Published in final edited form as:
J Clin Child Adolesc Psychol. 2021 ; 50(4): 443–449. doi:10.1080/15374416.2021.1929252.

Embracing Scientific Humility and Complexity: Learning "What Works for Whom" in Youth Psychotherapy Research

Michael C. Mullarkey, Ph.D., **Jessica L. Schleider, Ph.D.**
Department of Psychology, Stony Brook University

Abstract

Clinical psychological scientists have spent decades attempting to understand "what works for whom" in the context of youth psychotherapy, toward the longstanding goal of personalizing psychosocial interventions to fit individual needs and characteristics. However, as the articles in this Special Issue jointly underscore, more than 50 years of psychotherapy research has yet to help us realize this goal. In this introduction to the special issue, we outline how and why "aspiration-method mismatches" have hampered progress toward identifying moderators of youth psychotherapy; emphasize the need to embrace etiological complexity and scientific humility in pursuing new methodological solutions; and propose individual and structural strategies for better-aligning clinical research methods with the goal of personalizing mental health care for youth with diverse identities and treatment needs.

Keywords

Treatment moderators; psychotherapy; clinical trials; youth mental health; research methods

For more than half a century, a single empirical question has remained a key driver for psychotherapy research: "What treatment, by whom, is most effective for this individual with that specific problem, and under which set of circumstances?" (Paul, 1967, p. 111). Practitioners and researchers alike acknowledge that no single intervention, however evidence-based, will benefit all individuals equally—even within diagnostic categories, socio-demographic groups, or developmental stages. Yet, progress toward a "personalized medicine" for psychotherapy, whereby individuals may be matched with interventions tailored to their needs and circumstances, has stayed virtually stagnant (Simon & Perlis, 2010). The collection of articles in this special issue synthesizes past attempts to answer Paul's "what treatment, for whom" question, specifically by reviewing work on moderators of outcome in youth anxiety disorders (Norris & Kendall, 2020), obsessive-compulsive disorder (Kemp et al., 2020), substance use disorder (Bachrach & Chung, 2020), depression (Meyer & Curry, 2020), post-traumatic stress disorder (Danzi & La Greca, 2020), autism spectrum disorder (Klinger et al., 2020), and pediatric bipolar spectrum disorder treatments (Roley-Roberts & Fristad, 2020). Jointly, they highlight a longstanding, critical challenge within clinical intervention research: a frequent and fundamental mismatch between *aspirations* for empirical clarity and *methods* used to meet those aspirations. For example, we *aspire* as clinical psychologists to learn which interventions work best for whom—but our *study designs* and *analytic approaches* are often ill-equipped to reveal the answers. Each paper in this special issue admirably catalogues this aspiration-method mismatch. The authors survey their respective fields and reach a near unanimous conclusion: There are no reliably detected moderators of youth psychotherapy for any disorder. In this introduction to the special issue, we explore how and why this aspiration-method has—despite the best of intentions—poisoned the well of treatment heterogeneity research. We also highlight feasible, promising paths toward better aligning our science with our goal of improving personalized patient care.

Aspiration-Method Mismatch #1: Testing one moderator at a time will not reveal “what works for whom” in youth psychotherapy.

Innovations from other scientific disciplines help contextualize why “what works for whom” has been such a challenge for psychotherapy researchers to unpack. For instance, cancer researchers recently found that testing all candidate treatments for all patient subpopulations would take *90 years* via standard clinical research practices. Accordingly, they re-designed their investigative pipeline to increase speed while maintaining methodological rigor (Hobbs et al., 2018). For cancer scientists, “all candidate treatments” then included just *ten* investigational drugs. In psychotherapy research, there are *hundreds* of candidate moderators that may help us to personalize care.

Treatment moderators, particularly well characterized by Norris and Kendall (2021, p. 1), are “pre-randomization characteristics that identify which treatments work for whom under what circumstances.” No part of this definition suggests a “one-at-a-time” approach to moderator testing, nor does it require the use of interactions within traditional linear regression approaches. Yet, nearly all moderator tests reviewed by articles in this special issue exemplify this very approach. For reasons outlined below, over-reliance on this “one-at-a-time” approach may be hampering scientific progress in several respects.

