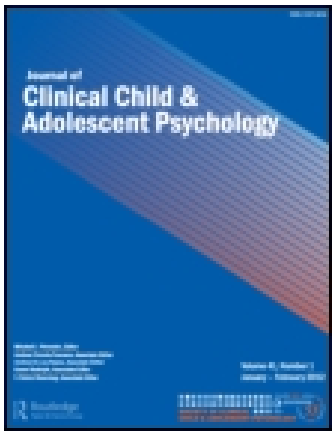

Journal of Clinical Child & Adolescent Psychology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hcap20

Commentary: Moving Toward Cost-Effectiveness in Using Psychophysiological Measures in Clinical Assessment: Validity, Decision Making, and Adding Value

Eric A. Youngstrom [a] & Andres De Los Reyes [b]
[a] Department of Psychology , University of North Carolina at Chapel Hill
[b] Department of Psychology , University of Maryland at College Park
Published online: 09 Feb 2015.

Journal of Clinical Child & Adolescent Psychology, 44(2), 352–361, 2015

ISSN: 1537-4416 print/1537-4424 online
DOI: 10.1080/15374416.2014.913252

Routledge
Taylor & Francis Group

Commentary: Moving Toward Cost-Effectiveness in Using Psychophysiological Measures in Clinical Assessment: Validity, Decision Making, and Adding Value

Eric A. Youngstrom
Department of Psychology, University of North Carolina at Chapel Hill

Andres De Los Reyes
Department of Psychology, University of Maryland at College Park

Psychophysiological measures offer a variety of potential advantages, including more direct assessment of certain processes, as well as provision of information that may contrast with other sources. The role of psychophysiological measures in clinical practice will be best defined when researchers (a) switch to research designs and statistical models that better approximate how clinicians administer assessments and make clinical decisions in practice, (b) systematically compare the validity of psychophysiological measures to incumbent methods for assessing similar criteria, (c) test whether psychophysiological measures show either greater validity or clinically meaningful incremental validity, and (d) factor in fiscal costs as well as the utilities that the client attaches to different assessment outcomes. The statistical methods are now readily available, along with the interpretive models for integrating assessment results into client-centered decision making. These, combined with technology reducing the cost of psychophysiological measurement and improving ease of interpretation, poise the field for a rapid transformation of assessment practice, but only if we let go of old habits of research.

Technology has changed radically in the space of our own lifetimes. The first cell phone was introduced in 1973 and weighed 2 pounds. The modern Internet launched in the early 1980s. Now smartphones have melded telephony and computing, creating new opportunities for information exchange, data capture, and interpretation. There are similar changes in the costs and efficiency of many other technologies, such as the advent of wireless, ambulatory devices for assessing heart rate and blood pressure that markedly reduce the costs, client burden, and training needed to assess cardiovascular functioning. It is now possible to reengineer our assessment methods to take advantage of these technologies. There is the potential for the switch to go far beyond a merely ''paperless'' version of gathering the same information that we currently collect using traditional clinical tools, such as symptom checklists and diagnostic interviews. Instead, new types of information, such as monitoring social media activity or using the smartphone as an actimeter, and new forms of decision support may transform clinical evaluation (Kazdin & Blase, 2011).

The articles in this special issue represent a step in deploying psychophysiological methods in the context of psychological research. They show sophistication in the implementation of techniques and in the statistical analyses. There are at least three additional steps that will help to fully evaluate and realize the potential of these methods to improve clinical care at the level of the individual client. These steps are (a) to appraise critically the validity of these psychophysiological methods for specific clinical purposes, (b) to analyze the results in a way that explicitly focuses on individual decision

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

making, and (c) to appraise whether these methods demonstrate added value when competing with or complementing other available assessment methods. This commentary elaborates each of these ideas next, with some examples as well as recommendations for further reading about related concepts.

CLINICAL VALIDITY: PREDICTING CRITERIA, PRESCRIBING TREATMENT, MEASURING PROCESS OR PROGRESS

When working with a client, the clinician engages in a form of hypothesis testing that has many similarities to the social science research process but also some important differences. A clinical formulation involves hypotheses about the factors causing or maintaining a problem. Assessment findings support some hypotheses and decrease the chances of others. Good assessment systematically considers some plausible rival hypotheses. Clinical hypothesis testing also struggles with the challenges of false positive (Type I) and false negative (Type II) errors in assessment. After arriving at a formulation, the clinician and client work, ideally collaboratively, to decide the course of treatment and the definitions of successful outcome, similar to planning the next steps in a program of research and establishing operational definitions.

