Journal of Clinical Child & Adolescent Psychology, 0(0), 1–21, 2012

ISSN: 1537-4416 print/1537-4424 online
DOI: 10.1080/15374416.2012.736358

Routledge
Taylor & Francis Group

Future Directions in Psychological Assessment: Combining Evidence-Based Medicine Innovations with Psychology's Historical Strengths to Enhance Utility

Eric A. Youngstrom

Departments of Psychology and Psychiatry, University of North Carolina at Chapel Hill

Assessment has been a historical strength of psychology, with sophisticated traditions of measurement, psychometrics, and theoretical underpinnings. However, training, reimbursement, and utilization of psychological assessment have been eroded in many settings. Evidence-based medicine (EBM) offers a different perspective on evaluation that complements traditional strengths of psychological assessment. EBM ties assessment directly to clinical decision making about the individual, uses simplified Bayesian methods explicitly to integrate assessment data, and solicits patient preferences as part of the decision-making process. Combining the EBM perspective with psychological assessment creates a hybrid approach that is more client centered, and it defines a set of applied research topics that are highly clinically relevant. This article offers a sequence of a dozen facets of the revised assessment process, along with examples of corollary research studies. An eclectic integration of EBM and evidence-based assessment generates a powerful hybrid that is likely to have broad applicability within clinical psychology and enhance the utility of psychological assessments.

What if we no longer performed psychological assessment? Although assessment has been a core skill and a way of conceptualizing individual differences central to psychology, training and reimbursement have eroded over a period of decades (Merenda, 2007b). Insurance companies question whether they need to reimburse for psychological assessment (Cashel, 2002; Piotrowski, 1999). Educational systems have moved away from using ability-achievement discrepancies as a way of identifying learning disability and decreased the emphasis on individual standardized tests for individual placement (Fletcher, Francis, Morris, & Lyon, 2005). Several traditional approaches to personality assessment, such as the various interpretive systems for the Rorschach, have had their validity challenged repeatedly (cf. Meyer & Handler, 1997; Wood, Nezworski, & Stejskal, 1996). Many graduate-level training programs are reducing their emphasis on aspects of assessment (Belter & Piotrowski, 2001; Childs & Eyde, 2002; Stedman, Hatch, & Schoenfeld, 2001) and psychometrics (Borsboom, 2006; Merenda, 2007a) in their curricula, and few undergraduate programs offer courses focused on assessment or measurement. Efforts to defend assessment have been sometimes disorganized and tepid, or hampered by a lack of data even when committed and scholarly (Meyer et al., 1998).

Is this intrinsically a bad thing? Training programs, systems of care, and providers all have limited resources. Assessment might be a luxury in which some could afford to indulge, paying for extensive evaluations as a way to gain insight into themselves. However, arguments defending assessment as a major clinical activity need to appeal to utility to be persuasive (Hayes, Nelson, & Jarrett, 1987). Here, "utility" refers to adding value to individual care, where the benefits deriving from the assessment procedure clearly outweigh the costs, even when the costs combine fiscal expense with other factors such as time and the potential for harm (Garb, 1998; Kraemer, 1992; Straus, Glasziou, Richardson, &

Thanks to Guillermo Perez Algorta for comments and suggestions
Correspondence should be addressed to Eric A. Youngstrom, Department of Psychology, University of North Carolina at Chapel Hill, CB #3270, Davie Hall, Chapel Hill, NC 27599-3270. E-mail: eay@unc.edu

Haynes, 2011). Although utility has often been described in contexts of dichotomous decision making, such as initiating a treatment or not, or making a diagnosis or not, it also applies to situations with ordered categories or continuous variables. Conventional psychometric concepts such as reliability and validity are prerequisites for utility, but they do not guarantee it. Traditional evaluations of psychological testing have not formally incorporated the concept of costs in either sense—fiscal or risk of harm.

Using utility as an organizing principle has radical implications for the teaching and practice of assessment. Assessment methods can justify their place training and practice if they clearly address at least one aspect of prediction, prescription, or process—the "Three Ps" of assessment utility (Youngstrom, 2008). *Prediction* refers to association with a criterion of importance, which could be a diagnosis, but also could be another category of interest, such as adolescent pregnancy, psychiatric hospitalization, forensic recidivism, graduation from high school, or suicide attempt. For our purposes, the criterion could be continuous or categorical, and the temporal relationship could be contemporaneous or prospective. The goal is to demonstrate predictive validity for the assessment procedure by any of these methods and to make a compelling case that the effect size and cost/benefit ratio suggest utility. *Prescription* refers more narrowly to the assessment providing information that changes the choice of treatment, either via matching treatment to a particular diagnosis or by identifying a moderator of treatment. Similarly, *process* refers to variables that inform about progress over the course of treatment and quantify meaningful outcomes. These could include mediating variables, or be measures of adherence or treatment response. Each of the Three Ps demonstrates a connection to prognosis and treatment. These are not the only purposes that could be served by psychological assessment, but they are some of the most persuasive in terms of satisfying stakeholders that the assessment method is adding value to the clinical process (Meehl, 1997). Many of the other conventional goals of psychological assessment (Sattler, 2002) can be recast in terms of the Three Ps and utility: Using assessment as a way of establishing developmental history or baseline functioning may have predictive value or help with treatment selection, as can assessment of personality (Harkness & Lilienfeld, 1997). Case formulation speaks directly to the process of working effectively with the individual. Gathering history for its own sake is much less compelling than linking the findings to treatment and prognosis (Hunsley & Mash, 2007; Nelson-Gray, 2003).

It was surprising to me as an educator and a psychologist how few of the commonly taught or used techniques can demonstrate any aspect of prediction, prescription, or process—let alone at a clinically significant level (Hunsley & Mash, 2007). Surveys canvassing the content of training programs at the doctoral and internship level (Childs & Eyde, 2002; Stedman et al., 2001; Stedman, Hatch, Schoenfeld, & Keilin, 2005), as well as evaluating what methods are typically used by practicing clinicians (Camara, Nathan, & Puente, 1998; Cashel, 2002), show that people tend to practice similar to how they were trained. There is also a striking amount of inertia in the lists, which have remained mostly stable for three decades (Childs & Eyde, 2002). Content has been set by habits of training, and these in turn have dictated habits of practice that change slowly if at all.

When I first taught assessment, I used the courses I had taken as a graduate student as a template and made some modifications after asking to see syllabi from a few colleagues. The result was a good, conventional course; but the skills that I taught had little connection to the things that I did in my clinical practice as I pursued licensure. Much of my research has focused on assessment, but that created a sense of cognitive dissonance compared to my teaching and practice. One line of research challenged the clinical practice of interpreting factor and subtest scores on cognitive ability tests. These studies repeatedly found little or no incremental validity in more complicated interpretive models (e.g., Glutting, Youngstrom, Ward, Ward, & Hale, 1997), yet they remained entrenched in practice and training (Watkins, 2000). The more disquieting realization, though, was that my own research into assessment methods was disconnected from my clinical work. If conventional group-based statistics were not changing my own practice, why would I put forth my research to students or to other practitioners? Why was I not using the assessments I taught in class? When I reflected on the curriculum, I realized that I was teaching the "same old" tests out of convention, or out of concern that the students needed to demonstrate a certain degree of proficiency with a variety of methods in order to match at a good internship (Stedman et al., 2001).

What was missing was a clear indication of utility for the client. Reviewing my syllabi, or perusing any of the tables ranking the most popular assessment methods, emphasized the disconnect: Does scoring in a certain range on the Wechsler tests make one a better or worse candidate for cognitive behavioral therapy? Does verbal ability moderate response to therapies teaching communication skills? How does the Bender Gestalt test do at predicting important criteria? Do poor scores on it prescribe a change in psychological intervention? . . . or or tell about the process of working with a client? . . . - What about Draw a Person? Our most widely used tools do not have a literature establishing their validity in terms of individual prognosis or treatment, and viewed through the lens of utility they look superfluous. Yet these are all in the top 10 most widely used for assessing

psychopathology in youths, according to practitioner surveys (Camara et al., 1998; Cashel, 2002), even though they do not feature prominently in evidence-based assessment recommendations (Mash & Hunsley, 2005).

Evidence-based medicine (EBM) is rooted in a different tradition, grounded in medical decision making and initially advocated by internal medicine and other specialties bearing little resemblance to the field of psychology (Guyatt & Rennie, 2002; Straus et al., 2011). EBM has grown rapidly, however, and it has a variety of strengths that could reinvigorate psychological assessment practices if there were a way to hybridize the two traditions (Bauer, 2007). The principles of emphasizing evidence, and integrating nomothetic data with clinical expertise and patient preferences, are consistent with the goals of "evidence-based practice" (EBP) in psychology (Spengler, Strohmer, Dixon, & Shivy, 1995; Spring, 2007). Indeed, the American Psychological Association (2005) issued a statement endorsing EBP along the lines articulated by Sackett and colleagues and the Institute of Medicine. However, this is more agreement about a vision; and there is a fair amount of work involved in completing the merger of the different professional traditions. In much of what follows, I refer to EBM instead of EBP when talking about assessment, because EBM has assessment-related concepts that have not yet been discussed or assimilated in EBP in psychology. Key components include a focus on making decisions about individual cases, and knowing when there is enough information to consider something "ruled out" of further consideration or "ruled in" as a focus of treatment. EBM also has a radical emphasis on staying connected to the research literature, including such advice as "burn your textbooks—they are out of date as soon as they are published" (Straus et al., 2011). The emphasis on scientific evidence as guiding clinical practice seems philosophically compatible with the Boulder Model of training, and resonates with recent calls to further emphasize the scientific components of clinical psychology (McFall, 1991).

EBM's focus on relevance to the individual puts utility at the forefront: Each piece of evidence needs to demonstrate that it is valid and that it has the potential to help the patient (Jaeschke, Guyatt, & Sackett, 1994). However, most discussions of EBP in psychology have focused on therapy, with less explication of the concepts of evidence-based assessment (see Mash & Hunsley, 2005, for comment). Despite the shared vision of EBM and the American Psychological Association's endorsement of EBP, most of the techniques and concepts involved in assessment remained in distinct silos. For example, the terms "diagnostic likelihood ratio," "predictive power," "wait-test" or "test-treat threshold," or even "sensitivity" or "specificity" are not included as index terms in the current edition of *Assessment of Children and Adolescents* (Mash & Barkley, 2007; these terms are defined in the assessment context later in this article). A hand search of the volume found five entries in 866 pages that mentioned receiver operating characteristic analysis or diagnostic sensitivity or specificity (excluding the chapter on pediatric bipolar disorder, which was heavily influenced by the EBM approach). Of those five, one was a passing mention of poor sensitivity for an autism screener, and the other four were the exceptions among a set of 77 trauma measures reviewed in a detailed appendix. Discussions of evidence-based assessment have focused on reliability and classical concepts of psychometric validity but not application to individual decision making in the ways EBM proposes (Hunsley & Mash, 2005; Mash & Hunsley, 2005).

Conversely, treatments of EBM barely mention reliability and are devoid of psychometric concepts such as latent variables, measurement models, or differential item functioning (Guyatt & Rennie, 2002; Straus et al., 2011), despite the fact that these methods are clearly relevant to situations where the "gold standard" criterion diagnosis is missing or flawed (Borsboom, 2008; Kraemer, 1992; Pepe, 2003). Similarly, differential item functioning, tests of structural invariance, and the frameworks developed for testing statistical moderation would advance EBM's stated goals of understanding the factors that change whether the research findings apply to the individual patient (i.e., what are the moderating factors?; Cohen, Cohen, West, & Aiken, 2003) and understanding the process of change (i.e., the mediating variables; MacKinnon, Fairchild, & Fritz, 2007).

