

Lack of improvements in youth psychotherapies or lack of investments in detecting improvements? Future directions in psychological assessment

Andres De Los Reyes^{a,*}, Elizabeth Talbott^b, Akram Yusuf^a, Nicole S.J. Dryburgh^{c,d},
Kimberly L. Goodman^e

^a Comprehensive Assessment and Intervention Program, Department of Psychology, University of Maryland, College Park, MD, USA

^b School of Education, William and Mary, Williamsburg, VA, USA

^c Department of Psychiatry and Behavioral Neurosciences, Offord Centre for Child Studies, McMaster University, USA

^d Department of Psychology, Harvard University, USA

^e Washington, DC, USA

ARTICLE INFO

Keywords:

Implementation science
Measurement-based care
Operations Triad Model
Psychotherapy
Situational specificity
Validity

ABSTRACT

Youth were experiencing mental health crises before the onset of the COVID-19 pandemic. Following this onset, their needs for mental health services have only increased. Yet, researchers encounter barriers to confronting these crises. The effects of therapies tested in controlled trials in the present day appear to be no more potent than those of their predecessors tested in trials conducted decades ago. Across these decades of scholarly work, researchers have invested far more of their efforts toward improving technologies for therapies than they have toward improving technologies for the assessment tools used to estimate therapeutic effects. The tools used today look a lot like those used in the 1970s—mainly surveys and interviews—and our strategies for integrating the data these tools produce focus on the sliver of their data that converge or yield the same results about youth mental health. Decades of work reveal that these integration strategies are incompatible with the data conditions that typify youth mental health assessments. We must invest in innovative assessment tools and integration strategies that capitalize on all the valid data produced by these conditions. This paper details pioneering directions in future research about psychological assessment. We describe the conceptual foundations underlying these research directions and highlight recent work by the authors and others supporting this pursuit. If we empower the assessment researchers of today to develop technologically innovative assessment tools and integration strategies, then we equip the therapy researchers of tomorrow to demonstrate that investing in therapy technologies pays off.

Children and adolescents (i.e., hereafter referred to as *youth* unless otherwise specified) were experiencing mental health crises before the onset of the COVID-19 pandemic, and following this onset, their needs for mental health services have only increased (COVID-19 Working Group, 2024). Imagine you are a therapist working with a youth who is experiencing a common need for mental health services, such as anxiety. How would you diagnose their anxiety? Which assessment tools might help guide the therapy they receive? Which tools might help you monitor their response to therapy? To address these questions, you might think about surveys administered via paper-and-pencil or a tablet device, or diagnostic interviews administered by trained assessors.

These might come to mind for you because they, along with behavioral tasks, are *traditional assessment tools*—the tools researchers and therapists use to assess mental health conditions and estimate responses to therapies designed to treat these conditions (Groth-Marnat & Wright, 2016; Hunsley & Mash, 2018; Weisz, Doss, & Hawley, 2005).

Based on your assessments, we want you to think about the therapies this youth might receive. Among a host of state-of-the-art therapy technologies, which might you consider? Might you consider the suite of pharmaceutical technologies (Cortese et al., 2024) or approaches to brain stimulation (Gallop et al., 2024)? Perhaps you might consider psychological therapies that infuse virtual reality (Blanco, Roberts,

* Corresponding author at: Comprehensive Assessment and Intervention Program, Department of Psychology, University of Maryland, Biology/Psychology Building, Room 3123H, College Park, MD 20742, USA.

E-mail addresses: adlr@umd.edu (A. De Los Reyes), ehtalbott@wm.edu (E. Talbott), ayusuf8@umd.edu (A. Yusuf), nicoledryburgh@fas.harvard.edu (N.S.J. Dryburgh).

<https://doi.org/10.1016/j.cpr.2026.102705>

Received 1 May 2025; Received in revised form 22 December 2025; Accepted 18 January 2026

Available online 19 January 2026

0272-7358/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Gannoni, & Cook, 2024), interactive media and gaming (Chen, Ou, Li, & Luo, 2024), teletherapy or mobile health (Zhang et al., 2024), or artificial intelligence into service delivery (Beg, Verma, & Verma, 2024)? All these therapies have two things in common. First, they were the byproduct of 50 years of carefully controlled randomized trials that estimated their effects, made possible by billions of dollars of global investments. Second, throughout these 50 years, estimates of therapy effects have largely come from traditional assessment tools, namely surveys and interviews (De Los Reyes, 2024; Weisz et al., 2005). Global investments have produced a considerable array of therapies over the last 50 years. In contrast, comparatively few investments have been made to improving assessment tools, which means the tools used to estimate therapy effects have undergone little change. The outcome data from surveys and interviews used to identify evidence-based therapies largely come from tools built with features from the 1970s. Even if the delivery methods have changed over this time (e.g., paper-and-pencil in 1970s vs. tablet or smartphone in the current day), the contents of assessment tools—i.e., the items they contain, their response options, and the informants serving as data sources (e.g., youth, parents, teachers)—have remained largely unchanged for five decades.