Despite these similarities, there are also major differences in the type of evidence admitted, the analytic models, and the standards for appraising the findings. What constitutes "acceptable" reliability changes when the context shifts from large group research to making high-stakes decisions about individuals. Whereas a reliability of .50 might be adequate if using a short scale to measure a construct in a large sample (Nunnally, 1967) or with a large number of repetitions (Bakeman, McArthur, Quera, & Robinson, 1997), making a treatment or classification decision about an individual based on a single measurement might warrant a reliability coefficient higher than .92 (Kelley, 1927). In group data, lower reliability attenuates the observed effect sizes, but large samples may compensate enough that statistical power is adequate still to arrive at the correct overall interpretation. For the individual client, it matters a great deal whether their score is accurate. Seemingly minor quantitative differences between scores on a clinical instrument (e.g., seven vs. eight symptom endorsements on a structured interview) may result in qualitatively different choices about intervention or placement (e.g., eight endorsed symptoms results in a diagnosis and treatment referral, whereas seven symptoms would not).

The realm of clinical applications also alters relative importance of different types of validity. Two major thresholds in the clinical enterprise guide decisions about when to start or suspend assessment and when to initiate different types of treatment. Evidence-based medicine (EBM) describes these as the Wait-Test Threshold, a probability threshold below which a diagnosis or concern is considered functionally ruled out and above which more assessment is needed; and the Test-Treat Threshold, above which a diagnosis is sufficiently likely that it becomes a focus of treatment (Straus, Glasziou, Richardson, & Haynes, 2011). For these purposes, the most valuable forms of validity are predictive validity, discriminative validity (the ability to discriminate between categories) and prescriptive validity (Youngstrom & Frazier, 2013). If the assessment predicts high-stakes criteria, such as suicide attempt or recividism, then it shows clinically relevant predictive validity. Distinguishing between those with or without a particular diagnosis, or probable treatment responders versus nonresponders, would be examples of discriminative validity. Prescriptive validity encompasses matching the treatment to the target problems, as well as evaluating potential moderators of treatment effects. Once treatment is initiated, then assessment moves to a monitoring role, gauging whether the intervention is producing the desired effects, as well as watching for potential adverse responses. These core roles form the "3 Ps" of clinical assessment: Predicting important criteria, Prescribing a treatment, and informing the Process or Progress of outcome (Youngstrom & Frazier, 2013).

The 3 P principles facilitate parsimony in assessment. If the combination of information available is enough to move the probability of a diagnosis below the Wait-Test Threshold, then we have enough evidence to make a decision, and we do no further assessment (unless treatment response is poor, or some new finding emerges that changes the picture). If the probability is high enough to clear the Test-Treat Threshold, then we suspend assessment related to diagnosis, and focus on process and progress measures. The EBM algorithm makes assessment focus on what is necessary to guide the next clinical action and stops assessment as soon as sufficient information is accumulated. Parsimony also derives from minimizing redundancy. If it would be helpful to know cognitive ability, then one valid test administration is usually enough, and little incremental value would come from completing a second or third ability battery. When several different candidate measures are available, then one evidence-based approach for selecting among them would be to pick the one that has the strongest validity coefficient. Another would be to continue using whatever is the convenient incumbent measure (e.g., whatever we already are comfortable using) until some other test demonstrates significantly greater validity, based either on meta-analysis (Hasselbad & Hedges, 1995) or head-to-head comparison (Youngstrom & Frazier, 2013).

Thus the first hurdles for psychophysiological assessment to clear are basic psychometric ones: Is the result reproducible, with sufficient accuracy to guide individual decision making? Does the measure show statistically significant criterion validity? Does it correlate with established measures of clinical constructs or Research Domain Criteria (Sanislow et al., 2010) dimensions? Does it discriminate between groups with different diagnoses or longitudinal trajectories? Is it sensitive to treatment effects? If a measure fails to accomplish these at a statistically significant level, we can probably rule it out as having a role in clinical contexts. That is a strong pronouncement, but it follows from the fact that whereas statistical power increases with larger sample sizes, the clinician works at the level of the individual case. If a method fails to demonstrate a significant signal with a total sample size of 30 or 300 drawn from the relevant population, it is unlikely to deliver enough power and precision to help steer choices about a single person.

As researchers, we can do a better job picking our statistical methods to match more closely how clinicians need to apply the information (Cumming, 2014). For decades, researchers have sorted cases into groups, such as those with versus without depression, and then used group-based statistics such as the *t* test or analysis of variance (or nonparametric alternatives; e.g., S. Cohen, Masyn, Mastergeorge, & Hessl, this issue) to see whether the average scores on the assessment tool differ between groups on average.

Of necessity, clinicians work in the opposite direction. They do not first assign diagnoses to all of their cases, then given them all a psychophysiological measure, and finally test whether the scores differ on average between diagnostic groups. They need to give the assessment first, and understand how it informs the probability that a case belongs to a particular category. Receiver Operating Characteristic (ROC) analysis is a much better model of how a clinician proceeds: it directly evaluates the performance of a measure at discriminating between groups (McFall & Treat, 1999; Swets, Dawes, & Monahan, 2000).

