Running Head: TOO GOOD TO BE TRUE

Is the Finding Too Good to Be True?

Moving from "More Is Better" to Thinking in Terms of Simple Predictions and Credibility

Eric A. Youngstrom Stephanie Salcedo

University of North Carolina at Chapel Hill

Thomas W. Frazier

John Carroll University and Autism Speaks

Guillermo Perez Algorta

Lancaster University

Author Note

Eric A. Youngstrom, Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill; Stephanie Salcedo, Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill; Thomas W. Frazier, Department of Psychology, John Carroll University, and Autism Speaks; and Guillermo Perez Algorta, Division of Health Research, Faculty of Health and Medicine, Lancaster University, UK. Correspondence concerning this article should be addressed 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. Email: eay@unc.edu

Abstract

In 2018, De Los Reyes and Langer expanded the scope of the Evidence Base Updates series to include reviews of psychological assessment techniques. In keeping with the goal of offering clear "take-home messages" about the evidence underlying the technique, experts have proposed a rubric for evaluating the reliability and validity support. Changes in the research environment and pressures in the peer review process, as well as a lack of familiarity with some statistical methods, have created a situation where many findings that appear “excellent” in the rubric are likely to be “too good to be true,” in the sense that they are unlikely to generalize to clinical settings or are unlikely to be reproduced in independent samples. We describe several common scenarios where published results are often too good to be true, including internal consistency, inter-rater reliability, correlation, standardized mean differences, diagnostic accuracy, and global model fit statistics. Simple practices could go a long way towards improving design, reporting, and interpretation of findings. When effect sizes are in the “excellent” range for issues that have been challenging, scrutinize before celebrating. When benchmarks are available base on theory or meta-analyses, results that are moderately better than expected in the favorable direction (i.e., Cohen’s $q \geq +.30$) also invite critical appraisal and replication before application. If readers and reviewers pull for transparency and do not unduly penalize authors who provide it, then change in research quality will be faster and both generalizability and reproducibility are likely to benefit.

Keywords: Psychometrics, reliability, validity, fit statistics, prediction

Is the Finding Too Good to Be True?

Moving from "More Is Better" to Thinking in Terms of Simple Predictions and Credibility

Many results are too good to be true. By this, we mean that they should not be accepted uncritically; even more, we advocate a mindset that combines curiosity with gentle skepticism. Results are too good to be true if they are unlikely to replicate, or if they will not generalize to situations with implications in clinical practice or policy. They could be exaggerated by clerical error, *p*-hacking, aspects of the research design, or simply assuming that "more is better" with all of our psychometric statistics, as opposed to thinking in terms of trade-offs and balancing of competing goals.

Existing conventions evolved for a reason, in much the same manner as our taste preferences served an adaptive function during evolutionary history. The $p < .05$ criterion grew out of a dialog between Fisher and colleagues as he was evaluating the effects of independent variables on agricultural production. The pace of research was slow. It took a season to grow a crop, and there were physical constraints on the size of the fields and number of plants. Results were calculated and checked by hand and compared to published tables of critical values (so .05 and .01 might be the only options, if those were all that was published in a reference work). Cohen's conventions for small, medium, and large effect sizes were based in large part on reviewing a year's worth of published articles in a leading journal of social psychology and another from clinical psychology in the 1970s. It is difficult to trace origin of the rule of thumb for Cronbach's alpha of .80 or higher being "good," but any reasonable effort to find a source discovers instead that there are a range of nuanced and informed opinions about it (Cronbach & Shavelson, 2004; Feldt, 1969; Nunnally, 1967).

But the research climate in which these conventions evolved is very different from the environment in which we are using them now. Standards from the era of farming—with analyses

done by hand, and figures and manuscripts generated by typewriters (e.g., stem and leaf plot) (Tukey, 1977) –- now guide our consumption of results in a world with M-Turk, big data, and statistical learning algorithms that will run staggering permutations on variable sets orders of magnitude larger than anything Fisher, Pearson, or Tukey saw in their lifetimes (James, Witten, Hastie, & Tibshirani, 2013). The shifting research environment does not make all the conventions obsolete or maladaptive. Just as perceived bitterness evolved to protect us from alkali toxins, and still protects us from contaminated food today, many of the statistical principles still function. Others may need some adaptation, though. Salt and sugar taste so good because they were vital but difficult nutrients to get for millions of years, so reward circuits evolved to motivate days' worth of hunting and gathering now impel us to binge on salty, fatty, sugary junk food that exploits our preferences. Psychometric conventions that were tuned in a bygone era, combined with systemic incentives to get significant and surprising results, are contributing to the proliferation of a junk food quality of science. Results that seem superficially tasty lack sustenance.

Therefore, we need to learn some healthy habits for quickly appraising research findings, whether it is as producers of the literature or consumers of it. The goals of this paper include reviewing examples where results that might conventionally be considered excellent (Hunsley & Mash, 2018) are instead likely to be too good to be true, inviting deeper inquiry rather than celebration as a first response. We first look at four types of psychometric coefficients: reliability statistics, effect sizes, model fit statistics, and meta-analytic summaries, and we explore instances when high coefficients warrant suspicion more often than enthusiasm. Next, we offer ways of developing and specifying predictions and expectations using rules of thumb, standardized checklists, as well as formal statistical tests, all to help decide whether results are credible. Things that are beyond the scope of this paper are various ethical issues, such as deliberately

falsifying data, *p*-hacking and ways of detecting it (for review, see Head, Holman, Lanfear, Kahn, & Jennions, 2015; Ioannidis, 2005), or matching the wrong statistical procedure with the research question ("Type III Error"). Even when we assume good faith and appropriate choice of model, there is still a surprising amount of room for junk food results. Our closing recommendations aim to train our sense of taste to promote a healthier information diet. The core idea can be distilled into a single sentence, even a single equation, that would lead to big progress.

Reliability

Reliability refers to the reproducibility of a measurement, which is essential to the reproducibility of the results and conclusions based on it. There are different facets of reliability, including reproducibility over sets of items (internal consistency, such as split-half, Cronbach's alpha, omega), over time (retest stability), and over judges (inter-rater reliability) (Hunsley & Mash, 2018). Generalizability theory points out that other facets are also possible and provides a unifying framework of variance decomposition, dividing the score variance "pie" into slices attributable to different factors. Item Response Theory (IRT) approaches (including Rasch, graded response, and other models) permit a fine grained look at reliability as a function of trait level; for example, telling whether the reproducibility of scores is similarly good at low, average, or high score ranges.

In all metrics, a higher value of the coefficient (closer to 1.0) indicates more reliable variance in the measurement. The convention is to treat more as better, and rubrics typically proceed in a linear fashion from "poor" to "adequate," "good" and "excellent" (Hunsley & Mash, 2007). An uncritical focus on maximizing this metric has a variety of unintended consequences.

Internal Consistency: Rethink Alpha Coefficients >.90

The downside of maximizing the reliability coefficient is perhaps best known with

Cronbach's alpha. The coefficient is not only a function of the typical correlation between items (which conceptually what we want it to measure), but also the length of the scale, and the variation between cases included in the sample. Other things being equal, the longer scale will have the higher alpha (Cronbach & Shavelson, 2004). Using the alpha as the guiding criterion for selecting a measure will thus be at odds with goals such as reducing the length of a battery or rater burden, which are key considerations in progress tracking and measurement-based care (Streiner, Norman, & Cairney, 2015). See Table 1 for an illustration.

