CLINICAL PSYCHOLOGY
SCIENCE AND PRACTICE

Evidence-Based Assessment as an Integrative Model for Applying Psychological Science to Guide the Voyage of Treatment

Eric A. Youngstrom, Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill
Anna Van Meter, Ferkauf Graduate School of Psychology, Yeshiva University
Thomas W. Frazier, Center for Autism, Cleveland Clinic
John Hunsley, School of Psychology, University of Ottawa
Mitchell J. Prinstein, Mian-Li Ong, and Jennifer K. Youngstrom, Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill

Evidence-based assessment (EBA) streamlines literature reviewing and organizing clinical assessment by targeting the vital few topics, "satisficing," and focusing on three major phases of clinical activity: prediction of diagnoses or other criteria, prescription of treatment or moderating factors, and process measurement. EBA is an organizing framework for applying a dozen steps to guide treatment. Technology is changing clinical assessment by increasing the efficiency and accuracy of scoring and feedback, as well as innovations that make more intensive assessment feasible. Fully implementing EBA suggests changes in training and requires a practice overhaul in exchange for greater efficiency, more accurate decisions, incrementally better outcomes, and increased service accessibility that could enable psychological science to help more people.

Address correspondence to Eric A. Youngstrom, Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, CB #3270, Davie Hall, Chapel Hill, NC 27599-3270. E-mail: eay@unc.edu.

doi:10.1111/cpsp.12207

Key words: **clinical decision-making, diagnosis, evidence-based assessment, outcome evaluation, psychological assessment.** ***[Clin Psychol Sci Prac 24: 331–363, 2017]***

The client is sitting in the office, looking emotionless as she describes in a flat voice why she is here. She broke up with her partner, she has been struggling at university, all she sees is failure; she has been contemplating suicide. She feels ready to leave this world.

What is the best way of understanding what brought her to this point? What are the key processes to address in order to keep her safe, to get her emotionally stable? How do we find the path toward a better-looking world that she could envision, move toward, and inhabit? Clinical assessment is supposed to provide these answers. The journey is a frequent metaphor for therapy. If clients are depressed, painfully anxious, or so impulsive that their clumsy social interactions make others fly away instead of gravitating closer, then the idea that therapy could take them to a place where they felt effective, and people liked them, sounds like travel to a different world.

But space travel also requires substantial preparation. This is akin to a comprehensive assessment. Research offers a body of knowledge about how things work, and data on how individuals might change in new circumstances. Research findings need to be combined with a practical understanding of where to aim the rocket, what to pack, how to chart progress, how to decide when you have landed safely, how to set up camp and get grounded for the longer term, and how to let people know of our arrival. Yet, this is rarely how clinical assessment is conducted. The metaphor exposes how clunky and dated many of our practices are, as well as how cumbersome the research base can be. Despite the pervasiveness of technology in other aspects of our lives, assessment still mostly uses paper-and-pencil measures or informal interviews. Major advances in psychological science are not well integrated in current practice. Getting the first person to the moon might take a library of studies and data, but we cannot lug the library along on every single voyage; rather, it is important to have a process for picking the essential tools for each individual journey.

Professional organizations, policymakers, and payers all advocate evidence-based practice (American Psychological Association, 2005; Hunsley, 2007). But what constitutes evidence that could guide clinical actions? Evidence-based medicine (EBM) is perhaps the most radical revisioning of the clinical enterprise, developing a set of values and skills to support the rapid search and appraisal of research to find information that directly guides the care of the individual patient (Straus, Glasziou, Richardson, & Haynes, 2011). It is a set of techniques to search the electronic library and identify key things that would change the trip or improve its chances of success. As a field, psychology focused first on evaluating treatments (Chambless & Hollon, 1998), and initially adopted a framework of listing "empirically validated" or "empirically supported" treatments. The focus was more on comparing rockets, less on customizing them for clients and particular mission objectives. More recent iterations have started to focus on teaching principles and skills to guide decision-making (Spring, 2007), and to extend the idea of empirical support to assessment (Hunsley & Mash, 2007). This is a welcome shift, as assessment is a foundational component of clinical psychology as a discipline, and a core competency that is also a distinct niche for psychology among health-care providers (Hunsley, 2007). At the same time, it is challenging to build a whitelist of empirically supported assessments, because context matters: Valid for what purpose? Validated in what group (Thompson & Vacha-Haase, 2000)? The current Standards for Educational and Psychological Testing emphasize that it is not the test itself that is valid or not; it is the inference drawn from the scores based on performance in a particular sample (American Educational Research Association, 2014). When working with youth, the age group becomes important because of the rapid change in cognition, verbal ability, meta-emotion, social skills, and relationships. A tool appropriate for a one age group will be too challenging for a younger one, and too easy, boring, or irrelevant for an older cohort. Collateral perspectives are also salient in couples work, and yet again in late life, as dementia and other cognitive impairment may arise. Gender and cultural identity also matter (Meyer et al., 2001).

In a clinical context, assessment is not an end of its own, but a means toward helping the client. What would be the essentials to pack for a client's successful voyage to that better world? A measure should get added to a battery only if its scores potentially address one of the three Ps (Youngstrom & Frazier, 2013) in anticipated scenarios: Does it *predict* an important criterion? Does it *prescribe* a type of treatment, or change in plan? Does it inform the *process* of working with the client, by measuring mediators, tracking progress, or documenting outcomes? Traditional psychometrics map to these roles, but the three Ps heuristic sharpens the pragmatic focus. Scores on an instrument could be highly reliable, but uninformative with regard to clinical care. Many measures may be promising clinically, but have little research that speaks directly to their application for clinical roles such as diagnosis or outcome measurement (Youngstrom & Van Meter, 2016). The heuristic organizes our thinking about assessment, our review of the literature, our choice of tools, and our interpretation of results. If the assessment helps pick the destination, plan the best vehicle and route for the journey, or provides a dashboard for tracking progress or vital processes, then it is worth adding to the

packing list. Otherwise, it is dead weight, adding to the expense and burden of the venture, and increasing the risk of failure. Our goal in this article is to apply an evidence-based assessment (EBA) model (Youngstrom, 2013) as a lens to constructively and critically evaluate some common methods, and to illustrate how less familiar methods may plug some gaps to help provide coverage that supports the clinical process from intake to termination. The EBA model provides a framework for reconnecting basic research in psychological science with clinical practice—it can be an integrative rather than polarizing approach. This article differs from prior work in that (a) it is the single most comprehensive treatment to date by our group, (b) it provides the most depth about the progress and process measures, and (c) it provides links to extensive supplemental material and open-teaching resources. We briefly review current practices, acknowledge the challenges in adopting a rigidly evidence-based approach, and then propose some tactics that promote a realistic synthesis into an integrated EBA approach, which we then detail from initial installation to treatment termination and long-term monitoring. We also include links at each section to online supplemental materials that include additional details, annotated bibliographies, and illustrations of application to cases.

CURRENT ROUTINE ASSESSMENT PRACTICES

Surveys of practitioners and training programs show high consistency in measures used (Table 1 in Youngstrom & Van Meter, 2016, compiles a half dozen surveys). The lists have been stable for decades, with a trend toward decreasing emphasis on sentence completion, the Bender-Gestalt test, and some projective techniques in clinical training programs (Ready & Veague, 2014). Given the long incumbency of many of these techniques, much research is available on each of them.

What is surprising is how little of that research focuses on specific clinical applications of assessment measures. It is as if we have spent decades learning how to build and use components (a radar, a booster, a bay . . . a Wechsler, a Bender, a Beck) without looking at whether or how they actually support the therapy mission. The scarcity worsens if we narrow the scope to focus on well-designed studies. Both the Cochrane Collaborative and IBM estimate that <2% of all published medical research is both valid and clinically relevant (Straus et al., 2011). EBM has developed sets of guidelines and criteria to be able to quickly evaluate the quality of design and reporting for studies of accuracy for diagnostic assessment (Bossuyt et al., 2003), and the most recent APA style guide includes checklists for improving the design and reporting of both research reports and meta-analyses (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008).

Such guidelines serve two ends: improving the quality of research, and also equipping the reader to rapidly triage reports and focus on the few that might actually influence care for a particular client. For their assessment handbook, Hunsley and Mash (2008) created a rubric for grading the evidence of reliability and validity for instrument scores. Chapter authors reported the information for scales in a consistent format across chapters, bringing a new degree of organization and clear terminology to the review of assessments. The consistent "look and feel" made it easy for readers to know where to find information. The structure also exposed gaps in the literature, where certain key parameters were understudied or unknown for a given tool in key groups or situations. We extend their rubric in Table 1, focused on norms, standardization, and reliability, and Table 2, focused on different aspects of validity, adding facets more tightly connected to the three Ps, such as discriminative validity or treatment sensitivity. We integrate ideas from the concept of "pragmatic measures," analogous to the effectiveness (versus efficacy) sphere in treatment research, emphasizing that the assessment should be important to stakeholders, present a low burden for respondents and staff, be sensitive to treatment, and be *actionable*—with results having clear implications for the next clinical action (Glasgow & Riley, 2013).

We also formally add the possibility of a result being "too good," creating an additional column in the conceptual matrix. There are at least two major ways of being too good: *intrinsic trade-offs*, where different parameters may be at odds, so that maximizing one characteristic impinges on the others; and *biased designs* inflating results. The intrinsic trade-offs may be unavoidable constraints, requiring thoughtful choices about which parameter is more important for the

Table 1. Rubric for evaluating norms and reliability for assessments (extending Hunsley & Mash, 2008; *indicates new construct or category)

	Rubric			
Criterion	**Adequate**	**Good**	**Excellent**	**Too Good***
Norms	*M* and *SD* for total score (and subscores if relevant) from a large, relevant clinical sample	*M* and *SD* for total score (and subscores if relevant) from multiple large, relevant samples, at least one clinical and one nonclinical	Same as "good," but must be from representative sample (i.e., random sampling, or matching to census data)	Not a concern
Internal consistency (Cronbach's alpha, split half, etc.)	Most evidence shows alpha values of 0.70–0.79	Most reported alphas 0.80–0.89	Most reported alphas ≥0.90	Alpha is also tied to scale length and content coverage—very high alphas may indicate that scale is longer than needed, or that it has a very narrow scope
Inter-rater reliability	Most evidence shows kappas of 0.60–0.74, or intraclass correlations of 0.70–0.79	Most reported kappas of 0.75–0.84, ICCs of 0.80–0.89	Most kappas ≥0.85, or ICCs ≥0.90	Very high levels of agreement often achieved by rerating from audio or transcript
Test–retest reliability (stability)	Most evidence shows test–retest correlations ≥0.70 over period of several days or weeks	Most evidence shows test–retest correlations ≥0.70 over period of several months	Most evidence shows test–retest correlations ≥0.70 over a year or longer	Key consideration is appropriate time interval; many constructs would not be stable for years at a time
*Repeatability	Bland–Altman (Bland & Altman, 1986) plots show small bias, and/or weak trends; coefficient of repeatability is tolerable compared to clinical benchmarks (Vaz, Falkmer, Passmore, Parsons, & Andreou, 2013)	Bland–Altman plots and corresponding regressions show no significant bias, and no significant trends; coefficient of repeatability is tolerable	Bland–Altman plots and corresponding regressions show no significant bias, and no significant trends; established for multiple studies; coefficient of repeatability is small enough that it is not clinically concerning	Not a concern

clinical task at hand. For a rocket, maximizing speed or range might require compensating decreases in the amount of living space or cargo carried. In psychometrics, internal consistency reliability provides a clear example of an intrinsic trade-off. Cronbach's alpha is probably the most widely reported coefficient in many areas of measurement (Ponterotto & Ruckdeschel, 2007). Scale length, expressed as the number of items, is part of the formula (Streiner, Norman, & Cairney, 2015). All else being equal, adding more items increases reliability. Greater reliability reaps benefits in terms of better precision (smaller standard errors). However, it also imposes costs of greater length, time, and burden. These work at direct cross-purposes with the goals of repeated measurement for the tracking process, and they also may reduce the measure's attractiveness to both clinicians and clients. Internal consistency also can be inflated by concentrating on a subset with more homogeneous content. This gimmick increases the coefficient at the cost of narrow coverage (Streiner et al., 2015). Would we have a better sense of our patient's status if we asked about her mood, her appetite, her sleep and cognition … or used items asking, does she feel sad? Down? Blue? Gloomy? Scores on the second scale have the shinier-looking alpha, but they are laser-focused on a single facet of a much more complex phenomenon. The scale with the broader coverage actually has higher content validity for the assessment of depression and is more likely to show incremental validity, despite having a lower alpha.

Biased designs are the more pernicious source of inflated estimates. Intraclass correlations (of which Pearson's *r* is a special case) are ratios of the between-case variance compared to variance attributable to other factors or errors. Designs that increase the between-case variance will produce larger coefficients than those selecting more homogeneous cases—the "restriction of range" problem (Streiner et al., 2015). Samples that combine severely distressed and well cases will yield different estimates than clinical samples where everyone

Table 2. Rubric for evaluating validity and utility (extending Hunsley & Mash, 2008; *indicates new construct or category)

Criterion	Adequate	Good	Excellent	*Too Excellent
Content validity	Test developers clearly defined domain and ensured representation of entire set of facets	Same as "adequate," plus all elements (items, instructions) evaluated by judges (experts or pilot participants)	Same as "good," plus multiple groups of judges and quantitative ratings	Not a problem; can point out that many measures do not cover all of the *DSM* criteria now
Construct validity (e.g., predictive, concurrent, convergent, and discriminant validity)	Some independently replicated evidence of construct validity	Bulk of independently replicated evidence shows multiple aspects of construct validity	Same as "good," plus evidence of incremental validity with respect to other clinical data	Not a problem
*Discriminative validity	Statistically significant discrimination in multiple samples; AUCs <0.6 under clinically realistic conditions (i.e., not comparing treatment seeking and healthy youth)	AUCs of 0.60 to <0.75 under clinically realistic conditions	AUCs of 0.75 to 0.90 under clinically realistic conditions	AUCs >0.90 should trigger careful evaluation of research design and comparison group; more likely to be biased than accurate estimate of clinical performance
*Prescriptive validity	Statistically significant accuracy at identifying a diagnosis with a well-specified matching intervention, or statistically significant moderator of treatment	Same as "adequate," with good kappa for diagnosis, or significant treatment moderation in more than one sample	Same as "good," with good kappa for diagnosis in more than one sample, or moderate effect size for treatment moderation	Not a problem with the measure or finding, per se, but high predictive validity may obviate need for other assessment components; compare on utility
Validity generalization	Some evidence supports use with either more than one specific demographic group or in more than one setting	Bulk of evidence supports use with either more than one specific demographic group or in multiple settings	Bulk of evidence supports use with either more than one specific demographic group AND in multiple settings	Not a problem
Treatment sensitivity	Some evidence of sensitivity to change over course of treatment	Independent replications show evidence of sensitivity to change over course of treatment	Same as "good," plus sensitive to change across different types of treatments	Not a problem
Clinical utility	After practical considerations (e.g., costs, respondent burden, ease of administration and scoring, availability of relevant benchmark scores, patient acceptability), assessment data are likely to be *clinically actionable**	Same as "adequate," plus published evidence that using the assessment data confers clinical benefit (e.g., better outcome, lower attrition, greater satisfaction), *in areas important to stakeholders***	Same as "good," plus independent replication	Not a problem

Note. **Pragmatic measures (Glasgow & Riley, 2013) place particular emphasis on being important to stakeholders, low in burden, clinically actionable, and sensitive to treatment effects.

has some degree of symptoms and impairment. The problem is not limited to reliability. Estimates of Cohen's *d*, or diagnostic accuracy measures (area under the curve, sensitivity, specificity) all vary depending on the average scores and variances in the target and comparison groups, and they also are susceptible to bias from other research design features. The reporting guidelines (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008; Bossuyt et al., 2003) and critical appraisal tools (Straus et al., 2011; Whiting et al., 2011) all warn about potential design defects, and meta-analyses show these can have large biasing effects (Youngstrom, Genzlinger, Egerton, & Van Meter, 2015). Test authors and publishers have a potential conflict of interest, where it is rational for them to present the best possible results for their product. Weaker designs that exaggerate coefficients are not fraudulent, but they are not describing performance in groups similar to the intended application. It would be prudent for researchers in addition to the publisher to evaluate designs (Buros, 1965). For all these reasons, it is worth thinking about whether a coefficient is "too good" for the task at hand. Alpha values in the 0.90s, rather than

eliciting a smile, should trigger reflection: Was the design flawed? Was the sample different in important ways from the patient with whom I am working now? Is the test optimized for a parameter less relevant for the clinical decision we are making? If the alpha is a 0.95, is there a shorter (or broader) alternative that would accomplish our goal better or more efficiently? With some coefficients, bigger is not always better. Item response theory and adaptive testing capitalize on this, optimizing item selection to deliver information efficiently (Streiner et al., 2015).