ROC examines the balance between diagnostic sensitivity (the rate of detection among true cases) versus specificity (the rate of correctly identifying cases that do not have the target condition) across all observed scores on the assessment variable. It is a nonparametric procedure, and the test of significance of the area under the ROC curves is equivalent to a Mann-Whitney *U* test. It is possible to convert Cohen's *d* effect size values into an Area Under the Curve (AUC; Hasselbad & Hedges, 1995), reinforcing the point that ROC is a viable form of analysis in any situation where researchers have commonly been doing *t* tests and other analyses comparing two groups. However, ROC is much more informative about how the assessment does at classification. The conventions for interpreting effect sizes also reveal the heightened challenge involved in classifying individuals correctly: Table 1 lists common benchmarks for *d* and AUC and converts them into each other. What is commonly considered a "large" *d* is only a mediocre AUC value.

Similarly, logistic regression offers an alternative model that lets the dependent variable be a pair of categorical options, such as "initiate treatment" versus "wait," or "diagnosis present" versus "absent." Logistic regression places few restrictions on the independent variables, and it affords the range of block entry models and tests for interaction effects, covariates, and suppressors that are familiar from ordinary least squares regression (Hosmer & Lemeshow, 2000). Logistic regression makes it possible to evaluate sets of predictors, adjusting for demographic and clinical covariates, and examining incremental validity of combinations of variables. The predicted probability score from logistic regression can also be used as the input in a ROC analysis, turning logistic regression into a method for integrating multiple predictors into a single classification function.

What are the implications for psychophysiological research? If we want the psychophysiological tools to play more of a role in diagnosis and treatment planning, then we should (a) switch to ROC analyses and logistic regression as primary tools for evaluating psychophysiological measures, as a way of gauging accuracy in clinically relevant terms, as well as testing potential interaction effects; (b) start directly comparing the validity of these tools to other available assessment methods, either via meta-analysis or direct comparisons in the same sample; and (c) test whether psychophysiological methods can provide incremental validity when deployed after using other less expensive or more familiar methods. There are well-established statistical approaches for testing whether the correlation or regression weight found in one sample differs significantly from that in a different sample, as would be the case if benchmarking against values reported in a

TABLE 1
Common Benchmarks for Interpreting *d* (J. Cohen, 1988) and Area Under the Curve from Receiver Operating Characteristic Analyses (Swets et al., 2000)

Cohen's d	*AUC*
.000	.500 = chance performance
.200 = small	.556
<.358	<.600 = poor
.500 = medium	.638
<.742	<.700 = fair
>.742	>.700 = good
.800 = large	.714
1.812	>.900 = excellent

TABLE 2
Summary of Design Features and Conversion of Largest Effect Size Into a Clinical Decision-Making Metric for Articles in the Special Issue

	Largest Effect Size						
Article	*Analytic Method*			*N*	*Physiological Measure*	*Criterion*	*AUC*
Bress, Meyer, & Hajcak, 2014	$r = .54$ (FN, CDI)	Correlation, regression	Differentiating Anxiety & Depression	25	FN of Evoked Response Potential	SCARED, CDI (youth and parent)	.82
S. Cohen et al., 2014	Fragile X vs. Autism; Autism vs. Typical comparisons (nonparametric, no effect size)	Kruskal-Wallis (4groups)	Autism & Fragile X	52	Electrodermal activity, potentiated startle, vagal tone	Diagnostic category (mean differences in ranks)	[a]
De Los Reyes et al., 2014	Context effect (giving a speech) on mean heart rate: $B = 3.12$ ($SE = .76$); Wald $\chi^2 = 16.62$, $p < .001$	Generalized estimating equation	Context effect (giving a speech) on participant's mean heart rate	22	Change in mean heart rate value	Giving a speech in front of live audience	b
Gatzke-Kopp, Greenberg, & Bierman, 2014	Parasympathetic reactivity to specific emotions moderates response to intervention for aggression; change in R^2 of .11 in post hoc subgroup analyses	Multiple regression	Incremental value of RSA measures after controlling for teacher report and other covariates	101 for teacher report (or 69 for peer nomination)	RSA– baseline and response to emotional film clip	Teacher ratings of emotion regulation & externalizing behavior problems; peer nominations of aggression	b
Leitzke et al., 2014	Maltreated youth display a blunted blood pressure response to an acute interpersonal stressor, partial $\eta^2 = .07$	Repeated measures ANOVA	Abuse status (and indirectly, risk of adverse response to stressor)	111, $n = 34$ with documented history of abuse	Change in Systolic Blood Pressure × Abuse Status interaction	Abuse Status (treated as factor in ANOVA)	.65
McLaughlin, Rith-Najaran, Dirks, & Sheridan, 2014	Low Vagal tone magnifies association between psychosocial stress exposure and YSR internalizing (but not externalizing) problems; beta = −.52 for an interaction term	Multiple regression with covariates, interaction terms, and probing	Interaction term between RSA, psychosocial stressor (and sometimes gender) in predicting internalizing subscores	157	RSA	YSR anxious/ depressed score	b
Moser, Durbin, Patrick, & Schmidt, 2014	$r = .52$ between Child Behavior Questionnaire total (parent completed) and parietal asymmetry	Correlation; preliminary study with no correction for Type I error	Correlation between Child Behavior Questionnaire total (parent completed) and parietal asymmetry	31	Parietal asymmetry		.81