The two traditions have much to offer each other (Bauer, 2007). Because the guiding visions are congruent, it is often straightforward to transfer ideas and techniques between the EBM and psychological assessment EBP silos. The ideas from EBM have reshaped how I approach research on assessment, and reorganized my research and teaching to have greater relevance to individual cases. Our group has mostly applied these principles to the assessment of bipolar disorder (e.g., Youngstrom, 2007; Youngstrom et al., 2004; Youngstrom, Freeman, & Jenkins, 2009), but the concepts are far more broad. In the next section I lay out the approach to assessment as a general model and discuss the links to both EBM and traditional psychological assessment. This is not an introduction to EBM; there are comprehensive resources available (Guyatt & Rennie, 2002; Straus et al., 2011). Instead, I briefly describe some of the central features from the EBM approach to assessment and then lay out a sequence of steps for integrating these ideas with clinical psychology research and practice. The synthesis defines a set of new research questions and methods that are highly clinically relevant, and it reorganizes assessment practice in a way that is pragmatic and patient focused (Bauer, 2007). The

combination of EBM and psychological assessment also directly addresses the "utility gap" in current assessment practice and training (Hunsley & Mash, 2007). Sections describing research are oriented toward filling existing gaps, not reinforcing any bifurcation of research from practice.

A BRIEF OVERVIEW OF ASSESSMENT IN EBM

EBM focuses on shaping clinical ambiguity into answerable questions and then conducting rapid and focused searches to identify information that addresses each question (Straus et al., 2011). Rather than asking, "What is the diagnosis?" an EBM approach would refine the question to something like, "What information would help rule in or rule out a diagnosis of attention deficit/hyperactivity disorder (ADHD) for this case?" EBM references spend little time talking about reliability and almost no space devoted to traditional psychometrics such as factor analyses or classical descriptions of validity (cf. Borsboom, 2006; Messick, 1995). Instead, they concentrate on a Bayesian approach to interpreting tests, at least with regard to activities such as screening, diagnosis, and forecasting possible harm. The core method involves estimating the probability that a patient has a particular diagnosis, or will engage in a behavior of interest (such as relapse, recidivism, or self-injury), and then using Bayesian methods to combine that prior probability with new information from risk factors, protective factors, or test results to revise the estimate until the revised probability is low enough to consider the issue functionally "ruled out," or high enough to establish the issue as a clear target for treatment (Straus et al., 2011).

Bayes' Theorem, a way of combining probabilities, is literally centuries old (Bayes & Price, 1763). There are two ways of interpreting Bayes' Theorem: A Bayesian interpretation focuses on the degree to which new evidence should rationally change one's degree of belief, whereas a frequentist interpretation connects the inverse probabilities of two events, formally expressed as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

In this formula, P(A) is the prior probability of the condition, before knowing the assessment result; P(A|B) is the posterior probability, or the revised probability taking into account the information value of the assessment result; and P(B|A)/P(B) conveys the degree of support that the assessment result provides for the condition, by comparing the probability of observing the result within the subset of those that have the condition, P(B|A), to the overall rate of the assessment result, P(B). For example, if 20% of the cases coming to a clinical practice have depression—base rate = P(A) = 20%—and the client scores high on a test with 90% diagnostic sensitivity to depression—P(B|A) = 90%, %, or 90% of cases with depression scoring positive—then Bayes' Theorem would combine these two numbers with the rate of positive test results regardless of diagnosis to generate the probability that the client has depression conditional upon the positive test result. If 30% of cases score positive on the test regardless of diagnosis (what Kraemer, 1992, called the "level" of the test, to distinguish it from the false alarm rate), then the probability that the client has depression rises to 60%. Conversely, if the client had scored below threshold on the same test, then the probability of depression drops to less than 3%. The example shows the potential power of directly applying the test results to the individual case but also illustrates the difficulty of combining the information intuitively, as well as the effort involved in traditional implementations of the Bayesian approach.

Luminaries in clinical psychology such as Paul Meehl (Meehl & Rosen, 1955), Robyn Dawes (Dawes, Faust, & Meehl, 1989), and Dick McFall (McFall & Treat, 1999) have advocated incorporating it into everyday clinical practice. Some practical obstacles have delayed the widespread adoption of the method, including that it requires multiple steps and some algebra to combine the information, and the posterior probability is heavily dependent on the base rate of the condition. An innovation of the EBM approach is to address these challenges by offering online calculators or a "slide rule" visual approximation, a probability nomogram (see Figure 1), avoiding the need for computation, albeit at the price of some loss in precision (Straus et al., 2011). The nonlinear spacing of the markers on each line geometrically accomplishes the same effect as transforming prior probabilities (the left-hand line of the nomogram) into odds, then multiplying by the change in the diagnostic likelihood (plotted on the center line) to extrapolate to the posterior probability (the right-hand line), again avoiding the algebra to convert the posterior odds back into a probability (see the appendix, or Jenkins, Youngstrom, Washburn, & Youngstrom 2011, for a worked illustration).

A second, more conceptual innovation developed by EBM is to move past dichotomous "positive test/negative test result" thinking and to suggest a multitiered way of mapping probability estimates onto clinical decision making. In theory, the probability estimate of a target condition could range from 0% to 100% for any given case. In practice, almost no cases would have estimated probabilities of exactly 0% or 100%, and few might even get close to those extremes given the limits of currently available assessment methods. The pragmatic insight is that we do not need such extreme probability levels in order to make most clinical decisions (Straus et al., 2011). If the revised probability

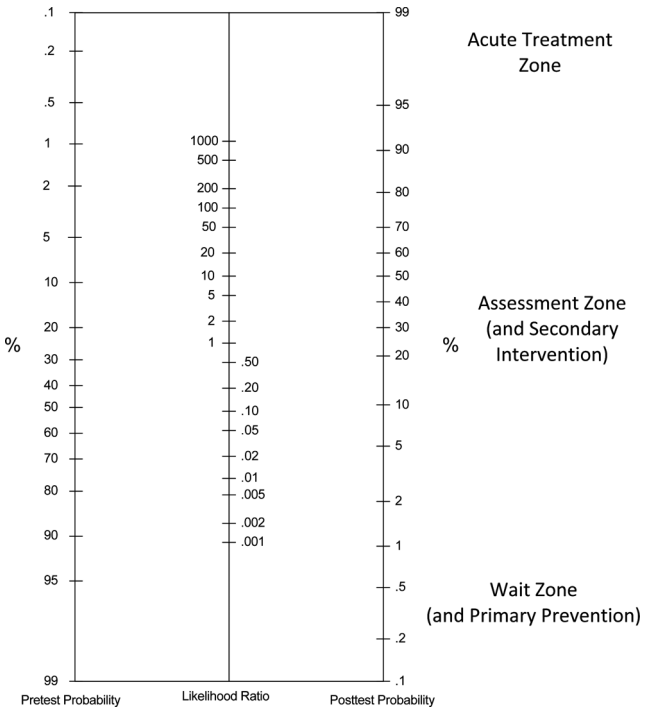

FIGURE 1 Probability nomogram for combining probability with likelihood ratios. *Note:* Straus et al. (2011) provided the rationale and examples of using the nomogram. Jenkins et al. (2011) illustrated using it with a case of possible pediatric bipolar disorder, and Frazier and Youngstrom (2006) with possible attention deficit/hyperactivity disorder.

is high enough, then it makes sense to initiate treatment, in the same way that if the weather forecast calls for a 95% chance of showers, then we would do well to dress for rain. EBM calls the threshold where it makes sense to initiate treatment the ''test-treat threshold''—probabilities above that level indicate intervention, whereas below that same point suggest continued assessment (Straus et al., 2011). Similarly, there is a point where the probability is sufficiently low to consider the target condition ''ruled out'' even though the probability is not zero. Below this ''wait-test'' threshold, EBM argues that there is no utility in continued assessment, nor should treatment be initiated. The two thresholds divide the range of probabilities and map them onto three clinical actions: actively treat, continue assessing, or decide that the initial hypothesis is not supported—and either assess or treat other issues (Guyatt & Rennie, 2002; Straus et al., 2011).

A third innovation in EBM is not to specify the exact locations for the wait-test and test-treat thresholds a priori. Instead, EBM provides a framework for incorporating the costs and benefits attached to the diagnosis, the test, and the treatment, and then using them to help decide where to set the bars for a particular case (Straus et al., 2011). Even better, there are ways of engaging the patient and soliciting personal preferences, including them in the decision-making process. For effective, low-risk, low-cost interventions, the treatment threshold might be so low that it makes sense to skip the assessment process entirely, as happens with routine vaccinations, or with the addition of fluoride to drinking water (Youngstrom, 2008). Conversely, for clinical issues where the treatment is freighted with risks, it makes sense to reserve the intervention until the probability of the target diagnosis is extremely high. For many families, atypical antipsychotics may fall in that category, given the serious side effects and the relative paucity of information about long-term effects on development (Correll, 2008). The EBM method creates a process for collaboratively weighing the costs, benefits, and preferences. This has the potential to empower the patient and customize treatment according to key factors, and it moves decision making from a simple, dichotomous mode to much more nuanced gradations. For the same patient, the test-treat thresholds might be more stringent for initiating medication than therapy, and so based on the same evidence it may make sense to start therapy, and wait to decide about medication until after additional assessment data are integrated.

These three innovations of (a) simplifying the estimation of posterior probabilities; (b) mapping the probability onto the next clinical action; and (c) incorporating the risks, benefits, and patient preferences in the decision-making process combine to restructure the process of assessment selection and interpretation. Assimilating these ideas has led to a multistep model for evaluating potential pediatric bipolar disorder (Youngstrom, Jenkins, Jensen-Doss, & Youngstrom, 2012). This model starts with estimates of the rate of bipolar in different settings, combines that with evidence of risk factors such as familial history of bipolar disorder, and then adds test results from either the Achenbach (Achenbach & Rescorla, 2001) or more specialized mood measures. Our group has published some of the needed components, such as the ''diagnostic likelihood ratios'' (DLRs; Straus et al., 2011) that simplify using a probability nomogram (Youngstrom et al., 2004), and vignettes illustrating how to combine test results and risk factors for individual cases (Youngstrom & Duax, 2005; Youngstrom & Kogos Youngstrom, 2005). We have tested whether weights developed in one sample generalize to other demographically and clinically different settings (Jenkins, Youngstrom, Youngstrom, Feeny, & Findling, 2012). These methods have large effects on how practicing clinicians interpret information, making their estimates more accurate and consistent, and eliminating a tendency to overestimate the risk of bipolar disorder (Jenkins, et al., 2011).

The methods are not specific to bipolar disorder: The core ideas were developed in internal medicine and have generalized throughout other medical practices (Gray, 2004; Guyatt & Rennie, 2002). These ideas define a set of clinically relevant research projects for each new content area, sometimes only involving a shift in interpretation, but other times entailing new statistical methods or designs. Adopting these approaches redirects research to build bridges to clinical practice and orients the practitioner to look for evidence that will change their work with the patient, thus spanning the research–practice gap from both directions.

TWELVE STEPS FOR EBM, AND A COROLLARY CLINICAL RESEARCH AGENDA

The process of teaching and using the EBA model in our clinic has augmented the steps focused on a single disorder, and no doubt there will be more facets to add in the future. A dozen themes is a good start for outlining a near-future approach to evidence based assessment in psychology. Table 1 lists the steps, a brief description of clinical action, and the corresponding clinical research agenda—reinforcing the synthesis of research and practice in this hybrid approach. Figure 2 lays out a typical sequence of working through the steps, and also maps them onto the clinical decision-making thresholds from EBM and the next clinical actions in terms of assessment and treatment. All of these steps presume that the provider has adequate training and expertise to administer, score, and interpret the assessment tools accurately, or is receiving appropriate supervision while training in their use (Krishnamurthy et al., 2004).