The Astronomical Challenge of Attaining EBA Perfection

When deciding how to conduct an assessment, from the clinician's point of view, there are three dimensions to consider: (a) the target content for the assessment, such as the diagnosis or construct of interest; (b) the function that the assessment could serve in the treatment context—which we organize into the three Ps; and (c) the different candidate assessments available. These dimensions could map out a gargantuan cube. Crossing the more than 365 diagnoses in the current *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5*; American Psychiatric Association, 2013) with the nine major parameters for norms, reliability, and validity in Hunsley and Mash's (2008) EBA tables, and with the 3,500+ measures listed in the Mental Measurements Yearbook (MMYB; http://buros.org/test-reviews-information) implies 11.5 million combinations (=365 * 9 * 3,500), a staggering number of permutations. One could argue that this is a conservative estimate. *DSM* does not consider all the psychosocial issues, personality traits, and interpersonal processes that could be within the scope of clinical assessment. There are additional psychometric parameters that could also be tracked, as well as considerations about utility (see Tables 1 and 2). The number of published tests far outnumbers what MMYB has reviewed. The Standardized Reporting of Diagnostic (STARD) assessments guidelines offer a 25-point checklist for critiquing the design and reporting of each assessment study (Bossuyt et al., 2003). There are competing measures available for many topics—more than 200 measures published for depression alone (Nezu, Ronan, Meadows, & McClure, 2000). Feel nauseated yet? . . .Vertigo? It looks like a task of astronomical proportions. The apparent enormity contributes to the implementation gap—if the problem is so enormous, where to start?

MAKING EBA MANAGEABLE

Luckily, our job is not to study the entire permutational expansion (see Youngstrom & Van Meter, 2016; Table 2, or Google "EBA Wikiversity"). Rather, our goal is to find "good enough" matches that could be used in practice. The pragmatic approach offered by EBM is to start with the problem, and combine it with the clinical objective to frame a potentially answerable clinical question that would guide care. With the problem and objective in focus, we can ask, "Do we have an assessment tool for that?" Is it valid for that purpose? Does it look appropriate for patients similar to the one with whom we are working (Straus et al., 2011)? If yes, then we can use it. If there are multiple contenders, then we can develop principled ways of comparing them and deciding which we should choose. If no assessment tool fits the bill, then we now have a clear definition of a clinical need for further research, a niche that needs filling.

Three strategies combine to bring the EBA task back down to Earth: (a) concentrating on test features most relevant to the task at hand, (b) focusing on the "vital few" clinical issues, and (c) "satisficing"—using a good enough method rather than hunting for the very best.

Concentrating on the Three Ps

Rather than comprehensively reviewing all psychometric features for all contending assessments, the three Ps provide a simplifying strategy. Select the clinical objective, and then focus on the few psychometric parameters that are most important for that objective. Table 3 crosses the traditional psychometric features with the *prediction*, *prescription*, and *process* phases of treatment, demoting parameters that are irrelevant and highlighting those that are core features. For example, discriminative validity—such as classifying cases into diagnoses or risk profiles—is the essential ingredient for clinical decisions at the prediction phase, and perhaps not relevant at the process phase.

Simplifying the scope also reduces the burden on the clinician. There are fewer things to juggle in memory; and it is possible to hone searches, leading to

Table 3. Crossing psychometric parameters with the three Ps of evidence-based assessment: Different properties are core features depending on the phase of assessment

Criterion	Prediction	Prescription	Process
Norms	Important for accurate prediction (cf. effects of distilled samples or healthy controls when estimating performance in clinical settings)	Not important for diagnostic criteria; nor for treatment moderators	Not relevant for idiographic measurement; but central to outcome benchmarking
Internal consistency (Cronbach's alpha, split half, etc.)	Not directly relevant	Not directly relevant	Could help improve precision for measures of change
Inter-rater reliability	For many assessments used at this stage, inter-rater reliability not a major facet	***Core feature.*** Inter-rater reliability is pivotal here for assigning diagnosis and treatment selection	Nomothetic approaches using checklists de-emphasize inter-rater reliability; idiographic assessment de-emphasizes
Test–retest reliability (stability)	Not directly relevant; prediction equations (especially prediction over time) already model this (albeit not always isolating component due to stability)	Not clinically relevant	Evidence of stability when not treated may provide a benchmark, but evidence of sensitivity to change in context of treatment much more directly relevant (see below)
Content validity	Accurate prediction may be possible using correlates and risk factors; thus content validity may be sufficient but not necessary	Treatment moderators may include demographic and clinical features not normally subsumed in psychometric measures	Important. The factor measured should be a core part of the construct for progress and outcome measures
Construct validity (e.g., predictive, concurrent, convergent, and discriminant validity)	Criterion validity more important: Good prediction could be achieved using other constructs and variables as predictors	***Core feature.*** The diagnosis should have established construct validity (e.g., Robins & Guze, 1970, framework)	***Core feature.*** Treatment sensitivity, and possibly predictive validity, are key features
Discriminative validity	***Core feature.*** Showing ability to improve classification or discriminate between trajectories	Conceptually important, but rarely measured structured diagnostic interviews or moderators. High rates of apparent comorbidity could be inflated by poor discriminative validity	Not directly relevant, unless using process measure to discriminate responder versus nonresponder
Prescriptive validity	Helpful, but diagnosis and formulation will need to be reviewed and finalized by clinician	***Core feature.*** This is the reason for establishing a diagnosis or a formulation. The level of granularity should be dictated by what moderates treatment or changes the prescription.	Secondary; if the person is not responding, then prescribes a re-evaluation of formulation
Convergent validity	Indirectly connected. Could be a high correlation with a criterion. This could be converted to an AUC or other measure of discrimination	Indirectly connected. Semistructured interviews that use the diagnostic criteria are formative assessments	Indirectly connected
Validity generalization	Valuable addition. Important for identifying moderators that would change application to individual client	Complicated. Diagnostic nosologies assume that same definitions mostly are invariant globally, and using consistent definition reveals differences in prevalence.	Rarely available, but valuable addition
Treatment sensitivity	Not directly relevant. Risk factors, correlates, and predictors may or may not be mutable. However, large Cohen's *d* separating clinical and nonclinical distributions would be desirable feature for ROC applications, as well as Jacobson benchmarks for clinically significant change	Not relevant. Treatment moderators may not be mutable. Loss of diagnosis could be a way of defining outcome, but alternate approaches will usually be more practical	***Core feature.*** Needs to be quantified different ways for progress and process measures (slope measures in mixed regressions, or generalizability facets of variance attributable to treatment) than for midterm and endpoint evaluations (where standard error of difference and normative benchmarks are key to making client-level change measures)
Clinical utility	Important if clinicians are going to use	Important if clinicians are going to use	Important if clinicians are going to use

Note. Reproduced from Youngstrom and Van Meter (2016).
Core feature denotes a psychometric property that is most important in the given assessment phase.

fewer hits and faster answers. When many instruments are available, comparing them on the key feature makes it easier to choose. It also reminds us that different tools may be better in different situations; rather than quality being an intrinsic feature of an instrument, it is always bound by context (American Educational Research Association, 2014). Is a hammer "valid"? It depends on whether the rocket needs work done on a nail, a rivet, or delicate electronics.

Focusing on "The Vital Few"

The second strategy focuses on the most common scenarios and prepares assessments to address them. Vilfredo Pareto coined the "Law of the Vital Few," also known as the 80:20 rule. As a first approximation, the roughly 20% of disorders that are most common account for approximately 80% of the cases at a clinic. Addressing common issues first offers the greatest return on effort invested because the work will apply to the largest number of cases. When the system is well tuned for them, then it becomes possible to gradually expand the coverage to address the specific needs of less common issues (Straus et al., 2011). This rule of thumb applies remarkably well: A review of more than 2,000 cases receiving services in Hawaii found that 89% had a primary diagnosis that had an identified empirically supported treatment (EST) available, and 70% of all their diagnoses—including comorbidities—matched to an EST option (Schiffman, Becker, & Daleiden, 2006). Building a core set of assessments and treatments for common issues will provide good coverage.

Identifying the "vital few" issues can take advantage of existing benchmarks (see Table 2 in Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009; Youngstrom, Choukas-Bradley, Calhoun, & Jensen-Doss, 2014). These are imperfect, as the clinics were not systematically sampled, and there is no information about referral patterns. Nevertheless, it is a start.

It is worth psychologists checking rates at their own clinics. Electronic medical records may make it easier to get a summary of the diagnoses, as it should be possible to have, in just a few seconds, annual summaries of the number of clients who were diagnosed with generalized anxiety disorder, panic disorder, major depression, insomnia, etc. Alternately, random sampling of charts can provide a good snapshot of services. Mapping to external benchmarks will take some effort. This need not be exact to inform the selection of treatment options. We can then compare the numbers from our clinic to external benchmarks. Epidemiological studies address the sampling weaknesses inherent in clinical data, but they also answer a slightly different question —"How common are these problems among youths in the community?" not "How common are the problems among families seeking psychological services?" Diagnostic rates may differ between clinical epidemiological studies for various reasons: Some problems are more impairing than others, and some may lead to legal involvement instead of mental health services, differing degrees of stigma, and personal and cultural attitudes toward conceptualizing the problem as a psychological issue, or psychological services as a valid option.

EBM purists will scoff or cringe at the idea of using case records to establish base rates (cf. Reynolds, 2016). Clinical diagnoses have imperfect validity, with kappas of 0.1–0.3 for agreement between clinical and research diagnoses (Rettew et al., 2009). However, we need to start somewhere, and we can calibrate our practice against external benchmarks. Similarities in diagnostic rates are reassuring, though not proof of accuracy. Discrepancies can be even more helpful: They should prompt reflection about whether there are real differences in the referral pattern, versus there being gaps in assessment coverage and practice leading to diagnostic errors. Both are certainly possible. Local rates can also reveal common referral questions that do not map neatly to formal diagnoses, such as dealing with parents' divorce ("adjustment disorder"?), conflict around developmental transitions such as dating or initiation of sexual activity ("parent–child conflict"?), or testing for giftedness or early kindergarten readiness (Weisz et al., 2011).

An alternate way of simplifying would be to focus on major dimensions of behavioral functioning. Factor analyses of major item sets in youths and adults tend to find fairly consistent structures, with one general factor of psychopathology, two underlying broad bands of internalizing and externalizing behaviors, and eight or more replicable factors underneath (Achenbach & Rescorla, 2001; Reynolds & Kamphaus, 2015) in a hierarchical or bifactor model (Snyder, Young, &

Hankin, 2016). A similar structure appears to roughly describe associations among *DSM* diagnoses in epidemiological data (e.g., Krueger, Chentsova-Dutton, Markon, Goldberg, & Ormel, 2003). Focusing only on one or two major dimensions would miss clinically important distinctions—between unipolar and bipolar mood disorder, or whether the behaviors were a response to trauma instead of endogenous processes or situational reinforcers, for instance—but fewer than a dozen would be enough to provide excellent coverage.

After listing our top diagnoses, constructs, and referral questions, and comparing them to external benchmarks, we are ready to match our assessments to our practice priorities. We can augment the list to include assessment of issues that are rare but so serious that they would require different management (Morrison, 2014). Suicide, being a threat to others, and abuse are examples of things that assessment should routinely evaluate because of their pathognomonic nature as well as legal and ethical obligations. Severity also guides decisions about whether things should be treatment targets. Milder conditions and presentations will tend to be more common, and may not cross the threshold to require treatment. Further clinical evaluation may also find rare but serious conditions, putting them on the radar even though they normally would not be a consideration. The vital few offer a menu of frequent clinical suggested itineraries—"favorite destinations" in the treatment GPS, but not the only ones possible.

Satisficing: Having a Good Enough Assessment Strategy

With our vital few problems clearly in focus, we now want to match our assessment tools to them and to the specific roles defined by the three Ps. To achieve good coverage of our practice, we might need assessment plans for 10 or more target problems; or if the clinic is highly specialized, then half a dozen problem areas might be enough to address the main focus of the clinic along with the common comorbidities or impostors that often masquerade as the specialty diagnosis (e.g., the similarity between anxiety and depression, or bipolar disorder and attention-deficit/hyperactivity disorder [ADHD]).

It is a lot of work to complete a systematic review, and yet more toil to arrive at a conclusion about what assessment methods and tools have the greatest validity for a particular construct. Pragmatically, "good enough" is a more attainable goal. Just as using any appropriate empirically supported treatment would be preferable to using something lacking empirical support, using assessment strategies with good evidence of validity would be an improvement over typical practice. We do not need the best rocket ship in the cosmos, just one that will get us safely to our chosen destination—particularly if finding the best ship would delay our journey.