Note. AUC = Area Under the Curve; CDI = Child Depression Inventory; FN = Feedback Negativity; RSA = respiratory sinus arrhythmia; ANOVA = analysis of variance; YSR = Youth Self Report.

[a] Eta-squared ranged from .03 to .31; all reported as 3 *df* so cannot estimate AUC.

[b] Coefficients from multiple regression cannot be converted into other effect size metrics, because they are contingent on the other predictors included in the equation (Lipsey & Wilson, 2001). Note that Franklin, Glenn, Jamieson, and Nock (2014) was a review paper that did not report any statistical analyses.

technical manual or prior publication (J. Cohen & Cohen, 1983), and similar methods are available for comparing AUC estimates (DeLong, DeLong, & Clarke-Pearson, 1988; Hanley & McNeil, 1983). The statistical power of these tests increases when measures are compared head-to-head in the same sample, and direct comparisons also control for a wide variety of factors that complicate comparisons based on different samples, such as changes in comorbidity or diagnostic interview. Appraising the psychophysiological measures through this series of lenses should help reveal whether (a) the methods show high-enough validity coefficients to justify the transfer from purely research to clinical contexts; (b) whether they are superior to existing measures in terms of validity—ignoring costs for the moment, and focusing first on validity; and (c) whether the psychophysiological measures might show promise as incremental measures in a more circumscribed role in clinical assessment.

Table 2 lists the studies in the special issue that report data on psychophysiological measures, along with the primary statistical analytic method and the largest effect size reported. The table also converts the largest effect size reported in the study to an estimated AUC value, where possible. These offer an estimate of the "best-case scenario" of using the psychophysiological measure as a primary method of discriminating clinical groups or making decisions based on the scores. They are best-case estimates both because we are cherry-picking the largest of a set of coefficients, but also because we are assuming that the criterion has good clinical validity (Youngstrom, Meyers, Youngstrom, Calabrese, & Findling, 2006). Examining the table suggests several things. One is that the conventional way of analyzing and reporting results does not make it easy to consider clinical utility; even armed with all the formulae for converting effect sizes, a motivated meta-analyst would not be able to include all of the studies reported here. A second observation is that many of the data sets could be reanalyzed in ways that might make clinical applications more obvious. For example, Bress, Meyer, and Hajcak (this issue) reported an effect size of $r = .54$ between evoked response potential to negative feedback and self-reported depressive symptoms (see Table 2). This translates to an estimated area under the curve of .82 for discriminating cases of depression. Using another formula (#13 in Hasselbad & Hedges, 1995), we can project what combination of diagnostic sensitivity and specificity might be attainable based on the overall discriminating power of the test. A curve with an AUC of .82 indicates that a threshold delivering specificity of .90 would have an estimated sensitivity of .54, whereas .80 specificity would yield .72 sensitivity, and .70 specificity would achieve .82 sensitivity. Of course, actual results with real data would vary around these projections, but they provide a clear sense of the potential clinical application of the measure. A third general conclusion based on the effect sizes reported in this issue is that even the best-case estimates of diagnostic validity fall in a range where they will be unlikely to supplant other available methods. The AUC values for parent checklists for bipolar disorder (Youngstrom et al., 2004), anxiety (Van Meter et al., in press), or attention deficit hyperactivity disorder (Warnick, Bracken, & Kasl, 2008) tend to hover in the .7 to .8 range, even under clinically realistic conditions with high rates of comorbidity and significant rates of diagnoses and impairment in the comparison group. If these are considered the incumbent measures, then the psychophysiological challengers look unlikely to depose them.