1. Identify the Most Common Diagnoses and Presenting Problems in Our Setting

Before concentrating on the individual client, it is important to take stock of our clinical setting. What are the common presenting problems? What are the usual diagnoses? Are there any frequent clinical issues, such as abuse, custody issues, or self injury?

After making the short list of usual suspects, then it is possible to take stock of the assessment tools and practices in the clinic. Are evidence-based assessment tools available for each of the common issues? Are they routinely used? What are the gaps in coverage, where fairly common issues could be more thoroughly and accurately evaluated? Recent work on evidence-based assessment in psychology has anthologized different instruments and reviewed the evidence for the reliability and validity of each (Hunsley & Mash, 2008; Mash & Barkley, 2007). These can help guide selection. Tests with higher reliability and validity will provide greater precision and more accurate scores for high-stakes decisions about individuals (Hummel, 1999; Kelley, 1927). Factor analyses also help explicate how different scales relate to underlying constructs and to each other, allowing for more parsimony in test selection.

Pareto's "rule of the vital few" is a helpful approximation: It is not necessary to have the resources to address every possible diagnosis or contingency, and pursuing comprehensiveness would yield sharply diminishing returns. Instead, approximately 80% of cases in most clinics will have the same ~20% of the possible clinical issues. Organizing the assessment methods to address the common diagnoses will focus limited resources to address the routine referrals and presenting problems. Making the list of typical issues more explicit also helps trainees and new clinicians to consider their work context, and it turns descriptive data into institutional wisdom that can improve the assessment process through the steps described next. Tests that do not have adequate reliability or evidence of validity cannot have utility for individual decision making. The heuristic of "is this test valid, and will it help with the patient?" (Straus et al., 2011) provides a way of identifying tests that we do not want to use, and should not continue to teach, without new evidence that shows sufficient validity. Thinking about the common presenting problems and the reliable and valid tests that assess them also would help organize a "core battery" if a clinic decides to implement a standardized intake evaluation.

Clinical research agenda. One research approach to identifying the common clinical issues is to conduct clinical epidemiological studies, looking at the rates of diagnoses and key behavioral indicators across a range of service settings. Most epidemiological research focuses on the general population, regardless of treatment status. More relevant to clinicians would be the distributions of diagnoses in outpatient practice, in special education, in residential treatment, and the other settings where we provide services.

A second research project would be to map the relatively short list of families' typical presenting concerns (Garland, Lewczyk-Boxmeyer, Gabayan, & Hawley, 2004) onto the much larger list of diagnostic possibilities. If a family comes in worried about aggression, what is the shortlist of hypotheses to consider? What are the cultural factors and beliefs about causes of behavior that change how families seek help and engage with different treatments (Carpenter-Song, 2009; Yeh et al., 2005)?

2. Know the Base Rates of the Condition in Our Setting

Meehl (1954) advocated "betting the base rate" as a simple strategy to improve the accuracy of clinical

TABLE 1
Twelve Steps in Evidence-Based Assessment and Research

Assessment Step	*Rationale*	*Clinical Research Agenda*
1. Identify most common diagnoses in our setting	Planning for the typical issues helps ensure that appropriate assessment tools are available and routinely used	Clinical epidemiology; mapping presenting problem and cultural factors onto diagnoses and research labels.
2. Know base rates	Base rate is an important starting point to anchor evaluations and prioritize order of investigation	Clinical epidemiology; meta-analyses of rates across different settings and methods.
3. Evaluate relevant risk and moderating factors	Risk factors raise "index of suspicion," enough combined elevate probability into assessment or possibly treatment zones	Compare rates of risk factors in those with versus without target diagnosis; reexpress as DLRs; meta-analyses to identify moderators.
4. Synthesize broad instruments into revised probability estimates	Already widely used; know what the scores mean in terms of changing probability for common conditions	Analyses generating DLRs for popular broad coverage instruments for different clinical targets.
5. Add narrow and incremental assessments to clarify diagnoses	Often more specific measures will show better validity, or incremental value supplementing broad measures	Test incremental validity, or superiority based on cost/benefit ratio.
6. Interpret cross-informant data patterns	Pervasiveness across settings/informants reflects greater pathology. Important to understand typical patterns of disagreement, and not overinterpret common patterns.	Test diagnostic efficiency of each informant separately; test incremental value of combinations.
7. Finalize diagnoses by adding necessary intensive assessment methods	If screening and risk factors put revised probability in the "assessment zone," what are the evidence-based methods to confirm or rule out the diagnosis in question? (e.g., KSADS, neurocognitive testing . . .)	Evaluate tests in sequence in different settings to develop optimal order and weights. Develop highly specific assessments to help rule in diagnoses.
8. Complete assessment for treatment planning and goal setting	Rule out general medical conditions, other medications; Family functioning, quality of life, personality, school adjustment, comorbidities	Develop systematic ways of screening for medical conditions and medication use. Test family functioning, personality, comorbidity, socioeconomic status and other potential moderators of treatment effects.
9. Measure processes ("dashboards, quizzes and homework")	Life charts, mood and energy checkups at each visit, medication monitoring, therapy assignments, daily report cards, three-column and five-column charts . . .	Demonstrate treatment sensitivity; meditational analyses; dismantling studies examining value added.
10. Chart progress and outcome ("midterm and final exams")	Repeat assessment with main severity measures—interview and/or parent report most sensitive to treatment effects	Jacobson and Truax (1991) benchmarks and reliable change metrics; comparison of effect sizes in same trial for different methods; develop low burden methods generalizable across patients, settings, systems. If poor response, revisit diagnoses.
11. Monitor maintenance and relapse	Discuss continued life charting; review triggers, critical events and life transitions	Event history analyses (predictors of relapse, durable recovery), key predictors, recommendations about next action if roughening.
12. Solicit and integrate patient preferences	Patient beliefs and attitudes influence treatment seeking and engagement. Possible to use these preferences to adjust wait-test and test-treat thresholds or utilities.	Qualitative analyses to identify key themes, cultural factors, preferences; studies of how to quantify preferences and add to decision making.

Note: DLR = diagnostic likelihood ratio; KSADS = Kiddie Schedule for Affective Disorders and Schizophrenia.

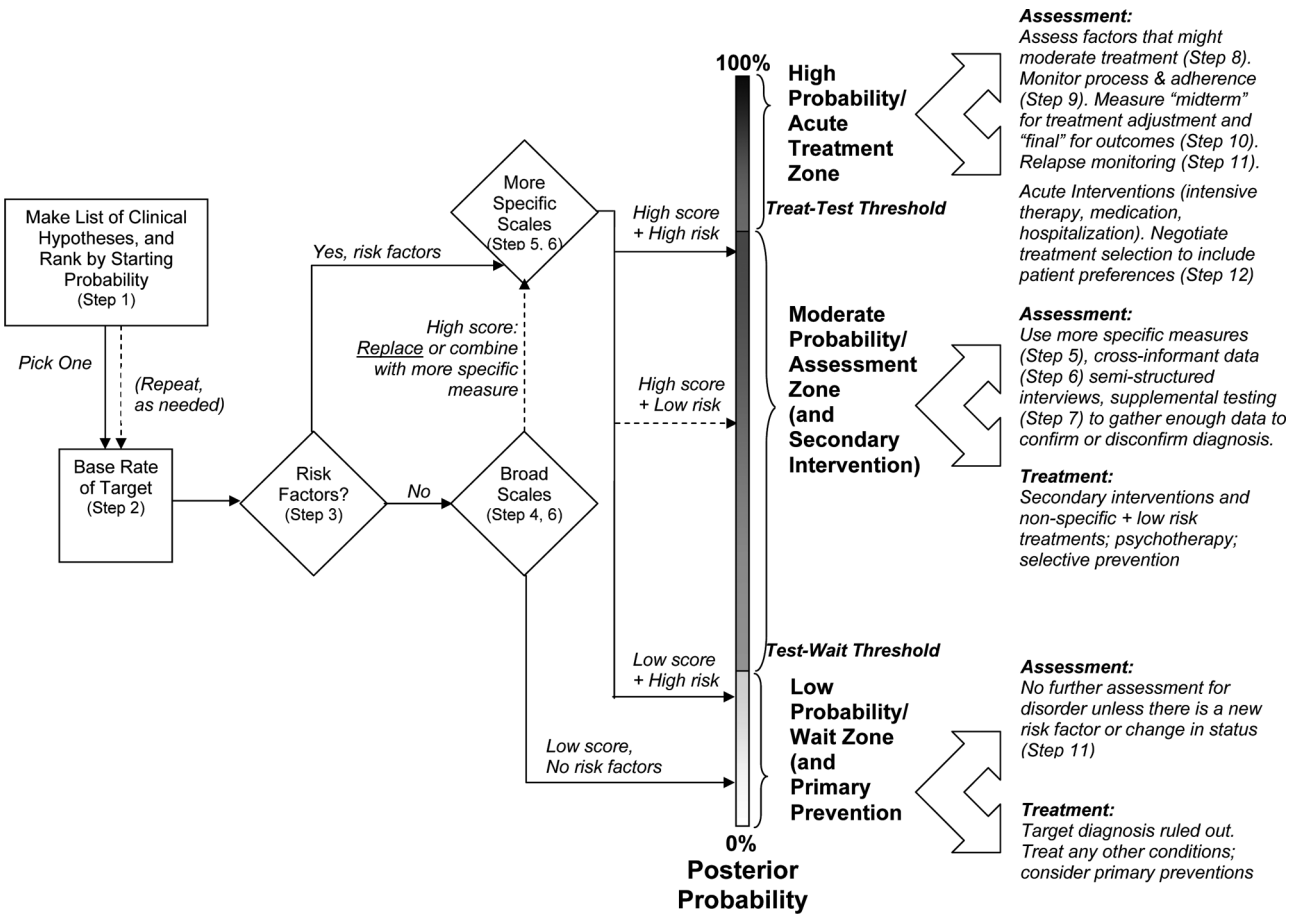

FIGURE 2 Mapping assessment results onto clinical decision making.

assessment, using the base rate as the Bayesian prior probability before adding assessment findings. When the same constellation of symptoms could be explained by an exotic or a quotidian illness, wager on the common cause. A stomachache and fever are more likely to be due to a cold virus than ebola hemhorragic fever, unless there are many other risk factors and signs that point toward the more rare explanation. The clinical epidemiological rates provide a helpful starting point for ranking the potential candidates in terms of probability before considering any case-specific information, organizing a set of potential clinical hypotheses. The prevalence of different conditions also provides a good starting estimate, taking advantage of what cognitive psychologists call the "anchoring heuristic" (Croskerry, 2003; Gigerenzer & Goldstein, 1996). Rather than interpreting case information intuitively, formally thinking about the base rates as a starting point helps increase the consistency of decision making across clinicians (Garb, 1998). Psychology has contributed both to the research about decision making and cognitive heuristics and to descriptive studies of prevalence in different settings.

Clinical research agenda. As more clinical epidemiology studies are published, then meta-analyses could describe general patterns across levels of service and identify moderating variables that change referral patterns. Studies using semistructured or structured interviews provide valuable benchmarks against which to compare local patterns. For example, if studies of urban community mental health centers find that roughly 50% of referrals meet criteria for a diagnosis of ADHD but only 20% of youths at a local center receive clinical diagnoses, or 80% for that matter, then the benchmark raises important questions about whether local assessment practices could benefit from upgrading the evidence based components.