It is important not to get sucked into the black hole of chasing perfect assessments: Although there may not be a tool that has demonstrated reliability and validity for your exact combination of problem and objective, there could be options that would be useful if adapted somewhat. For example, do you need to read it aloud to a young child? Or get it translated for a Spanish-speaking older adult patient? These would clearly be departures from the way that the instrument was developed and validated, and must be noted in any assessment report. They are likely to change the performance to some degree. However, if the test showed good validity under the other conditions, and there is no better alternative available, then the adapted administration may suffice. The research may also provide some guidance about the potential degree of bias introduced by the moderating factor (e.g., reading aloud, or translation). If the validity coefficient is medium or large, and the bias is small, it should still do the job. It is likely to be superior to flying blind or relying on clinical judgment alone. We should keep using a "good enough" measure until something beats it on both validity and cost (Youngstrom & Van Meter, 2016). This is not excusing slipshod practices (cf. Reynolds, 2016), but rather suggesting that implementing recommendations from handbooks and reviews can build a platform for rapid launch that can continue to be reused until an upgrade is available.

EBA IN PRACTICE: THE ARC OF EBA ACROSS TREATMENT

There are a dozen steps in evidence-based assessment, starting with preparatory work to lay the foundation and have the right tools in place, and ending with post-termination monitoring (see Table 4 for list). The steps have a typical sequence, but the order should be flexible and guided by clinical judgment and the

Table 4. Twelve steps in implementing evidence-based assessment in clinical evaluation

Assessment step	Rationale	Steps to put in practice
Preparatory Work Before Seeing Client		
A. Plan for most common issues in our setting	Think about what common clinical hypotheses are; have assessment methods ready that could identify each	Review our practice; select a sample of cases (six months, random draw from past year); make list of most common diagnoses and clinical issues
B. Benchmark base rates for issues	Know the "top ten" issues for our practice; have plan for detecting	Compare local base rate to external benchmarks—other practices and published rates
Prediction Phase		
C. Evaluate risk and protective factors and moderators	Risk factors can trigger more detailed assessment; protective factors (e.g., coping skills, parental warmth) and moderators may inform treatment planning	Make short checklist of key risk factors; make second list of factors that might change treatment selection or moderate outcome; develop plan for how to routinely assess them
D. Revise probabilities based on intake assessments	EBA approach gives appropriate weight to each finding; eliminates geographic and training variations in interpretation (Mulley & Wennberg, 2011); provides clear guidance about next action	Make a table matching assessment tools with common presenting problems. Identify gaps in coverage. Make cheat sheet with key information about assessment for each application
E. Gather collateral, cross-informant perspectives	Impairment across settings is core to establishing some diagnoses. Discriminative validity of caregiver, teacher, or youth report may differ for specific conditions. Collateral may be crucial if client is young, in denial, or using substances	Gather collateral information to revise case formulation; consider parent, spouse, roommate; also behavioral traces such as Facebook postings. Anticipate typical level of agreement
Prescription Phase		
F. Add focused, incremental assessments	Broad measures will detect problems; focused measures helpful for differentiation among competing hypotheses	Have follow-up tests available and criteria for when they should be used. Organize so that key information is easy to integrate
G. Add more intensive methods to finalize diagnoses and case formulation	If revised probability falls in the "assessment zone," what are evidence-based methods to confirm or rule out the diagnosis in question?	Do (semi-)structured interview or review checklist with client to confirm sufficient criteria; supplement with other methods as needed to cross treatment threshold. Cognitive ability, achievement, and diagnostic reading tests might be used here for addressing learning disability hypothesis
H. Assess for treatment plan and goal setting	Rule out general medical conditions, other medications; family functioning, quality of life, personality, work/school adjustment; comorbidities. Rating scales also can establish baseline severity benchmarks against which to measure treatment response	Develop systematic ways of screening for medical conditions and medication use. Assess family functioning, personality, comorbidity, SES, and other potential treatment moderators. Verbal ability might also be a treatment moderator, changing the appropriateness of cognitive versus behavioral components
X. Learn and use client preferences	Client and caregiver beliefs and attitudes affect treatment seeking and engagement, and are vital for discussing balance of risks and benefits of treatment	Assess client concordance with treatment plan; ask about cultural factors that might affect treatment plan and engagement; shared decision-making programs
Process Phase		
I. Goal setting: Milestones and outcomes ("midterm and final exams")	Select key targets for treatment; define interim and outcome goals. Periodically repeat main severity measures; if poor response, revisit diagnoses	Make cheat sheet with Jacobson and Truax (1991) benchmarks for measures routinely used; track homework, progress on skills
J. Progress and process measures ("dashboards, quizzes and homework")	Check learning and use of therapy skills; benchmark early treatment response—progress, or need to tweak treatment?	Track homework, session attendance, life charts, symptom check-ins at each visit, medication monitoring, therapy assignments, daily report cards
K. Wrapping up and maintaining gains	Excellent termination planning celebrates and consolidates gains, and plans for maintenance (e.g., knowing and managing anniversaries and trigger situations)	Develop list of key predictors, recommendations about next action if starting to worsen

Note. Steps use letters instead of numbers to reinforce the idea that there is not a strict order. Considering patient preferences and cultural factors is ideally infused throughout the assessment and treatment process. It is an x-factor for enhancing rapport and engagement; hence, we label it step "X" even though it is mentioned in the middle of the table. Adapted from Youngstrom and Van Meter (2016).

client's needs. The steps can loosely be parsed into chunks that correspond with the three Ps as a flexible scaffolding. There is a set of open-science, open-teaching pages on Wikiversity with additional material that illustrates the application of these ideas and gathers supporting information about measures and interpretation. Google "EBA Wikiversity" to find the home page. Links to specific sections are included in the electronic version of the article; and the site organization follows the outline of this section of the article.

Prelaunch Phase: Installing EBA Into a Practice Setting

Several tactics can help structure practice to use more EBA quickly. Using prevalence rate benchmarks to focus on the vital few diagnoses and clinical problems establishes the core set of constructs that will need to be assessed. First, therefore, the psychologist should evaluate the common assessments against the roles defined by crossing these vital few problems with the key three P functions.

Does the clinic already have tools in place to address the common patient problems? When considering options for evidence-based treatments, there are numerous websites that could be consulted to obtain details on possible psychological treatments and their level of research support. When we consider psychological assessment, however, no such sites exist. Therefore, when there are gaps in a clinic's coverage of assessment tools, the next step is to look for systematic reviews, meta-analyses, or expert recommendations about assessments for the poorly covered niches. In EBM, expert recommendations are lower-tier evidence (Straus et al., 2011), but the Cochrane systematic reviews do not have many entries related to psychological assessment, and most meta-analyses are organized around an instrument, not comparing different instruments for a particular niche. Some of the best options are special issues (Hunsley & Mash, 2005; Mash & Hunsley, 2005), edited handbooks on assessment (Hunsley & Mash, 2008; Mash & Barkley, 2007), and elements of the AACAP practice parameters (King & The Work Group on Quality Issues, 1997). The next best option would be to do focused electronic literature searches, using the clinical topic AND the assessment function as terms in combination to narrow the set of hits into a manageable number that also are more likely to answer the clinical question. Crossing searches with terms that narrow the focus to specific clinical functions slashes the number of hits down to a manageable size (cf. Table 2 in Youngstrom & Van Meter, 2016). Learning how to do a targeted search in order to update clinical-based practices in real time is a core skill for training and practice (Hunsley, 2007).

The time spent on the searches to identify the best assessment tools for diagnosis, progress monitoring, and outcome measurement for the common disorders in your setting is an investment in the success of future missions (i.e., better client outcomes). Building a library of effective tools will ultimately save time while making practice more efficient and effective.

Prediction Phase—Defining the Focus of the Mission

Identify Common Issues and Benchmark Base Rates (Steps A and B). The first phase of assessment involves making rapid decisions about contending hypotheses, and then deciding which to evaluate further to build a case formulation and a treatment plan. These hypotheses often include potential diagnoses, which organize constellations of information about associated features, prognosis, and matching treatments. Predictions also more broadly encompass things such as risk of self-harm or threatening others; and the methodology would extend to other predictions, such as level of risk in a forensic setting (e.g., Archer, Wheeler, & Vauter, 2016). Although our discussion focuses on diagnosis as an example, the methods are more general. The criterion could just as easily be latent growth trajectories, grade retention, or premature dropout from treatment, rather than diagnosis per se (cf. Sellbom & Hopwood, 2016). Benchmarking the base rates for the most frequent problems is the preamble, creating a shortlist of hypotheses that will be worth considering precisely because they are commonplace (Meehl, 1954). As a debiasing strategy, we should look for disconfirming evidence as well as confirmatory evidence. The top panel of Figure 1 illustrates a graphical way of viewing the common issues as leading initial hypotheses that warrant assessment.

Studies of clinical decision-making find that when we use unstructured interviews, we tend to formulate one hypothesis based on the presenting problem (usually in the first few minutes of the interview!), and then we do an excellent job of searching for

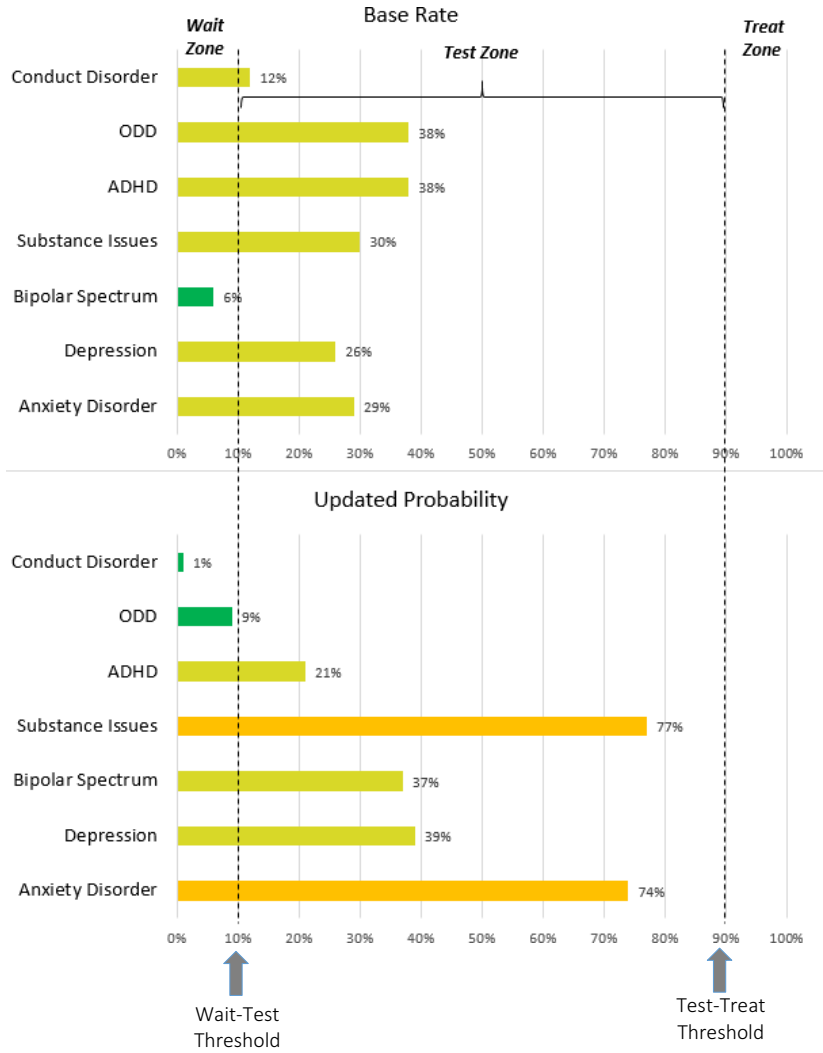

Figure 1. Using base rates and Bayesian updated probabilities to create a dashboard of probable clinical hypotheses and next actions. The Wait-Test and Test-Treat thresholds are concepts from evidence-based medicine (Straus et al., 2011). The exact location of the threshold depends on the risks and benefits involved, as well as patient preferences (Step X in the EBA model). The updated probabilities integrate the information from test results and risk factors, expressed as diagnostic likelihood ratios. Reproduced from Youngstrom (2016), http://ericyoungstrom.web.unc.edu/files/2016/02/EBA-support-materials-v1.0.docx

confirmatory data (Croskerry, 2003). We tend not to look for disconfirming evidence, and we rarely consider competing or augmenting hypotheses (Garb, 1998). These dynamics play into our tendency to underestimate comorbidity and to have "favorite" diagnoses that we identify at high rates (Jenkins & Youngstrom, 2016). Cognitive heuristics can be particularly error-prone when working with minority groups, who may use different language to describe the presenting problem—leading to a different starting hypothesis (see example in Wikiversity EBA page about base rates). The benchmarks remind us that these disorders are equally common in both demographic groups and deserve equal initial consideration (Alegria et al., 2008).

EBM Decision-Making and Zones of Clinical Action. Prevalence rates for the common local

diagnoses provide a set of starting diagnostic probabilities (i.e., if the base rate of depression is 30%, absent specific assessment data, any new patient has a 30% chance of meeting criteria for depression). Whereas probability can range anywhere from 0 to 100%, there basically are three clinical actions with regard to any hypothesis, separated by two thresholds (Straus et al., 2011; see also Figure 1). The Wait-Test threshold demarcates when a probability is low enough to consider a hypothesis ruled out—sufficiently unlikely that it is not worth considering further unless new information later raises the probability back above the threshold. Probabilities above the Wait-Test threshold get active consideration and are the main focus of the evaluation process. If new information raises the probability yet higher, then it may cross the second Test-Treat threshold, above which the hypothesis is considered ruled in and becomes a main focus of acute intervention. The threshold location depends on the risks and benefits attached to the diagnosis and the treatment, and also should consider patient preference. If the treatment is risky, or the patient has reservations, then the bar is higher for starting treatment. Conversely, if the treatment is low risk and inexpensive, then the treatment threshold could be so low that assessment becomes unnecessary, as is the case with vaccination or iodization of table salt to prevent thyroid problems. If the risks associated with *not* treating are high (e.g., suicidality, psychosis), then the threshold for starting treatment may be low, whereas less serious problems might lend themselves more to a "wait-and-see" approach (see Wikiversity EBA prediction phase overview for elaboration about zones of clinical action).

Some of the base rates will be high enough to start in the Assessment Zone (Figure 1). At typical outpatient clinics, the base rates of ADHD, disruptive behavior, anxiety, and mood disturbance will be high enough that they routinely should be considered in evaluation. The probabilities define routine evaluation targets. A core battery should match the measures to the Assessment Zone issues. If we use a semistructured interview, we want to make sure that the modules cover the Assessment Zone topics, as well as less common ones that could get promoted to Assessment Zone levels of probability via screening or identification of risk factors.

Routine Evaluation (Steps C, D, and E). The next layer of assessment consists of brief key factors from developmental history (Step C), screens and broad measures (Step D), and gathering information from collateral informants' perspectives (Step E). The screening measures can include instruments with broad content coverage, such as the Achenbach checklists for both youth and adults (Achenbach & Rescorla, 2001, 2003), the Strengths and Difficulties Questionnaire (SDQ; Goodman, Ford, Simmons, Gatward, & Meltzer, 2003), or the Symptom Checklist-90 (Derogatis, 1977). Their subscales address symptoms associated with many of the most common issues: Internalizing or emotional problems scores inform about whether anxiety or mood disorder might be present; externalizing scores scout for disruptive behavior disorders; and attention problems provide data related to ADHD or learning issues. Compared to interviews, checklists are relatively brief, and they provide systematic assessment of a range of possibilities. They also can be completed ahead of the initial session: in the waiting room, mailed ahead, or increasingly, as an online evaluation or via computer-assisted administration. Electronic formats offer the advantage of scoring and interpretation before or during the first session, thus letting the intake session focus on more targeted diagnostic assessment and treatment planning.