Any individual treatment moderator is unlikely to have large effects.

The likelihood of finding individual predictors that substantially impact treatment response —that is, in a *clinically significant* respect—is exceptionally low (Sherman & Pashler, 2019). Many interaction effects identified as statistically significant provide negligible benefit in explaining how well individuals respond to a given treatment (Cohen et al., 2019). In other words, per a large portion of treatment moderator results reported in scientific manuscripts, a “better” treatment is often *just 0.1% better* than the alternative based on a single moderator, in terms of predicting total symptom reductions for individuals or groups.

There are proofs outlining a piranha problem (Tosh et al., 2020), whereby any given variable in a large set (e.g., a set of possible moderators) is unlikely to have a large effect on an outcome (e.g., treatment effectiveness), unless (1) the variables also all exert large effects on one another (i.e., they are all highly correlated); or (2) each variable’s effect, in reality, is smaller than it first appears. Either scenario is problematic if we hope to find strong individual moderators of treatment. In scenario one, a large effect observed in any one clinical trial would be impacted heavily by the presence of other moderators, measured and unmeasured. Therefore, we cannot assume the effect of that single moderator will remain large in other circumstances for other patients. In scenario two, individual moderator effects may appear large due to underpowered samples or questionable research practices (Leichsenring et al., 2017), but actually be too small to help meaningfully guide our understanding of what works for whom.

A one-at-a-time approach will produce slow patient benefits, at best.

Even if theory and research *did* support the one-at-a-time approach, testing individual treatment moderators within traditional regression frameworks would, optimistically, take *decades* to practically improve patient care—even if strong individual moderators could be identified—given that doing so for several decades has yielded almost *no* consistent moderators to date. Indeed, even research on *overall* treatment effectiveness, let alone effectiveness for diverse subgroups, can take up to 17 years to inform real-world clinical practice (Morris et al., 2011). A clinical research pipeline optimized to yield robust, relatively rapid benefit to patients, by way of personalizing clinical care, would look little like our current practices.

Ultimately, testing treatment moderators one-at-a-time is misaligned with our field's long standing goal: improving and personalizing patient care as rigorously and rapidly as possible. The one-at-a-time approach, despite its frequent application, facilitates continued focus on *straightforward, simplistic* explanatory models of whom psychotherapies benefit most. The ubiquity of the "one-at-a-time" approach might also be symptomatic of academia's broader incentive structure, which has historically rewarded "salami-slicing" of data to produce multiple papers from a given dataset (e.g., by publishing a series of manuscripts on various moderator tests; Hilgard et al., 2019; Leichsenring et al., 2017). Our current system for identifying treatment moderators may therefore select for the methods less well matched to our aspirations of identifying what works for whom and more well matched to producing the maximum number of scientific papers (Smaldino & McElreath, 2016).

Aspiration-Method Mismatch #2: Most clinical trial samples are too small and non-representative to reveal what works for whom.

One source of this effect overestimation is samples that are almost always underpowered for interaction effects in the linear regression framework. The average sample size of youth psychotherapy clinical trials over the past 50 years is 68.69 participants (Weisz et al., 2017). An optimistic estimate of the effect size of treatment (e.g., not excluding trials at high risk of bias) is $d = 0.46$. Therefore, the average youth psychotherapy trial over the past 50 years only has 47.5% power to detect the average *main effect* of treatment assuming independent and equally sized treatment groups. Youth psychotherapy trials would have to include 152 total participants to have 80% power to detect main effect differences of $d = 0.46$. Most interactions require at least four times the sample size necessary to detect a main effect (Though see https://aaroncaldwell.us/SuperpowerBook/ for open-source tools to determine power for particular interaction effects). Therefore, assuming there was a single interaction effect with a similar effect size as the main effect of treatment ($d = 0.46$), a youth psychotherapy trial would optimistically need 608 total participants to detect that interaction. The average youth psychotherapy clinical trial conducted in the past 50 years would need nearly 900% more participants to reliably detect an interaction effect as large as the main effect of treatment.