An alternative to alpha (or any other coefficient that includes the number of items in the formula) would be to focus on the *average* inter-item correlation, or the average corrected item-total correlation (Streiner et al., 2015). These take scale length out of the reliability equation. They are only a partial solution, though. Maximizing the internal consistency may result in narrow coverage of the desired construct. Consider two sets of items focused on depression: *Does the person feel sad? ...feel down? ...feel depressed?* Versus: *Does the person feel down? ...have less energy than usual? ...have more trouble sleeping?* The second set has the lower inter-item correlations (because each item assesses a distinct symptom), but it also has the better coverage of the construct. In a clinical setting, the second measure would provide a better sense of the severity of depression, and also whether treatment was helping. If the scale were brief enough to be tolerated for repeated assessment, though, then the combination of shorter scale length and breadth of coverage typically results in a modest looking internal consistency estimate. A scale with items that correlate .35 with each other on average would have an alpha of .73 in a 5 item version, and .84 in a 10 item version, whereas a five-item set with average inter-item r=.50 would have an alpha of .83. Researchers or reviewers focused on maximizing alpha would be prone to pick the narrower scale, or push for longer scales that might raise response burden to levels that increase biased response sets or missing data (Dillman, Smyth, & Christian,

2014).

A more complete solution would be to use a pair of competing criteria to balance each other out. Pairing internal consistency with the correlation with the full length scale is a strategy advocated when developing short forms (Smith, McCarthy, & Anderson, 2000). Rearranging the Spearman-Brown prophecy formula makes it possible to project the reliability of a typical short form based on the alpha of the full length version; we can then look for a scale that has strong coverage (high R or R^2 with the longer version, or with a criterion variable such as diagnosis or interviewer-rated severity) while also maintaining at least that threshold of internal consistency (see Youngstrom, Van Meter, Frazier, Youngstrom, & Findling, 2018, for an example). If the intended application is progress or outcome measurement, then sensitivity to change (quantified as an omega-squared in a generalized linear model) would be a good counterbalancing metric.

Simply focusing on maximizing alpha risks picking scales that are too good to be true when used in many contexts. Reliability estimates are a ceiling for validity, but not an estimate of it. Overly narrow coverage will attenuate the correlation with the construct, meaning that the validity may actually be much lower than the tasty-looking coefficient implies. The scale with broader coverage could have an equal or higher validity coefficient, despite the lower internal consistency. In our own work, the full length General Behavior Inventory provides extremely high values for Cronbach's alpha (e.g., alpha>.92 to .96) in parent, youth, and teacher report for both depression and hypomanic/biphasic scores. However, one common critique was that the scale was onerously long for clinical use, motivating the development of various short forms and carved versions that showed identical or improved clinical utility despite more modest internal consistency estimates (.88 to .91; Youngstrom, A. Van Meter, et al., 2018). As a result, the short forms have been translated into the most languages, used in the most clinical trials, endorsed by the PhenX Tool Kit, have the largest effect sizes in meta-analyses of diagnostic accuracy

(Youngstrom, Egerton, et al., 2018; Youngstrom, Genzlinger, Egerton, & Van Meter, 2015) and are most often requested for use by clinicians. The lower alpha was not a consideration.

Inter-Rater Reliability: >.85 Is Often Too Good to Be True

Inter-rater reliability is often the more relevant aspect of reliability for clinical applications. If two different interviewers evaluated the case, would they agree about the diagnosis? How closely would their estimate of the severity of the problems match? Cohen's kappa is the most widely used metric for agreement about categorical variables such as diagnosis or dichotomous estimates of treatment response. Intra-class correlations are the typical metric for dimensional scores (McGraw & Wong, 1996a, 1996b). Again, higher coefficients indicate better performance, indexed as agreement better than chance, or as variance attributed to differences between cases instead of between raters or random error. The prevailing rubrics typically suggest that values >.80 are excellent (e.g., Landis & Koch, 1977). Reviewers often are critical of papers that report values lower than this, suggesting that the reliability was subpar or worse. Investigators are pressed to document that the reliability exceeds that threshold (Brennan & Prediger, 1981).

Fortunately for authors focused on the short term, due to pragmatic issues like tenure and promotion, or getting a grant renewed, there are many ways to whip up a batch of tasty looking coefficients without resorting to fraud. One is by judiciously selecting the choice of statistic. There is a family of intra-class correlations that use different definitions of the numerator (the desirable variance) and the denominator (the error variance). Two conceptual differences are the fixed versus random effects estimation, and consistency versus absolute agreement (McGraw & Wong, 1996a, 1996b). Fixed effects are appropriate when we have observed all of the possible levels of the variable (e.g., both biological sexes, those with or without a particular treatment exposure). Random effects are the better conceptual choice when we are sampling from a larger

universe of possibilities (e.g., gender identity, or clinical sites or therapists): They generalize beyond our specific raters to the larger population of potential interviewers. The random effects estimate will almost always be smaller than the corresponding fixed effect estimate; they could be tied when the variance attributable to the random factor is precisely zero. The random effects model would usually be the more realistic match to our research designs: We have not often comprehensively represented all possible variations in clinic or therapist, for example. However, the fixed effect model produces the larger coefficient. The typical practice is to report it, without clearly labeling the intraclass correlation as a fixed effect model. Ambiguously labeled ICCs are almost always the larger consistency value (Gruber & Weinstock, 2018).

The agreement versus consistency distinction hinges on whether we want to track differences in calibration as well as differences in how we rank cases. Consistency coefficients focus on whether the raters rank the cases in the same order, ignoring whether there is a discrepancy in the average scores across raters. Spearman's rho and Pearson's *r* are examples of consistency metrics, and one variant of consistency intraclass correlation is identical to *r*. In contrast, absolute agreement measures include the variance between raters in the denominator, penalizing the coefficient for the raters being calibrated differently. Again, the best-case scenario would be when the variance between raters is exactly nil, and then the absolute agreement and consistency coefficients would be identical. Otherwise, the consistency coefficient will always be higher, and that is why it is the one almost always reported instead.

If the statistic is calculated and reported accurately, is there any harm in using consistency instead of absolute agreement? Imagine students taking two sections of a class. The grades assigned by two teaching assistants have a consistency coefficient of .95 – excellent! But one TA's scores average 10 points lower than the other. The consistency coefficient ignores this as an uninteresting source of variance. The students are not likely to agree. They would have a

vigorous preference for a measure of absolute agreement instead. In clinical trials, having raters not well calibrated contributes to differences in whom gets enrolled across raters or sites, reducing power to detect treatment effects by increasing error variance, and potentially adding to placebo response rates. When test authors only report consistency coefficients, they are implicitly describing a best-case scenario where differences in rater anchoring and calibration are nonexistent. These are rarely realistic assumptions, and they underestimate the challenges involved in using the assessment with new raters or in new settings.