These tools often are diagnostically sensitive to the condition of interest, meaning that most cases with a disruptive behavior disorder will have elevated externalizing scores, for example. They tend to be less diagnostically specific, though. Poor concentration and attention problems could be due to a wide variety of factors, including anxiety, mood disorder, a learning disability, or substance misuse—not just the ADHD that the scale name would seem to imply. Paradoxically, the combination of good sensitivity and lower specificity results in a tool that usually is more helpful ruling a condition out than in. Low internalizing scores are more decisive at reducing the probability of anxiety or depression than high scores are at confirming a single diagnosis. High externalizing scores cannot establish a particular diagnosis, although they increase the chances of several contending hypotheses; but low externalizing scores would be decisive at ruling out the same set of contenders. EBM has a mnemonic to try to

help improve interpretation of assessments with this high-sensitivity, lower-specificity profile: SnNOut reminds us that on a Sensitive test, a Negative result rules the diagnosis Out (Straus et al., 2011).

We can go a step further and quantify the information value by looking at how common the result would be among those with versus without the condition of interest. A "diagnostic likelihood ratio" (DLR) uses the rate among those with the condition (e.g., the diagnostic sensitivity) as the numerator, and the rate in those without the condition (e.g., the false alarm rate, or 1—specificity) as the denominator. The DLR can range from almost zero (meaning that the assessment finding is vanishingly unlikely to occur in someone with the target condition) to near infinity (virtually guaranteeing the presence of the target condition). A DLR of 1 would mean that the finding is equally likely in those who have versus those who do not have the condition. DLRs close to 1 mean that the assessment result is ambiguous or indeterminate; it has not changed the probability of the diagnosis. Values much lower than 1 decrease the odds of diagnosis, and values greater than 1 increase the odds. As a rough guide, values greater than 10 (or smaller than 0.10) are frequently clinically decisive, around 5 (or 0.2) are helpful, and smaller than 2 (or between 0.5 and 1) are not clinically informative. A DLR is an effect size, and it integrates the influences of reliability and convergent and discriminant validity for a particular score under the conditions in the sample (AERA, 2014).

The most precise way of interpreting the DLR combines it with the prior probability, often the base rate of the target condition, using a form of Bayes' Theorem. EBM aficionados have created online probability calculators or smartphone apps that handle the mechanics to yield the revised probability. The probability nomogram (Google "probability nomogram") is a graphical alternative that trades some accuracy for convenience: It is easy to connect the dots and read the results in real time with a patient, but it adds a bit of interpolation and human error. Still, the improvements over unaided, impressionistic interpretation are substantial, including large increases in the absolute accuracy, big gains in the consistency of interpretation, and protection from cognitive biases that often contribute to overdiagnosis of rare conditions (Jenkins, Youngstrom, Washburn, & Youngstrom, 2011). The bottom panel of Figure 1 illustrates updated probabilities ("posterior predictive values") based on DLRs from family history and screening test results (also see Youngstrom et al., 2014, for a detailed example of application to a case). IBM's medical decision support package in Watson uses this type of approach to simultaneously revise multiple probabilities of different conditions at the same time, building a dashboard that shows a panel of updated estimates so that the clinician can easily see what the current leading hypotheses are, and whether more assessment or acute treatment is the best next action. Based on the integration of the information so far (as indicated in the bottom panel of Figure 1), externalizing and impulse control disorders have been demoted as hypotheses, and an anxiety or mood disorder with comorbid substance misuse (perhaps as an attempt at self-medication?) are the contenders that we would want to probe more.

Collateral Informants (E). When working with children, it is usually the parent who buys the ticket for the therapeutic journey—the parent initiates the referral, schedules the appointment, transports the child, and decides whether to continue with treatment. Authorities agree that routinely gathering data from multiple informants is important to understand the context of the child's behavior (De Los Reyes et al., 2015). However, the agreement between perspectives is only moderate, with meta-analyses and cross-national studies finding correlations of 0.2 to 0.3 between parent, teacher, and youth reports on the same measures (Rescorla et al., 2013), and 0.4 between adults and a collateral about internalizing or externalizing problems, rising to 0.68 for substance use (Achenbach, Krukowski, Dumenci, & Ivanova, 2005). Clinically, the common scenario is unimpressive levels of agreement. If the average level of self-reported concerns at a clinic had a *T*-score of 70 (two standard deviations above the mean, commonly considered "clinically elevated"), then the average level of partner-reported concerns on the same scale would be *T*-scores of 58—well within the normal range (Youngstrom, Meyers, Youngstrom, Calabrese, & Findling, 2006). Given the high internal consistency of each informant report, some interpret the data as indicating a high degree of situational

specificity in behavior (Achenbach, 2006), although this is challenging to reconcile with models that formulate cases in terms of the presence or absence of a diagnosis. Others have interpreted findings as indicating that some informants are valid for one diagnosis, and not others. Clinicians often consider youth report more accurate for internalizing problems, and not accurate about attention problems, for example (see De Los Reyes & Kazdin, 2004, for a review). Rater validity also may change by setting: Parents may be better informants than teachers about sleep disturbance, for example, because teachers will not observe bedtime behavior or waking. More subtly, "teacher report" is not all based on the same context: Children spend most of the school day in the same room in elementary school, and start shifting between rooms and teachers for different subjects in middle and high school. Different informants may notice or care about different symptoms (Freeman, Youngstrom, Freeman, Youngstrom, & Findling, 2011). Agreement also changes longitudinally (van der Ende, Verhulst, & Tiemeier, 2012).

How does EBA make sense of the different perspectives? The DLR approach provides a middle way, using separate DLRs to accord each perspective an appropriate amount of weight, rather than a yes/no decision about whether to include each score with regard to a particular hypothesis. In EBA, the different informants are also useful in formulating the treatment package. If multiple informants see the same problems, that often indicates greater severity (Carlson & Youngstrom, 2003), justifying more assertive intervention. In contrast, if teacher report shows few concerns, then treatment may be more focal and not need to address functioning in the school environment (Vollmer & Northup, 1996). If the youth scores are lower than the corresponding scores from collateral adults, it is often a warning that the clients will not be as motivated for treatment because they do not see themselves as having the problem. Artful therapy then frames treatment goals in ways that engage clients yet also address the referral in ways that keep the adults motivated to continue bringing the child to treatment.

Prescription Phase

If the core assessments are well matched to the vital few presenting problems, then the patient completing measures ahead of the first visit may accomplish most of the process described above. Equipped with the data, the clinician can hone in on a few hypotheses for detailed exploration, while avoiding some of the snares of confirmation bias and other heuristics (Davidow & Levinson, 1993). The next step is to evaluate critically the leading hypotheses and decide the focus and form for the optimal individualized treatment plan.

Add Focused, Incremental Assessments (Step F). For many clinical issues, there are focused rating scales available that have greater diagnostic accuracy than broad measures can provide. Scales that cover manic symptoms will be much more helpful than broad measures for discriminating cases with bipolar disorder, for example (Youngstrom et al., 2015). Because none of these measures have perfect diagnostic accuracy, they can give high rates of false-positive results, especially when used in low base rate settings. For this reason, they are not well suited for universal screening, and they may not be a good choice for a routine intake core battery (Straus et al., 2011). However, the DLR approach lets us combine the information from these more specialized tests with the base rate, risk factors, and other information, producing an accurate revised probability estimate that integrates all of the information. Incremental validity is key here. When the same informant fills out multiple rating scales, they are very highly correlated with each other due to the shared source variance. Correlations below 0.3 do not introduce much bias compared to the gains in accuracy using a Bayesian framework, but correlations above 0.5 start to feed redundant information into the assessment process and exaggerate the apparent confidence of the results (Youngstrom, 2013). Thus, it makes sense to "take the best" and only include the most valid rating scale from each informant for each hypothesis. There tend to be brief and inexpensive options, adding little burden or cost to the evaluation process (see Wikiversity for example). They also often provide helpful information about the severity of core symptoms,

establishing a benchmark for evaluating treatment progress.

As costs come down and technology makes it more feasible to interpret the information rapidly, physiological measures are likely to play an increasing role at this stage of the evaluation process (De Los Reyes & Aldao, 2015). Although they would need to show large effect sizes to be useful in isolation for diagnosis, it is more likely that they will be able to offer meaningful incremental validity (Haynes & Lench, 2003).

Add More Intensive Methods to Finalize Diagnoses (Step G). The time elapsed in evaluating the case so far may be less than 15 minutes of face-to-face contact if the routine assessment and follow-up measures are selected and completed ahead of the session (Jenkins, Youngstrom, Youngstrom, Feeny, & Findling, 2012). Integrating the results provides a dashboard that ranks the front-running hypotheses in order of probability, as is getting done with several applications using machine learning with medical record data. The clinician can then pick more intensive assessment strategies that will often be decisive in establishing the presence or absence of a diagnosis.

Semistructured and Structured Diagnostic Interviews. Structured diagnostic interviews were developed to improve inter-rater reliability of diagnoses. Fully structured diagnostic interviews (F-SDIs) are scripted to eliminate the need for clinical acumen; they can be computer-administered (e.g., Mundt et al., 2013; Wasserman, McReynolds, Lucas, Fisher, & Santos, 2002; Wasserman et al., 2004) as expert systems that embed the diagnostic algorithms into the decision tree and branching logic (Susskind & Susskind, 2015). F-SDIs typically deliver inter-rater kappas of 0.8+, whereas inter-rater agreement between clinicians conducting unstructured interviews usually hovers closer to zero (Garb, 1998). Limitations of F-SDIs include concerns about validity, potential gaps in coverage, and acceptability. The patient might misinterpret a computer-administered question, but a clinician could probe to determine whether an infatuation or an excessive interest actually rose to the level of a clinical obsession. Narrow coverage is another potential issue. Acceptability concerns are the potentially fatal flaw: F-SDIs only work if professionals use them.

Compared to fully structured interviews, semistructured diagnostic interviews (S-SDIs) offer more latitude for phrasing the questions and probes, and they allow more clinical judgment in scoring responses. The structured aspect keeps the element of considering multiple diagnostic possibilities—not just the first or the favorite hypothesis; and structure also maps to diagnostic criteria, increasing the inter-rater reliability of the resulting diagnoses (Garb, 1998). A well-designed S-SDI balances improvements in validity—due to greater clinical judgment and probing—with the trade in lower inter-rater reliability due to idiosyncratic decisions creeping back.

Comparing the typical clinical interview to SDIs reveals several patterns. Agreement about some diagnoses is modest overall, and significantly worse about many (Rettew et al., 2009). Some gaps are likely due to fiscal considerations—many payers will not reimburse treatment for conduct disorder, creating an incentive to label the behaviors as something else reimbursable. Variations in training contribute to wide discrepancies in diagnosing pediatric bipolar disorder or autism. Clinicians also tend to diagnose fewer conditions than more structured interviews detect. Sometimes F-SDIs detect cases without enough impairment to warrant clinical concern; but clinicians also tend to call off the search as soon as they have found a (reimbursable) diagnosis (Jensen & Weisz, 2002), underestimating comorbidity rates. Clinicians often assign a diagnosis without having documented enough symptoms to meet formal criteria for a diagnosis (Miller, 2002). Adding SDIs improves diagnostic reliability across a gamut of clinical settings (see Wikiversity EBA page about interviews), improves psychosocial outcomes, and provides equally good symptom reduction with less clinician time—allowing more patients to be seen, or reduced expenses per patient (Hughes et al., 2005). When the clinician's diagnosis of record is confirmed by an SDI, families engage more with treatment, have fewer no-shows or cancellations, and have significantly better internalizing outcomes (Jensen-Doss & Weisz, 2008).

Despite these advantages, clinicians rarely choose to use SDIs (Jensen-Doss & Hawley, 2010). They see them as infringing on their professional autonomy and

potentially damaging rapport with clients (Suppiger et al., 2009). Ironically, clients actually prefer SDIs, believing that the evaluation gives clinicians more comprehensive understanding of the client's needs and situation (Bruchmuller, Margraf, Suppiger, & Schneider, 2011; Suppiger et al., 2009). Development and validation of free SDIs directly address concerns about cost (Beidas et al., 2015).

Why not just use the SDI as the basis for diagnosis, skipping the preamble of the rating scales and risk factors? First, no single interview covers all of the diagnoses and clinical topics that a clinician might encounter, and surprisingly, some of the omissions can be common issues. Older interviews did not ask about ADHD in adults, nor about mania or hypomania in youths. Training requirements and rater drift also create implementation obstacles. Finally, it may be easier for our client to first broach her drinking, her thoughts of self-harm, or her confusion about her sexuality in the privacy of a questionnaire or computer session than face-to-face (Shaffer, Fisher, Lucas, Dulcan, & Schwab-Stone, 2000).

A hybrid approach, using rating scales and risk factors to generate the contending hypotheses, can guide the selection of which interview modules to use for follow-up. It becomes possible to "take the best," picking sections of interviews that are best for each particular purpose. The hybrid also lets the clinician streamline the interview, skipping modules without having to ask the screening questions when the prior data already move the probability of the diagnosis into the green zone.

Cognitive and Achievement Testing. Cognitive and achievement testing may also play an important role at this stage for some cases. If the referral question focuses on academic performance, or if the contending hypotheses include a learning disorder, then cognitive testing and achievement testing clearly provide valuable data (Fletcher, Francis, Morris, & Lyon, 2005). Sometimes the referral question focuses on possible ADHD or learning disability, and cognitive testing may reveal that the person has average or low average ability—rather than having some other disorder—and has been functioning in a fast-paced or challenging environment by dint of hard work. Similarly, developmental disorder evaluations need to gauge whether social deficits are out of line with cognitive ability. Brief ability tests would be well suited to recognizing these scenarios, and verbal ability estimates may also help gauge the suitability of more cognitive components in therapy versus emphasizing behavioral interventions.

Despite decades of speculation and clinical lore about subtest scatter or profile analysis being associated with clinical conditions such as ADHD, the EBA framework makes clear that the associations are too weak to be informative for individual decision-making (e.g., Watkins, Kush, & Glutting, 1997). Deficits in processing speed or working memory are nonspecific, meaning that many different clinical conditions can show them, and the pattern does not help to pick between the hypotheses. Ipsative analysis has technical psychometric challenges making it unlikely that apparent strengths and weaknesses will be stable over time or show incremental predictive value (Canivez, 2013). The validity coefficients for checklists are significantly higher for predicting diagnoses such as ADHD or autism than those found with cognitive profile analysis—and checklists can usually be administered and scored much more quickly and at much lower cost. Even those who would benefit from assessment of cognitive ability usually will not need the traditional full battery; in most cases, a four-subtest estimate of general ability will provide the most robust and empirically supported information. Specific questions about neuropsychological deficits also would be better addressed by tailored, high-validity batteries, rather than using omnibus cognitive ability tests that attempt to provide neuropsychological Swiss Army knife tools that are not adequate for high-stakes use (Canivez, 2013).