Instead, research should concentrate on how these measures could augment and complement other assessment strategies, perhaps investigating how psychophysiological assessment could be useful in subgroups or as a moderator of other environmental risk factors (Kraemer, 1992). For instance, adolescents meeting criteria for social anxiety disorder and healthy community controls both show physiological habituation to stressful social interactions (e.g., one-on-one interactions and public speaking) when assessed using ambulatory heart rate monitors (Anderson & Hope, 2009). Where the patient and control groups differed was that patients were more likely to believe that they had stable and high levels of physiological arousal throughout the social interactions. The physiological measures told a different story, making it possible to use the therapy session to train patients to recognize when their bodies physiologically habituate to socially stressful scenarios. In the context of mental health assessment, physiological measures may yield the most clinically valid information when used in conjunction with (i.e., not in replacement of) existing, well-established clinical tools. Logistic regression provides an analytic framework for gauging incremental validity, as well as modeling the discriminatory power of combinations of variables for making clinical decisions.

DECISION MAKING ABOUT INDIVIDUAL CASES

ROC, logistic regression, and other forms of classification analyses are a big step in the direction of clinical application, but there is even more that could be done to fully realize the promise of psychophysiological measures in the clinic. Ideally, assessment should inform about the status of the individual client. EBM has developed a framework for taking assessment findings and integrating them with other information about base

rates, risk factors, and prior testing to generate a revised probability of diagnosis (Straus et al., 2011). The mechanics are based on Bayes's Theorem, and EBM has developed tools such as probability nomograms and web or smartphone applications that eliminate the need for the clinician to do any algebra to integrate the information.

The raw material required as an input for the individual decision-making process is a "diagnostic likelihood ratio" (DLR), which is the rate at which the result would be observed in those with the target diagnosis, divided by the rate at which the same result would be observed in those without the diagnosis. For example, if a hypothetical level of vagal tone function was observed in 20% of cases with clinically severe aggression versus only 5% of cases without aggression, then the DLR would be $20\%/5\% = 4.0$, and if a clinician observed similar levels of vagal function in a case at the clinic, then the odds of that case having severe aggression would be four times more likely. That information becomes most helpful when combined with other facts about the case, such as their prior level of risk for aggression, whether they had high teacher-reported externalizing problems, and so forth. When all of these pieces of information can be reexpressed in the metric of DLRs, then the EBM framework makes it easy to synthesize all of the information into a personalized estimate for that case (Jaeschke, Guyatt, & Sackett, 1994a, 1994b).

There are a variety of ways that researchers can estimate DLRs, or that clinicians could estimate them by converting other information. Diagnostic sensitivity and specificity can be converted into two DLR estimates—one for those who test in the positive range, and the others who test negative. If the raw data are available, then scores can be grouped into quintiles (e.g., Van Meter et al., in press), and DLRs estimated separately for each band, potentially retaining more information from the measure (Jaeschke et al., 1994b). If there are normative data for the test in both clinical and nonclinical population, then it is possible to use the percentiles to estimate DLRs (Frazier & Youngstrom, 2006). Relative to other analytic methods, it is pleasantly simple to calculate the DLRs, and they greatly enhance the application of results to clinical decision making.

There has been similar discussion about the difference between group-based effect sizes versus individualized measures of clinical outcome, as well as diagnostic accuracy. The literature on clinically significant change has articulated a variety of methods for evaluating individual progress and outcomes in therapy. One of the most well known is the model from Jacobson and colleagues, defining clinically significant change as involving two components: reliable improvement, combined with shifting scores past one of three benchmarks defined by normative data (Jacobson, Roberts, Berns, & McGlinchey, 1999). Jacobson's reliable change index divides the individual's change score by the standard error of the difference. Ratios exceeding 1.65 would be 90% likely to reflect real change, and values greater than 1.96 would exceed 95% confidence. The necessary ingredient to make the reliable change index is the standard error of the difference—easy to calculate with the data, and often impossible to estimate from the published reports. Adding the standard error of the difference would immediately make it easier for clinicians to use psychophysiological measures in the context of evaluating individual treatment response (Youngstrom & Frazier, 2013).

The second part of the Jacobson definition, the benchmarking against reference distributions, involves having access to samples where the range of scores is known in a large group of cases with the condition of interest as well as a nonclinical reference group. Jacobson et al. proposed three benchmarks: moving the case *Away from the clinical range of functioning*—with a suggested operational definition of being more than 2 standard deviations away from the clinical average, moving the case *Back into the normal range of functioning*—operationalized as being within 2 standard deviations of the nonclinical average, or crossing *Closer to the nonclinical than clinical distribution*—traversing the weighted average of the two group means (Jacobson et al., 1999). Again, these benchmarks are simple to generate with the raw data yet are rarely reported. Adding benchmarks to research reports, particularly in instances that have large samples, would make it easier for clinicians to apply the psychophysiological measures in practice.