There are ways that investigators can craft the research design to yield more optimistic reliability estimates, as well. Two examples include selecting extreme cases for the reliability analysis and minimizing sources of error variance much more than would be feasible in typical settings. Stacking the sample with extreme cases maximizes the variance between cases, maximizing the numerator in the reliability estimate. Judges will have an easier time distinguishing between severely depressed cases and healthy controls than trying to grade degrees of depression among a set of cases all drawn from an outpatient clinic. Similarly, if the goal is to maximize the reliability estimate, then having two judges code the written transcript of an interview will yield a higher estimate of agreement than watching a video recording, which in turn would be higher than if the two judges independently interviewed the same person. In all three scenarios, the judges would be considering the information provided verbally, but the video adds variance due to nonverbal behavior, and the re-interview adds variance due to differences in phrasing of questions, as well as interpersonal dynamics, changes over time (re-interview is confounded with retest stability), and a plethora of other facets. It is easier to publish the estimate based on coding transcripts, even though the re-interview scenario is probably the more helpful benchmark for how the interview would perform when generalized to another client or setting. Published reports do not clearly disclose the design choices, as we have learned in our efforts to

code these features in several meta-analyses: Fewer than 20% of published reports included the reliability of the scale or of the diagnostic interview (Youngstrom et al., 2015). It is safest to assume that when we read an inter-rater reliability coefficient value >.85 without clear details to the contrary, it is probably not "excellent," but rather based on a fixed-effects, consistency model, and using transcripts, case notes, or audio recording rather than richer inputs, perhaps with selected cases. If the researchers did something else, they probably are aware that the model is more conservative, and definitely would know if it was more work (as a re-interview or video coding would be), and they would be sure to make the reader aware accordingly.

Item Response Theory Reliability, Information Values and Theta

Item Response Theory (IRT) models are gaining popularity because of several technical and practical advantages for scale building and evaluation. However, presenting and interpreting IRT presents its own set of challenges. Depending on the purpose of the measure and how it was developed, IRT reliability (information) estimates might be considered low or unacceptable by many researchers or reviewers. Knowing the purpose of the measure matters: Measures of psychopathology are particularly likely to show a pattern of fit that does not match generic expectations. If the measure is intended to assess the construct across the entire population and generate fine gradations in ability or severity, then it may be realistic to expect reliability levels to be >.80 from theta of -3 to +3 (where theta is the level of the underlying trait, scaled roughly as a *z*-score). However, for diagnostic measures this standard would be unrealistically harsh. Diagnostic measures need to have strong reliability in the region of the latent trait where the clinical group meets/overlaps the non-clinical group. This is often at theta ≥+1.5 to theta ≤+2.5 or 3 (assuming higher scores indicate more pathology). Reliability levels in the .4s or .5s at thetas of -3 to -0 will not significantly change measure performance measure for its intended purpose. When the goal is detecting pathological anxiety, precise measurement of "degree of

relaxation" at the low end is not crucial. Similarly, if a measure is designed to differentiate at average levels of the trait – say a normal personality measure like the Big 5 – drops in reliability at the ends of the distribution may not impact utility enough to warrant lengthening. A contrasting scenario might occur when an author is trying to develop a scale for measuring change and reliability drops precipitously at theta=0 or lower. For measures of pathology, this would correspond to low accuracy about levels of the trait as the person nears remission.

Adaptive testing approaches that use computers to choose calibrated items from a larger pool can obviate some of these problems. However, some measures are not suited for adaptive testing. For others, the necessary resources may be unavailable for development or implementation of adaptive frameworks. IRT reliability based estimates and changes in reliability across the latent trait need to be interpreted with reference to the intended use.

Validity and Effect Sizes

Effect sizes provide a helpful way of thinking about findings. They move us away from the dichotomous thinking of null hypothesis significance testing, changing the question from yes or no significance to "how big is the effect?" Focusing on the size of the effect immediately makes things less abstract, and we are more likely to consider whether the size is plausible, and whether it has practical significance. The plausibility of the effect size becomes a sort of face validity for the finding, and often will quickly raise questions about the appropriateness of the research design, analyses, or reporting. We use three common effect sizes – correlation, standardized mean differences (SMD), and diagnostic accuracy – to illustrate when results might be too good to be true (Kelley & Preacher, 2012).

Correlation

Correlation coefficients are among the most widely used effect sizes in social and clinical psychology, and they are widespread in other areas as well. It is well understood that most forms

of correlation coefficient are bounded by values of 1.0 and -1.0, with a coefficient of zero indicating no association between the variables. Cohen (1988) suggested values of *r* ~.1, .3, and .5 as rough benchmarks for small, medium, and large Pearson correlations based in part on reviewing a year's worth of articles in a leading journal. He stated that these were descriptive, not value judgments or indicators of practical significance. They were pegged to perception thresholds, with small effects being at the limit of what might be perceived (such as the difference in average height between 15 and 16 year old girls, about half an inch, expressed as a point-biserial correlation, p. 27), medium being .3 (about the difference between height in 14 versus 18 year olds, or the taste of name brand products and the "no frills" substitutes my mother kept trying to sneak past us as kids), and large being obvious (*r*~.5, such as the difference in average intelligence between those starting college and those finishing doctorates, or the height again – 2 inch difference between ages 13 and 18; Cohen, 1988). He also noted that on the one hand, he was feeding his perceptions via a diet of peer reviewed articles that had been through the kitchen of peer review before appearing on the menu of the *Journal of Abnormal and Social Psychology* – most research fare would not be so carefully prepared or refined through such rigorous critic reviews. On the other hand, the majority of the studies still had inadequate statistical power, even though they had been published (Cohen, 1962). Despite Cohen publishing this exposé about power in psychological research, and the paper being cited more than 1,500 times, power remained essentially unchanged decades later (Cohen, 1992).

"Big data" has accomplished what Cohen's exhortations could not. Survey Monkey, Qualtrics, and REDCap have automated survey delivery and scoring, much as agribusiness automated farming. Survey panels and M-Turk took the "psychology subject pool" recipe and scaled it to nation-sized and always in season. Data archiving and open data policies made curated datasets worth hundreds of millions of dollars available for secondary analysis. Google,

Facebook, and web scraping changed the scale of the data by several orders of magnitude yet again, and the Internet of Things and wearables are adding yet more huge and deep data streams to the broth. With *N*=10,000, roughly the size of a typical epidemiological study in one of the repositories, power would be 90% to reject the null if the true correlation were .032 or larger (Faul, Erdfelder, Buchner, & Lang, 2009). Google is making data freely available with *N*>100,000 or millions of observations (Stephens-Davidowitz, 2017). These sample sizes make statistical significance trivial. Machine learning is making it possible to test larger sets of variables to identify the most interesting sets of predictors. This is stepwise model building raised on steroids, making nominal *p* values meaningless. Gourmands of statistical learning refer to this as "the curse of dimensionality," where the computer can obligingly search through store-fulls of model ingredients, dutifully reporting the best fit (often now operationalized as predictive accuracy or bias reduction). The challenge is deciding which predictors are robust and likely to work again in other samples (showing low variance in the coefficient, in the parlance of statistical learning; James et al., 2013).

Does the size of the correlation match what was found in a prior study with the same variables? With data less expensive and available on unprecedented scales, a healthy research regimen needs to build on a pyramid of validated measurement ingredients, combined in recipes that make conceptual sense, with results compared to expectations at least qualitatively, although there are a variety of formal tests available. Steiger (1980) provided the formula for direct tests of two correlations or regression coefficients, either drawn from the same sample or two independent samples. The inputs are the two coefficients, the sample size, and the nuisance correlation if the coefficients are based on the same rather than the independent sample. Significant differences indicate that the coefficients are further apart than would likely be explained by sampling error. If the new coefficient is significantly higher, that would motivate

some careful contemplation about factors that might make it too good to be true. Differences in the reliability of the measures (Schmidt & Hunter, 1996), restriction of range attenuating the correlations (or over-dispersion amplifying them), and variations in sample composition all are worthy candidates that can quickly be tested via applications of simple formulae or examination of the enrollment procedure and sample demographics. Steiger's test and similar methods are powerful complements to Cohen's enjoinder to look for published effect sizes as the basis for interpretation.