Assess for Treatment Planning and Goal Setting (Step H). By this point, the clinician has one or more hypotheses that have accrued sufficient supporting evidence that they are in the treatment zone. The focus now shifts to gathering information that would change the way treatment is provided. In research parlance, the variables of interest would be conceptualized as confounds or moderators. It also is crucial to negotiate treatment goals that the client finds important and motivating.

Confounds would be alternate explanations for the apparent diagnosis, especially those that would warrant

a different type of intervention. Common examples include symptoms due to a general medical condition or side effects of prescription or illicit drugs. A simple checklist covering common concerns for the age groups with which we work would help to systematically evaluate the possible confounds ahead of time, rather than revisiting the issue when our otherwise well-formulated intervention fails to get traction.

The moderators could be general factors changing likely outcomes across a broad swathe of interventions, or they might be more specific to our first-choice treatment. Comorbid anxiety and aggression problems may have a different prognosis than either in isolation. Personality traits are likely to influence engagement with treatment, including completion of homework between sessions and keeping appointments (Bagby, Gralnick, Al-Dajani, & Uliaszek, 2016). In services for children, greater parental distress or impairment may make adherence more difficult and missed appointments or premature termination more likely. Good reviews of EST options and more recent randomized trials are likely to include discussion of moderators, helping clinicians see what things may alter the course of treatment (see Wikiversity EBA page about treatment planning for example).

Goal setting also improves the efficacy of therapy (Marshall, Haywood, & Fitzpatrick, 2006; Tryon & Winograd, 2011). Good goals are those that are shared by the therapist and client, communicated clearly, motivating to the client, and measurable. Framing matters: Consider the difference between "reduce irritability" versus "decrease conflict" or "get my spouse off my back" from the client's point of view (Freeman et al., 2011). These goals might be functionally identical from the therapist's point of view, all indicating that coping skills, communication strategies, and conflict resolution might be treatment components; but the client's willingness to try learning them will vary markedly depending on the frame. It matters whether the goals are "approach" oriented, increasing positive behaviors or positive functioning, versus "avoidance" oriented, focusing on symptom reduction (Wollburg & Braukhaus, 2010): Approach goals may actually produce larger symptom reductions. Good goals also connect with the client's strengths.

A psychometrically driven goal-setting method could be to review the scores on the measures gathered so far, and then pick one as a primary outcome measure, or pick a shortlist of key targets. Potential advantages of this method are recycling information already gathered as the basis for a treatment plan—avoiding additional burden on the client, and ensuring a tight connection between the formulation and the treatment goals. Good choices of measure would be sensitive to treatment effects and have normative data to establish benchmarks.

A complementary approach in services for children would be to ask the child and the parent what their top goals are for treatment (Weisz et al., 2011). These often are more specific and contextualized than a summary score from a psychometric scale. Often they overlap with the content of checklists, but not always, and often the highest priorities for the parent or client focus on a behavior that falls outside of the content of the most elevated scales. Focusing only on scale scores adds the virtue of quantifiability, but risks becoming abstract and failing to connect with the most important goals, which runs the risk of undermining the youth's engagement or the caregiver's support of the treatment. Low-burden, idiographic assessment can customize the goals, yet scale them consistently for tracking across sessions (Weisz et al., 2011).

Assessing Functioning and Quality of Life. Level of functioning is a major driver of service utilization—greater impairment leads to more frequent and more expensive services (Bickman, 1996; Butt, 2016). Improved functioning can divert clients from needing emergency treatment, and can prevent escalation or relapse.

Assessing functioning and quality of life also shifts attention from symptoms and problems to positive aspects of life (Frisch, 1998). Clients do not just want to eliminate the negative; they want to accentuate the positive. They do not just want symptoms to go away; they want their lives back. Successful treatment should lead to measurable gains in friendships, positive communication at home, better academic performance, and a higher sense of subjective well-being or quality of life. There is a growing evidence base for scores on scales assessing these important constructs (see

Wikiversity EBA page about goal setting, quality of life section), and this is a major growth area for psychometrically informed tool development (Butt, 2016).

Learn and Use Client Preferences (Step X). Ideally, the case formulation and treatment planning involve an ongoing discussion with the client and family. We emphasize the ongoing nature of the discussion by labeling this the "X factor" rather than isolating it to a single step in a series. EBM recognizes the import of discussing patient values by not specifying universal, fixed locations for the Wait-Test and Test-Treat thresholds. Instead, they are negotiated on a case-by-case basis considering the treatment's risks and benefits (Straus et al., 2011). This reflects the growing emphasis on shared decision-making, where the patient is an informed consumer whose values and preferences are the basis for an equal partnership (Susskind & Susskind, 2015).

Patient beliefs about the causes of problems, their attitudes toward biological or spiritual explanations, the past experiences of relatives and friends with treatment, and other cultural and idiographic factors all shape perceptions of treatment (Yeh et al., 2005). Better to start the dialogue early and give their opinions weight at the beginning, or clients will vote with their feet later in the form of low motivation and engagement, and early dropout. Clients often put higher value on spiritual perspectives than the therapist does, and some cultures are less likely to conceptualize things in biological terms than the research community does. Cognitive behavioral therapy (CBT) may be less effective with people from Asian cultures, who may be less likely to focus on cognitive symptoms of depression than Europeans and European Americans (Kalibatseva & Leong, 2011). Similarly, interpersonal psychotherapy (IPT) may be more attractive and effective than other ESTs for Hispanic families due to the greater emphasis on family relationships (Mufson, Dorta, Olfson, Weissman, & Hoagwood, 2004). A core portion of the values discussion needs to be idiographic, though. What is important to the outcome for individuals is their own constellation of beliefs and attitudes, not stereotypes and averages that may or may not match them. Collaborative care models and evidence-based practice recognize this (Norcross, Beutler, & Levant, 2006).

As we move toward collaborative care models and shared decision-making, we need to also attend to what information our clients have, how they interpret it, and how we can help them navigate the information effectively. Researchers can give more thought to packaging results in a way that is digestible by a lay audience, and we can do more to teach the public critical-thinking skills to separate hype from evidence (Gigerenzer & Muir Gray, 2011). Effective communication will involve a blend of attention to details such as graphical formats and concrete ways of describing probabilities, as well as a greater willingness to engage with popular and online media (Becker, Spirito, & Vanmali, 2015). We need to understand where the client is coming from and how things look to him or her before we can engage in a conversation about change.

Process Phase

Once the formulation has clarified target issues and expected therapeutic change mechanisms, the role of assessment changes to defining measurable goals, monitoring processes, and evaluating progress toward goals. Measuring the rate of progress tells us when we are making good headway, or when we are stuck; it also helps see when a detour or new route—a modification of treatment plan—is in order. It is here that the traditional assessment toolkit has been least developed.

Goal Setting: Milestones and Outcomes (Step I). The steps of the EBA process so far have focused on clarifying the goals of the journey and selecting the best route and mode of transportation to get there. For treating depression, the EBA process maps an initial vague description (e.g., "I want to feel better!") to a measurable definition of outcomes (e.g., no longer experiencing significant depression symptoms, enjoying your relationships with friends and with family as much as other kids your age) and milestones along the way (e.g., getting back into a nonclinical range of functioning). Making them measurable adds value to the therapy process. It provides an explicit, shared definition of the goal. It avoids black-and-white, all-or-none thinking—it reveals changes in degree. If therapy were an academic class, then these clinical assessments would be the quizzes and exams. It is also possible to show

progress graphically, allowing the patient to process the visual image in a different way than abstract or verbal information.

Measurement also exposes whether treatment is helping or not. Measuring change is complicated. Reliability is a precondition for quantifying change, but it is not enough. If the measure is unreliable, then improvement cannot be distinguished from variation due to error and random fluctuation. However, measures can show high retest stability and be impervious to change, defeating use as an outcome measure. Correlation coefficients can be high even when absolute agreement is poor (Bland & Altman, 1986). Change scores answer conceptually different questions than regression model residuals, so they need not support the same conclusions (see Rogosa, 1995). Change scores, though easier to estimate in clinical practice, are only appropriate when the true treatment effect should be larger than the daily or weekly fluctuations on the measure, or the signal will disappear in the noise. The measure's reliability should exceed 0.5 before looking at change scores. A pragmatic way of evaluating a test would be to create a "sensitivity to change coefficient," a generalizability theory coefficient comparing the variance explained by the time-by-treatment interaction to a denominator combining it and its associated error term (Streiner et al., 2015). This focuses on the demonstrated ability of the test score to detect change, but it could confound scale performance with the type of treatment. Sensitivity to a psychotherapy effect need not be the same as sensitivity to an antidepressant medication (DeRubeis, Siegle, & Hollon, 2008). Pragmatically, clinicians should pick outcome measures that have shown ability to measure change at a group level on processes and goals that the treatment targets. It is hard to infer sensitivity to change based on conventionally reported psychometrics, such as Cronbach's alpha. There is no substitute for empirical demonstrations here. Picking the same measures that have proved sensitive in treatment trials also will create synergy in terms of benchmarking, as we describe later.

Clinically Significant Change. Jacobson and colleagues developed a psychometrically informed system for defining clinically significant change at the level of the individual client (Jacobson, Roberts, Berns, & McGlinchey, 1999; Jacobson & Truax, 1991). Although others have proposed technical refinements, the original, simpler model produces similar results and remains the dominant approach to thinking about clinical significance in the domain of psychotherapy trials. The framework combines two elements: reliable change—pegged to the stability and precision of the outcome measure—and normative benchmarks that compare the client's scores to clinical and nonclinical reference groups. Reliable change is defined by the standard error of the difference score. Jacobson and colleagues suggested using the scale's retest reliability coefficient as the basis for the reliable change index (RCI). In practice, retest reliability is rarely available, and it also becomes thorny to decide what retest interval should be used. For all these reasons, the tendency is to use Cronbach's alpha instead as the basis (Ogles, 1996). Jacobson and colleagues also suggested dividing the observed change by the standard error, which would convert the change to a z-score metric. Then standard rules of thumb would apply: 1.96 or greater would be 95% likely to reflect real change (i.e., occurring by chance less than 5% of the time), and 1.65 would be 90% likely.

Minimally Important Difference (MID). Some areas of medicine are using patient or clinician-anchored definitions of the smallest change in outcome that more than 50% would agree was a subjectively meaningful shift. Recent hybrid methods are using item response theory to calibrate items against anchored ratings as a way of unifying psychometric and scale judgment approaches (Thissen et al., 2016). Whereas the RCI approach makes sure that the observed change clears a rigorous threshold based on the stability of scores, anchored MID emphasizes criterion validity, operationalized as patient or clinician judgments about meaningful change. Rather than asking the individual, "What is the smallest number of points on this scale that things would need to change for you to feel like there was a meaningful difference?" (a purely idiographic approach), MID uses group data to set thresholds, thus building nomothetic benchmarks. Because MID grew from other areas of medicine, the NIH PROMIS outcome measures are

one of the first places where researchers are publishing MID benchmarks (e.g., Thissen et al., 2016). Surprisingly, MID benchmarks tend to be smaller than the RCI would be for the same measure, suggesting that patients may notice and value relatively small amounts of change. MID benchmarks often are close to an effect size of $d \sim 0.5$ across a wide variety of constructs, perhaps because this approximates where the "just noticeable difference" is in a variety of cognitive and psychophysical tasks (Streiner et al., 2015).

Normative Benchmarks. The second part of Jacobson and colleagues' definition of clinical significance is the comparison to normative benchmarks, adding the component that MID approaches bundle together in a single metric. In the RCI method, there are three normative benchmarks, labeled A, B, C; others have added the mnemonics Away, Back, and Closer. *Away* is defined by a threshold two standard deviations away from (usually, below) average for a clinical reference group. If there were a sample of clinical cases, such as 5,000 treatment-seeking youths, or a group of 5,000 with diagnoses of ADHD, then the threshold would be two standard deviations below the mean in that sample. *Back* is set two standard deviations above the nonclinical distribution's mean. *Closer* is based on the weighted average of the two means; passing the weighted average shows that the score has moved closer to the nonclinical than the clinical distribution (Jacobson & Truax, 1991).

There are technical and practical obstacles to using this sort of clinical significance model. The necessary ingredients—retest stability, a standard error of the difference, and norms for both clinical and nonclinical groups—are rarely reported or available. Most clinical measures only have descriptive statistics published for convenience samples from clinical settings and research studies, not nationally representative samples. Age norming is another potentially major omission. Development of cognitive abilities is so rapid in childhood and adolescence that standardized tests develop norms with 3-month age cohorts. The rate of change in cognitive functioning slows at other ages, and shifts in emotion regulation or personality may be less pronounced, but they are not adequately reflected in the scoring of existing measures. Importantly, if the clinical and nonclinical distributions overlap substantially, then some of the benchmarks may be nonsensical or impossible. For many widely used, well-normed measures, the *Away* definition would require raw scores below zero on some of the syndrome scales, and the *Back* definition would define *T*-scores of 70—normally considered clinically elevated—as within the normal range of functioning.

Convenience is another barrier. It is not realistic that busy clinicians would hunt down technical parameters and then perform calculations on a case-by-case basis in real time. For clinical significance to be feasible to measure in practice, the benchmarks and thresholds would need to be calculated ahead of time for the measures likely to be used (Step B), and ideally programmed into an Excel spreadsheet or a Google Doc so that the clinician could just enter the score and get the results. Then, the baseline scores could be used to generate the change targets.