ADDING VALUE TO THE ASSESSMENT PROTOCOL

For psychophysiological measures, or any other new assessment method, to earn a place in the clinical toolkit, it should show added value. This is a higher hurdle than simple statistical significance. Instead of comparing the effect size to a null hypothesis of r or $d = 0$ (or $AUC = .50$), the focus shifts to comparing the performance of the new method to what can be accomplished by existing alternatives (see also Cumming, 2014). Table 3 lays out some heuristics for critically appraising the new method and deciding whether it is worth incorporating into clinical practice. One rule of thumb is that effect sizes need to be at least medium in magnitude under clinically realistic, generalizable conditions if the tool is going to influence clinical decision making about individual cases. Even clearing

TABLE 3
Some Simple Heuristics for Considering Costs When Potentially Adding a Measure to an Assessment Protocol

Research Evidence	*Clinical Decision*	*Rationale*
Not statistically significant	Do not add new measure	Not valid for purpose
Statistically significant, but small effect size	Do not add . . .	Small effects not adequate for making decisions about individuals
Statistically significant, but less valid than another test that costs same or less	Do not add . . .	More cost efficient alternative available
Statistically significant, medium or large effect, but sample less valid than design for other measures with comparable effect sizes	Do not add . . .	Competing measures produce similar or better results under conditions with greater external validity; new measure likely to have shrinkage of valid in more clinically representative designs
Similar or larger effect size than incumbent measures in a head to head comparison or meta-analytic comparison adjusting for design quality, but costs more than incumbent	Do not add, *unless* can demonstrate incremental value that is sufficient to justify increased costs	Incumbent measure is a more cost efficient way of achieving same assessment validity
Shows incremental validity–under clinically representative conditions–that is large enough to justify costs, at least for subset of cases	Adopt as an augmentation for indicated cases	By sequencing the tests in an optimal order, can limit the costs and errors associated with testing everyone, but keep some of the benefits of incremental value for subset of cases
Shows superior validity over existing measures even after adjusting for costs and design features	Switch to new measure	We have a new champion!

that hurdle, if the new method shows smaller validity coefficients than the incumbent methods under similar conditions, then there is no practical reason to adopt the new method. If the new method can demonstrate incremental value, then it is worth exploring further. Incremental validity often is quantified as a partial *r* in a regression equation, but it also could be useful to add psychophysiological measures for specific subgroups. Classification trees and related methods can help examine whether variables have value in subgroups or as statistical moderators (Kraemer, 1992; Strobl, Malley, & Tutz, 2009).

If the new method demonstrates superior validity, or clinically meaningful incremental validity, then the last hurdle is cost comparison. There are several considerations with regard to cost. One is the fixed cost associated with purchasing the equipment and materials needed for the assessment. Psychophysiological measures involve a substantial initial investment, and they also may require training to administer and interpret accurately (cf. De Los Reyes et al., this issue, for an innovative approach to reducing the costs associated with interpretation). There also are costs per administration, which include the clinician or psychometrician's time as well as any consumable material.

Costs and benefits are a matter not just of money but also of values. As procedures become more invasive or uncomfortable, the balance shifts against using them if less burdensome methods could accomplish similar validity. In addition, there are costs and benefits attached to the diagnostic outcomes. The perceived value of an accurate diagnosis (true positive result) versus accurately ruling out a diagnosis (true negative result) may not be the same. The costs of a false positive diagnosis include unwarranted worry, potential stigma, and all the risks and expenses associated with treatment, whereas the costs of a false negative result involve failing to intervene, or selecting a suboptimal or mismatched treatment. These perceived utilities are rarely equal across all four scenarios, and they probably vary across individuals depending on personal beliefs and cultural mores.

Although complicated, these issues are not insurmountable. In fact, there are multiple quantitative frameworks available that can integrate these costs into the decision-making model. Swets and colleagues have a model that considers the utilities attached to the false positives, false negatives, true positives, and true negatives to determine the optimal threshold for scoring a measure in a ROC framework (Swets et al., 2000). Kraemer (1992) has a model that goes a step further, synthesizing the fixed costs and unit costs of a test with the utilities of each outcome, provided that these can be reexpressed in terms of dollars. Kraemer's model is one of the most comprehensive, providing a unified research framework for integrating base rates with costs and benefits to determine locally optimized decision thresholds on a test. Both the Swets and the Kraemer approach depend on having access to raw data on a validation sample, and both involve changing the decision threshold on a measure based on specific considerations that could change across clinical settings. In this regard, both approaches are options only for researchers, not for practicing clinicians; and generalizability needs to be carefully appraised. EBM offers a third model that is both feasible for a clinician without access to raw group

data and shifts the focus back on the individual client. Effect sizes such as the Number Needed to Treat and Number Needed to Harm can be combined into a Likelihood of Help versus Harm, which in turn can be weighted by client preferences for each of the four potential diagnostic outcomes (Straus et al., 2011). Limitations of the EBM approach are that it is not geared toward considering the fiscal costs associated with testing and that it assumes dichotomous outcomes. However, it avoids the problems of locally specific cut scores, and it dovetails well with the rest of the mechanics of the EBM clinical decision-making process.