Benchmarking Results. Some topics have an extensive literature available. Agreement across informants is an example. Two large meta-analyses compared agreement between parents, youths, and teachers about the youths' emotional and behavioral problems (Achenbach, McConaughy, & Howell, 1987; De Los Reyes et al., 2015), summarizing almost two thousand effect sizes from almost 500 samples—as ubiquitous as pizza! —*both* found average r=.28 across all informants. Informant type moderates agreement. Youth ratings tended to correlate r~.2 with parent and teacher ratings of internalizing and about .3 about externalizing problems; parents and teachers agree in a similar range, and two parents rating the same child agree at the .5 to .6 level. The covariation across raters is a small fraction of the reliable variance in all of the ratings, which suggests that there might be a lot of situational specificity, dyadic patterns, or other systematic differences in perspective.

The robustness of the pattern – it holds across measures, countries, and decades – means it provides a helpful baseline prediction. It is not a physical constant: It is possible for an observed correlation to be higher, especially if the scales involved are more reliable (longer) and or similar in terms of situation and behaviors observed. However, when correlations fall outside of sampling error from these benchmarks, that should spark a search for possible explanations. Readers and reviewers often forget how low typical agreement is about behavior, and

coefficients matching naïve expectancies actually would be too good to be true (see regression prediction formula). Conceptualizing parent-youth agreement as an inter-rater reliability task might lead to expecting an *r*~.8, or thinking that agreement should meet Cohen's rule of thumb for a large effect, would create a sense of cognitive dissonance when confronted with empirical data. Not remembering or understanding the implications of the modest correlation has led to heated debates, such as discounting one perspective or another as invalid or wrong (cf. De Los Reyes & Kazdin, 2005), or proposing that clinical diagnoses should only be made when clinically significant symptoms are observed by different informants in multiple settings (Carlson & Dyson, 2012).

Cohen (1988) went so far as to suggest a new effect size specifically to compare correlations. Cohen's *q* is the difference between the Fisher's *z'* transformation of two correlations. The *z'* transformation "stretches" the correlation so that it is not bounded at ±1.0, and makes it so that differences between *z'* values have an interval level scaling. Cohen (1988) devoted a chapter in his power recipe book entirely to *q*.

Method variance. When two measurements are made using the same method, then the scores will be correlated even if they were evaluating different constructs. Some of the variance in each score is due to the method of assessment, and because the two scores share the method, they have "shared method variance" (Campbell & Fiske, 1959; Podsakoff, MacKenzie, & Podsakoff, 2012). The multitrait-multimethod matrix provides a framework for thinking through the variance sources, which often have an additive effect. If two scales both measure depression, they should share variance; if they both are measured by caregiver report, then they will have a second helping of shared variance. When correlations look surprisingly large, shared method variance is often the culprit. It often is possible to ignore the labels, look at the pattern of correlations, and tell which informants provided subsets of scores. If the table is arranged by

informant, then there will be a block diagonal pattern. In situations where converging measures of the same trait show only moderate correlations, as is the case with cross-informant agreement about youth behavior, or neurocognitive and physiological measures with behavior ratings (Owens, Evans, & Margherio, 2020; Youngstrom & De Los Reyes, 2015), it is possible for the shared-method contribution to be larger than the shared-trait portion. This creates substantial challenges in confirmatory factor analysis, especially because large numbers of indicators are required for identification models specifying both method and trait factors.

Standardized Mean Difference (Cohen's *d*)

Cohen's *d*, or the standardized mean difference (SMD) between two groups, is the natural effect size for *t*-tests, single degree of freedom contrasts in ANOVA, and other ways of comparing two groups on a continuous measure. It is scaled as a *z*-score, which brings with it a certain set of conventions. A *d* of zero indicates that the means were identical. A *d* of 1 would be a one standard deviation difference between the groups. Cohen used this logic to develop a set of non-overlap alternate effect sizes--*U1, U2, U3*. A *d* value of 1 means that the average score in one group fell at the 68th percentile for the other group, for example.

By extension, this means that *d* values of 2 would be unusually large (2.5% of a normal distribution would be that extreme or more), and $d>3$ would be a one in a thousand event by chance. When we are studying psychological treatments or psychopathology correlates, effect sizes this large are more likely to signify a computational error, or perhaps falsification, than a replicable finding. We have seen these published, though. A common scenario is for authors to accidentally use the standard error of the mean in place of the standard deviation of the scores. The configuration of the output tables in SPSS make this an easy mistake to make. In a meta-analysis, we had a couple studies reporting effect sizes of $d>5.55$ in a content area where the average effect size was 1.05 (Youngstrom et al., 2015). Even without having a meta-analysis

available, the *d* values were a red flag on their own. Checking the descriptive statistics in the articles revealed that they were reporting "*SD*s" of 3.9 and 4.9 on Achenbach *T*-score scales, where the population *SD* is 10. Quick algebra using the sample size confirmed that they, too, were reporting *SE* values as *SD*s. A quick rule of thumb: If the metric is a *T* (*SD*=10) or standard score (*SD*=15), and that effect size is greater than what would be expected from the literature, double check the source of the *SD* values.

Cohen provided benchmarks for *d*, with values of .2, .5, and .8 suggested as small, medium, and large. These correspond to Pearson correlations of .1, .3, and .5 (after applying a correction for attenuation to point-biserial correlations; Rice & Harris, 2005). These were intended as a first approximation when published benchmarks are not available. When prior work provides a meaningful estimate, we can compare results in the present study to prior work either via confidence intervals, or a direct formal test. The *N*, *M*, and *SD* are sufficient statistics, meaning that readers and reviewers can apply the test using free online calculators. In many cases, getting a significant result may be cause for reflection; after analytic error and design artifacts are ruled out, then potential moderating variables become interesting contenders.

Diagnostic Accuracy

Tests of diagnostic accuracy represent a special case of bivariate statistics where the predictor is often continuous and the criterion is categorical. Logistic regression and receiver operating characteristic (ROC) analysis are methods of choice for analyzing these data, and the Area Under the Curve (AUC) from ROC analysis is a frequently used effect size, calibrated so that .50 indicates chance performance and 1.0 is the maximum, combining perfect accuracy for true cases (sensitivity) with perfect accuracy for true non-cases (specificity). A common rubric is that AUC values of .90 or higher are "excellent," .80 or higher are "good," .70+ are "fair", and .60+ are poor, and below .60 is a "fail" (Swets, Dawes, & Monahan, 2000). However, especially

in the area of clinical assessment, values greater than .90 are usually too good to be true when the goal is generalizing to a clinical setting.