A disparity exists in the investments made to improve therapies, relative to the investments made to improve the tools used to estimate therapy effects. This is important, because the therapies designed to treat youth mental health, and the tools used to gather evidence about therapy effects are *complementary variables* (see Kiukas, Lahti, Pellonpää, & Ylinen, 2019). They might have been developed by distinct research teams, but their existence is intertwined: You cannot have evidence-based therapies without high-quality evidence. To illustrate the importance of this problem, consider the state of the science about

psychological therapies for youth mental health. Fig. 1 depicts the last 50 years of estimates of youth therapy effects. The effects of therapies tested in controlled trials now appear to be no more potent than those of their predecessors tested in trials decades ago; in many cases, the effects are trending downward (Weisz et al., 2019). Do these data indicate that therapies, in fact, have not improved in their potency over time? The answer to this question depends on your confidence in the quality of the evidence.

Consider Heisenberg's (1927) Uncertainty Principle. The Uncertainty Principle holds that prioritizing the quality of the data used to understand one complementary variable comes at a cost. Prioritizing one variable diminishes the quality of the data used to understand the other variable. Over the last 50 years, investing in therapies may very well have translated into improving the potency of the therapies that youth receive. Yet, during this time, assessment tools also needed investments. Creating this much of an imbalance between investments in therapies on the one hand, and assessment tools on the other, creates barriers to not only accurately estimating therapy effects but also optimizing their effects for the youth who receive them.

The lack of investments in assessment tools have produced three barriers to accurately estimating therapy effects and optimizing the benefits of therapies youth receive. First, the individual studies of youth therapies meta-analyzed and depicted in Fig. 1 estimated therapy effects using a variety of outcome measures, measurement modalities, and informants (Hunsley & Mash, 2018; Weisz et al., 2005). This approach robustly produces *discrepant results*—distinct estimates about a mental health domain (e.g., therapy effects) in terms of its form, function, level, or relations with measures of other domains (e.g., moderators of therapy effects; De Los Reyes & Ekins, 2023). These discrepancies have been