Psychophysiological measures may also have a niche in measuring process or progress during treatment. Biofeedback interventions are a compelling example, where the physiological measure becomes a central tool for implementing the intervention and gauging response. These applications do not map neatly into either the diagnostic decision threshold or the Jacobsonian clinically significant change model, but that does not imply that they have less worth. Instead, progress measurement is another growth opportunity for clinically relevant research.

Overall, these sorts of cost–benefit analyses represent a major area for new research, and one where psychology is particularly suited to make contributions due to the discipline's expertise in assessing values and attitudes. Researchers could employ the Swets or Kraemer methods to work out assessment algorithms optimized for certain contexts, and then evaluate how much these algorithms change under different conditions—such as changes in base rate due to referral patterns, or big decreases in the fixed cost of an assessment method. Clinicians can explore integrating client values and preferences into the assessment process using the EBM framework. In the interim period, before the next wave of cost-oriented research is published, clinicians can also use the heuristics in Table 3 to critically appraise the literature. It is a relatively narrow conceptual space where new methods might actually warrant clinical attention. Using fMRI as an example, most research has focused on comparing a defined patient population to healthy controls. Viewed from the perspective of a practicing clinician, comparisons of ill versus well are clinically trivial; and methodologists have long been aware that these designs produce exaggerated estimates of validity (Bossuyt et al., 2003; Zhou, Obuchowski, & McClish, 2002). Clinicians and consumers can pay more attention to imaging studies when the designs include clinically realistic comparison groups; and even then, ties in terms of validity will favor the incumbent assessment methods because of the markedly greater expense attached to fMRI, leaving aside client attitudes toward being in a scanner. It is only when fMRI demonstrates greater accuracy, or clinically meaningful incremental value, under generalizable conditions, addressing ecologically important criteria (Berkman & Falk, 2013) that questions of cost require scrutiny. Viewed through the same heuristics, the lower cost and relative ease of administration of newer methods for assessing blood pressure in real time (Leitzke, Hilt, & Pollak, this issue) make them promising candidates for clinical utility, provided they demonstrate meaningful criterion or incremental validity. The research design and statistical issues remain the same, but the lower costs and burden create an advantage for many psychophysiological measures to show an extra surge close to the finish line of clinical utility.

DISCUSSION

Psychophysiological measures have many conceptual advantages. They offer a distinct source of information, not sharing method variance with self-report or other collateral informants. Consequently, they provide a powerful convergent measure of latent constructs (Campbell & Fiske, 1959). They also potentially bypass problems of social desirability, malingering, lack of insight, or other artifacts that beset other information sources. More recent studies are also providing a nuanced understanding of how psychophysiological measures provide a window into susceptibility, resilience, and interactions with environmental factors. Technological advances are reducing the costs and burden associated with gathering psychophysiological data, bringing us closer to the goal of integrating these methods into clinical practice.

At the same time, much work needs to be done to accelerate and promote the uptake of these methods in clinical assessment. Research designs need to include clinically representative samples, not solely relying on "distilled" designs that enhance internal validity at the expense of generalizability—directly analogous to the role of effectiveness versus efficacy studies of treatment effects. The statistical analyses and reporting need to shift to models—such as ROC, logistic regression, and classification trees—that better approximate the ways that clinicians need to apply assessments to individual cases. Researchers can also compare their assessment validity coefficients to benchmarks for incumbent measures using meta-analytic methods, and they can use the same heuristics in Table 3 to guide how they present findings. Rather than clinical applicability being a vague promise, often verging on cliché in Discussion sections, researchers can offer more detailed descriptions of the situations and subgroups where a psychophysiological measure could add value to the assessment protocol as well as making concrete suggestions about how scores could be interpreted.

Clinicians also can help by continuing to search for clinically relevant and valid research, and critically appraising new methods compared to and combined with established tools of our trade. EBM offers practical tips for how to optimize searches to find helpful research quickly (Straus et al., 2011). However, having worked in the roles of researcher, teacher, and clinician, it is clear that researchers have the responsibility to do the best possible work and present it in a way that makes it easier for clinicians to do the "right thing" and adopt evidence based methods. Key ingredients, such as diagnostic likelihood ratios and standard errors of the difference score, are simple for researchers to calculate, yet they are conspicuous by their absence from the published literature. They involve much more work for clinicians to try to estimate from the parameters that are frequently published instead, and using the alternate formulae requires comfort with quantitative methods, as well as time and effort, that are all likely to be more scarce commodities in many clinical settings. It is far easier for researchers to adjust the design, analysis, and reporting of assessment strategies to provide scaffolding, support, and "off the shelf solutions" for our clinical partners. The pay-off will be greater use of evidence-based methods, better assessment, improved clinical decisions, and ultimately more people helped.