While several papers note the potential for false negatives due to low power (Bachrach & Chung, 2020; Meyer & Curry, 2020; Norris & Kendall, 2020), low power increases the likelihood of false positives and overestimates of effect size as well (Forstmeier et al., 2017). It is tempting to imagine that tests for treatment moderators in underpowered samples give us an approximation of reality – not the whole puzzle but at least some of the pieces. However, to advance our understanding of treatment moderators we have to let go of this idea. Investigations of treatment moderators in underpowered samples do not provide us an approximation; they provide us an illusion (Simmons et al., 2011).

Many trials also do not assess structural factors that impact multiple levels of people's day to day lives, including their psychological health (Hatzenbuehler, 2016). For example, only one study of treatments for pediatric bipolar spectrum disorders assessed socioeconomic status (Roley-Roberts & Fristad, 2020). Although we agree the identification of modifiable treatment moderators is needed (Bachrach & Chung, 2020), we also agree with other authors that identifying structural moderators as well will ultimately lead to more robust decisions about what works best for whom under what circumstances (Klinger et al., 2020).

Further, we cannot achieve a broad understanding of "what works for whom" if nearly all the "whom" studied identify as White. Papers in this special issue highlight that individual trials can have >90% White participants (Kemp et al., 2020), and across entire fields the modal representation of certain minoritized groups (e.g., individuals who identify as Latinx) is often zero (Norris & Kendall, 2020). Many trials do not include enough people of Color (PoC) to even perform overly-aggregated comparisons between PoC and White participants' overall treatment response, much less examine the differential effects within individual racial groups.

Aspiration-Method Mismatch #3: We prioritize group-level design and statistics over approaches that directly assess what works for individuals.

Overly aggregated could arguably also characterize the predominant approach of using between-subjects designs and methods to identify treatment moderators. As noted in one review (Norris & Kendall, 2020), there is evidence that findings based on between-subjects designs may not generalize to individual people (i.e., treatment moderators may be non-ergodic). By comparison, designs and methods that use person-level/idiographic methods to identify treatment moderators, while rising, are still less common.

What now? Embracing humility and complexity in our hearts and in our methods

The articles in this special issue equip us with keen knowledge of the challenges, problems, and still-outstanding questions in youth-focused treatment heterogeneity research. Likewise, the aspiration-method mismatches highlighted may seem difficult to practically address. However, through a series of *individual and structural solutions,* there are promising paths toward improving past approaches.

The first step in this process likely involves embracing humility about how little we know, as psychological scientists, about "what works for whom" in the domain of youth psychotherapy. It is sobering to reckon with the present articles' uniform message: a large body of well-intentioned, painstaking clinical research has yielded few practical improvements for personalized youth psychotherapy. Facing this reality is central to pursuing more productive, clinically useful paths moving ahead.

Embracing humility may also allow us to more openly consider alternative approaches (Whitcomb et al., 2017). Some methodological reforms that seem too radical or far-fetched if the status quo remains unscrutinized can seem urgent and necessary when the status quo is critically examined by the field. Given that our current approaches to identifying moderators have yielded few benefits for patients, humility should remain central in crafting our efforts moving forward. For example, we agree with authors who call for widening the scope of variables that are considered as potential treatment moderators (Danzi & La Greca, 2020).

The task of embracing humility should not be a goal for individual researchers alone. Institution-level humility seems warranted, as well. Journal editorial boards, grant review panels, and funding agencies all play key roles in shaping standard research methods and practices. By acknowledging our lack of progress in identifying what works for whom, these institutions can support new directions and standards in our attempts to personalize mental health care. For instance, these institutions could prioritize funding, publishing, and publicizing projects that ambitiously, and rigorously, aim to fix our aspiration-methods mismatches. We will now provide recommendations that can guide individual researchers and institutions toward better understanding what works for whom in youth psychotherapy.

Rethinking data collection, data sharing, and analytic frameworks to maximize our chances of discovering what works for whom

Recommendation 1: Streamline and Normalize Clinical Trial Data Sharing.— We echo the calls from many papers in this issue to make clinical trial data more accessible, and more practically useful, for identifying treatment moderators. We agree with the several papers (Kemp et al., 2020; Meyer & Curry, 2020; Norris & Kendall, 2020) that call for pooling trial data to test treatment moderators using individual patient data meta-analysis. We also agree that structures should be put into place, such as a national database, to make this data broadly accessible across researchers. This level of accessibility will accelerate our understanding of treatment moderators, and already has excellent templates available in the National Clinical Trials Database and the open science approaches of the ABCD study (see: www.abcdstudy.org)—from which *hundreds* of papers on child development and psychopathology have already been published by scientists across the globe.