3. Evaluate the Relevant Risk and Moderating Factors

Within the EBM framework, risk factors become data to integrate into the formal assessment process. The DLR central to the EBM method is a ratio of the diagnostic sensitivity to the false alarm rate. Put another way, the DLR compares how often the test result or risk

factor would occur in those with the diagnosis (i.e., sensitivity) versus its rate in those without the diagnosis (i.e., false alarm rate). If low birth weight was present in 3% of youths with ADHD but only 1% of those without ADHD, then the DLR attached to low birth weight would be 3.0 for ADHD. The DLR is the factor by which the odds of diagnosis change in Bayesian analysis. For clinical purposes, the conceptual status of low birth weight changes from an empirically identified "risk factor" to a variable contributing a specific weight to decision making about a particular individual case. EBM suggests that risk factors or tests producing DLRs of less than 2 are rarely worth adding to the evaluation process, whereas values around 5 are often helpful, and values greater than 10 frequently have decisive impact on an evaluation (Straus et al., 2011).

Clinical research agenda. Extensive developmental psychopathology research has focused on identifying risk and protective factors. However, these are primarily reported in terms of statistical significance and group-level effect sizes (Kraemer et al., 1999). The next step is to convert these findings into a metric amenable to idiographic assessment and decision making. The necessary statistics to generate DLRs for risk factors are simple. A chi-squared test comparing the presence or absence of the risk factor in those with or without the diagnosis is sufficient to test the validity of the risk factor (Kraemer, 1992). The next step, rarely taken in psychology to date, is to report the percentages: How common is the risk factor in those with the diagnosis versus without? Those constitute the numerator and denominator of the DLR.

4. Synthesize Broad Instruments into Revised Probability Estimates

Many clinics and practitioners use a broad assessment instrument as a standard element of their intake (e.g., Child Behavior Checklist, Behavior Assessment System for Children; Achenbach & Rescorla, 2001; Reynolds & Kamphaus, 2004). Broad instruments have a variety of strengths, including providing norm-referenced scores that compare the level of problems to what would be age- and gender-typical levels, as well as systematically assessing multiple components of emotional and behavior problems regardless of the particular referral question. This breadth prevents some cognitive heuristics that otherwise plague unstructured clinical assessments, such as concentrating only on one hypothesis, or "search satisficing" and stopping the evaluation as soon as one plausible diagnosis is identified (Croskerry, 2003; Spengler et al., 1995). The next step in an evidence-based assessment approach is to incorporate the test results and see how they raise or lower the posterior probability of the contending diagnoses. In the Bayesian EBM framework, the test score ranges have DLRs attached, and these get combined with the prior probability and risk factor DLRs to generate a revised probability estimate. It is worth noting that broad measures will not cover all possible conditions, despite their breadth. Problems that are rare in the general population may not have enough representation to generate their own "syndrome scale." This does not invalidate the use of broad measures in an EBA approach, but rather reminds us to be aware of the limits of content coverage and not unwittingly exclude clinical hypotheses outside of the scope of coverage.

Clinical research agenda. There have been a smattering of studies using Receiver Operating Characteristic (ROC) analyses to evaluate the diagnostic efficiency of broad instruments with regard to specific diagnoses such as ADHD (e.g., Chen, Faraone, Biederman, & Tsuang, 1994) and anxiety (e.g., Aschenbrand, Angelosante, & Kendall, 2005). The next step would be to calculate multilevel likelihood ratios attached to low, moderate, and high scores on the test (Guyatt & Rennie, 2002). The multilevel approach preserves more information from continuous measures, and it also is likely to be more generalizable and less sample dependent than approaches focused on picking the "optimal" cut scores (Kraemer, 1992). The approach can be simple yet still highly informative: Samples could be divided into thirds or quintiles on the Externalizing or Internalizing scale, and then the percentage of cases with the diagnosis compared to the percentage without the diagnosis in each score stratum to determine the diagnostic likelihood ratio (e.g., Youngstrom et al., 2004). As the research literature becomes more rich, then it would be possible for meta-analyses to test the generalizability of results and document moderating factors (Hasselbad & Hedges, 1995).

5. Add Narrow and Incremental Assessments to Clarify Diagnoses

At some clinics a common referral issue may not be adequately assessed by broad instruments. Pervasive developmental disorders, eating disorders, bipolar disorders, and other topics all may require the addition of more specialized measures or checklists (Mash & Hunsley, 2005). Again, a good survey of the common issues at a particular setting guides rational additions to the assessment battery. Some important issues may only be addressed by a single item or omitted entirely from broad assessment measures: The Achenbach instruments do not have scales for mania, eating

disorders, or autism, per se, for example. Psychological research has also made advances in terms of documenting incremental validity of combinations of tests (Johnston & Murray, 2003) as well as statistically testing what factors moderate the performance of tests (Cohen et al., 2003; Zumbo, 2007). The best candidates for addition to the assessment protocol will be tools that have demonstrated validity for the target diagnosis, and ideally have DLRs available so that the scores can be translated directly into a revised probability.

Clinical research agenda. Validating more narrow tests for diagnostic efficiency involves several steps. At early stages, studies performing receiver operating characteristic analyses would establish the discriminative validity of the assessment (McFall & Treat, 1999). Ideally the study design would follow the recommendations of the Standardized Reporting of Diagnostic tests guidelines (Bossuyt et al., 2003), and it would use clinically generalizable comparison groups to develop realistic estimates of performance (Youngstrom, Meyers, Youngstrom, Calabrese, & Findling, 2006a). Later steps in the research process could include comparing the ROC performance of multiple tests either in the same sample (using procedures developed by Hanley & McNeil, 1983), or meta-analytically (Hasselbad & Hedges, 1995). Logistic regression models, using diagnosis as the dependent variable, could test whether there is incremental value in combining different tests. Logistic regression also offers a flexible framework for testing potential moderators of assessment performance, such as gender, ethnicity, culture (Garb, 1998), or credibility of the informant (Youngstrom et al., 2011). EBM teaches us to ask, "Do these results apply to this patient?" (Straus et al., 2011). The psychometric tradition has developed powerful tools to answer the question of whether results generalize, versus the validity changing due to demographic or clinical characteristics (Borsboom, 2006). When appropriate samples are available, then generating multilevel likelihood ratios for the narrow instrument also would be crucial to facilitate clinical application.

6. Interpret Cross-Informant Data Patterns

A stock recommendation in clinical assessment of youths is to gather data from multiple informants, including parents, teachers, and direct observations, as well as self-report or performance measures from the youth. However, it is well-established that these different sources of information show only modest to moderate convergence, usually in the range of $r = .1$–$.4$ (Achenbach, McConaughy, & Howell, 1987). Additional data can actually degrade the quality of clinical decision making, especially when the new data have low validity for the criterion of interest or when suboptimal strategies are used to synthesize information. Context and diagnostic issue moderate the validity of data across informants (De Los Reyes & Kazdin, 2005). Self-report of attention problems, or teacher report of manic symptoms, are examples of information with validity that is significantly lower than could be gleaned by asking the same questions of other sources. Adding more tests to a battery always increases the time, cost, and complexity, but it does not always improve the output (Kraemer, 1992). Cross-informant data often add considerably to the time and expense of an assessment. The psychological assessment literature has developed to a point where we can decide when the additional assessment is worth the effort, and when it would be more efficient to forego. A related point is that we can anticipate common patterns of disagreement: Whoever initiates the referral will usually be the most worried party. Low cross-informant correlations and regression to the mean will combine so that the typical scenario often looks unimpressive in terms of agreement: If the average level of parent-reported problems has a T score of 70, the expected level of youth or teacher reported problems would be in the range of 54 to 56 (Achenbach & Rescorla, 2001; Youngstrom, Meyers, Youngstrom, Calabrese, & Findling, 2006b). Recognizing and thinking through common scenarios will help avoid misinterpreting patterns in the cross-informant data (Croskerry, 2003). When different informants have shown incremental validity, then integrating the different scores into a revised probability makes sense. Even when incremental validity for diagnostic purposes may be poor, there is still value in assessing cross-informant agreement with regard to motivation for treatment (Hunsley & Meyer, 2003).

Clinical research agenda. The ideas of cross-informant data and validity are well developed in psychological assessment and virtually unknown in the traditional EBM literature. ROC and logistic regression again provide an analytic framework for evaluating the diagnostic efficiency of each informant's perspective and testing whether there is significant incremental value added by combining different informants' perspectives.

7. Finalize Diagnoses by Adding Necessary Intensive Assessment Methods

One of the goals in sequencing the assessment steps is to try to set up a "fast and frugal" order that maximizes the information value of instruments already widely used (Gigerenzer & Goldstein, 1996) and that minimizes the additional time and expense used in the first wave of assessment for a case. Based on the initial findings,

many clinical hypotheses will be "ruled out." However, few of our assessment tools are sufficiently specific to a diagnostic issue or accurate enough to confirm a diagnosis on their own. After conducting the initial evaluation, clinicians will often find that the revised probability estimate falls in the middle "assessment zone," and additional assessment is needed to confirm or disconfirm the diagnosis. More intensive and expensive tests are justified for contending diagnoses at this stage: The prior steps have screened out low probability cases so that the more expensive methods are not being used indiscriminately (Kraemer, 1992). Reserving some procedures until there are documented risk factors and suggestive findings helps establish "medical necessity" for added assessment.

One good option would be to perform a structured or semistructured diagnostic interview, or at least the modules that are relevant to the diagnostic hypotheses for the particular case at hand. Structured interviews are more reliable and valid than unstructured clinical interviews, and they do a better job of detecting comorbid diagnoses if the full version is administered (Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009). However, they are not a panacea: They do not have perfect validity themselves, and they can take more time than unstructured interviews (Kraemer, 1992). Also, none of them include all possible diagnoses, and any given protocol may omit at least one diagnosis that might be common at a particular setting. Until the most recent version, for example, the Kiddie Schedule for Affective Disorders and Schizophrenia (Kaufman et al., 1997) did not include a module for pervasive developmental disorders; and many interviews designed for use with youths omit bipolar disorder, eating disorders, nonsuicidal self-injury, or other conditions that have become a concern since the interviews were written or validated.

Of interest, structured approaches may be more popular with clients than with the practitioners, who cite concerns about damaging rapport as well as loss of professional autonomy as objections to routine use of more structured approaches (Suppiger et al., 2009). Structured approaches may put more administrative burden on the clinician as well as taking more time with the client (Ebesutani, Bernstein, Chorpita, & Weisz, 2012). By placing semistructured approaches at Step 7, I advocate a "combined" approach, where we consider the findings from our setting (e.g., base rates), any risk factors that might modify initial hypotheses, and the results from any checklists or rating scales *before* beginning an interview. Although Step 7 sounds late in the process, it actually falls in the first 5 to 15 min of working with an individual case. Equipped with the context and data from the prior steps, it becomes possible to decide whether to change interviews or augment with other modules or tests to cover gaps in the default interview. It also might be possible to omit modules from a semistructured interview based on revised probabilities falling below the "wait-test" threshold, although the time savings will be modest if the interview already was structured to "skip out" after a few negative responses to screening questions.

Other strategies that make sense to invoke at this stage include any other procedure that has shown incremental validity for the question of interest (Johnston & Murray, 2003) but might be too expensive or burdensome to use more generally. Essentially, this stage is a "selected or targeted" zone of assessment, analogous to selected, secondary interventions in the parlance of the International Institute of Medicine and of community mental health (Mechanic, 1989). Neurocognitive testing, daily mood charting, and soon various forms of brain imaging all might fit in this category.