Kraemer also pointed out that our criterion diagnoses are not perfectly reliable, and this imposes a ceiling on the AUC values that we can expect to observe in data. If the diagnostic interview is not perfectly accurate, then a good predictor will identify cases that the diagnostician missed, and it will correctly rule out some cases that the diagnostician mistakenly labeled. Students will remember the frustration of having errors on the key when their exam was marked. Kraemer provides the formula for estimating the ceiling. For diagnoses with kappa of .85 (an optimistic but plausible scenario), the upper bound AUC will be .925. Once again, observed AUC values above .90 look too good to be true – they would be performing near the upper limit of feasible accuracy given the reliability of our diagnostic tools. When seeing values in this range, three design features often are involved: (a) stacking the deck for a large reliability estimate, using extreme cases, transcripts, and consistency models as described above; (b) using fully structured interviews, maximizing reliability at the potential cost of clinical validity, and (c) maximizing the similarity of the predictor and criterion in terms of content coverage and source variance. A paper that had patients read the BDI questionnaire and then compared it to a doctor reading them similar questions (the “structured interview”) found >98% accuracy (Steer, Cavalieri, Leonard, & Beck, 1999), which obviously is too good to be true as an estimate of how the questionnaire administration would predict more clinically valid and generalizable diagnoses.

Design issues that inflate the effect size are often the culprit. Including healthy controls in the sample will add a lot of cases with extremely low scores on measures of psychopathology. These cases will all score below a reasonable threshold on the measure, boosting the diagnostic specificity with cases that are easy to identify (Youngstrom et al., 2015). A similar source of bias would be exclusion of comorbid cases or diagnostic groups that would produce overlapping

neurocognitive or behavioral patterns. Studies that compared cases with bipolar disorder to cases with ADHD but no comorbid mood disorder or healthy age-mates produce much larger AUCs estimates (well in excess of .90; e.g., Tillman & Geller, 2005) for many scales, only to shrink precipitously when applied in more diagnostically mixed samples where everyone is seeking treatment (Youngstrom, Meyers, Youngstrom, Calabrese, & Findling, 2006). The more conservative scenario is the better representation of how the measure would fare at a clinic where everyone is seeking help, and there are many different presenting problems and variations of comorbidity all in the sample. Most papers currently touting imaging, gene, or blood tests as diagnostic measures hinge on comparing healthy controls to well defined target groups, not a clinically generalizable design (e.g., Rocha-Rego et al., 2014; Woodruff, El-Mallakh, & Thiruvengadam, 2011; see Zeier et al., 2018 for review). The simple heuristics are to be suspicious when we see AUC > .90, and to ask, does the sample look like the people with whom I would want to use the measure (Jaeschke, Guyatt, & Sackett, 1994)? The more clinically complex the setting, and the more different the demographics, the more that we should expect the effect size to shrink (Konig et al., 2007). Internal cross-validation is not a substitute for finding data that closely resemble where we will need to use the measure (Youngstrom, Halverson, Youngstrom, Lindhiem, & Findling, 2018).

Comparing AUCs. With time and motivation, we can use more formal methods to decide whether the result differs from expectation. There are methods for comparing the AUC to published results, as well as more powerful methods for head-to-head comparison when the data are available. Web sites and R packages make it possible to use many of these tests even when the raw data are not available. It also is easy to convert AUC to Cohen's *d*, and vice versa. Converting Cohen's *d* benchmarks into AUC values provides a more realistic rubric for evaluating test performance in applied contexts (Rice & Harris, 2005). In addition, there are

techniques to test the difference in the AUC values drawn from different samples, as well as more powerful options when comparing predictors in the same sample (e.g., (DeLong, DeLong, & Clarke-Pearson, 1988; Hanley & McNeil, 1983; Venkatraman, 2000). It also is simple to convert *d* to *r*, and thence to *z'*. These are available in free software, including *R* packages (e.g., Robin et al., 2011) and web sites with programmed spreadsheets (https://en.wikiversity.org/wiki/Evidence_based_assessment/ROC_Party/Ready_to_ROC). The AUC and the standard error are sufficient statistics to be able to get a formal comparison of the new estimates with the old benchmark.

Reflections on Quality Checklists. There are a variety of checklists that are now available for assessment, treatment, and various other research designs, as well as corresponding lists to evaluate the quality of the reporting in published reports. The first are intended for the chef who wants to prepare a competent offering that will pass inspection. The second type is more built for the reviewer, auditor, or a food critic to check systematically that the standards are met. Both tend to be more detailed than could easily be used by general consumers, who need something more concise and focused on the information that would change validity to the point that it changes choices. The STAndardized Reporting of Diagnostic tests (STARD) guidelines list 25 items (Bossuyt et al., 2003), one of which we have not yet seen reported in any psychology article, and many of which are rarely reported. In our meta-analyses to date, total quality score has been unrelated to the effect size (Youngstrom, Egerton, et al., 2018; Youngstrom et al., 2015). However, several of the key ingredients independently do predict the validity coefficient. Using "distilled" samples that include artificially purified test groups has been a robust predictor tainting the results (Youngstrom, Egerton, et al., 2018; Youngstrom et al., 2015; Youngstrom et al., 2006), much like finding a bug in the soup. Rather than expecting readers to routinely conduct a 25 step review on every meal, focusing attention on some key

indicators may lead to faster improvements. We also have noted no change in the quality of designs or reporting in the decades before versus after the introduction of the STARD guidelines, with average scores hovering in the high 70s (passing, but unimpressive), suggesting that there may be a problem of implementation.

Model Fit Statistics

Confirmatory models offer an opportunity to compare the fit between the parameter estimates implied by the model versus what is observed in the data. Much effort has been focused on creating and evaluating different fit statistics (Maydeu-Olivares, 2013). There are absolute fit measures, such as the Goodness of Fit Index and Adjusted GFI, which are scaled so that 1.0 is the maximum; we can think of these as multivariate analogs to *R*-squared in terms of describing covariance reproduced by the model. The RMSEA is another measure of absolute fit, albeit scaled so that lower values indicate better fit, and zero would be perfect (Kelley & Preacher, 2012). There also are model comparison fit statistics, either anchored to a conceptual null model (e.g., CFI, TLI), or designed to compare empirical models to each other in terms of fit and parsimony (e.g., Akaike Information Criterion, Bayesian Information Criterion, and derivatives; see Raftery, 1995).

We need these to compare competing models, and to decide whether a model provides a good balance between parsimony and fit. Because of sampling variability, even a correctly specified model that holds in the population will not fit perfectly in a sample (Burnham & Anderson, 2016; Preacher & Merkle, 2012). It is helpful to remember that global fit statistics indicate only average model fit, and they do not also indicate the model's explanatory power, nor person-level fit and the accuracy of predictions for individual cases (Preacher & Merkle, 2012).

Often we focus almost exclusively on global model fit, treating the statistics as if there were a cutoff for good fit (e.g., CFI>.95; Hu & Bentler, 1995), or we report a suite of fit statistics

and focus the attention on the ones most favorable to our preferred model. We develop "fit statistic tunnel vision," where we do not go outside staring at the fit indices to see where the model is miss-specified, or even whether the parameter estimates make sense (Kline, 2016). When authors or reviewers focus first on fit, that can create conundrums where reasonable models get rejected in favor of those that are overfit – introducing bias in parameter estimates, and also increasing the likelihood of future replication efforts failing. A high profile example of this is the WISC-5, where the model selected based on fit indices produces a factor loading higher than 1.0 – as reported in the technical manual! (Wechsler, 2014).