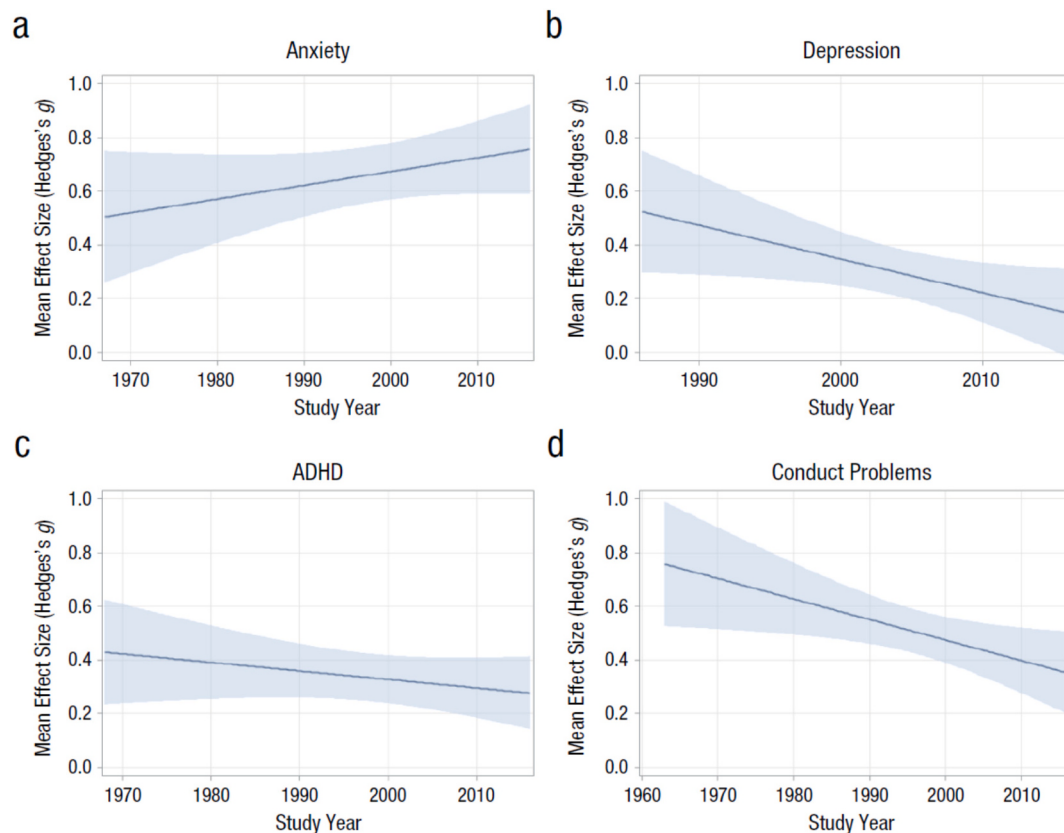


Fig. 1. Estimated change in mean effect size over time for treatment of (a) anxiety, (b) depression, (c) attention-deficit/hyperactivity disorder (ADHD), and (d) conduct problems in the mixed-effects model. The lines indicate the mean Hedges's g , and the blue shading around lines represents the 95% confidence interval. There was a significant interaction between study year and target problem. There was no significant change in mean effects across the years for anxiety and ADHD, but there was a significant decline for depression and conduct problems. (Reproduced from Weisz et al., 2019). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

observed for decades (see Achenbach, 2020; Kraemer et al., 2003), and yet all throughout this time the strategies researchers have most often used to integrate their data focus on the sliver of results that converge (e.g., De Los Reyes, 2024; Eid, Geiser, & Koch, 2024). Greater investments may produce assessment tools and integration strategies that facilitate extracting valid data from both discrepant and converging results.

Second, evidence-based practices extend beyond the use of data to select therapies. Evidence should also inform the tailoring of services. In theory, evidence-based practices include the repeated use of data during service delivery, to inform the tailoring of therapy for individual youth on such issues as therapy planning, construction of therapeutic activities (e.g., role-plays, exposures), and monitoring therapeutic responses (i.e., *measurement-based care*; McLeod, Jensen-Doss, Lyon, Douglas, & Beidas, 2022). Yet, in practice therapists often make decisions about tailoring therapy without data, and largely based on a combination of intuition and experience (see Venturo-Conerly, Reynolds, Clark, Fitzpatrick, & Weisz, 2023). This is another long-known and persistent problem in service delivery (see Dawes, 1994). Greater investments may produce assessment tools and integration strategies that infuse evidence into the process of tailoring therapies for individual youth and their unique mental health needs.

Third, the lack of progress in innovating assessment tools to estimate therapy effects, assessment tools to tailor therapies, and strategies for integrating the data produced by these tools creates a vacuum in approaches to data optimization. Without progress in innovating assessment tools and data integration strategies, researchers lack an impetus to design laboratory studies to validate the scores that assessment tools produce or test approaches to implementing assessment tools and data integration strategies in applied settings such as schools, clinics, and hospitals. Relatedly, assessment tools are only as useful as their ability to reflect how youth experience mental health and the communities and social environments in which they live (e.g., Bettis et al., 2024; Giusto, Triplett, Foster, & Gee, 2024). Yet, it is rare to see researchers develop and test assessment tools in partnership with youth, their parents, and other partners in service delivery (e.g., Dryburgh et al., 2025; Hausman et al., 2013). Taken together, we require new directions in research that are conceptually grounded, forward-thinking, and address barriers to both validating scores taken from assessment tools and implementing these tools in applied settings.

1. Aims and scope

The aims and scope of our paper necessitate understanding *contemporary technologies* with implications for psychological assessment and their status in therapy research and service delivery. These include neurobiological technologies (e.g., task-based functional magnetic resonance imaging [task-based fMRI], event-related potentials via electroencephalography [EEG/ERP]; Beauchaine & Hinshaw, 2020; Insel et al., 2010), and experience sampling and real-time monitoring technologies (e.g., location data, heart rate variability, skin conductance, sleep health; Kleiman, Glenn, & Liu, 2019; Stone, Schneider, & Smyth, 2023). The state of the science of psychological assessments via these contemporary technologies differs from that of therapies. We previously mentioned the value of using evidence to inform tailoring therapies. This tailoring includes such features of therapies as the dosage of medications to suit youth of varying ages or body weights, the intensity of brain stimulation techniques to accommodate youth of varying ages or developmental periods, and the use of age-appropriate media for digital therapies. Tailoring therapies facilitates (a) rigorously estimating their effects (e.g., randomized controlled trials; Kazdin, 2024) and (b) classifying these effects within systems for identifying evidence-based therapies (e.g., Chambless & Ollendick, 2001; Southam-Gerow & Prinstein, 2014).