Clinical research agenda. The field has been doing a good job of validating assessment strategies. The next step needed is to evaluate these tools embedded in assessment sequences tailored for distinct settings. Test consumers should not accept the developers' descriptions of test performance uncritically but rather think about how characteristics in the target and comparison group affect test performance (Bossuyt et al., 2003; Youngstrom et al., 2006a).

8. Refine Assessment for Case Formulation, Treatment Planning, and Goal Setting

There are a large number of general medical conditions and medication-related side effects that can masquerade as psychological issues. These often are measured in haphazard fashion, rather than via structured review of systems. Similarly, there are many potential treatment targets or outcome modifiers—such as personality or temperament traits, school adjustment, family functioning, parental education level—that also could be valuable to assess as part of case conceptualization and treatment selection. As we learn more about moderators of outcome, and factors that make people better matches for some treatments than others, organizing assessment to rapidly evaluate these relevant moderators will be an excellent opportunity to integrate research and practice. Assessing quality of life and functioning also is pivotal in establishing treatment goals beyond symptom reduction (Frisch, 1998).

Clinical research agenda. Much more needs to be done in terms of systematizing the evaluation of treatment moderators and also "Axis III" factors (American Psychiatric Association, 2000), such as medications and general medical conditions that have psychological

effects. Here, the initial research can move from descriptive studies to examining these variables as moderators of treatment response or predictors of optimal treatment match.

9. Measure Processes ("Quizzes, Homework, and Dashboards")

Once treatments are started, then the role of assessment changes from diagnosis to monitoring treatment progress, including mediators, process variables, and outcomes. Sometimes the intervention itself will generate products that can be used for progress checks. Examples would include behavior tracking charts, reward calendars, daily report cards, three-column and five-column charts from cognitive-behavioral therapy, and daily mood charts (Youngstrom, 2008). Many aspects of functional behavioral analysis fit well in this context, too (Vollmer & Northup, 1996). Activities completed outside of the therapy session are frequently described as "homework" to promote skill generalization. Extending the metaphor, skill assessments during sessions could be likened to "quizzes" to evaluate learning. All of these can be ratcheted toward enhancing outcome by tracking and plotting them systematically (Cone, 2001; Powsner & Tufte, 1994). Weight loss programs all measure weight repeatedly, and they have demonstrated added value of written records of food consumption and exercise on producing greater and more lasting change (Grilo, Masheb, Wilson, Gueorguieva, & White, 2011). Process measurement is much more elaborated in psychological assessment than in most of EBM, which has concentrated on diagnosis, treatment selection, and likelihood of help versus harm as the primary assessment activities (Straus et al., 2011). If the patient is failing to progress as anticipated, and especially if there are complications, we should also use this as an opportunity to reassess our case formulation and diagnoses.

Clinical research agenda. Much could be done looking at human factors that promote the uptake of some tracking methods over others. Does a smartphone application improve utilization compared to pencil and paper (e.g., Chambliss et al., 2011)? Does better utilization lead to better outcome or more durable effects? Augmentation or dismantling studies, adding or subtracting different elements of process tracking, can be embedded within other trials or routine care at clinics, helping to identify what forms of tracking are most helpful. Another promising line of work would be examining how to package these assessments into "dashboards" that provide a clear summary of progress easily interpreted by family and therapist alike (Few, 2006; Powsner & Tufte, 1994).

10. Chart Progress and Outcome ("Midterm and Final Exams")

Continuing with the education metaphor, outcome evaluation can be cast as the "final exam," measuring the amount of change over the course of treatment. There are several operational definitions of outcome, including loss of diagnosis, percentage reduction of symptoms on a severity measure, or more complex definitions of "clinically significant change" that combine information about the precision of the measure—such as the "reliable change index"—with comparisons to normative benchmarks based on distributions in clinical and nonclinical samples (Jacobson & Truax, 1991). All of these involve more lengthy and comprehensive evaluation than the "process" measures just described, and so these panels of assessment methods are used more episodically. In clinical practice, outcome evaluation is more likely to be informal, based on the view that it is obvious when people are improving, and the belief that clients and payers will not accept the additional assessment involved (Suppiger et al., 2009). Contrary to expectation, clients are likely to view thorough assessments positively (Suppiger et al., 2009), and payers are more likely to reimburse assessments that are clearly linked to treatment (Cashel, 2002). Services databases consistently show modest rates of improvement and great heterogeneity in outcomes for treatment as usual, with some cases improving markedly, and others actually deteriorating. Meehl and others have argued that the slow progress in psychological treatment is due in large part to our failure to measure outcomes and get corrective feedback about when our interventions help, are inert, or even harm (Christensen & Jacobson, 1994; Meehl, 1973).

Research about patterns of treatment response also indicates potential value in having a scheduled "midterm," where more intensive evaluation is done to quantify early response to treatment. Early response to intervention, both psychotherapy and pharmacological (Curry et al., 2011), often predicts long-term response (Howard, Moras, Brill, Martinovich, & Lutz, 1996). If a person does not show improvement over the first 4 to 8 weeks or sessions, then it makes sense to either augment or change the modality of treatment (Lambert, Hansen, & Finch, 2001). Careful assessment of early response is also crucial to monitoring side effects and potential treatment-emergent changes in mood or behavior that should trigger alterations in the treatment plan (Joseph, Youngstrom, & Soares, 2009). Outcome evaluation is another area where psychological assessment has developed more sophisticated models for evaluating individual change compared to the metrics commonly used in EBM. Number needed to treat (the number of people who would need exposure to the treatment for

one more case to have a good outcome), number needed to harm (the number of people who would need exposure to the treatment for one more case to experience harmful side effects or iatrogenic outcomes), and similar indices are all measures of probabilistic efficacy based on groups of cases and dichotomous outcomes (Guyatt & Rennie, 2002). Psychological assessment offers much in terms of benchmarking against typical ranges of functioning, looking at change on continuous measures, and considering the precision of measurement when evaluating individual outcomes.

Clinical research agenda. There are a variety of methods worth investigating, including trials examining whether the addition of assessment at the "midterm" or end of acute treatment changes engagement, adherence, and acute or long-term outcomes (e.g., Ogles, Melendez, Davis, & Lunnen, 2001). A second line of work could optimize instruments for outcome evaluation by demonstrating sensitivity to treatment effects, developing shorter versions that retain sufficient precision to guide individual treatment decisions, and establishing meaningful benchmarks for "clinically significant change" approaches.

11. Monitor Maintenance and Relapses

Many disorders of childhood and adolescence carry a high risk of relapse, such as mood disorders; others are associated with an elevated risk of developing later pathology, perhaps as forms of heterotypic continuity. Anxiety often augurs later depression (Mineka, Watson, & Clark, 1998), and ADHD often presages substance issues or conduct problems (Taurines et al., 2010). More could be done in terms of educating families around signs of relapse or cues of early onset of later problems. Creative work is being done with mood disorders, helping patients identify signs of "roughening" and changes in energy or behavior that might offer early warning of relapse (Sachs, 2004), and then planning ahead of time for strategies that can help restabilize mood or promote earlier intervention to minimize the effects of recurrence. Given what we know about the epidemiology of mental health problems and developmental changes through adolescence and early adulthood, a combination of general screening and brief, targeted evaluations of warning signs could accomplish much good. This aspect of assessment has not received much attention from either the EBM or psychological assessment traditions yet, and represents a major growth area.

Clinical research agenda. It would be intriguing to evaluate how customized assessment strategies might predict shorter lag to seeking treatment, increased utilization of prevention or early intervention services, or diversion from more acute and tertiary treatments. Similarly, it would be important to know whether brief, broad coverage measures might have a role in primary care or other settings as predictors of relapse or progression in youths who have previously benefitted from treatment. Advances in technology make a variety of "smart" applications feasible as methods for monitoring behavior for cues of relapse.

12. Solicit and Integrate Patient Preferences

The placement of the wait-test and treat-test thresholds is flexible in EBM (Straus et al., 2011) (see also Figure 2). Their location is supposed to be guided by the costs and benefits attached to the diagnosis or treatment, as well as patient preferences. For dichotomous outcomes, such as recovery or remission, there is a developed framework combining the number needed to treat with the number needed to harm, yielding a Likelihood of Help versus Harm that can be further adjusted based on patient preferences (Straus et al., 2011). There are other formal mathematical approaches to synthesizing costs, benefits, and assessment parameters to optimize decision thresholds (Kraemer, 1992; Swets, Dawes, & Monahan, 2000), too. The EBM approach is attractive because it is simple enough that it could be done in session with families, potentially working through several "what if..." scenarios together to help explore a range of options and guide consensual decisions.

There is a rich layer of additional information that could be added here, using surveys and interviews to solicit beliefs about causes of emotional and behavioral problems, differences in what is perceived as problematic, and attitudes toward help-seeking and different services. Beliefs about medication and therapy have great influence over treatment seeking and engagement (Yeh et al., 2005). The effects of culture on decisions to seek or continue treatment are likely to be as big or bigger than culture's moderating effects on the accuracy of assessments or intervention efficacy. This aspect of assessment is one of the most promising places to combine psychological assessment's sophistication about measuring beliefs, attitudes, and preferences with the mathematical framework and decision aids offered by EBM.

Clinical research agenda. Qualitative methods as well as quantitative interviews and surveys have much to add in terms of knowledge about patient preferences. There also is a great deal that could be done integrating preferences into the decision-making framework, adjusting the test score thresholds for screening programs at a policy level (Swets et al., 2000) or negotiating personalized decision making with individual cases (Straus et al.,

2011). The algorithms have been available for decades, but it is only recently that technology has made it convenient for families and practitioners to use the tools. Recent developments understanding the role of culture in service selection, stigma, and attitudes to treatment also provides more rich inputs into the decision-making process (Hinshaw & Cicchetti, 2000; Yeh et al., 2005). Although last in the ''steps'' listed here, understanding patient attitudes is something we could profitably weave through the entire assessment process.

DISCUSSION

When it convened more than a dozen years ago, the Psychological Assessment Work Group of the American Psychological Association concluded there was surprisingly little published data to document the value of conventional psychological assessment in terms of better outcomes (Eisman et al., 1998; Meyer et al., 1998). The situation has improved only modestly in subsequent years (Hunsley & Mash, 2007). Our failure to measure things that matter to families and for treatment still contributes to the slow progress of our interventions (Meehl, 1973; Nelson-Gray, 2003).

EBM lacks the psychometric sophistication that has characterized the best traditions of psychological assessment. Psychological assessment has developed a wide range of instruments, and psychometric models could provide sophisticated techniques for honing the analytical underpinnings of EBM (Borsboom, 2008). What EBM offers, though, is a pragmatic focus on understanding and helping the individual case. EBM ties assessment to clinical decision making with a directness and clarity that has been missing in much of psychological assessment. Integration is possible, keeping the psychometric and conceptual strengths of psychological assessment but incorporating them into the decision-making framework articulated in EBM. The fit is not seamless, but it is patient centered, clinically relevant, and compelling. Some of the looser connections will be promising areas of investigation in their own right. EBM has historically emphasized dichotomous outcomes (e.g., recovery, death), whereas psychology has focused more on continuous measures. It is possible to convert dimensional effect sizes, such as Cohen's *d* or a correlation coefficient, into other effect sizes such as risk ratios (Hasselbad & Hedges, 1995), making it possible to reexpress outcomes in metrics that fit within the EBM decision-making framework, but it also would be intriguing to develop parallel approaches that capitalize on the greater information intrinsic to continuous measures.