Recommendation 2: Develop a "Best Practice" Battery of Psychotherapy Research Moderator Variables.—Merely collecting more data and sharing it more effectively will not independently yield clinically-useful knowledge of youth treatment moderators. Certainly, standardization of at least some outcome measures and potential treatment moderators is necessary to more effectively harmonize future clinical trial data. Systems to standardize outcome assessments in mental health care have already been achieved at a country-wide level and could serve as a roadmap for future attempts to identify treatment moderators (Ludlow et al., 2020). While there are potential drawbacks to mandating inclusion of only specific, pre-selected outcome variables (e.g. could artificially narrow the scope of symptoms assessed for a given disorder; Patalay & Fried, 2020), a solution that might apply across clinical settings treatment types might involve suggesting standard outcomes that capture overall functional improvement, inherently relevant and important to patients regardless of presenting problem type (Chevance et al., 2020; e.g., quality of life; Cuijpers, 2020). Further, the standard inclusion of certain candidate treatment moderators beyond demographic information—particularly those that enable inclusion of multi-level and structural factors that may relate to treatment response—may spur far faster identification of promising moderators than ad-hoc, "one at a time" testing.

The devil certainly resides in the details of selecting those standardized moderators sets. Thus, we propose prioritizing inclusion of variables that are minimally burdensome to participants *and* researchers (e.g., inexpensive to administer and requiring minimal participant time) while carrying high potential for information density (e.g., a single variable that may yield additional information). For example, asking a participant's zip code is inexpensive and low-burden, yet it can yield tens of thousands (if not more) potential predictor variables of interest. Crucially, zip code also allows us to derive structural variables such as health provider shortages, estimates of explicit racism, and estimates of explicit homophobia that may impact which treatment is more effective for whom (Price et al., 2020). Although these candidate moderators may not be modifiable, assessing them will help us get a fuller, more accurate picture of for whom and *under what circumstances* one treatment is likely to be more helpful than another. Importantly, when collecting these information-dense variables, special attention should be paid to informed consent and steps should be taken to minimize any potential harm (e.g., reidentification) that could occur due to collecting this kind of data.

Recommendation 3: Capitalize on Passive Sensing Technology to Collect Low-Burden Behavioral Data.—Low-burden, high-information-density variables need not be self-reported. Passive sensing data collected via smartphones before randomization to treatment could serve as an additional low-burden, information-dense source of idiographic treatment moderators. Passive sensing data is somewhat higher burden for researchers due to cost, but it provides a wealth of information that may be inaccessible through self-report alone. Thus, the value of the rich information passive sensing data offers can often outweigh the additional cost to researchers – especially given the low burden on patients' time and attention. This idiographic information could also help us directly test whether certain kinds of idiographic information improves our ability to predict who responds to which treatment above and beyond predictors that can be collected in a single baseline session. For example, can idiographic information collected intensively pre-treatment help match patients to a more helpful psychotherapy more effectively than one-time, self-report data alone? Further, within randomized clinical trials, the relative benefit of instituting additional, pre-randomization assessment periods (e.g., to allow for a passive data collection window spanning days or weeks) should be weighed against any increased risk for participant attrition, as well as ethical concerns around delaying clinical care. We agree with at least one paper in this special issue that ergodicity – or the generalizability of findings from the group level to the individual level – is not an on-off switch, but a continuum (Norris & Kendall, 2020). Using passive data collection approaches to gather information-dense idiographic data across trials – perhaps during a relatively short, pre-randomization period – could help us determine where on that ergodicity continuum treatment moderators fall.