Although many measures are too cumbersome to repeat every session, there is definitely value in doing them periodically, and regular client feedback also reduces premature termination (Swift, Greenberg, Whipple, & Kominiak, 2012). Figure 2 shows what this might look like in practice. An interim report is valuable in terms of telling whether the client is making expected progress. Early response predicts overall response for many interventions. If there is no signal of improvement after four to six sessions (Howard, Moras, Brill, Martinovich, & Lutz, 1996), that is a cue to take stock. How is rapport? Is the client engaged? Are they following through with skills and homework? If not, why not? Are the goals motivating to them, or are they not in agreement with the treatment plan? If the client is worsening, then it is definitely worth revisiting the diagnosis and formulation—ESTs are unlikely to be iatrogenic when matched with the appropriate problem. However, treatments can harm (Lilienfeld, 2007), so it is crucial to have an early warning if well-intentioned methods are having unintended consequences. Maltreatment, trauma, substance misuse, and hypomania are all things that are often missed during intake due to stigma, a lack of insight, or other factors—and all can undermine the effectiveness of interventions. Using the 90% definition of reliable change is probably a reasonable yardstick to use

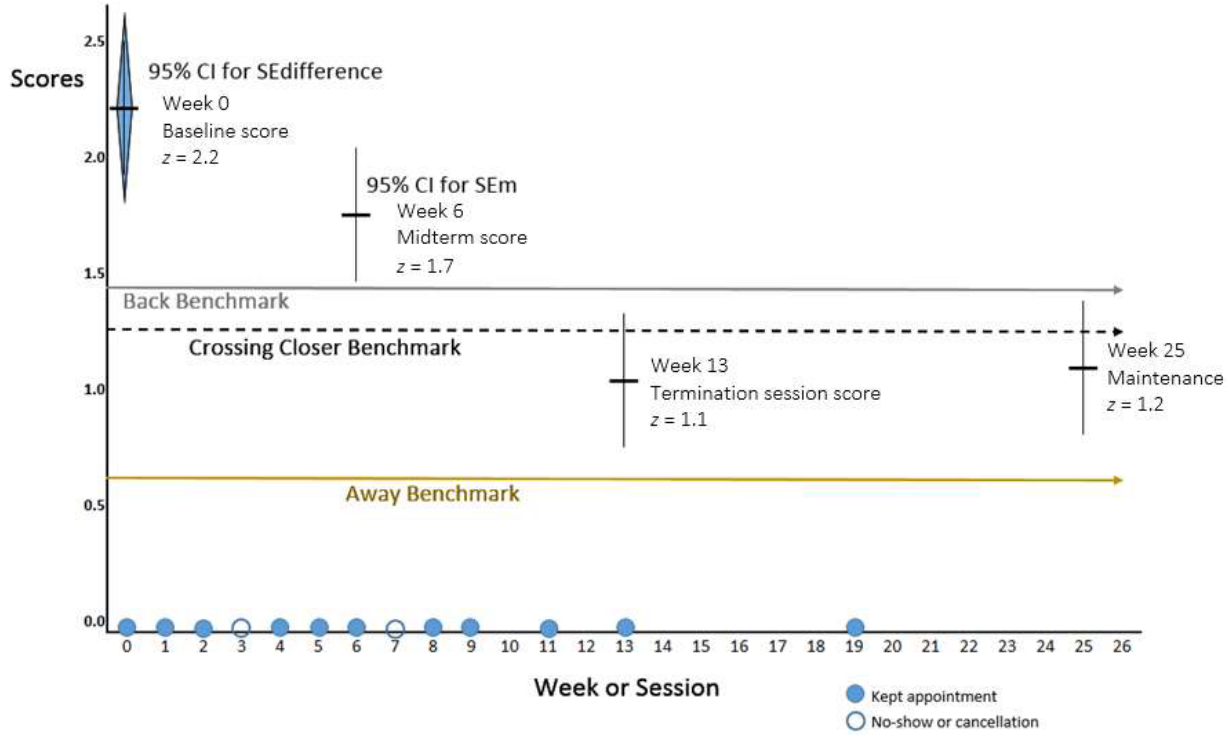

Figure 2. Using the reliable change index and normative benchmarks to look for clinically significant change at the individual patient level, using repeated "midterm" and final assessments. SEdifference = standard error of the difference score; SEm = standard error of the measure. Away, Back, and Closer thresholds based on Jacobson and Truax (1991) definitions. Juxtaposing the session attendance may reveal associations between engagement and progress. Reproduced from Youngstrom and Van Meter (2016).

for benchmarking interim response—it is a high enough bar to be a strong signal, but still a goal within grasp for many cases. It also is a way for guiding course corrections during treatment, creating a dialectic between assessment and therapy (Weisz, 2004). For our client, repeating one of the mood measures at session 6 would provide a formal check on whether treatment was getting traction. Scores lower than the critical value for reliable change (the blue diamond around baseline score in Figure 2) would indicate clear progress, even if they had not yet passed one of the other benchmarks. Any increase in score would signal a need to reevaluate treatment, and perhaps diagnosis or case formulation as well.

Program Evaluation and Benchmarking. Few clinicians report doing this sort of outcome measurement routinely. Practical and systemic barriers may be even more important than attitudes (Hatfield & Ogles, 2007). Low-cost, low-burden, valid measures are key to successful implementation. Recent reviews have anthologized lists of measures showing good psychometric properties that are also free to use clinically (Beidas et al., 2015), and professional societies and advocacy groups are starting to aggregate these on websites. When clinicians get both quantitative data and narrative description about client progress, the quantitative information has a larger impact on decisions about changing or continuing treatment. This defines an easy opportunity for improvement in clinical practice: Reducing the barriers and making it convenient to use outcome assessment would increase uptake and improve treatment decisions. Clinicians and systems of care can benchmark their clients' progress compared to what is found in published studies using evidence-based assessments too (Spilka & Dobson, 2015; Weersing & Weisz, 2002). This can help shape expectations for the length of treatment and typical

response, as well as providing cues when progress is slower than anticipated.

Progress and Process Measures (Step J)

The most fine-grained level of assessment tracks change in between sessions. In the journey of therapy, this is the GPS measuring progress toward goals (how many miles covered?). Daily or session-level measures act as the speedometer, providing a sense of how rapidly things are progressing, or whether they are stalled or moving backwards. Practicing without systematic assessment is akin to barreling blind into space—unlikely to help you arrive at your destination on time and without significant detours. Mapping out your route before you leave and checking progress against that plan allows for recalculating when obstacles are encountered, and for an ongoing conversation between therapist and client about what the ultimate destination is and how best to get there. EST treatment manuals often include some form of progress tracking, whereas treatment as usual often tracks progress informally and impressionistically, if at all (Hunsley, 2007). The techniques used here vary widely across interventions, from daily report cards completed by teachers, to sticker charts and token economies, or three- and five-column charts in CBT, diary cards in dialectical behavior therapy, and daily life or mood charts for mood disorder, including several free examples that have been customized for children. These regular progress measures increase the effectiveness of the treatment (Guo et al., 2015), the same as how weigh-ins give a diet plan more traction. Streamlined assessment tracking a few top problems on a session-by-session basis is both feasible and highly sensitive to treatment effects with youths and families (Weisz et al., 2011). The more frequent measurement schedule makes these data even more useful as harbingers of progress or setbacks, making rapid course correction possible. These data also can signal when to transition to tapering and planned termination, as well as continue monitoring to maintain gains.

Tracking process is where the traditional assessment measures are weakest. None of the commonly taught and used measures are suitable for weekly, let alone daily, repetition. Scores on such measures can decrease simply due to repeated test administration (Longwell & Truax, 2005). The niche is ripe for innovation. This also is the place where technology is most rapidly transforming assessment (see progress and process Wikiversity EBA page for example).

Technology is not just changing the type or amount of data, but also the convenience. Downloading an app or putting on a wearable and then going about your life is much easier than actively recording details in a diary. Wearables skip the step of data entry and potentially provide feedback in session or in real time. Convenience plus rapid feedback are likely to increase use. These advances directly address old barriers to implementation (Hall et al., 2014). There will be a hiccup where the technology churns out a surfeit of data that requires a few iterations to shape into formats that clinicians can read quickly and understand the implications for treatment. It is a safe bet that this will happen relatively swiftly.

Graphical and Time-Series Analyses. Applied behavioral analysis has a long tradition of graphing measures of behavior in single-subject designs (Kazdin, 1992). More recently, increases in computing power have made it feasible to apply models such as interrupted time-series analyses to individual data, using either auto-regressive integrated moving average (ARIMA) models or nonparametric randomization alternatives (Hopwood et al., 2016; Houle, 2009). Data captured at a daily or even more frequent level—such as diary, ecological momentary assessment, or wearable device data—provide the density of sampling for these methods to come into their own. Mixed regression or hierarchical linear models provide a framework for integrating changes in trajectories across groups of cases, for research or program evaluation purposes (e.g., Speer, 1992, 1993), whereas graphs remain the most user-friendly way of tracking progress.

Measuring Mediators. Mediators are the process variables, conveying the mechanism of impact for the intervention. If clinicians are not routinely measuring outcomes in clinical practice, it is an even bigger request to suggest measuring mediators as well. However, if the costs and burden are low, and the information is available rapidly, then there could be value added. First, for interventions that are skill-based,

measuring competence or mastery would provide valuable feedback guiding movement through treatment. From time out to improved communication in couples, efficacious treatments often evaluate new skills in session and then track use in between sessions, praising success and debugging problems in subsequent sessions.

Second, measures may help track signs of early response or precursors of progress. Some studies of pharmacological depression treatments find that somatic symptoms change before the cognitive symptoms do, suggesting that there could be signs of progress before the patient thinks that she is doing better (DeRubeis et al., 2008). Conversely, changes in sleep precede worsening of depression; they might precede improvement in mood or be a valuable endpoint in their own right for health reasons. It is possible that other interventions that target somatic aspects, such as exercise, phototherapy, behavioral activation, or nutritional supplements, could produce measurable gains in symptom clusters that precede subjectively reported progress.

Third, there are variables that promote good outcomes even though they are not core specific therapeutic components, such as the therapeutic alliance and providing psychoeducation about aspects of the disorder. Attending to therapeutic alliance (Horvath, Del Re, Flückiger, & Symonds, 2011; Shirk, Karver, & Brown, 2011), detecting ruptures in the relationship, and repairing them quickly all promote retention in treatment (Safran, Muran, & Eubanks-Carter, 2011), and often better outcomes (Castonguay, Constantino, & Holtforth, 2006). This could be a simple Likert-type scale that the client and therapist complete independently in between sessions. Any disagreement or drop in either party's rating could be a cue to discuss directly in the next session. These ratings are likely to be less reactive and more candid if done separately rather than face-to-face. This could be built into the therapy progress note for the clinician and the homework for the client, or automated in a smartphone application. Psychoeducation has shown large effects in terms of improving knowledge about illnesses such as bipolar disorder, and these have mediated improvement in mood symptoms (Fristad, Verducci, Walters, & Young, 2009). Measuring knowledge gains provides a context to correct misunderstandings and fill gaps in much the same way that quizzes do in teaching and coaching settings.

Wrapping Up and Maintaining Gains (Step K)

Good assessment helps decide when we have arrived at an acceptable destination. Various definitions and assessment strategies are possible, including strings of success on daily or weekly tracking measures, documented improvement in functioning, or reduction or remission of symptoms, and attainment of other treatment goals. Toward the conclusion of active treatment, we can repeat the panel of outcome measures and compare progress against the clinical significance benchmarks (Step I). The clinically significant change thresholds provide a candidate definition for arrival at the desired outcome. They can offer an objective stopping rule, signaling a transition to planned termination. At a systems level, this can lead to more efficient resource allocation, as therapy slots open up for newer and higher acuity cases. Clinicians getting regular feedback about progress on a session-by-session basis have fewer patients drop out, and they reach good outcomes faster (Lambert, Hansen, & Finch, 2001). This assessment also can be an opportunity to renegotiate the therapeutic contract: Here is how much progress we made on our goals; do you want to commit to another set of sessions to aim for an even more assertive target? Is there another issue that you would like to address?

If our client is happy with her progress and ready to conclude, then the focus shifts to termination planning. First, celebrate the successes! Especially if she surpassed the benchmarks for clinically significant change, that is a grand slam—not a minor win. Even when using interventions with moderate or large average effect sizes, clinically significant change is not yet routine (Youngstrom et al., 2013).

Next, review what skills worked and what she found most helpful. Also briefly consider what she did not find effective and why. This also is a form of assessment, similar to an exit interview at a job or a post-trip review. It is worthwhile to write down high-level conclusions, but it also would be worthwhile to build a list of what techniques were favorites as a reminder for any routine use for maintenance of gains. It is reviewing and updating the packing list, should it be necessary to take another trip.

It is also crucial to have a frank discussion about the potential for relapse or situations that might trigger new problems. Discuss what would be cues and ways

to monitor for them. If the client enjoyed using a wearable device or got into a habit with one of the other process measures, that could continue as a wellness promotion tracking tool. Define what would be warning signs (e.g., three nights of bad sleep in a row, or her mood chart scores getting more extreme for three or more days). List what trigger situations might be. Develop a plan for booster sessions, or restarting a brief course of therapy to nip relapse in the bud. Ongoing surveillance is crucial for mood disorders and substance problems, where relapse is the norm. In the context of conduct problems, this has been called the "dental model" of routine checkups designed to maintain treatment gains (Kazdin & Weisz, 1998). Assessment can play a key role in preserving health as well as reaching it.

FUTURE DIRECTIONS: CLINICAL SCIENCE, OR SCIENCE FICTION?

Technology is changing assessment and treatment in two broad ways: *automation* and *innovation* (Susskind & Susskind, 2015). Automation is taking things that humans used to do and computerizing them. This has been happening with assessment scoring software, computerized administration of neurocognitive tests, and now diagnostic interviews. The automation often makes the professional functions faster, sometimes more accurate as well, but it need not transform them.

Technology's innovative impact is fundamentally transformative. A computerized administration of a structured diagnostic interview not only improves reliability, but also replaces the human in that component of the assessment process. The diagnostic algorithm gets implemented by the software, with perfect reproducibility. Where humans struggle to find the likelihood ratios and integrate them correctly, this would be a basic operation for a computer system. Machine learning paired with the big data accessible via the Internet, and the "Internet of things" as we connect more devices on our bodies and environment to the web, will create systems that have greater predictive power than anything psychological science has developed in its first century (Kaplan, 2015). It is assessment meets Amazon—it will have recommendations for us based on our past behaviors and current situation. More disconcertingly, artificial intelligence can also deliver therapy. The ELIZA program, an early example from the first decades of AI research, provided startlingly engaging nondirective therapy decades ago (Weizenbaum, 1984). Current versions would obviously be more flexible and powerful, and future systems yet more so (Lucas, Gratch, King, & Morency, 2014).

What about empathy? Machine learning has already moved from face recognition to emotion recognition (Bartlett et al., 2005), and cell phone apps are even now monitoring our emotional responses to advertising. Link the statistical context recognition engines from voice recognition software—Siri or Dragon™—with the emotion detection software, connect them with "ELIZA 3.0," and feed it our search history from Google and our FitBit. ELIZA 3.0 will hear subtle tones in our voice, read the emotion in our face, and make nuanced inferences based on probabilistic searches of the gigabytes of data we have generated just going about our lives. ELIZA 3.0 will not have feelings, yet will deliver remarkably personalized, manufactured empathy using a fundamentally different process. ELIZA 3.0 won't "feel" any more than a submarine can swim, yet the machine clinician may similarly exceed many of the performance parameters of the human clinician.

From the perspective of the professional, this is a tectonic shift that could spell extinction and certainly propels evolution (Susskind & Susskind, 2015). Why pay for a human professional to make routine diagnoses when options that are less expensive and more convenient are also more accurate and lead to better results? Agile and evolving professionals will have a range of options, including incorporating these technological advances into their professional services, specializing in the frontiers that the machines do not yet cover, offering boutique services to the affluent who prefer handcrafted work, shifting into an interpretation and decision support role that is heavy on empathy and cultural sensitivity, or becoming knowledge engineers who help build and refine the systems (Susskind & Susskind, 2015).