Remedies include having a strong foundation of exploratory models before moving to confirmatory mode, testing across different sets of indicators as well as samples to improve understanding of the concept space as well as the measurement models. Newer approaches to evaluating models also look at piecewise fit, rather than focusing solely on global fit, and emphasize the interpretability of the parameter estimates. This is true in IRT (Maydeu-Olivares, 2013; Thissen, 2013) as well as covariance structure modeling (Kline, 2016).

Meta-Analysis and Credibility

Meta-analysis is well-suited for identifying results that are too good to be true. Cohen (1992) called it one of the few bright spots he had seen develop in the field during his career. The simplest versions gather the effect sizes, convert them to a consistent metric, and then test them for homogeneity. Cochran's *Q* statistic and funnel plots are examples of well-established statistical and graphical ways of looking for outliers. Estimates that fall outside the confidence interval for the meta-analytic summary are outliers likely to have different factors influencing their result.

Meta-analysis also can model variables that might explain heterogeneity in observed effect sizes. The most general model would be meta-regression, which can incorporate

continuous or categorical predictors (referred to as “moderators” in the meta-analytic parlance, because they are changing the size of effect sizes that usually summarize a relationship between two other variables). When doing a meta-regression, sample values that fall outside the residualized funnel plot, or that have large Studentized deleted residuals, would be outliers that warrant detailed scrutiny (Viechtbauer, 2010).

In our own efforts to test moderators, we generate two sets of candidates. One is a list of substantively interesting, often hypothesis driven variables, such as differences in informant (Youngstrom et al., 2015) or content coverage (Youngstrom, Egerton, et al., 2018). The other is a set of design features, including ones mentioned above. Using distilled samples and having shared source variance between the predictor and the criterion are two that often have a big impact on the flavor of the result. In contrast, the influence of *p*-hacking (Head et al., 2015) or differences in reference time period for scales tends to be relatively subtle.

When conducting a meta-analysis, we encourage the authors to review and discuss the outliers and identify potential contributing factors. At a minimum, such speculation could inform future studies and reviews. It may be possible to find enough similar studies to code the suspicious variable as a new candidate moderator for a supplemental or exploratory analysis. For consumers of the literature, meta-analyses provide a valuable sense of the typical distribution of effect sizes, helping us recognize when new results fall more in the realm of skepticism than credibility.

Recommendations

In keeping with the Evidence Based Updates series, we provide recommendations for next actions as well as re-calibrated expectations for evaluating assessment tools and practices (De Los Reyes & Langer, 2018).

Researchers

When we are in the mode of designing a new project, we can select high quality ingredients. These include investing in better assessments – picking the ones likely to have high validity, investing as much as possible in training and adherence, and using planning checklists to look for opportunities to enhance quality and report it accurately. The results are likely to be more sustaining when built around a simple recipe with *a priori* goals. Cohen's (1992) admonition to focus on fewer variables is worth remembering as a counterbalance to present enthusiasm for statistical learning models. Curating the candidate predictors on the basis of prior literature, theory, and psychometrics will be a good fusion of styles.

Authors can choose their effect size to improve consumption by the intended audience. Effect sizes can be converted between each other. The choice should combine familiarity (e.g., *r* and *d* are staples in psychology; NNT and LHH are more exotic imports from evidence-based medicine; Straus, Glasziou, Richardson, & Haynes, 2011), match with purpose (e.g., AUC for diagnostic studies, SMD for group comparisons, and r for regression-type analyses), and ease of interpretation. Effect sizes that have an asymptote and non-interval scale properties can be hard to compare. Odds ratios are well known case in point (e.g., 0.1 and 10 are of the same magnitude, and the distance from 2.0 to 4.0 means considerably more than from 102.0 to 104.0), but correlation and AUC also are harder to compare as they get larger, and it also can be tough to judge the practical importance of smaller values (Rosenthal, 1991). Cohen's *d* has many advantages as a metric, and may often be worth using as the primary or supplementary presentation format (Rice & Harris, 2005).

Authors can also be clear when doing effectiveness work or selecting more conservative and generalizable models. It is legitimate to suggest that the reader apply a different rule of thumb, e.g., "Because we are using clinically realistic comparison groups, we believe an AUC of .80 (or d=1.2) is a good target, rather than a .90 that usually has only been achieved by studies

with distilled cohorts in this content area" (e.g., Salcedo et al., 2017). This is not the same thing as lazily using convenience samples and whatever variables are laying around. There can be skill applied to archival and big data; a good chef can use principles to work with the ingredients at hand to deliver a memorable and satisfying meal. Efficacy paradigms are testing whether we could get the hypothesized result with the premium ingredients and intensive resources; effectiveness is adapting and improvising based on principles, and picking ingredients that are essential and can scale. Like a good army cook, dissemination and implementation research pushes for thinking about how to feed hundreds quickly, keeping them working under challenging conditions.

Researchers should consider generalizability and replication, even when doing exploratory research. The better the documentation and the more that the methods are selected with an eye towards reproducibility, the better the odds of replication. The Open Science Framework (OSF.io) is a free option for posting the code to run the analyses, or a detailed supplement with the technical specifications of the models (see also Code Ocean, and badges for open materials, code, or data from the Association for Psychological Science). The resulting research is likely to have more utility, as validity is a prerequisite, and it will have a longer shelf life and citation history.

Peer Reviewers

Embracing the perspective that we advocate also has several implications for peer reviewers. Reviewers are positioned to push authors for better reporting. Ask for more details about the reliability methods and analyses. Nudge authors to report more clinically useful effect sizes, and to compare them to benchmarks based on prior work, meta-analyses, or reasonable predictions from conceptual models. Consider the implications of design features such as shared method variance or reliance on healthy controls as a comparison group, and encourage authors to

mention the implications in the discussion section. Know rules of thumb for statistical analyses (van Belle, 2002) and benchmark values for the relevant literature, or make a habit of looking for them (e.g., search for meta-analyses). Fact check the results, especially if they seem improbable, using free software to estimate power or effect sizes, or to recalculate statistics.

At the same time, progress will likely be faster if reviewers complement the push for better reporting with recalibrated standards that acknowledge the trade-offs inherent in more generalizable designs. Effectiveness work fundamentally involves less internal validity than efficacy designs, and it also provides a more realistic sense of how the findings are likely to translate into practice. It would be helpful to explicitly match the calibration of the review with the intended audience. It would be absurd to conduct a James Beard review of the food preparation and presentation of a corner stand burger, whereas a simple five-star rating system, perhaps combined with key indicators about cleanliness and price, are often enough to make informed decisions. From a dissemination perspective, more people will get fed via burgers than Michelin-rated meals, too. Methodologists will be able to offer more balanced and useful critiques when they consider generalizability, calibrate their review appropriately, and nudge authors to be equally frank and realistic in evaluating the generalizability.

Requiring conventional rules of thumb paradoxically creates incentives to use weaker designs and statistical methods with unrealistic assumptions (e.g., consistency and fixed effects models). When sketchy results flood the market, it becomes harder for stronger designs and more accurately labeled reports to get accepted. Reviewers can help by not penalizing papers that are using more conservative and generalizable methods. Remembering that reliability is necessary for validity, do not penalize a paper for reporting a lower reliability coefficient based on a more generalizable model, especially if the paper still produced significant and meaningful results. Reliability is a prerequisite, not the end goal in most studies.