Recommendation 4: Prioritize (Very) Large-Scale Clinical Trials.—This ease of administration at scale is crucial, because one place where we depart from several papers included in this issue is whether we should conduct new, massive clinical trials. While we agree leveraging the data we already have as effectively as possible should be done, there are inherent limitations in data that presently exist. The lack of standardization in outcome variables and moderators limits what we can test. More importantly, the systematic exclusion of anyone not identifying as White—and the lack of assessment in a large portion of youth psychotherapy research of gender identity or sexual orientation—unacceptably limits the capacity of previously collected data to serve the needs of diverse patient populations. We acknowledge the logistics for undertaking new, massive RCTs are formidable. However, we argue the costs of not attempting to do so are far higher. Further, there are models of more truly massive, coordinated data collection. For example, the Psychological Science Accelerator has collected data from tens of thousands of people in over 40 countries with minimal outside funding support (Forscher & Ijzerman, 2021). Certainly, extending this kind of massive data collection into clinical trials for psychotherapy will require innovation. However, it is not an impossible goal—at least for treatments amenable to remote delivery. “Massive Open Online Intervention (MOOI)” trials have been successfully conducted for digital mental health interventions (Muñoz et al., 2016; Schleider et al., 2020; Schleider, Mullarkey, Fox, Dobias, Shroff, Hart, & Roulston, 2021). Moving forward, the MOOI framework could be applied to trials of telehealth-compatible psychotherapies. Behaviors betray priorities, and behaving in line with “how do we conduct massive, new RCTs to identify standardized treatment moderators of a standardized, patient valued outcome?” is far more patient-centered than “conducting massive, new RCTs is logistically very difficult and therefore not worth trying.”

Recommendation 5: Use models that can robustly test many high-dimensional moderators simultaneously.—Just how massive should these new RCTs be? We propose they should be “powered” such that models investigating many high-dimensional moderator effects simultaneously could be used to forecast treatment response. Initial estimates indicate these kinds of prediction-focused models would ideally have at least 500 participants per treatment arm (Luedtke et al., 2019), and there are open-source tools available for trial-specific sample size determination (Riley et al., 2020). Taking a more prediction-focused approach in designing and planning clinical trials will likely yield more useful leads to identifying treatment moderators than post-hoc, explanatory approaches (Yarkoni & Westfall, 2017). In addition assessing many high-dimensional moderators at once, these prediction-focused models, even so called “black box” models, can also gauge the relative importance of all included moderators (Nemesure et al., 2021). These models can therefore evaluate not just whether a candidate moderator has a non-zero effect on the outcome, but how much more or less that candidate moderator affects the outcome compared to other candidate moderators. Taking this prediction-focused approach *from the start*, if paired with rigorous internal and external validation on a "held out" test set (as described in several papers in this issue), will complement rather than impede later explanatory approaches. These prediction-focused approaches are hardly a panacea, but there are early indications that fitting models that accommodate the multiply determined nature of psychological outcomes can improve our understanding (Fox et al., 2019; Wang et al., 2021). Psychopathology and the therapy process itself are complex systems, and matching our methods to that complexity will help us match our methods to our aspirations.

No part of this process will be easy, and the benefits are worth the difficulty.—In summary, we have argued there is a fundamental mismatch between our aspirations of discovering what works for whom and our current methods. The articles in this issue underscore the consistency of this mismatch, across diverse youth problem and treatment types. Taking steps to center our goal of identifying "what works for whom" should take precedence over continuing to rely on standard approaches. Identifying robust, reliable treatment moderators is devastatingly hard; even so, the steps we have outlined here, and the points highlighted throughout this issue, may catalyze real strides towards identifying what works for whom. Better aligning our methods with our aspirations will benefit our research, our knowledge, and ultimately, the many youths in need of personalized psychotherapy.

Disclosures.

JLS receives grant support from the National Institutes of Health (DP5OD28123), the Klingenstein Third Generation Foundation, the American Psychological Foundation, the Society for Clinical Child and Adolescent Psychology, Limbix, Inc., and the Upswing Fund for Adolescent Mental Health. JLS and MCM have co-authored and receive royalties from sales of a therapeutic workbook for adolescents, published by New Harbinger. JLS is under contract with Oxford University Press to co-edit a book on low-intensity mental health interventions for youth. The authors report no other financial conflicts.

References

Bachrach RL, & Chung T (2020, in press). Moderators of Substance Use Disorder Treatment for Adolescents. Journal of Clinical Child and Adolescent Psychology, 1–12.