Investing in therapy technologies means designing services for individuals, and this means these technologies must be built from the “ground up” for tailoring. Historically, the neurobiological and

experience sampling technologies cited previously were not, from their foundations, built to assess individual differences in psychological phenomena—a prerequisite for estimating therapy effects in research or making clinical decisions in service delivery. Estimating therapy effects requires tools that meet thresholds for reliable, valid assessment of individual differences (e.g., Hunsley & Mash, 2018; Youngstrom et al., 2017). These thresholds also ensure that therapists can use these same tools to make accurate clinical decisions (American Educational Research Association [AERA], 2014). Researchers might use these tools to precisely estimate how individual youth in a therapy study change in their functioning over the course of their therapy. Therapists might use these tools to precisely detect whether a given score from a tool distinguishes youth in discrete and valid ways, such as meeting or not meeting criteria for a diagnosis or responding or not responding to a therapy they received.

To justify their implementation in therapy research and service delivery, assessment tools require evidence that the scores they produce meet established thresholds for psychometric soundness. As of this writing, scores taken from contemporary technologies require a great deal more psychometric evidence to meet these thresholds. Consider meta-analytic work that reveals low test-retest reliability estimates for psychological indices produced via task-based fMRI (Elliott et al., 2020). Such issues surrounding reliability estimates extend to psychological indices produced via EEG/ERP (cf. Hajcak, Meyer, & Kotov, 2017; Infantolino, Luking, Sauder, Curtin, & Hajcak, 2018; Szenczy, Levinson, Hajcak, Bernard, & Nelson, 2021). Variable reliability also characterizes scores taken from experience sampling technologies (e.g., Kleiman et al., 2019; X. Hu et al., 2024; Stone et al., 2023). Issues regarding measurement reliability must be addressed before considering whether the scores these tools produce index psychological phenomena relevant to estimating therapy effects. The lack of a strong psychometric evidence base for scores taken from contemporary technologies plays a significant role in why our current understanding of therapy effects in youth mental health relies on the same technologies that therapy researchers used decades ago (e.g., surveys and interviews; Weisz et al., 2005; Weisz, Venturo-Conerly, Fitzpatrick, Frederick, & Ng, 2023). If surveys and interviews predominate the suite of tools that produce scores meeting psychometric thresholds (i.e., see Hunsley & Mash, 2018), then we should not be surprised that traditional assessment tools are what researchers use to characterize therapy effects in research and classify evidence-based therapies, and what therapists use to guide their decision-making when delivering mental health services (AERA, 2014; Weisz et al., 2017).

Investing in technologies that might capture individual differences in psychological phenomena is not synonymous with investing in research about psychological assessment. This paper focuses on investing in technologies with the *specific intent* of assessing mental health in psychometrically sound ways. As such, we focus on research about assessment tools with demonstrated psychometric support for the scores they produce. This is a conscious decision we made for this paper, given the current state of science funding (Gensler, Johnson, Panizza, & di Maurobc, 2025; Leifer, Liu, & Nagel, 2025; Steen, 2025). It would be impractical to chart directions in future research that require significant investments to, for instance, build an evidence base for instrumentation that produces scores with as-of-yet uncertain psychometric properties (e.g., studies requiring thousands of participants to arrive at clinical or normative scores based on data from task-based fMRI or EEG/ERP). Thus, this paper describes research directions that emphasize tools with an evidence base for the psychometric properties of the scores they produce, that are open access, and that are brief and thus can be feasibly implemented in both research and applied settings. In these ways, the future directions we describe are sensitive to (a) current funding conditions, (b) the broader movement toward scientific transparency, and (c) the need for equitable access to quality assessment services across settings and diverse populations of youth (e.g., Becker-Haimes et al., 2020; Beidas et al., 2015; Howard, Neal, Kirtley, Urry, & Tackett, 2025;

Sanchez et al., 2022; Tackett et al., 2019a).

This paper articulates innovative and actionable directions in future research about psychological assessment. We briefly outline the conceptual foundations underlying these directions in future research. Heisenberg's (1927) Uncertainty Principle highlights the value in considering imbalances between complementary variables, such as the investments made to therapies versus the tools used to estimate their effects. Thus, the future directions that we propose address barriers produced by imbalances in the investments made in therapies versus assessment tools. We also report findings from recent studies that support investing in research about assessment tools and integration strategies. We expect efforts in designing, validating, and implementing tools for assessing mental health, along with strategies for integrating the data they produce, will optimize our ability to estimate therapeutic effects. Catalyzing these efforts now will equip the therapy researchers of tomorrow to demonstrate that investing in advancements in therapy technologies, in fact, results in improving the potency of therapeutic effects.

