doctoral dissertation). New York, NY: Fordham University.

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

Tenney, E. R., Small, J. E., Kondrad, R. L., Jaswal, V. K., & Spellman, B. A. (2011). Accuracy, confidence, and calibration: How young children and adults assess credibility. *Developmental Psychology, 47,* 1065–1077. http://dx.doi.org/10.1037/a0023273

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison Wesley.

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

Walfish, S., McAlister, B., O'Donnell, P., & Lambert, M. J. (2012). An investigation of self-assessment bias in mental health providers. *Psychological Reports, 110,* 639–644. http://dx.doi.org/10.2466/02.07.17.PR0.110.2.639-644

*Walker, E., & Lewine, R. J. (1990). Prediction of adult-onset schizophrenia from childhood home movies of the patients. *The American Journal of Psychiatry, 147,* 1052–1056. http://dx.doi.org/10.1176/ajp.147.8.1052

*Walters, G. D., White, T. W., & Greene, R. L. (1988). Use of the MMPI to identify malingering and exaggeration of psychiatric symptomatology in male prison inmates. *Journal of Counseling and Clinical Psychology, 56,* 111–117. http://dx.doi.org/10.1037/0022-006X.56.1.111

White, M. J., Nichols, C. N., Cook, R. S., Spengler, P. M., Walker, B. S., & Look, K. K. (1995). Diagnostic overshadowing and mental retardation: A meta-analysis. *American Journal on Mental Retardation, 100,* 293–298.

*Wittemann, C. L., & van den Bercken, J. H. (2007). Intermediate effects in psychodiagnostic classification. *European Journal of Psychological Assessment, 23,* 56–61. http://dx.doi.org/10.1027/1015-5759.23.1.56

Wood, W., & Eagly, A. H. (2009). Advantages of certainty and uncertainty. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 455–472). New York, NY: Russell Sage Foundation.

Woodman, T., & Hardy, L. (2003). The relative impact of cognitive anxiety and self-confidence upon sport performance: A meta-analysis. *Journal of Sports Sciences, 21,* 443–457. http://dx.doi.org/10.1080/0264041031000101809

*Young, R. C. (1972). Clinical judgment as a means of improving actuarial prediction from the MMPI. *Journal of Consulting and Clinical Psychology, 38,* 457–459. http://dx.doi.org/10.1037/h0032915

Received August 25, 2014
Revision received June 30, 2015
Accepted June 30, 2015 ■