Chevance A, Ravaud P, Tomlinson A, Le Berre C, Teufer B, Touboul S, Fried EI, Gartlehner G, Cipriani A, & Tran VT (2020). Identifying outcomes for depression that matter to patients, informal caregivers, and health-care professionals: qualitative content analysis of a large international online survey. The Lancet Psychiatry, 7(8), 692–702. [PubMed: 32711710]

Cohen ZD, Kim TT, Van HL, Dekker JJM, & Driessen E (2019). A demonstration of a multi-method variable selection approach for treatment selection: Recommending cognitive-behavioral versus psychodynamic therapy for mild to moderate adult depression. Psychotherapy Research: Journal of the Society for Psychotherapy Research, 1–14. [PubMed: 29254460]

Cuijpers P (2020). Measuring success in the treatment of depression: what is most important to patients? Expert Review of Neurotherapeutics, 20(2), 123–125. [PubMed: 31906736]

Danzi BA, & La Greca AM (2020, in press). Treating Children and Adolescents with Posttraumatic Stress Disorder: Moderators of Treatment Response. Journal of Clinical Child and Adolescent Psychology, 1–7.

Forscher P, & Ijzerman H (2021). How should we fund the PSA? Psychological Science Accelerator. https://psysciacc.org/2021/01/11/how-should-we-fund-the-psa/

Forstmeier W, Wagenmakers E-J, & Parker TH (2017). Detecting and avoiding likely false-positive findings - a practical guide. Biological Reviews of the Cambridge Philosophical Society, 92(4), 1941–1968. [PubMed: 27879038]

Fox KR, Huang X, Linthicum KP, Wang SB, Franklin JC, & Ribeiro JD (2019). Model complexity improves the prediction of nonsuicidal self-injury. Journal of Consulting and Clinical Psychology, 87(8), 684–692. [PubMed: 31219275]

Hatzenbuehler ML (2016). Structural stigma: Research evidence and implications for psychological science. American Psychologist, 71(8), 742–751. 10.1037/amp0000068

Hilgard J, Sala G, Boot WR, & Simons DJ (2019). Overestimation of action-game training effects: Publication bias and salami slicing. Collabra. Psychology, 5(1), 30.

Hobbs BP, Chen N, & Lee JJ (2018). Controlled multi-arm platform design using predictive probability. Statistical Methods in Medical Research, 27(1), 65–78. [PubMed: 26763586]

Kemp J, Barker D, Benito K, Herren J, & Freeman J (2020, in press). Moderators of Psychosocial Treatment for Pediatric Obsessive-Compulsive Disorder: Summary and Recommendations for Future Directions. Journal of Clinical Child and Adolescent Psychology*,* 1–8.

Klinger LG, Cook ML, & Dudley KM (2020, in press). Predictors and Moderators of Treatment Efficacy in Children and Adolescents with Autism Spectrum Disorder. Journal of Clinical Child and Adolescent Psychology, 1–8.

Leichsenring F, Abbass A, Hilsenroth MJ, Leweke F, Luyten P, Keefe JR, Midgley N, Rabung S, Salzer S, & Steinert C (2017). Biases in research: risk factors for non-replicability in psychotherapy and pharmacotherapy research. Psychological Medicine, 47(6), 1000–1011. [PubMed: 27955715]

Ludlow C, Hurn R, & Lansdell S (2020). A Current Review of the Children and Young People's Improving Access to Psychological Therapies (CYP IAPT) Program: Perspectives on Developing an Accessible Workforce. Adolescent Health, Medicine and Therapeutics, 11, 21–28.

Luedtke A, Sadikova E, & Kessler RC (2019). Sample Size Requirements for Multivariate Models to Predict Between-Patient Differences in Best Treatments of Major Depressive Disorder. Clinical Psychological Science, 2167702618815466.

Morris ZS, Wooding S, & Grant J (2011). The answer is 17 years, what is the question: understanding time lags in translational research. Journal of the Royal Society of Medicine, 104(12), 510–520. 10.1258/jrsm.2011.110180 [PubMed: 22179294]

Meyer AE, & Curry JF (2020, in press). Moderators of Treatment for Adolescent Depression. Journal of Clinical Child and Adolescent Psychology, 1–12.