2. Conceptual foundations

2.1. Situational specificity

The directions in future research that we describe in this paper rest on conceptual foundations that trace back decades. As such, we briefly summarize this conceptual work, to set up a larger discussion about these future directions. We begin with a theory posed in the 1980s. The theory—*situational specificity*—stems from a robust observation in assessments of youth mental health that extends to assessments of psychological states among adults (see Mischel & Shoda, 1995). When assessing youth mental health, researchers have long known that their scientific observations cannot be accurately captured with the data produced by a single assessment tool (e.g., Lapouse & Monk, 1958; Richters, 1992). Consequently, over decades researchers have developed many tools for assessing youth mental health domains—tools that traverse various modalities and informants (see Hunsley & Mash, 2018; Meyer et al., 2001). The informants most typically sought out by researchers are everyday people in the lives of youth—people who spend significant amounts of time interacting with youth and, presumably, significant time observing their behavior. These people include parents and teachers, who are often solicited as informants because they not only observe youth for significant periods of time, but they also observe youth within distinct social contexts (e.g., parent at home vs. teacher at school). Scholars in Quantitative Psychology define these people as *structurally different informants*—people who harbor unique abilities to observe youth in contexts pertinent to assessing and understanding their behavior (Eid et al., 2008, 2024).

Prior meta-analytic work documents that use of structurally different informants results in data that often correlate in the small-to-moderate range (e.g., Pearson r correlations in the 0.20-to-0.30 range; Achenbach, McConaughy, & Howell, 1987; De Los Reyes et al., 2015). These low correlations among informants' reports exemplify a literature defined by the discrepant results described previously. Studies about youth mental health frequently produce distinct estimates about anything from the rate of a diagnostic condition (e.g., conduct disorder), to identifying a risk or protective factor for the condition (e.g., parenting), to the conclusions one draws of the effect of a therapy designed to treat the condition (e.g., behavioral parent training for conduct disorder; De Los Reyes, 2024).

Situational specificity facilitates characterizing the discrepant results produced by studies about youth mental health. Specifically, let us consider that, in order for researchers to rely on a given informant, like a teacher, they must spend a great deal of time (a) developing tools that allow the informant to provide precise and accurate information about mental health domains, (b) designing studies that allow for these tools to be thoroughly vetted for the data they produce, and (c) integrating

findings across studies to determine whether that informant provides sound data via these assessment tools (Hunsley & Mash, 2018). Let us also consider that the use of structurally different informants is predicated on the notion that youth may behave differently within and across the social contexts that typify their day-to-day lives (e.g., Kraemer et al., 2003; Makol et al., 2020). These are what we refer to in this paper as *data conditions*. By “data conditions” we mean those features that characterize the information that researchers collect, using the assessment tools they developed and the informants to whom they administer the tools. When researchers encounter discrepant results, they do so within data conditions typified by use of assessments and informants for which their use is supported by psychometric evidence.

Within the data conditions we just described, let us consider an example of situational specificity. Consider a therapist who collects a report from a parent who says their child displays oppositional behavior, but the child's teacher does not corroborate the parent's impressions. This discrepancy between parent and teacher reports may represent a true difference in how the child behaves at home versus at school. Situational specificity allows researchers to test whether this discrepancy is a marker of contextual changes in the child's behavior. Discrepant results that reflect situational specificity contain *domain-relevant information*—valid data about the mental health domain being measured.

2.2. How situational specificity grounds frameworks in psychological assessment

Situational specificity and prior meta-analytic work serve as useful starting points for characterizing discrepant results. Yet, they are insufficient in one key respect. Let us return to the example of a child's oppositional behavior reported by their parent and not their teacher. This may signal that the child displays oppositional behavior at home to a greater degree than they do at school. Clearly, not all children behave in this way. Some children might display greater oppositional behavior at school than home. Other children might display a high degree of oppositional behavior at both school and home. Situational specificity compels us to consider why social contexts are not created equal. Social contexts like home and school might vary in their *contingencies*. By “contingencies” we mean the scenarios that precede, precipitate, elicit, or otherwise trigger both the behaviors on which researchers rely to measure mental health domains, and the consequences that maintain these behaviors (see Dunlap & Kern, 2018). In this respect, we can think of contingencies as “if-then” scenarios. Examples of these contingencies include unfamiliar scenarios (e.g., Avoidance when introduced to new people, resulting in difficulties with building positive interpersonal relationships), unexpected scenarios (e.g., Irritability and anger when faced with a school task, resulting in decreased academic performance), and confrontational scenarios (e.g., Aggression in response to conversations with parents, resulting in increased family discord or conflict).