REFERENCES

Anderson, E. R., & Hope, D. A. (2009). The relationship among social phobia, objective and perceived physiological reactivity, and anxiety sensitivity in an adolescent population. *Journal of Anxiety Disorders*, *23*, 18–26. doi:10.1016/j.janxdis.2008.03.011

Bakeman, R., McArthur, D., Quera, V., & Robinson, B. F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods*, *2*, 357–370.

Berkman, E. T., & Falk, E. B. (2013). Beyond brain mapping: Using neural measures to predict real-world outcomes. *Current Directions in Psychological Science*, *22*, 45–50. doi:10.1177/0963721412469394

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

Bress, J. N., Meyer, A., & Hajcak, G. (2015). Differentiating anxiety and depression in children and adolescents: Evidence from event-related brain potentials. *Journal of Clinical Child & Adolescent Psychology*, *44*, 238–249.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, S., Masyn, K., Mastergeorge, A., & Hessl, D. (2015). Psychophysiological responses to emotional stimuli in children and adolescents with autism and fragile X syndrome. *Journal of Clinical Child & Adolescent Psychology*, *44*, 250–263.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. Advance online publication. doi:10.1177/0956797613504966

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*, 837–845.

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 & Adolescent Psychology*, *44*, 221–237.

De Los Reyes, A., Augenstein, T. M., Aldao, A., Thomas, S. A., Daruwala, S., Kline, K., & Regan, T. (2015). Implementing psychophysiology in clinical assessments of adolescent social anxiety: Methods for those without psychophysiological backgrounds. *Journal of Clinical Child & Adolescent Psychology*, *44*, 264–279.

Franklin, J. C., Glenn, C. R., Jamieson, J. P., & Nock, M. (2015). How developmental psychopathology theory and research can inform the Research Domain Criteria (RDoC) Project. *Journal of Clinical Child & Adolescent Psychology*, *44*, 280–290.

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

Gatzke-Kopp, L. M., Greenberg, M. T., & Bierman, K. (2015). Children's parasympathetic reactivity to specific emotions moderates response to intervention for early-onset aggression. *Journal of Clinical Child & Adolescent Psychology*, *44*, 291–304.

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

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

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York, NY: Wiley.

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.

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

Jaeschke, R., Guyatt, G. H., & Sackett, D. L. (1994b). Users' guides to the medical literature: III. How to use an article about a diagnostic test: B: What are the results and will they help me in caring for my patients? *Journal of the American Medical Association*, *271*, 703–707.

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. doi:10.1177/1745691610393527

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

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

Leitzke, B. T., Hilt, L. M., & Pollak, S. D. (2015). Maltreated youth display a blunted blood pressure response to an acute interpersonal stressor. *Journal of Clinical Child & Adolescent Psychology*, *44*, 305–313.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

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

McLaughlin, K. A., Rith-Najaran, L., Dirks, M. A., & Sheridan, M. A. (2015). Low vagal tone magnifies the association between psychosocial stress exposure and internalizing psychopathology in adolescents. *Journal of Clinical Child & Adolescent Psychology*, *44*, 314–328.

Moser, J. S., Durbin, C. E., Patrick, C. J., & Schmidt, N. B. (2015). Combining neural and behavioral indicators in the assessment of internalizing psychopathology in children and adolescents. *Journal of Clinical Child & Adolescent Psychology*, *44*, 329–340.

Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.

Sanislow, C. A., Pine, D. S., Quinn, K. J., Kozak, M. J., Garvey, M. A., Heinssen, R. K., … Cuthbert, B. N. (2010). Developing constructs for psychopathology research Research domain criteria. *Journal of Abnormal Psychology*, *119*, 631–639. doi:10.1037/a0020909

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.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*, 323–348. doi:10.1037/a0016973

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

Van Meter, A., Youngstrom, E. A., Youngstrom, J. K., Ollendick, T., Demeter, C., & Findling, R. L. (in press). Clinical decision-making about child and adolescent anxiety disorders using the Achenbach System of Empirically Based Assessment. *Journal of Clinical Child & Adolescent Psychology*.

Warnick, E. M., Bracken, M. B., & Kasl, S. (2008). Screening efficiency of the Child Behavior Checklist and Strengths and Difficulties Questionnaire: A systematic review. *Child and Adolescent Mental Health*, *13*, 140–147. doi:10.1111/j.1475-3588.2007.00461.x

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

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

Zhou, X.-H., Obuchowski, N. A., & McClish, D. K. (2002). *Statistical methods in diagnostic medicine*. New York, NY: Wiley.