Exploring the potential for synthesis reorganized my approach to assessment research, teaching, and supervision. Viewing assessment through an EBM tinted lens defines a set of clinical research topics that comprise a thematic program of investigation. The research designs and statistical methods are readily available and not complex. Adopting these methods need not add to the expense of the assessment process: Better decisions can be made by using the same tools but interpreting them differently. For example, we have found that there can be pronounced changes in clinical decisions about vignettes, with increased accuracy and consistency, and an elimination of a tendency to overdiagnose bipolar disorder, based on identical assessment data combined with brief training in the probability nomogram as a way of interpreting scores (Jenkins et al., 2011). The value of these methods is not limited to bipolar disorder, any more than it would be limited to any single area within medicine (Guyatt & Rennie, 2002). The hybridization of psychological assessment with EBM ideas produces ideas with vigor and clinical relevance to rejuvenate assessment and ultimately improve outcomes for families (Bauer, 2007).

REFERENCES

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implication of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*, 213–232. doi:10.1037/0033-2909.101.2.213

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles*. Burlington: University of Vermont.

Algorta, G. P., Youngstrom, E. A., Phelps, J., Jenkins, M. M., Youngstrom, J. K., & Findling, R. L. (2012). An inexpensive family index of risk for mood issues improves identification of pediatric bipolar disorder. *Psychological Assessment*. Advance online publication. doi:10.1037/a0029225

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

American Psychological Association. (2005). Policy statement on evidence-based practice in psychology. Retrieved from http://www.apa.org/practice/resources/evidence/evidence-based-statement.pdf

Aschenbrand, S. G., Angelosante, A. G., & Kendall, P. C. (2005). Discriminant validity and clinical utility of the CBCL with anxiety-disordered youth. *Journal of Clinical Child and Adolescent Psychology*, *34*, 735–746. doi:10.1207/s15374424jccp3404_15

Bauer, R. M. (2007). Evidence-based practice in psychology: implications for research and research training. *Journal of Clinical Psychology*, *63*, 685–694. doi:10.1002/jclp.20374

Bayes, T., & Price, R. (1763). An essay towards solving a problem in the doctrine of chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418. doi:10.1098/rstl.1763.0053

Belter, R. W., & Piotrowski, C. (2001). Current status of doctoral-level training in psychological testing. *Journal of Clinical Psychology*, *57*, 717–726.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425–440. doi:10.1007/s11336-006-1447-6

Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology*, *64*, 1089–1108. doi:10.1002/jclp.20503

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., ... de Vet, H. C. W. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *British Medical Journal*, *326*, 41–44. doi:10.1136/bmj.326.7379.41

Camara, W., Nathan, J., & Puente, A. (1998). *Psychological test usage in professional psychology: Report of the APA practice and science directorates* (p. 51). Washington, DC: American Psychological Association.

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*, 61–85. doi:10.1007/s11013-008-9120-4

Cashel, M. L. (2002). Child and adolescent psychological assessment: Current clinical practices and the impact of managed care. *Professional Psychology: Research and Practice*, *33*, 446–453. doi:10.1037//0735-7028.33.5.446

Chambliss, H. O., Huber, R. C., Finley, C. E., McDoniel, S. O., Kitzman-Ulrich, H., & Wilkinson, W. J. (2011). Computerized self-monitoring and technology-assisted feedback for weight loss with and without an enhanced behavioral component. *Patient Education and Counseling*, *85*, 375–382. doi:10.1016/j.pec.2010.12.024

Chen, W. J., Faraone, S. V., Biederman, J., & Tsuang, M. T. (1994). Diagnostic accuracy of the Child Behavior Checklist scales for attention-deficit hyperactivity disorder: A receiver-operating characteristic analysis. *Journal of Consulting and Clinical Psychology*, *62*, 1017–1025. doi:10.1037/0022-006X.62.5.1017

Childs, R. A., & Eyde, L. D. (2002). Assessment training in clinical psychology doctoral programs: what should we teach? What do we teach? *Journal of Personality Assessment*, *78*, 130–144. doi:10.1207/S15327752JPA7801_08

Christensen, A., & Jacobson, N. S. (1994). Who (or what) can do psychotherapy: The status and challenge of nonprofessional therapies. *Psychological Science*, *5*, 8–14. doi:10.1111/j.1467-9280.1994.tb00606.x

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Cone, J. D. (2001). *Evaluating outcomes: Empirical tools for effective practice*. Washington, DC: American Psychological Association.

Correll, C. U. (2008). Antipsychotic use in children and adolescents: Minimizing adverse effects to maximize outcomes. *Journal of the American Academy of Child & Adolescent Psychiatry*, *47*, 9–20. doi:10.1097/chi.0b013e31815b5cb1

Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine*, *78*, 775–780. doi:10.1097/00001888-200308000-00003

Curry, J., Silva, S., Rohde, P., Ginsburg, G., Kratochvil, C., Simons, A., ... March, J. (2011). Recovery and recurrence following treatment for adolescent major depression. *Archives of General Psychiatry*, *68*, 263–269. doi:10.1001/archgenpsychiatry.2010.150

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

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin*, *131*, 483–509. doi:10.1037/0033-2909.131.4.483

Ebesutani, C., Bernstein, A., Chorpita, B. F., & Weisz, J. R. (2012). A transportable assessment protocol for prescribing youth psychosocial treatments in real-world settings: Reducing assessment burden via self-report scales. *Psychological Assessment*, *24*, 141–155. doi:10.1037/a0025176

Eisman, E. J., Dies, R. R., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., ... Moreland, K. L. (1998). *Problems and limitations in the use of psychological assessment in contemporary health care delivery: Report of the Board of Professional Affairs Psychological Assessment Workgroup, Part II* (p. 22). Washington, DC: American Psychological Association.

Few, S. (2006). *Information dashboard design: The effective visual communication of data*. Cambridge, MA: O'Reilly Press.

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*, 506–522. doi:10.1207/s15374424jccp3403_7

Frazier, T. W., & Youngstrom, E. A. (2006). Evidence-based assessment of attention-deficit/hyperactivity disorder: Using multiple sources of information. *Journal of the American Academy of Child & Adolescent Psychiatry*, *45*, 614–620. doi:10.1097/01.chi.0000196597.09103.25

Frisch, M. B. (1998). Quality of life therapy and assessment in health care. *Clinical Psychology: Science and Practice*, *5*, 19–40.

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

Garland, A. F., Lewczyk-Boxmeyer, C. M., Gabayan, E. N., & Hawley, K. M. (2004). Multiple stakeholder agreement on desired outcomes for adolescents' mental health services. *Psychiatric Services*, *55*, 671–676. doi:10.1176/appi.ps.55.6.671

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650–669. doi:10.1037/0033-295X.103.4.650

Glutting, J. J., Youngstrom, E. A., Ward, T., Ward, S., & Hale, R. (1997). Incremental efficacy of WISC–III factor scores in predicting achievement: What do they tell us? *Psychological Assessment*, *9*, 295–301. doi:10.1037/1040-3590.9.3.295

Gray, G. E. (2004). *Evidence-based psychiatry*. Washington, DC: American Psychiatric Publishing.

Grilo, C. M., Masheb, R. M., Wilson, G. T., Gueorguieva, R., & White, M. A. (2011). Cognitive-behavioral therapy, behavioral weight loss, and sequential treatment for obese patients with binge-eating disorder: A randomized controlled trial. *Journal of Consulting & Clinical Psychology*, *79*, 675–685. doi:10.1037/a0025049

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

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, *148*, 839–843.

Harkness, A. R., & Lilienfeld, S. O. (1997). Individual differences science for treatment planning: Personality traits. *Psychological Assessment*, *9*, 349–360.

Hasselbad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin*, *117*, 167–178. doi:10.1037/0033-2909.117.1.167

Hayes, S. C., Nelson, R. O., & Jarrett, R. B. (1987). The treatment utility of assessment: A functional approach to evaluating assessment quality. *American Psychologist*, *42*, 963–974.

Hinshaw, S. P., & Cicchetti, D. (2000). Stigma and mental disorder: Conceptions of illness, public attitudes, personal disclosure, and social policy. *Development & Psychopathology*, *12*, 555–598. doi:10.1017/S0954579400004028

Hodgins, S., Faucher, B., Zarac, A., & Ellenbogen, M. (2002). Children of parents with bipolar disorder. A population at high risk for major affective disorders. *Child & Adolescent Psychiatric Clinics of North America*, *11*, 533–553.

Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy: Efficacy, effectiveness, and

patient progress. *American Psychologist*, *51*, 1059–1064. doi:10.1037/0003-066X.51.10.1059

Hummel, T. J. (1999). The usefulness of tests in clinical decisions. In J. W. Lichtenberg & R. K. Goodyear (Eds.), *Scientist-practitioner perspectives on test interpretation* (pp. 59–112). Boston, MA: Allyn and Bacon.

Hunsley, J., & Mash, E. J. (2005). Introduction to the special section on developing guidelines for the evidence-based assessment (EBA) of adult disorders. *Psychological Assessment*, *17*, 251–255. doi:10.1037/1040-3590.17.3.251

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology*, *3*, 29–51. doi:10.1146/annurev.clinpsy.3.022806.091419

Hunsley, J., & Mash, E. J. (Eds.). (2008). *A guide to assessments that work*. New York, NY: Oxford University Press.

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment*, *15*, 446–455.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19. doi:10.1037/0022-006X.59.1.12

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: B: What are the results and will they help me in caring for my patients? *Journal of the American Medical Association*, *271*, 703–707.

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*, 121–129. doi:10.1037/a0022506

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

Johnston, C., & Murray, C. (2003). Incremental validity in the psychological assessment of children and adolescents. *Psychological Assessment*, *15*, 496–507.

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

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., ... Ryan, N. (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*, 980–988. doi:10.1097/00004583-199707000-00021

Kelley, T. L. (1927). *Interpretation of educational measurements*. Yonkers, NY: World Books.

Kovacs, M. (1992). *Children's Depression Inventory Manual*. North Tonawanda, NY: Multi-Health Systems.

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

Kraemer, H. C., Kazdin, A. E., Offord, D. R., Kessler, R. C., Jensen, P. S., & Kupfer, D. J. (1999). Measuring the potency of risk factors for clinical or policy significance. *Psychological Methods*, *4*, 257–271.

Krishnamurthy, R., VandeCreek, L., Kaslow, N. J., Tazeau, Y. N., Miville, M. L., Kerns, R., ... Benton, S. A. (2004). Achieving competency in psychological assessment: directions for education and training. *Journal of Clinical Psychology*, *60*, 725–739. doi:10.1002/jclp.20010

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: using patient outcome data to enhance treatment effects. *Journal of Consulting & Clinical Psychology*, *69*, 159–172. doi:10.1037/0022-006X.69.2.159

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, *58*, 593–614. doi:10.1146/annurev.psych.58.110405.085542

Mash, E. J., & Barkley, R. A. (Eds.). (2007). *Assessment in children and adolescents*. New York, NY: Guilford.

Mash, E. J., & Hunsley, J. (2005). Evidence-based assessment of child and adolescent disorders: Issues and challenges. *Journal of Clinical Child and Adolescent Psychology*, *34*, 362–379. doi:10.1207/s15374424jccp3403_1

McFall, R. (1991). Manifesto for a science of clinical psychology. *The Clinical Psychologist*, *44*, 75–88.

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessment with signal detection theory. *Annual Review of Psychology*, *50*, 215–241. doi:10.1146/annurev.psych.50.1.215

Mechanic, D. (1989). *Mental health and social policy*. Englewood Cliffs, NJ: Prentice-Hall.