Muñoz RF, Bunge EL, Chen K, Schueller SM, Bravin JI, Shaughnessy EA, & Pérez-Stable EJ (2016). Massive open online interventions: A novel model for delivering behavioral-health services worldwide. Clinical Psychological Science, 4(2), 194–205.

Nemesure MD, Heinz MV, Huang R, & Jacobson NC (2021). Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. Scientific Reports, 11(1), 1980. [PubMed: 33479383]

Norris LA, & Kendall PC (2020, in press). Moderators of Outcome for Youth Anxiety Treatments: Current Findings and Future Directions. Journal of Clinical Child and Adolescent Psychology*,* 1–14.

Patalay P, & Fried EI (2020). Editorial Perspective: Prescribing measures: unintended negative consequences of mandating standardized mental health measurement. Journal of Child Psychology and Psychiatry. 10.1111/jcpp.13333

Paul GL (1967). Outcome research in psychotherapy. Journal of Consulting Psychology*,* 31, 109–118. [PubMed: 5342732]

Price M, Weisz JR, McKetta S, Hollinsaid NL, Lattanner M, Reid A, & Hatzenbuehler M (2020). Are psychotherapies less effective for Black youth in communities with higher levels of anti-Black racism? 10.31219/osf.io/szu7v

Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, Moons KGM, Collins G, & van Smeden M (2020). Calculating the sample size required for developing a clinical prediction model. BMJ , 368, m441. [PubMed: 32188600]

Roley-Roberts ME, & Fristad MA (2020, in press). Moderators of treatment for pediatric bipolar spectrum disorders. Journal of Clinical Child and Adolescent Psychology, 1–14.

Schleider JL, Dobias ML, Sung JY, Mumper E, & Mullarkey MC (2020). Acceptability and utility of an open-access, online single-session intervention platform for adolescent mental health. Journal of Medical Internet Research: Mental Health, 7, e2013.

Schleider JL, Mullarkey MC, Fox K, Dobias M, Shroff A, Hart E, & Roulston CA (2021). Single-Session Interventions for Adolescent Depression in the Context of COVID-19: A Nationwide Randomized-Controlled Trial. 10.31234/osf.io/ved4p

Sherman RA, & Pashler H (2019). Powerful Moderator Variables in Behavioral Science? Don't Bet on Them (Version 3). 10.31234/osf.io/c65wm

Simmons JP, Nelson LD, & Simonsohn U (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science, 22(11), 1359–1366. [PubMed: 22006061]

Simon GE, & Perlis RH (2010). Personalized medicine for depression: can we match patients with treatments? The American Journal of Psychiatry, 167(12), 1445–1455. [PubMed: 20843873]

Smaldino PE, & McElreath R (2016). The natural selection of bad science. Royal Society Open Science, 3(9), 160384. [PubMed: 27703703]

Tosh C, Greengard P, Goodrich B, Gelman A, & Hsu D (2020). The piranha problem: Large effects swimming in a small pond. stat.columbia.edu. http://www.stat.columbia.edu/~gelman/research/unpublished/piranhas.pdf

Wang SB, Coppersmith DDL, Kleiman EM, Bentley KH, Millner AJ, Fortgang R, Mair P, Dempsey W, Huffman JC, & Nock MK (2021). A Pilot Study Using Frequent Inpatient Assessments of Suicidal Thinking to Predict Short-Term Postdischarge Suicidal Behavior. JAMA Network Open, 4(3), e210591. [PubMed: 33687442]

Weisz JR, Kuppens S, Ng MY, Eckshtain D, Ugueto AM, Vaughn-Coaxum R, Jensen-Doss A, Hawley KM, Krumholz Marchette LS, Chu BC, Weersing VR, & Fordwood SR (2017). What five decades of research tells us about the effects of youth psychological therapy: A multilevel meta-analysis and implications for science and practice. The American Psychologist, 72(2), 79–117. [PubMed: 28221063]

Whitcomb D, Battaly H, Baehr J, & Howard-Snyder D (2017). Intellectual humility: Owning our limitations. Philosophy and Phenomenological Research, 94(3), 509–539.

Yarkoni T, & Westfall J (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 12(6), 1100–1122. [PubMed: 28841086]