Although these scenarios may sound far-fetched, all are already happening to varying extents. From the view of the people consuming the services, this evolution is unalloyed progress. The trends are making information and services available to a broader audience

than ever before, at greater convenience, lower cost, and still higher consistency and quality (Susskind & Susskind, 2015). It is impossible to meet the needs of the many with the current service model that relies on the relatively few professionals (Kazdin & Blase, 2011). Creating a mini-utopia that provided an empirically supported treatment for anxiety to all the children in the United States who have an anxiety disorder, using an individual session fee-for-service model, would cost more than $35 billion. That is just for anxiety disorders, just for youths under age 18, and just the United States—where less than 5% of the people on the planet currently live. To achieve the promise of clinical science actually delivering its potential public health benefit requires a radically different model. Wikipedia, WebMD, and Watson are the early champions of this change (Kaplan, 2015). The seismic shift is happening, and our choice is not whether to allow it, but whether the new, better methods should be a premium service for the wealthy, or a commons that is shared for the good of the many (Kaplan, 2015; Susskind & Susskind, 2015). Evolving our practices to thrive in a world with EBA available to all is a powerful guiding vision.

GENERAL DISCUSSION

Changing the vision of assessment to something that happens through the entire arc of treatment, not just at intake, exposes a variety of roles that have not all been well developed in our psychological assessment traditions. If we were taking a long trek, we would not aim once at the beginning—no matter how carefully—and then never make adjustments again along the way. Yet the field has concentrated our assessment efforts on the front end, creating measures of personality traits, diagnostic constructs, and to a lesser extent symptom severity. Measures of treatment mediators and moderators, and tools for quantifying change, have been neglected. Despite the neglect, there are strong hints that better assessment leads to better treatment matching, better engagement, improved fidelity and adherence, and enhanced outcomes.

Implementing the EBA approach takes work to install. Like physically remodeling our ship to make it more appealing to our clients, it is a significant overhaul to redesign our assessment process to more tightly connect with the goals and needs of our client. However, once the remodeling is done, it does not hinder our work with the client. Just the opposite—it makes the ongoing work more engaging and efficient. Once implemented in a clinical setting, the first half dozen steps in the model (Table S1 from Youngstrom & Van Meter, 2016) add less than five minutes of face-to-face time, and they can rely entirely on measures in the public domain (Beidas et al., 2015). If the client uses an app to track daily mood or sleep, the best ones cost less than $5. For less than five minutes and five dollars, we can have large gains in terms of accuracy, consistency of our clinical decisions, and protection from common sources of bias and error in our work. With the client entering our office at the beginning of the article, we already would know not just about the suicide attempt, but also have preliminary information about whether mood symptoms, impulsivity, substance misuse, or environmental stressors might be major contributors. We would be able to do appropriate risk management as well as customizing the treatment plan to address need and goals that she would find motivating.

The upgraded assessment approach is also an additive that boosts the average outcome for treatment a moderate amount, enhancing the rocket fuel of our evidence-based treatments. The regular monitoring and feedback strategy uses good behavioral principles to increase the chances that the new skills learned in therapy become healthy habits that generalize to the world outside of the session. Measures of progress and process, traditionally the Achilles' heel of our assessment portfolio, are rapidly improving with new technology. Intriguingly, youths tend to be more technophilic than their parents, creating a dynamic where innovations are likely to happen faster and promote engagement even more with youths (Boydell et al., 2014). In addition to making the treatment process more efficient, an evidence-based approach to assessment at intake and over the course of therapy is also likely to result in a moderate-sized boost to client outcomes, and increase the chances that they will develop healthy habits and insight into their difficulties that will promote continued wellness after treatment concludes.

After decades of incremental advances, evidence-based assessment is ready for liftoff. It has been hard to gather the tools and the information about how to

interpret them, let alone apply them to an individual case in real time. Experts have argued that the complexity of the clinical enterprise, along with the challenges in changes systems of training and service delivery, is frankly harder than "rocket science" (Bickman, 2008). Now smartphone software, cloud-based solutions built by major test publishers, and online communities (e.g., http://feelingkindablue.ning.com) and systems are opening new worlds of possibility. The change need not be intimidating. We are going to be able to work smarter, not harder, in ways that help us understand our clients faster, define goals that matter to them, and measure progress in ways that they find rewarding. Evidence-based assessment will help therapists and their clients pick good destinations, reach them more often, and enjoy the trip more along the way. It may also become possible to reach far more people than ever before, and have psychological science open new worlds of potential for them.

ACKNOWLEDGMENTS

Dr. E. Youngstrom has consulted with Pearson, Janssen, Otsuka, Lundbeck, Joe Startup Technologies, and Western Psychological Services about psychological assessment.

REFERENCES

Achenbach, T. M. (2006). As others see us: Clinical and research implications of cross-informant correlations for psychopathology. *Current Directions in Psychological Science*, *15*, 94–98. https://doi.org/10.1111/j.0963-7214.2006.00414.x

Achenbach, T. M., Krukowski, R. A., Dumenci, L., & Ivanova, M. Y. (2005). Assessment of adult psychopathology: Meta-analyses and implications of cross-informant correlations. *Psychological Bulletin*, *131*, 361–382. https://doi.org/2005-04167-003 [pii]10.1037/0033-2909.131.3.361

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles*. Burlington, VT: University of Vermont.

Achenbach, T. M., & Rescorla, L. A. (2003). *Manual for ASEBA adult forms & profiles*. Burlington, VT: University of Vermont.

Alegria, M., Chatterji, P., Wells, K., Cao, Z., Chen, C. N., Takeuchi, D., … Meng, X. L. (2008). Disparity in depression treatment among racial and ethnic minority populations in the United States. *Psychiatric Services*, *59*, 1264–1272. https://doi.org/10.1176/appi.ps.59.11.1264

American Educational Research Association. (2014). *The standards for educational and psychological testing*. Washington, DC: Author.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). 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

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*, 839–851. https://doi.org/10.1037/0003-066X.63.9.839

Archer, R. P., Wheeler, E. M. A., & Vauter, R. A. (2016). Empirically supported forensic assessment. *Clinical Psychology: Science and Practice*, *23*, 348–364. https://doi.org/10.1111/cpsp.12171

Bagby, R. M., Gralnick, T. M., Al-Dajani, N., & Uliaszek, A. A. (2016). The role of the five-factor model in personality assessment and treatment planning. *Clinical Psychology: Science and Practice*, *23*, 365–381. https://doi.org/10.1111/cpsp.12175

Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J. (2005, June). *Recognizing facial expression: Machine learning and application to spontaneous behavior*. Paper presented at the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

Becker, S. J., Spirito, A., & Vanmali, R. (2015). Perceptions of 'evidence-based practice' among the consumers of adolescent substance use treatment. *Health Education Journal*, *75*, 358–369. https://doi.org/10.1177/0017896915581061

Beidas, R. S., Stewart, R. E., Walsh, L., Lucas, S., Downey, M. M., Jackson, K., … Mandell, D. S. (2015). Free, brief, and validated: Standardized instruments for low-resource mental health settings. *Cognitive & Behavioral Practice*, *22*, 5–19. https://doi.org/10.1016/j.cbpra.2014.02.002

Bickman, L. (1996). A continuum of care: More is not always better. *American Psychologist*, *51*, 689–701. https://doi.org/10.1037/0003-066X.51.7.689

Bickman, L. (2008). A measurement feedback system (MFS) is necessary to improve mental health outcomes. *Journal of the American Academy of Child and Adolescent Psychiatry*, *47*, 1114–1119. https://doi.org/10.1097/CHI.0b013e3181825af8

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, *1*, 307–310.

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. https://doi.org/10.1136/bmj.326.7379.41

Boydell, K. M., Hodgins, M., Pignatiello, A., Teshima, J., Edwards, H., & Willis, D. (2014). Using technology to deliver mental health services to children and youth: A scoping review. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, *23*, 87–99.

Bruchmuller, K., Margraf, J., Suppiger, A., & Schneider, S. (2011). Popular or unpopular? Therapists' use of structured interviews and their estimation of patient acceptance. *Behavior Therapy*, *42*, 634–643. https://doi.org/10.1016/j.beth.2011.02.003

Buros, O. K. (1965). Foreword. In O. K. Buros (Ed.), *The mental measurements yearbooks* (6th ed., p. xxii). Lincoln, NE: University of Nebraska.

Butt, Z. (2016). In pursuit of empirically supported assessment for use in medical settings. *Clinical Psychology: Science and Practice*, *23*, 382–402. https://doi.org/10.1111/cpsp.12176

Canivez, G. L. (2013). Psychometric versus actuarial interpretation of intelligence and related aptitude batteries. In D. H. Saklofske, V. L. Schwean, & C. R. Reynolds (Eds.), *The Oxford handbook of child psychological assessments* (pp. 84–112). New York, NY: Oxford University Press.

Carlson, G. A., & Youngstrom, E. A. (2003). Clinical implications of pervasive manic symptoms in children. *Biological Psychiatry*, *53*, 1050–1058. https://doi.org/10.1016/s0006-3223(03)00068-4

Castonguay, L. G., Constantino, M. J., & Holtforth, M. G. (2006). The working alliance: Where are we and where should we go? *Psychotherapy*, *43*, 271–279. https://doi.org/10.1037/0033-3204.43.3.271

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, *66*, 7–18. https://doi.org/10.1037/0022-006X.66.1.7

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

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

De Los Reyes, A., & Aldao, A. (2015). Introduction to the special issue: Toward implementing physiological measures in clinical child and adolescent assessments. *Journal of Clinical Child and Adolescent Psychology*, *44*, 221–237. https://doi.org/10.1080/15374416.2014.891227

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin*, *141*, 858–900. https://doi.org/10.1037/a0038498

De Los Reyes, A., & Kazdin, A. E. (2004). Measuring informant discrepancies in clinical child research. *Psychological Assessment*, *16*, 330–334. https://doi.org/10.1037/1040-3590.16.3.330

Derogatis, L. (1977). *SCL-90: Administration, scoring, and procedures manual for the R(evised) version*. Baltimore, MD: Johns Hopkins University School of Medicine.

DeRubeis, R. J., Siegle, G. J., & Hollon, S. D. (2008). Cognitive therapy versus medication for depression: Treatment outcomes and neural mechanisms. *Nature Reviews Neuroscience*, *9*, 788–796. https://doi.org/10.1038/nrn2345

van der Ende, J., Verhulst, F. C., & Tiemeier, H. (2012). Agreement of informants on emotional and behavioral problems from childhood to adulthood. *Psychological Assessment*, *24*, 293–300. https://doi.org/10.1037/a0025500

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. https://doi.org/10.1207/s15374424jccp3403_7

Freeman, A. J., Youngstrom, E. A., Freeman, M. J., Youngstrom, J. K., & Findling, R. L. (2011). Is caregiver-adolescent disagreement due to differences in thresholds for reporting manic symptoms? *Journal of Child and Adolescent Psychopharmacology*, *21*, 425–432. https://doi.org/10.1089/cap.2011.0033

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

Fristad, M. A., Verducci, J. S., Walters, K., & Young, M. E. (2009). Impact of multifamily psychoeducational psychotherapy in treating children aged 8 to 12 years with mood disorders. *Archives of General Psychiatry*, *66*, 1013–1021. https://doi.org/10.1001/archgenpsychiatry.2009.112

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

Gigerenzer, G., & Muir Gray, J. A. (Eds.). (2011). *Better doctors, better patients, better decisions*. Cambridge, MA: MIT Press.

Glasgow, R. E., & Riley, W. T. (2013). Pragmatic measures: What they are and why we need them. *American Journal of Preventive Medicine*, *45*, 237–243. https://doi.org/10.1016/j.amepre.2013.03.010

Goodman, R., Ford, T., Simmons, H., Gatward, R., & Meltzer, H. (2003). Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *International Review of Psychiatry*, *15*, 166–172. https://doi.org/10.1080/0954026021000046128

Guo, T., Xiang, Y. T., Xiao, L., Hu, C. Q., Chiu, H. F., Ungvari, G. S., … Wang, G. (2015). Measurement-based care versus standard care for major depression: A randomized controlled trial with blind raters. *American Journal of Psychiatry*, *172*, 1004–1013. https://doi.org/10.1176/appi.ajp.2015.14050652

Hall, C. L., Taylor, J., Moldavsky, M., Marriott, M., Pass, S., Newell, K., … Hollis, C. (2014). A qualitative process evaluation of electronic session-by-session outcome measurement in child and adolescent mental health services. *BMC Psychiatry*, *14*, 113. https://doi.org/10.1186/1471-244X-14-113

Hatfield, D. R., & Ogles, B. M. (2007). Why some clinicians use outcome measures and others do not. *Administration & Policy in Mental Health*, *34*, 283–291. https://doi.org/10.1007/s10488-006

Haynes, S. N., & Lench, H. C. (2003). Incremental validity of new clinical assessment measures. *Psychological Assessment*, *15*, 456–466. https://doi.org/10.1037/1040-3590.15.4.456

Hopwood, C. J., Thomas, K. M., Luo, X., Bernard, N., Lin, Y., & Levendosky, A. A. (2016). Implementing dynamic assessments in psychotherapy. *Assessment*, *23*, 507–517. https://doi.org/10.1177/1073191116649658

Horvath, A. O., Del Re, A. C., Flückiger, C., & Symonds, D. (2011). Alliance in individual psychotherapy. *Psychotherapy*, *48*, 9–16. https://doi.org/10.1037/a0022186

Houle, T. T. (2009). Statistical analyses for single-case experimental designs. In D. H. Barlow, M. K. Nock, & M. Hersen (Eds.), *Single case experimental designs: Strategies for studying behavioral change* (3rd ed., pp. 271–305). Boston, MA: Pearson.

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. https://doi.org/10.1037/0003-066X.51.10.1059

Hughes, C. W., Emslie, G. J., Wohlfahrt, H., Winslow, R., Kashner, T. M., & Rush, A. J. (2005). Effect of structured interviews on evaluation time in pediatric community mental health settings. *Psychiatric Services*, *56*, 1098–1103. https://doi.org/10.1176/appi.ps.56.9.1098

Hunsley, J. (2007). Training psychologists for evidence-based practice. *Canadian Psychology*, *38*, 32–42. https://doi.org/10.1037/cp2007_1_32

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. https://doi.org/10.1037/1040-3590.17.3.251

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

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology*, *67*, 300–307. https://doi.org/10.1037/0022-006X.67.3.300

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. https://doi.org/10.1037/0022-006X.59.1.12

Jenkins, M. M., & Youngstrom, E. A. (2016). A randomized controlled trial of cognitive debiasing improves assessment and treatment selection for pediatric bipolar disorder. *Journal of Consulting and Clinical Psychology*, *84*, 323–333. https://doi.org/10.1037/ccp0000070

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. https://doi.org/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. https://doi.org/10.1037/a0025775

Jensen, A. L., & Weisz, J. R. (2002). Assessing match and mismatch between practitioner-generated and standardized interview-generated diagnoses for clinic-referred children and adolescents. *Journal of Consulting and Clinical Psychology*, *70*, 158–168. https://doi.org/10.1037//0022-006x.70.1.158

Jensen-Doss, A., & Hawley, K. M. (2010). Understanding barriers to evidence-based assessment: Clinician attitudes toward standardized assessment tools. *Journal of Clinical Child and Adolescent Psychology*, *39*, 885–896. https://doi.org/10.1080/15374416.2010.517169

Jensen-Doss, A., & Weisz, J. R. (2008). Diagnostic agreement predicts treatment process and outcomes in youth mental health clinics. *Journal of Consulting and Clinical Psychology*, *76*, 711–722. https://doi.org/10.1037/0022-006x.76.5.711

Kalibatseva, Z., & Leong, F. T. (2011). Depression among Asian Americans: Review and recommendations. *Depression Research and Treatment*, *2011*, 320902. https://doi.org/10.1155/2011/320902

Kaplan, J. (2015). *Humans need not apply: A guide to wealth and work in the age of artificial intelligence*. New Haven, CT: Yale University Press.