Readers and Consumers

Balance is key: On the one hand, exceptional results require exceptional support, and we should need to be persuaded; on the other hand, well executed and generalizable clinical studies will require a different calibration for evaluation. If the results amount to a miracle cure, or an assessment that will bring crystal clarity where all was fog before, then the results are probably too good to be true. We need to be familiar with key ways of deciding quickly whether the research is valid and likely to apply to the cases where we would need to make choices based on the results. Not all that is gold will glitter, and a lot that appears shiny fades rapidly in practice. A wry joke in Evidence Based Medicine is, "hurry and use the new drug before the next research studies come out" and the effect size shrinks (Silverman, 1998). We need to retrain our tastes to prefer more humble but realistic results.

General Conclusion: Think in Terms of Prediction

A powerful heuristic that all stakeholders can use is to make a prediction about the effect size or result ahead of time, write it down, and then compare the prediction to the finding. The act of making the prediction and expressing it in the form of an effect size (or an expected value for the test statistic) refines our thinking into a precise operational definition of the expectancy. It organizes our consideration of design and sample characteristics, as well as external benchmarks from prior studies and reviews. Writing it down takes but a moment, and it prevents any lazy, "Oh, yes, that is about what I expected" HARKing (Hypothesizing After the Results are Known) (Kerr, 1998). Comparing the prediction and the observation provides feedback that helps us learn and calibrate for the future (Meehl, 1973), as well as stimulating critical thinking about the case in point. It is possible to formalize the process, using Bayesian methods to combine the prior estimate with the new information (Etz & Vandekerckhove, 2018; Kruschke, 2011); but even leaving that aside, making a quantitative prediction and then comparing it to potentially

confirming or disconfirming data is a fast and free cognitive heuristic that will improve our interpretation of findings and decision making in general (Croskerry, 2003; Jenkins & Youngstrom, 2016).

Because effect sizes are convertible, it is possible to combine multiple ones in the same report, pairing a familiar one with another that facilitates comparison. Thinking in terms of effect sizes also moves us away from black and white thinking and towards focusing on the application of the results and how context might influence generalization. Cohen laid out a formal process for testing for differences between effect sizes, where $q = z'_{(observed)} - z'_{(expected)}$. Our suggested rules of thumb would be that (a) if the effect size converted to a *z'* is greater than .8 (see Table 2), it is probably too good to be true for a diagnostic test (e.g., AUC ~.9), a clinical outcome (*d*~1.8), or a convergent correlation (*r*=.67); and (b) if the *q*>.30 comparing the observed result to a reasonable benchmark, then it again may be too good to be true. A discrepancy this big would mean finding a large effect when a medium would have been expected, or a medium effect when a small would be plausible. Cohen (1988, p. 115) chose *q* ~.3 as a "medium-sized" discrepancy, so our rule of thumb translates to "When an analysis shows a result that is better than expected to a medium or larger degree, pause and reflect."

Healthier research reporting will be more nourishing but will be an acquired taste. Bayesian enthusiasts often offer a radical reworking of how we approach statistical analysis built around prediction. It definitely would address many of the short comings of our current practices, but it is a vegan cleanse or a ketogenic diet, so different that most consumers are unlikely to be able to switch quickly and sustain the change, even though it delivers results. We offer a pragmatic emphasis on making smarter choices one meal at a time, and encourage peer-to-peer support and accountability. As we wean ourselves from sugary kappas, distilled AUCs >.9, and high-sodium alphas, we will need to remind ourselves and each other that the results will be

better for the field in terms of both generalizability and reproducibility.

References

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

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

Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa: Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement, 41*, 687-699. doi:10.1177/001316448104100307

Burnham, K. P., & Anderson, D. R. (2016). Multimodel Inference. *Sociological Methods & Research, 33*, 261-304. doi:10.1177/0049124104268644

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

Carlson, G. A., & Dyson, M. (2012). Diagnostic Implications of Informant Disagreement About Rage Outbursts: Bipolar Disorder or Another Condition? *Israel Journal of Psychiatry & Related Sciences, 49*, 44-51.

Cohen, J. (1962). The statistical power of abnormal-social psychological research : a review. *Journal of Abnormal and Social Psychology, 65,* 145-153.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational & Psychological Measurement, 64*, 391-418.

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

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. doi:10.1037/a0038498

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of

childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin, 131*, 483-509. doi:10.1037/0033-2909.131.4.483

De Los Reyes, A., & Langer, D. A. (2018). Assessment and the Journal of Clinical Child and Adolescent Psychology's Evidence Base Updates Series: Evaluating the tools for gathering evidence. *Journal of Clinical Child & Adolescent Psychology, 47*, 357-365. doi:10.1080/15374416.2018.1458314

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.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys*. Hoboken, NJ: Wiley.

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review, 25*, 5-34. doi:10.3758/s13423-017-1262-3

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

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika, 34*, 363-373. doi:10.1007/bf02289364

Gruber, J., & Weinstock, L. M. (2018). Interrater reliability in bipolar disorder research: current practices and suggestions for enhancing the best practices. *International Journal of Bipolar Disorders, 6*, 1. doi:10.1186/s40345-017-0111-7

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.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The Extent and Consequences of P-Hacking in Science. *PLOS Biology, 13*, e1002106. doi:10.1371/journal.pbio.1002106

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76-99). Thousand Oaks, CA: Sage.

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

Hunsley, J., & Mash, E. J. (2018). Developing criteria for evidence-based assessment: an introduction to assessments that work. In J. Hunsley & E. J. Mash (Eds.), *A guide to assessments that work* (2nd ed., pp. 3-16). New York, NY: Oxford UP.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2*, e124.

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

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*. New York, NY: Springer.

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 & Clinical Psychology, 84*, 323-333. doi:10.1037/ccp0000070

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17*, 137-152. doi:10.1037/a0028086

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and social psychology review, 2*, 196-217. doi:10.1207/s15327957pspr0203_4

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press.

Konig, I. R., Malley, J. D., Weimar, C., Diener, H. C., Ziegler, A., & German Stroke Study, C. (2007). Practical experiences on the necessity of external validation. *Statistics in Medicine, 26*, 5499-5511. doi:10.1002/sim.3069

Kruschke, J. K. (2011). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. New York, NY: Academic Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159.

Maydeu-Olivares, A. (2013). Goodness-of-Fit Assessment of Item Response Theory Models. *Measurement: Interdisciplinary Research & Perspective, 11*, 71-101. doi:10.1080/15366367.2013.831680

McGraw, K. O., & Wong, S. P. (1996a). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46.

McGraw, K. O., & Wong, S. P. (1996b). "Forming inferences about some intraclass correlations coefficients": Correction. *Psychological Methods, 1*, 390.

Meehl, P. E. (1973). Why I do not attend case conferences. In P. E. Meehl (Ed.), *Psychodiagnosis: Selected papers* (pp. 225-302). Minneapolis, MN: University of Minnesota Press.

Nunnally, J. C. (1967). *Psychometric theory* (1st ed.). New York, NY: McGraw-Hill.