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

Meehl, P. (1973). Why I do not attend case conferences. In P. Meehl (Ed.), *Psychodiagnosis: Selected papers* (pp. 225–302). New York, NY: Norton.

Meehl, P. E. (1997). Credentialed persons, credentialed knowledge. *Clinical Psychology: Science and Practice*, *4*, 91–98. doi:10.1111/j.1468-2850.1997.tb00103.x

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *55*, 194–216.

Merenda, P. F. (2007a). Psychometrics and psychometricians in the 20th and 21st centuries: How it was in the 20th century and how it is now. *Perceptual & Motor Skills*, *104*, 3–20. doi:10.2466/pms.104.1.3-20

Merenda, P. F. (2007b). Update on the decline in the education and training in psychological measurement and assessment. *Psychological Reports*, *101*, 153–155. doi:10.2466/pr0.101.1.153-155

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749. doi:10.1037/0003-066X.50.9.741

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., Moreland, K. L., ... Dies, R. R. (1998). *Benefits and costs of psychological assessment in health care delivery: Report of the Board of Professional Affairs Psychological Assessment Workgroup, Part I* (p. 90). Washington, DC: American Psychological Association.

Meyer, G. J., & Handler, L. (1997). The ability of the Rorschach to predict subsequent outcome: A meta-analysis of the Rorschach Prognostic Rating Scale. *Journal of Personality Assessment*, *69*, 1–38. doi:10.1207/s15327752jpa6901_1

Mineka, S., Watson, D., & Clark, L. A. (1998). Comorbidity of anxiety and unipolar mood disorders. *Annual Review of Psychology*, *49*, 377–412. doi:10.1146/annurev.psych.49.1.377

Nelson-Gray, R. O. (2003). Treatment utility of psychological assessment. *Psychological Assessment*, *15*, 521–531.

Ogles, B. M., Melendez, G., Davis, D. C., & Lunnen, K. M. (2001). The Ohio Scales: Practical outcome assessment. *Journal of Child & Family Studies*, *10*, 199–212.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. New York, NY: Wiley.

Piotrowski, C. (1999). Assessment practices in the era of managed care: Current status and future directions. *Journal of Clinical Psychology*, *55*, 787–796.

Powsner, S. M., & Tufte, E. R. (1994). Graphical summary of patient status. *The Lancet*, *344*, 368–389. doi:10.1016/S0140-6736(94)91406-0

Ravens-Sieberer, U., & Bullinger, M. (1998). Assessing health-related quality of life in chronically ill children with the German KINDL: First psychometric and content analytic results. *Quality of Life Research*, *7*, 399–407.

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*, 169–184. doi:10.1002/mpr.289

Reynolds, C. R., & Kamphaus, R. (2004). *BASC–2 Behavior Assessment System for Children*. Circle Pines, MN: American Guidance Service.

Sachs, G. S. (2004). Strategies for improving treatment of bipolar disorder: Integration of measurement and management. *Acta Psychiatrica Scandinavica*, 7–17. doi:10.1111/j.1600-0447.2004.00409.x

Sattler, J. M. (2002). *Assessment of children: Behavioral and Clinical Applications* (4th ed.). La Mesa, CA: Publisher Inc.

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. doi:10.1177/0011000095233009

Spring, B. (2007). Evidence-based practice in clinical psychology: What it is, why it matters; What you need to know. *Journal of Clinical Psychology*, *63*, 611–631.

Stedman, J. M., Hatch, J. P., & Schoenfeld, L. S. (2001). The current status of psychological assessment training in graduate and professional schools. *Journal of Personality Assessment*, *77*, 398–407. doi:10.1207/S15327752JPA7703_02

Stedman, J. M., Hatch, J. P., Schoenfeld, L. S., & Keilin, W. G. (2005). The structure of internship training: Current patterns and implications for the future of clinical and counseling psychologists. *Professional Psychology: Research and Practice*, *36*, 3–8. doi:10.1037/0735-7028.36.1.3

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, NY: Churchill Livingstone.

Suppiger, A., In-Albon, T., Hendriksen, S., Hermann, E., Margraf, J., & Schneider, S. (2009). Acceptance of structured diagnostic interviews for mental disorders in clinical practice and research settings. *Behavior Therapy*, *40*, 272–279. doi:S0005-7894(08)00088-9 [pii]

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

Taurines, R., Schmitt, J., Renner, T., Conner, A. C., Warnke, A., & Romanos, M. (2010). Developmental comorbidity in attention-deficit/hyperactivity disorder. *Attention Deficit and Hyperactivity Disorders*, *2*, 267–289. doi:10.1007/s12402-010-0040-0

Vollmer, T. R., & Northup, J. (1996). Some implications of functional analysis for school psychology. *School Psychology Quarterly*, *11*, 76–92.

Wagner, K. D., Hirschfeld, R., Findling, R. L., Emslie, G. J., Gracious, B., & Reed, M. (2006). Validation of the mood disorder questionnaire for bipolar disorders in adolescents. *Journal of Clinical Psychiatry*, *67*, 827–830. doi:10.4088/JCP.v67n0518

Watkins, M. W. (2000). Cognitive profile analysis: A shared professional myth. *School Psychology Quarterly*, *15*, 465–479. doi:10.1037/h0088802

Wood, J. M., Nezworski, M. T., & Stejskal, W. J. (1996). The comprehensive system for the Rorschach: A critical examination. *Psychological Science*, *7*, 3–10. doi:10.1111/j.1467-9280.1996.tb00658.x

Yeh, M., Hough, R. L., Fakhry, F., McCabe, K. M., Lau, A. S., & Garland, A. F. (2005). Why bother with beliefs? Examining relationships between race/ethnicity, parental beliefs about causes of child problems, and mental health service use. *Journal Consulting and Clinical Psychology*, *73*, 800–807. doi:10.1037/0022-006X.73.5.800

Youngstrom, E. A. (2007). Pediatric bipolar disorder. In E. J. Mash & R. A. Barkley (Eds.), *Assessment of childhood disorders* (4th ed., pp. 253–304). New York, NY: Guilford.

Youngstrom, E. A. (2008). Evidence-based strategies for the assessment of developmental psychopathology: Measuring prediction, prescription, and process. In D. J. Miklowitz, W. E. Craighead, & L. Craighead (Eds.), *Developmental psychopathology* (pp. 34–77). New York, NY: Wiley.

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., Calabrese, J. R., Gracious, B. L., Demeter, C., DelPorto Bedoya, D., & Price, M. (2004). Comparing the diagnostic accuracy of six potential screening instruments for bipolar disorder in youths aged 5 to 17 years. *Journal of the American Academy of Child & Adolescent Psychiatry*, *43*, 847–858. doi:10.1097/01.chi.0000125091.35109.1e

Youngstrom, E. A., Findling, R. L., Danielson, C. K., & Calabrese, J. R. (2001). Discriminative validity of parent report of hypomanic and depressive symptoms on the General Behavior Inventory. *Psychological Assessment*, *13*, 267–276.

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*, 353–390. doi:10.1016/j.chc.2008.12.002

Youngstrom, E. A., Jenkins, M. M., Jensen-Doss, A., & Youngstrom, J. K. (2012). Evidence-based assessment strategies for pediatric bipolar disorder. *Israel Journal of Psychiatry & Related Sciences*, *49*, 15–27.

Youngstrom, E. A., & Kogos Youngstrom, J. (2005). Evidence-based assessment of pediatric bipolar disorder, Part 2: Incorporating information from behavior checklists. *Journal of the American Academy of Child & Adolescent Psychiatry*, *44*, 823–828. doi:10.1097/01.chi.0000164589.10200.a4

Youngstrom, E. A., Meyers, O. I., Youngstrom, J. K., Calabrese, J. R., & Findling, R. L. (2006a). Comparing the effects of sampling designs on the diagnostic accuracy of eight promising screening algorithms for pediatric bipolar disorder. *Biological Psychiatry*, *60*, 1013–1019. doi:10.1016/j.biopsych.2006.06.023

Youngstrom, E. A., Meyers, O., Youngstrom, J. K., Calabrese, J. R., & Findling, R. L. (2006b). Diagnostic and measurement issues in the assessment of pediatric bipolar disorder: Implications for understanding mood disorder across the life cycle. *Development and Psychopathology*, *18*, 989–1021. doi:10.1017/S0954579406060494

Youngstrom, E. A., Youngstrom, J. K., Freeman, A. J., De Los Reyes, A., Feeny, N. C., & Findling, R. L. (2011). Informants are not all equal: predictors and correlates of clinician judgments about caregiver and youth credibility. *Journal of Child and Adolescent Psychopharmacology*, *21*, 407–415. doi:10.1089/cap.2011.0032

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*, 223–233.

APPENDIX

Case Example

Referral Question: Tandi is a 10-year-old girl living with her biological parents and older sister who is coming for an outpatient evaluation because her mother is concerned about her increasing "mood swings." Tandi is in

regular education at a public school, taking accelerated classes. Her mother describes her as having been outgoing and cheerful as a child, but recently seems to have become more quiet, irritable, and crabby, sometimes snapping at her family, and recently slamming doors and throwing things. According to her mom, the paternal aunt has been diagnosed with bipolar disorder, and her mom has heard that this runs in families. She wants to know if Tandi has bipolar disorder.

Steps 1 & 2. Identify the Most Common Diagnoses and Presenting Problems in Our Setting, and Know the Base Rates of the Condition in Our Setting. The clinic where Tandi's family presented uses an electronic medical record, so it is possible to produce a report listing the most frequent diagnoses. The most common diagnosis is adjustment disorder (~60% of cases), followed by attention deficit/hyperactivity disorder (ADHD; 40%), oppositional defiant disorder (ODD; 35%), and major depressive disorder (30%, but lower in younger children and higher postpubertally). Posttraumatic stress disorder (PTSD), conduct disorder, and bipolar spectrum disorders are all diagnosed in roughly 10% of cases. The clinician has compared these rates with published rates from other outpatient settings and knows that the rank order seems plausible compared to external benchmarks. The somewhat higher rates of externalizing problems and lower rates of anxiety disorders reflect logical patterns in local referral sources. Based on this, bipolar disorder is worth assessing to address the referral question, but it is not a leading candidate. The clinic has stocked rating scales and assessment tools for all of the diagnoses that occur in 10% or more of cases, so the resources are available to explore bipolar disorder further if warranted.

Step 3. Evaluate the Relevant Risk and Moderating Factors *(Also Illustrating the Use of the Probability Nomogram).* Family history of bipolar disorder is a well-established risk factor, based on decades of research and multiple reviews. A clear diagnosis of bipolar in a first degree relative is associated with a diagnostic likelihood ratio (DLR) of 5.0, indicating a fivefold increase in the odds of the youth having a bipolar disorder (Youngstrom & Duax, 2005). A second-degree relative, such as the paternal aunt, will share on average half as many genes with the person being assessed, and thus confer half as much risk. The clinician asks the mother for more details about the aunt. Per mother's report, the aunt has been psychiatrically hospitalized twice and treated with lithium as well as an atypical antipsychotic—all details that support a bipolar diagnosis. Conceptually, the aunt's history is a "yellow flag" increasing the index of suspicion for bipolar disorder. The clinician asks the mother to complete the half-page Family Index of Risk for Mood (Algorta et al., 2012) as a way of gathering information about other relatives. The aunt is the closest relative clearly affected by mood disorder, although other relatives have histories of substance use or depression. The clinician uses the probability nomogram (Figure 1) to estimate how the family history changes the probability of a bipolar disorder for Tandi. The clinician begins by plotting the base rate of bipolar spectrum disorder at the clinic on the left hand line of the nomogram, placing a dot at the 10%. The aunt's history of bipolar disorder would have a DLR of 2.5 (or half of the 5.0 attached to a first degree relative having bipolar disorder). The 2.5 is plotted on the middle line of the nomogram. Connecting the dots and extending across the right hand line yields an estimate of ~24% for the new, "posterior" probability of bipolar disorder. If the clinician used an online calculator instead of the nomogram, then he or she would generate a probability of 22%, not very different. The FIRM score of 8 also has a DLR of 2.5; plugging that DLR into the nomogram would lead to a probability of ~22 to 24%. Note that the clinician does not treat the FIRM score and the aunt's diagnosis as separate pieces of information. Instead, the clinician either chooses to focus on the one that seems more valid or uses each separately to generate two probabilities that "bracket" Tandi's risk in a form of sensitivity analysis that examines how sensitive the estimates are to changes in the inputs. Here, both results are close together. Both also are above the clinician's wait-test threshold. More assessment is needed to decide whether bipolar is present or absent for Tandi.