Kazdin, A. E. (Ed.). (1992). *Methodological issues and strategies in clinical research*. Washington, DC: American Psychological Association.

Kazdin, A. E., & Blase, S. L. (2011). Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspectives on Psychological Science*, *6*, 21–37. https://doi.org/10.1177/1745691610393527

Kazdin, A. E., & Weisz, J. R. (1998). Identifying and developing empirically supported child and adolescent treatments. *Journal of Consulting and Clinical Psychology*, *66*, 19–36.

King, R. A., & The Work Group on Quality Issues. (1997). Child and adolescent assessment practice parameter. *Journal of the American Academy of Child and Adolescent Psychiatry*, *36*, 4s–20s.

Krueger, R. F., Chentsova-Dutton, Y. E., Markon, K. E., Goldberg, D., & Ormel, J. (2003). A cross-cultural study of the structure of comorbidity among common psychopathological syndromes in the general health care setting. *Journal of Abnormal Psychology*, *112*, 437–447.

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

Lilienfeld, S. O. (2007). Psychological treatments that cause harm. *Perspectives on Psychological Science*, *2*, 53–70.

Longwell, B. T., & Truax, P. (2005). The differential effects of weekly, monthly, and bimonthly administrations of the Beck Depression Inventory-II: Psychometric properties and clinical implications. *Behavior Therapy*, *36*, 265–275. https://doi.org/doi: 10.1016/S0005-7894(05)80075-9

Lucas, G. M., Gratch, J., King, A., & Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, *37*, 94–100. https://doi.org/10.1016/j.chb.2014.04.043

Marshall, S., Haywood, K., & Fitzpatrick, R. (2006). Impact of patient-reported outcome measures on routine practice: A structured review. *Journal of Evaluation in Clinical Practice*, *12*, 559–568. https://doi.org/10.1111/j.1365-2753.2006.00650.x

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

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. https://doi.org/10.1207/s15374424jccp3403_1

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

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., … Reed, G. M. (2001). Psychological testing and psychological assessment. A review of evidence and issues. *American Psychologist*, *56*, 128–165. https://doi.org/10.1037/0003-066X.56.2.128

Miller, P. R. (2002). Inpatient diagnostic assessments: 3. Causes and effects of diagnostic imprecision. *Psychiatry Research*, *111*, 191–197. https://doi.org/10.1016/S0165-1781(02)00147-6

Morrison, J. (2014). *Diagnosis made easier: Principles and techniques for mental health clinicians* (2nd ed.). New York, NY: Guilford Press.

Mufson, L. H., Dorta, K. P., Olfson, M., Weissman, M. M., & Hoagwood, K. (2004). Effectiveness research: Transporting interpersonal psychotherapy for depressed adolescents (IPT-A) from the lab to school-based health clinics. *Clinical Child and Family Psychology Review*, 7, 251–261. https://doi.org/10.1007/s10567-004-6089-6

Mulley, A. G., & Wennberg, J. E. (2011). Reducing unwarranted variation in clinical practice by supporting clinicians and patients in decision-making. In G. Gigerenzer & J. A. Muir Gray (Eds.), *Better doctors, better patients, better decisions* (pp. 45–52). Cambridge, MA: MIT Press.

Mundt, J. C., Greist, J. H., Jefferson, J. W., Federico, M., Mann, J. J., & Posner, K. (2013). Prediction of suicidal

behavior in clinical research by lifetime suicidal ideation and behavior ascertained by the electronic Columbia-Suicide Severity Rating Scale. *Journal of Clinical Psychiatry*, *74*, 887–893. https://doi.org/10.4088/JCP.13m08398

Nezu, A. M., Ronan, G. F., Meadows, E. A., & McClure, K. S. (2000). *Practitioner's guide to empirically based measures of depression*. New York, NY: Kluwer.

Norcross, J. C., Beutler, L. E., & Levant, R. F. (Eds.). (2006). *Evidence-based practices in mental health*. Washington, DC: American Psychological Association.

Ogles, B. M. (1996). *Assessing outcome in clinical practice*. Boston, MA: Allyn and Bacon.

Ponterotto, J. G., & Ruckdeschel, D. E. (2007). An overview of coefficient alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. *Perceptual & Motor Skills*, *105*, 997–1014. https://doi.org/10.2466/pms.105.3.997-1014

Ready, R. E., & Veague, H. B. (2014). Training in psychological assessment: Current practices of clinical psychology programs. *Professional Psychology: Research and Practice*, *45*, 278–282. https://doi.org/10.1037/a0037439

Rescorla, L. A., Ginzburg, S., Achenbach, T. M., Ivanova, M. Y., Almqvist, F., Begovac, I., … Verhulst, F. C. (2013). Cross-informant agreement between parent-reported and adolescent self-reported problems in 25 societies. *Journal of Clinical Child and Adolescent Psychology*, *42*, 262–273. https://doi.org/10.1080/15374416.2012.717870

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. https://doi.org/10.1002/mpr.289

Reynolds, C. R. (2016). Contextualized evidence and empirically based testing and assessment. *Clinical Psychology: Science and Practice*, *23*, 410–416. https://doi.org/10.1111/cpsp.12181

Reynolds, C. R., & Kamphaus, R. (2015). *Behavior Assessment System for Children (BASC)* (3rd ed.). Bloomington, MN: Pearson Clinical Assessment.

Robins, E., & Guze, S. B. (1970). Establishment of diagnostic validity in psychiatric illness: Its application to schizophrenia. *American Journal of Psychiatry*, *126*, 983–986. https://doi.org/10.1176/ajp.126.7.983

Rogosa, D. (1995). Myths and methods: "Myths about longitudinal research" plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change* (pp. 3–66). Hillsdale, NJ: Erlbaum.

Safran, J. D., Muran, J. C., & Eubanks-Carter, C. (2011). Repairing alliance ruptures. *Psychotherapy*, *48*, 80–87. https://doi.org/10.1037/a0022140

Schiffman, J., Becker, K. D., & Daleiden, E. L. (2006). Evidence-based services in a statewide public mental health system: Do the services fit the problems? *Journal of Clinical Child and Adolescent Psychology*, *35*, 13–19. https://doi.org/10.1207/s15374424jccp3501_2

Sellbom, M., & Hopwood, C. J. (2016). Evidence-based assessment in the 21st century: Comments on the special series papers. *Clinical Psychology: Science and Practice*, *23*, 403–409. https://doi.org/10.1111/cpsp.12183

Shaffer, D., Fisher, P., Lucas, C. P., Dulcan, M. K., & Schwab-Stone, M. E. (2000). NIMH Diagnostic Interview Schedule for Children Version IV (NIMH DISC-IV): Description, differences from previous versions, and reliability of some common diagnoses. *Journal of the American Academy of Child and Adolescent Psychiatry*, *39*, 28–38. https://doi.org/10.1097/00004583-200001000-00014

Shirk, S. R., Karver, M. S., & Brown, R. (2011). The alliance in child and adolescent psychotherapy. *Psychotherapy*, *48*, 17–24. https://doi.org/10.1037/a0022181

Snyder, H. R., Young, J. F., & Hankin, B. L. (2016). Strong homotypic continuity in common psychopathology-, internalizing-, and externalizing-specific factors over time in adolescents. *Clinical Psychological Science*, *5*, 98–110. https://doi.org/10.1177/2167702616651076

Speer, D. C. (1992). Clinically significant change: Jacobson and Truax (1991) revisited. *Journal of Consulting and Clinical Psychology*, *60*, 402–408. https://doi.org/10.1037/0022-006X.60.3.402

Speer, D. C. (1993). "Clinically significant change: Jacobson and Truax (1991) revisited": Correction. *Journal of Consulting and Clinical Psychology*, *61*, 27. https://doi.org/10.1037/0022-006X.61.1.27

Spilka, M. J., & Dobson, K. S. (2015). Promoting the internationalization of evidence-based practice: Benchmarking as a strategy to evaluate culturally transported psychological treatments. *Clinical Psychology: Science and Practice*, *22*, 58–75. https://doi.org/10.1111/cpsp.12092

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. https://doi.org/10.1002/jclp.20373

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.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use* (5th ed.). New York, NY: Oxford University Press.

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. https://doi.org/S0005-7894(08)00088-9 [pii]

Susskind, R., & Susskind, D. (2015). *The future of the professions: How technology will transform the work of human experts*. New York, NY: Oxford University Press.

Swift, J. K., Greenberg, R. P., Whipple, J. L., & Kominiak, N. (2012). Practice recommendations for reducing premature termination in therapy. *Professional Psychology: Research and Practice*, *43*, 379–387. https://doi.org/10.1037/a0028291

Thissen, D., Liu, Y., Magnus, B., Quinn, H., Gipson, D. S., Dampier, C., … DeWalt, D. A. (2016). Estimating minimally important difference (MID) in PROMIS pediatric measures using the scale-judgment method. *Quality of Life Research*, *25*, 13–23. https://doi.org/10.1007/s11136-015-1058-8

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational & Psychological Measurement*, *60*, 174–195.

Tryon, G. S., & Winograd, G. (2011). Goal consensus and collaboration. *Psychotherapy (Chic)*, *48*, 50–57. https://doi.org/10.1037/a0022061

Vaz, S., Falkmer, T., Passmore, A. E., Parsons, R., & Andreou, P. (2013). The case for using the repeatability coefficient when calculating test-retest reliability. *PLoS ONE*, *8*, e73990. https://doi.org/10.1371/journal.pone.0073990

Vollmer, T. R., & Northup, J. (1996). Some implications of functional analysis for school psychology. *School Psychology Quarterly*, *11*, 76–92. https://doi.org/10.1037/h0088922

Wasserman, G. A., McReynolds, L. S., Ko, S. J., Katz, L. M., Cauffman, E., Haxton, W., & Lucas, C. P. (2004). Screening for emergent risk and service needs among incarcerated youth: Comparing MAYSI-2 and Voice DISC-IV. *Journal of American Academy of Child and Adolescent Psychiatry*, *43*, 629–639. https://doi.org/10.1097/00004583-200405000-00017

Wasserman, G. A., McReynolds, L. S., Lucas, C. P., Fisher, P., & Santos, L. (2002). The Voice DISC-IV with incarcerated male youths: Prevalence of disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, *41*, 314–321. https://doi.org/10.1097/00004583-200203000-00011

Watkins, M. W., Kush, J. C., & Glutting, J. J. (1997). Prevalence and diagnostic utility of the WISC-III SCAD profile among children with disabilities. *School Psychology Quarterly*, *12*, 235–248. https://doi.org/10.1037/h0088960

Weersing, V. R., & Weisz, J. R. (2002). Community clinic treatment of depressed youth: Benchmarking usual care against CBT clinical trials. *Journal of Consulting and Clinical Psychology*, *70*, 299–310. https://doi.org/10.1037/0022-006X.70.2.299

Weisz, J. R. (2004). Treatment dissemination and evidence-based practice: Strengthening intervention through clinician-researcher collaboration. *Clinical Psychology: Science and Practice*, *11*, 300–307. https://doi.org/10.1093/clipsy/bph085

Weisz, J. R., Chorpita, B. F., Frye, A., Ng, M. Y., Lau, N., Bearman, S. K., … Hoagwood, K. E. (2011). Youth Top Problems: Using idiographic, consumer-guided assessment to identify treatment needs and to track change during psychotherapy. *Journal of Consulting and Clinical Psychology*, *79*, 369–380. https://doi.org/10.1037/a0023307

Weizenbaum, J. (1984). *Computer power and human reason*. New York, NY: Penguin.

Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., … Bossuyt, P. M. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine*, *155*, 529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

Wollburg, E., & Braukhaus, C. (2010). Goal setting in psychotherapy: The relevance of approach and avoidance goals for treatment outcome. *Psychotherapy Research*, *20*, 488–494. https://doi.org/10.1080/10503301003796839

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 of Consulting and Clinical Psychology*, *73*, 800–807. https://doi.org/10.1037/0022-006X.73.5.800

Youngstrom, E. A. (2013). Future directions in psychological assessment: Combining evidence-based medicine innovations with psychology's historical strengths to enhance utility. *Journal of Clinical Child and Adolescent Psychology*, *42*, 139–159. https://doi.org/10.1080/15374416.2012.736358

Youngstrom, E. A., Choukas-Bradley, S., Calhoun, C. D., & Jensen-Doss, A. (2014). Clinical guide to the evidence-based assessment approach to diagnosis and treatment. *Cognitive and Behavioral Practice*, *22*, 20–35. https://doi.org/10.1016/j.cbpra.2013.12.005

Youngstrom, E. A., & Frazier, T. W. (2013). Evidence-based strategies for the assessment of children and adolescents: Measuring prediction, prescription, and process. In D. J. Miklowitz, W. E. Craighead, & L. Craighead (Eds.), *Developmental psychopathology* (2nd ed., pp. 36–79). New York, NY: Wiley.

Youngstrom, E. A., Genzlinger, J. E., Egerton, G. A., & Van Meter, A. R. (2015). Multivariate meta-analysis of the discriminative validity of caregiver, youth, and teacher rating scales for pediatric bipolar disorder: Mother knows best about mania. *Archives of Scientific Psychology*, *3*, 112–137. https://doi.org/10.1037/arc0000024

Youngstrom, E. A., Meyers, O. I., Youngstrom, J. K., Calabrese, J. R., & Findling, R. L. (2006). 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. https://doi.org/10.1017/S0954579406060494

Youngstrom, E. A., & Van Meter, A. (2016). Empirically supported assessment of children and adolescents. *Clinical Psychology: Science and Practice*, *23*, 327–347. https://doi.org/10.1111/cpsp.12172

Youngstrom, E. A., Zhao, J., Mankoski, R., Forbes, R. A., Marcus, R. M., Carson, W., … Findling, R. L. (2013). Clinical significance of treatment effects with aripiprazole versus placebo in a study of manic or mixed episodes associated with pediatric bipolar I disorder. *Journal of Child and Adolescent Psychopharmacology*, *23*, 72–79. https://doi.org/10.1089/cap.2012.0024

Received September 22, 2016; revised March 29, 2017; accepted May 3, 2017.