Owens, J. S., Evans, S. W., & Margherio, S. M. (2020). Assessment of attention deficit hyperactivity disorder. In E. A. Youngstrom, M. J. Prinstein, E. J. Mash, & R. Barkley (Eds.), *Assessment of Disorders in Childhood and Adolescence* (5th ed.). New York, NY: Guilford Press.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63*, 539-569. doi:10.1146/annurev-psych-120710-100452

Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods, 17*(1), 1-14. doi:10.1037/a0026804

Raftery, A. (1995). *Bayesian model selection in social research*. Cambridge: Blackwell.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's *d*, and *r*. *Law and Human Behavior, 29*, 615-620.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Muller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*, 77. doi:10.1186/1471-2105-12-77

Rocha-Rego, V., Jogia, J., Marquand, A. F., Mourao-Miranda, J., Simmons, A., & Frangou, S. (2014). Examination of the predictive value of structural magnetic resonance scans in bipolar disorder: a pattern classification approach. *Psychological Medicine, 44*, 519-532. doi:10.1017/S0033291713001013

Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD, and alternative indices. *American Psychologist, 46*, 1086-1087.

Salcedo, S., Chen, Y. L., Youngstrom, E. A., Fristad, M. A., Gadow, K. D., Horwitz, S. M., . . . Findling, R. L. (2017). Diagnostic efficiency of the Child and Adolescent Symptom Inventory (CASI-4R) Depression subscale for identifying youth mood disorders. *Journal of Clinical Child and Adolescent Psychology*, 1-15. doi:10.1080/15374416.2017.1280807

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199-223.

Silverman, W. A. (1998). *Where's the evidence: Debates in modern medicine*. New York, NY:

Oxford UP.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12*, 102-111.

Steer, R. A., Cavalieri, T. A., Leonard, D. M., & Beck, A. T. (1999). Use of the Beck Depression Inventory for primary care to screen for major depression disorders. *General Hospital Psychiatry, 21*, 106-111.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245-251.

Stephens-Davidowitz, S. (2017). *Everybody lies: Big data, new data, and what the Internet can tell us about who we really are*. New York, NY: Dey.

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 UP.

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

Thissen, D. (2013). The Meaning of Goodness-of-Fit Tests: Commentary on "Goodness-of-Fit Assessment of Item Response Theory Models". *Measurement: Interdisciplinary Research & Perspective, 11*, 123-126. doi:10.1080/15366367.2013.835205

Tillman, R., & Geller, B. (2005). A brief screening tool for a prepubertal and early adolescent bipolar disorder phenotype. *American Journal of Psychiatry, 162*, 1214-1216.

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

van Belle, G. (2002). *Statistical rules of thumb*. New York, NY: Wiley.

Venkatraman, E. S. (2000). A permutation test to compare receiver operating characteristic curves. *Biometrics, 56*, 1134-1138.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*, 1-48.

Wechsler, D. (2014). *Wechsler Intelligence Scale for Children - 5th Edition*. San Antonio, TX: NCS Pearson.

Woodruff, D. B., El-Mallakh, R. S., & Thiruvengadam, A. P. (2011). A potential diagnostic

blood test for attention deficit hyperactivity disorder. *Attention-Deficit/Hyperactivity Disorder, 3*, 265-269. doi:10.1007/s12402-011-0057-z

Youngstrom, E. A., & De Los Reyes, A. (2015). Commentary: moving toward cost-effectiveness in using psychophysiological measures in clinical assessment: validity, decision making, and adding value. *Journal of Clinical Child & Adolescent Psychology, 44*, 352-361. doi:10.1080/15374416.2014.913252

Youngstrom, E. A., Egerton, G. A., Genzlinger, J., Freeman, L. K., Rizvi, S. H., & Van Meter, A. (2018). Improving the global identification of bipolar spectrum disorders: Meta-analysis of the diagnostic accuracy of checklists. *Psychological Bulletin, 144*, 315-342. doi:10.1037/bul0000137

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. doi:10.1037/arc0000024

Youngstrom, E. A., Halverson, T. F., Youngstrom, J. K., Lindhiem, O., & Findling, R. L. (2018). Evidence-Based Assessment from simple clinical judgments to statistical learning: Evaluating a range of options using pediatric bipolar disorder as a diagnostic challenge. *Clinical Psychological Science, 6*, 234-265. doi:10.1177/2167702617741845

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

Youngstrom, E. A., Van Meter, A., Frazier, T. W., Youngstrom, J. K., & Findling, R. L. (2018). Developing and validating short forms of the Parent General Behavior Inventory Mania and Depression Scales for rating youth mood symptoms. *Journal of Clinical Child & Adolescent Psychology*, 1-16. doi:10.1080/15374416.2018.1491006

Zeier, Z., Carpenter, L. L., Kalin, N. H., Rodriguez, C. I., McDonald, W. M., Widge, A. S., & Nemeroff, C. B. (2018). Clinical Implementation of Pharmacogenetic Decision Support Tools for Antidepressant Drug Prescribing. *American Journal of Psychiatry, 175*, 873-886. doi:10.1176/appi.ajp.2018.17111282

Table 1

Association between scale length, average inter-item correlation, and Cronbach's alpha.

	Number of Items on Scale (*k*)							
Average Item r	**50**	**40**	**30**	**25**	**20**	**15**	**10**	**5**
.10	.85	.82	.77	.74	.69	.63	.53	.36
.15	.90	.88	.84	.82	.78	.73	.64	.47
.20	.93	.91	.88	.86	.83	.79	.71	.56
.25	.94	.93	.91	.89	.87	.83	.77	.63
.30*	*.96*	*.94*	*.93*	*.91*	*.90*	*.87*	*.81*	*.68*
.35	.96	.96	.94	.93	.92	.89	.84	.73
.40	.97	.96	.95	.94	.93	.91	.87	.77
.45	.98	.97	.96	.95	.94	.92	.89	.80
.50	.98	.98	.97	.96	.95	.94	.91	.83
.55	.98	.98	.97	.97	.96	.95	.92	.86
.60	.99	.98	.98	.97	.97	.96	.94	.88
.65	.99	.99	.98	.98	.97	.97	.95	.90
.70	.99	.99	.99	.98	.98	.97	.96	.92

Note. "Excellent" values of alpha >= .90 (bold line) could be achieved by long scales with even though they might include items measuring heterogeneous constructs, and would require inter-item correlations > .6 for short forms often used in applied settings.
*The italicized row shows how picking a moderate target for item correlation, such as average *r* ~.3, would produce a sliding scale of alpha values depending on scale length.

Table 2

Comparison of four common effect sizes and conventions.

r	*d*	AUC	Fisher *z′*
.000	.000	.500	.000
.100 **S**	**.200** **S**	.556	.100
.200	.408	.614	.203
.243	**.500** **M**	.638	.248
.300 **M**	.629	.672	.310
.350	.747	**.700** **S**	.365
.371	**.800** **L**	.714	.390
.400	.873	.731	.424
.500 **L**	1.155	.793	.549
.514	1.198	**.800** **M**	.568
.600	1.500	.856	.693
.670	1.805	**.900** **L**	.811
.700	1.960	.917	.867
.800	2.667	.970	1.099
.900	4.129	.998	1.472
.950	6.085	1.000	1.832
.990	14.036	1.000	2.647

Note. Boldfaced coefficients represent common rules of thumb for small, medium, and large effects. Coefficients above *z'* values of .8 may be too good to be true, and warrant critical evaluation in most clinical applications, unless they are reliability coefficients. Similarly, coefficients with *z'* values more than .3 higher than expected also deserve scrutiny of the design, analyses, and reporting.