Family history of bipolar disorder also increases the risk of depression, ADHD, and a variety of other conditions, typically with a DLR in the range of 1.5 to 3.0 based on a meta-analysis (Hodgins, Faucher, Zarac, & Ellenbogen, 2002). However, because it is Tandi's aunt, not a first-degree relative, the conferred risk would be half as high (falling in the 1.25 to 1.5 range). This is low enough that the clinician decides to concentrate on looking for more valid information rather than spending time combining these DLRs with the prior probabilities for the other diagnoses (Straus et al., 2011).

Step 4. Synthesize Broad Instruments into Revised Probability Estimates. Tandi's mother completed the Child Behavior Checklist (CBCL) as part of the core intake battery the clinic uses. The *T* scores are 63 for Externalizing, 67 for Internalizing, 70 for Anxious/Depressed, 67 for Withdrawn/Depressed, 51 for Attention Problems, 66 for Aggressive Behavior, and 53 for Rule Breaking. Impressionistically, the scores could be consistent with an adjustment disorder (which is still the leading hypothesis) or depression. The Externalizing scores look mild for ODD, and the low Attention Problems decreases suspicion of ADHD substantially. The low Rule Breaking

score also decreases the probability of conduct disorder, which already was uncommon at the clinic (base rate of 10%). The clinician considers conduct disorder "ruled out" unless there is new information that increases concern about it. Adjustment disorder, depression, ODD, ADHD, and bipolar are still the focus of assessment. The clinician does a PubMed search on "Child Behavior Checklist" AND "bipolar disorder" AND "sensitivity and specificity" and finds a paper that published DLRs for the CBCL Externalizing score compared to a semi-structured diagnostic criterion (Youngstrom et al., 2004). The *T* of 63 is actually in the low range for youths with bipolar disorder, and it is more than twice as likely for youth to score in this range if they do not have a bipolar diagnosis (DLR = 0.47). The clinician uses the probability of 24% (from Step 3) as the new starting point on the nomogram left hand line, and plots the DLR of 0.47 on the midline, producing a revised estimate of ~15%. If the clinician used a calculator instead for all of the steps, the probability estimate would be 12%. Using similar approaches, the clinician finds that the probability of depression is up to about 65%, ADHD is down to below 20%, and no information is readily available for predicting adjustment disorder with the CBCL. To this point, the clinician has neither added any extra assessment tools to the battery except the FIRM nor spent any additional time interviewing the family. The steps have made the list of hypotheses and the interpretation more systematic than would otherwise often be the case, and relying on base rates and published weights counteracts potential cognitive heuristics due to the family's description of the presenting problem.

Step 5. Add Narrow and Incremental Assessments to Clarify Diagnoses. Based on the current hypotheses and probability estimates, the clinician decides to add some mood rating scales evaluating both depressive and hypomanic/manic symptoms as well as gather a teacher report about Tandi's school functioning. The clinician opts for the Achenbach Teacher Report Form as a concise way of gathering data about attention problems (potentially ruling ADHD out if low, vs. indicating continued assessment if high) as well as the degree of pervasiveness of the aggressive behaviors (helpful for the ODD hypothesis). The literature suggests that the teacher report of mood symptoms is unlikely to be helpful for differential diagnosis but could be helpful for treatment planning.

The clinician has Tandi complete the Child Depression Inventory (CDI; Kovacs, 1992) and the Mood Disorder Questionnaire (MDQ; Wagner et al., 2006), which has the easiest reading level of the hypomania/mania rating scales having published data with youths (Youngstrom, 2007). The clinician asks the mom to complete the Parent General Behavior Inventory, which asks about both depressive and hypomanic symptoms (PGBI; Youngstrom, Findling, Danielson, & Calabrese, 2001). Because the mother is specifically concerned about the possibility of bipolar disorder, the clinician and mother agree to have her do the full-length version rather than one of the abbreviated ones, to provide the most comprehensive description even though there is no statistical advantage of the longer versus shorter versions. Mom's scores for Tandi on the PGBI are 16 on the Hypomanic/Biphasic Scale (28 items) and 39 on the Depression Scale (46 items). The Hypomanic/Biphasic score falls in the low range for bipolar disorder, with a DLR of .46. Using the nomogram, this reduces the probability of a bipolar disorder to ~ 7%. Tandi's scores come back moderately high on the CDI and below threshold on the MDQ. Using the sensitivity (38%) and specificity (74%) published by Wagner et al. (2006) yields a DLR of 0.84. This is close enough to 1.0 that the clinician could ignore it rather than feeding it into the nomogram or a calculator; impressionistically, it is revising the low probability of bipolar disorder to become slightly lower still. The scores on the CDI and PGBI Depression are both suggestive of depression, raising the probability to ~85%.

Step 6. Interpret Cross-Informant Data Patterns. The Teacher Report Form (TRF) comes back with all scores below a *T* of 60. Tandi's grades have been good (all 3 s and 4 s on a 4-point scale). The low score on Attention Problems from the teacher, combined with the other assessment data, reduces the probability of ADHD below 5%. The clinician considers it functionally ruled out, based on the probability and the absence of any "red flags" in the academic record. The low scores do not change the probability of a mood disorder. They slightly reduce the chances of ODD. Tandi's high self-report of depressive symptoms is consistent with her mom's report of internalizing concerns, suggesting that Tandi may be motivated for treatment working on internalizing issues.

Step 7. Finalize Diagnoses by Adding Necessary Intensive Assessment Methods. The clinician selects the depression module of the MINI as a brief, structured interview to formally cover the diagnostic criteria for major depression and dysthymic disorder, along with the ODD module. The clinician also asks about recent life events and potential stressors, looking for possible precipitants for an adjustment disorder. At this stage, the clinician also considers other rival hypotheses that could be consistent with the presentation. Before diagnosing depression, we are supposed to rule out the possibility of medication side effects or general medical conditions. The clinician explains the rationale for doing the interview and asks about medications, vitamins, or other

drugs that Tandi might be taking. Tandi has had regular pediatrician visits, and her health has been good. She is not taking any prescription medication, and to her mom's knowledge, neither her peer group nor her older sister's is using any illicit substances. The MINI results identify a sufficient number of symptoms and duration for a diagnosis of a major depressive episode, with impairment at home. The severity appears mild to moderate based on the rating scales as well as descriptions during the MINI and the clinician's observations of Tandi. Based on assessment findings, the clinician assigns a diagnosis of major depressive disorder, single episode, moderate severity. The ODD module does not pass threshold, and the clinician formulates the irritability as being a feature of the depression rather than a separate diagnostic issue.

Step 8. Refine Assessment for Treatment Planning and Goal Setting. Based on the information so far, depression seems to be a main concern. The CDI and CBCL Internalizing provide good baseline scores for severity of the problem. The clinician has charts indicating the number of points each measure needs to change to demonstrate improvement (Youngstrom, 2007), based on the reliable change index approach, as well as benchmarks for treatment targets for "clinically significant change" on those as primary outcome measures (Jacobson & Truax, 1991). The clinician supplements this with measures of quality of life to look at positive aspects of functioning (Frisch, 1998) and selects the KINDL as a brief, developmentally appropriate instrument with both parent- and youth-report forms available (Ravens-Sieberer & Bullinger, 1998). To help decide which therapeutic modality might be most helpful in reducing the depressive symptoms, the clinician considers Tandi's verbal ability educational level of the family, and cultural background, all of which suggest a good fit with cognitive behavioral or psychoeducational approaches. The clinician also decides to gather more information about family functioning to gauge the extent to which family dynamics and communication might helpful to address, perhaps indicating a greater emphasis on family-focused therapy.

Step 12. Solicit and Integrate Patient Preferences. As noted in the article, it makes sense to do "Step 12" whenever in the assessment sequence it would be helpful in making decisions about assessment or treatment. The clinician presents the initial formulation to the family, discussing how changes in Tandi's mood can offer a parsimonious explanation for the clinical picture emerging from the testing. During the discussion, the clinician is able to directly address the mother's concern about possible bipolar disorder, stating that the probability of bipolar disorder is currently quite low, and pointing to specific findings establishing the basis for that judgment. The clinician and family discuss several different options for treatment, ranging from "wait and see," through individual therapy for Tandi (involving supportive discussion combined with problem-solving and coping skills coaching), or family therapy, and antidepressant medication. Because no one in the immediate family has taken an antidepressant before, the clinician talks through the risks and benefits, providing the number needed to treat and the number needed to harm estimates for each approach. The family decides to try an approach combining some family psychoeducation with individual therapy for Tandi, holding the medication in abeyance because her depressive symptoms are still only mild to moderate, and thus the potential benefit seems lower compared to the potential for side effects and the family's hesitation about using medication.

Step 9. Measure Processes ("Dashboards, Quizzes and Homework"). Tandi and her mother download a mood charting app onto the mother's smartphone, and they use this to track both of their moods on a daily basis. This feeds directly into the mood monitoring and problem-solving skills that the clinician works to teach Tandi in individual sessions. The clinician also uses a sticker chart with Tandi to track the number of times each week that she tries new problem solving skills.

Step 10. Chart Progress and Outcome. In addition to regularly reviewing the mood charting and "homework" sticker chart, the clinician has Tandi and her mom repeat the CDI and CBCL after six sessions to see if there is measurable improvement on the primary outcomes. The family completes these a third time, along with repeating the quality of life measures, as they approach the termination session. The updated scores are compared to the "clinical significance" benchmarks as well as the baseline scores. Discussing the benchmarks helps the mother to reduce her sense of perfectionism, and allays her concerns that Tandi's moodiness might be a sign of bipolar disorder, by giving her a better appreciation for the behaviors that fall within typical functioning for Tandi's age.

Step 11. Monitor Maintenance and Relapse. During the termination session, the clinician and family review progress, celebrate their success, and plan for the future. This includes a discussion about the possibility of relapse. The clinician decides that this is important to discuss given the high rate of relapse for depression, and the fact that both early onset of depression and family history of mood disorder are risk factors that

increase Tandi's chances of remission. The clinician frames the potential for relapse as a possibility but emphasizes that Tandi and the family have mastered the skills to beat mood issues. The group discusses what would be warning signs of depression starting to recur, and they also make a list of situations that might increase stress and risk for relapse (such as getting a bad grade, losing a friend, getting very sick, or if the family were to relocate . . .). The list is framed as a set of ''reminders'' to check in on everyone's mood and coping when dealing with stressful situations. The clinician and mother also discuss warning signs that might raise concern about bipolar disorder, as both the family history and early onset suggest that if Tandi develops future mood issues, they are more likely to follow a bipolar spectrum course over the long term, even though she did not show signs of bipolar illness during this initial episode.