If two social contexts each contain contingencies that tend to elicit the same kinds of behavior (e.g., anxiety, oppositional behavior, hyperactivity), with consistent responses by others to that behavior (e.g., parents, teachers), then informants who observe behavior in those contexts may produce reports that agree or converge with one another. Conversely, if two social contexts contain distinct kinds of contingencies and varying responses by others, then informants who observe behavior in those contexts may produce reports that disagree or diverge with one another. Importantly, prior work conforms with these notions. For instance, youth exhibit considerable variability in where they display behaviors indicative of youth mental health domains, such as social anxiety, depression, autism, oppositional behavior, and conduct problems (for a review, see De Los Reyes, 2024). This contextual variability in how youth display behavioral indicators of mental health has profound implications for psychological assessment and by extension, the directions in future research described in this paper. Although assessments of youth mental health domains robustly produce discrepant

results, not all assessments disagree. Further, among those results that do disagree, they might not all mean the same thing. Consequently, the last dozen years has seen the emergence of three frameworks designed to facilitate interpreting discrepant results in youth mental health.

First, the data conditions produced by youth mental health assessments require concepts for characterizing them. In the early 2010s, these concepts were integrated within a framework known as the *Operations Triad Model* (De Los Reyes, Thomas, Goodman, & Kunder, 2013). Briefly, the framework includes a concept that reflects conditions in which social contexts contain enough similarities in terms of their contingencies, such that informants who observe behavior within distinct contexts make similar reports about the youth undergoing evaluation (*Converging Operations*). The Operations Triad Model also includes a concept that reflects conditions in which social contexts contain enough dissimilarities in terms of their contingencies, such that informants who observe behavior within distinct contexts produce discrepant reports about the youth undergoing evaluation (*Diverging Operations*). Both Converging and Diverging Operations reflect data conditions in which patterns of reports contain domain-relevant information. However, assessment tools are not perfect. Neither are the informants who make reports using these tools. Thus, the Operations Triad Model also includes a concept for those data conditions in which informants make discrepant reports that are best explained by measurement confounds, such as random error or rater biases (i.e., *Compensating Operations*).

Second, the Operations Triad Model reveals a problem with historical approaches to validating scores taken from assessment tools, namely that they compel their users to focus on one source of domain-relevant information. Historically, the strategies that researchers have most often used to integrate data (e.g., latent variable models, composite scores, AND/OR rules or combinatorial algorithms, a priori selection of a primary outcome measure; see De Los Reyes & Epkins, 2023; Eid et al., 2008, 2024), have been grounded in a validation paradigm that treats results that converge as the only source of valid data—the *multi-trait, multi-method matrix* (MTMM; Campbell & Fiske, 1959; Fiske & Campbell, 1992). The MTMM treats the discrepant results that characterize much of the youth mental health literature as Compensating Operations (e.g., random error, bias). This means that, within the MTMM, Diverging Operations do not exist. This feature of the MTMM required the development of a validation paradigm that distinguishes Diverging Operations from Compensating Operations. This paradigm—*CONTEXT* (Classifying Observations Necessitates Theory, Epistemology, and Testing; De Los Reyes et al., 2023)—contains a feature that is absent from the MTMM (see also Haefel et al., 2022; Watts, Greene, Bonifay, & Fried, 2024), namely the ability to *falsify* or conduct direct tests of the notion that Diverging Operations do not exist. In essence, *CONTEXT* guides scholars to ground their assessment practices in falsifiability, so that their practices align with the data conditions produced when assessing youth mental health. These practices include selecting informants whose reports are hypothesized to produce valid discrepant results, implementing analytic procedures that allow users to model or treat discrepant results as valid sources of data, and constructing study designs that allow researchers to test links between indices of discrepant results and validity criteria.

Third, the goal of validation studies is to provide researchers with certainty about how to interpret the data produced by assessment tools. Following validation, researchers also need guidance on how to use assessment tools in applied settings, namely the use of data produced by these tools to make valid decisions about mental health services (i.e., *implementation science*; Wiltsey Stirman & Beidas, 2020). Consider if an assessment tool produces discrepant results that reflect Diverging Operations. Results that reflect Diverging Operations have the potential to inform how therapists tailor interventions to the unique social contexts of youth who receive them. To do otherwise—to treat all discrepant results as Compensating Operations—is to prevent valid data from informing service delivery. Clearly, the possibility that researchers

might mischaracterize a set of discrepant results as error or bias when they, in fact, reflect valid data is not exclusive to the results of validation studies. Thus, just as *CONTEXT* extends the Operations Triad Model into validation studies, the *Needs-to-Goals Gap* framework extends the Operations Triad Model into implementation science (De Los Reyes et al., 2022a). The Needs-to-Goals Gap framework specifically focuses on measurement-based care. As such, the framework applies the concepts of Converging, Diverging, and Compensating Operations to the use of data to tailor therapies and make clinical decisions about such factors as diagnosis, case conceptualization, therapy planning, constructing therapeutic activities, and monitoring therapeutic responses.

2.3. How conceptual foundations ground future directions in psychological assessment

Taken together, situational specificity, the Operations Triad Model, *CONTEXT*, and the Needs-to-Goals Gap framework ground the directions in future research described in this paper. If assessment tools produce data that reflect situational specificity (i.e., Converging and Diverging Operations), then these tools produce data conditions that contain domain-relevant information about youth mental health. Although these tools can produce converging results, we must balance the possibility of these results with the reality that discrepant results typify youth mental health assessments. Discrepant results define the outcomes of youth mental health assessments because common assessment practices greatly increase the likelihood of these results. The strategic use of structurally different informants to assess mental health robustly produces discrepant results within assessments conducted in both research and applied settings. This use also makes it possible to estimate how youth behave within and across clinically relevant social contexts. The Operations Triad Model allows for these assessments to reveal discrepant results, and at the same time, produce uniquely valid data about youth mental health.

If assessment tools produce discrepant results that reflect situational specificity, and these results can be detected precisely and accurately, then we expect strategies designed to integrate discrepant results to boost the explanatory power of data produced by assessment tools. The Operations Triad Model facilitates conceptualizing discrepant results, *CONTEXT* facilitates constructing validation studies that accurately detect valid discrepant results, and the Needs-to-Goals Gap framework facilitates using the discrepant results observed in applied settings to make evidence-based and valid decisions about mental health services. These conceptual foundations equip researchers to develop, test, and implement the next generation of tools for assessing youth mental health as well as the strategies for integrating the data produced by these tools.

3. Future directions in research about psychological assessment

Fig. 2 depicts six directions in future research, which all extend from the Operations Triad Model. We grouped these directions in future research in sets of three. We grounded one set of three directions in *CONTEXT*, to guide validation studies. We grounded a second set of three directions in the Needs-to-Goals Gap framework, to guide implementation studies. Organizing our directions in future research in this way facilitates explaining how validation studies inform implementation studies. Further, we link each future direction to considerations surrounding (a) technology; (b) transparency; (c) community engagement; and (d) the representativeness of data produced by assessment tools, with a particular emphasis on the use of participatory action approaches to research (see Giusto et al., 2024).

The thesis underlying this paper is that investments in improving assessment tools and practices for integrating the data these tools produce are central to accurately estimating therapy effects and tailoring therapies to the unique needs of the youth who receive them. Along these lines, Table 1 aligns each future research direction depicted in Fig. 2 with (a) a current assessment practice it is designed to reform, (b)

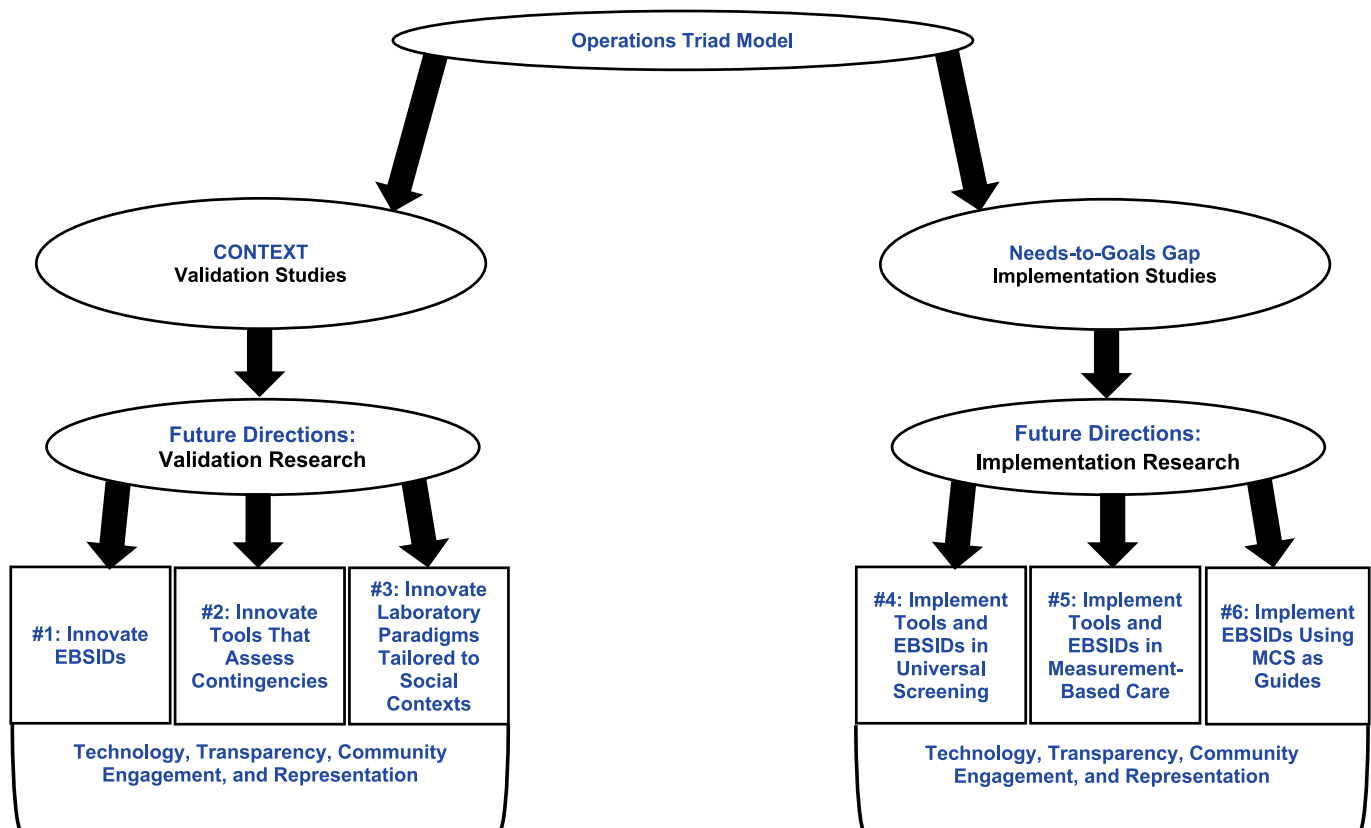


Fig. 2. Directions in future research informed by conceptual foundations for validation and implementation studies. Each future direction includes considerations regarding technology, transparency, community engagement, and representation. CONTEXT = Classifying Observations Necessitates Theory, Epistemology, and Testing; EBSIDs = evidence-based strategies for integrating data; MCS = Monte Carlo simulations.

the anticipated benefits of investing resources in reforming this practice, and (c) citations supporting the reform.

3.1. Future directions: validation studies

Validation studies provide researchers with certainty about what the data produced by assessment tools mean. These studies are time-intensive and incorporate use of multiple assessment modalities and informants. In this sense, validation studies inform research that seeks to implement assessment tools in applied settings. The validation studies we describe below include study design components that prioritize certainty—namely use of time-intensive laboratory procedures—but in the service of validating scores taken from the kinds of assessment tools that can be feasibly implemented in mental health service delivery. Importantly, these studies leverage the strengths of controlled observations, and in a way that ensures that they accurately simulate the social environments that surround youth and their experiences with the mental health domains that researchers seek to understand.

3.1.1. Future direction #1: innovate evidence-based strategies for integrating data (EBSIDs)

What if a colleague or trainee approached you for advice on assessing a mental health domain among youth, like their anxiety, hyperactivity, or depression? We previously mentioned that no single tool definitively measures any of these or other domains. Thus, you likely would advise your colleague or trainee to take a *multi-modal* approach—use multiple tools that measure the domain but in distinct and complementary ways. Let us say your colleague or trainee took your advice, used multiple tools, and now they need more advice. This is because this approach produced the discrepant results we previously described. By construction, discrepant results necessitate strategies for integrating data. What

integration strategy might you advise they use? Would you suggest they take an average of these data? Or construct a latent variable to estimate the *common variance* among these data (i.e., converging results)? Or should they pick the one score that “best represents” the domain?

As mentioned previously, the integration strategies we just described are those most often used by scholars to integrate youth mental health data, they logically flow from the MTMM, and this means these strategies require users to treat *unique variance* (i.e., discrepant results) as measurement confounds (see De Los Reyes & Epkins, 2023; Eid et al., 2008, 2024). When integrating data from individual samples of youth, prior work indicates that the scores produced by these integration strategies retain roughly 10% (or less) of the variance produced by multi-informant assessments (see De Los Reyes et al., 2015; Howe et al., 2019). Similarly, the data that we depicted in Fig. 1 were produced using meta-analytic procedures for computing mean effect sizes. These procedures also integrate data in ways that prioritize common variance. Thus, it should come as no surprise that in the same meta-analytic dataset used to produce the findings depicted in Fig. 1, the largest moderating factor of therapy effects, by far, is the informant completing the outcome measure (see Table 1 of Weisz et al., 2017). Yet, both authoritative reviews of youth therapies and classification systems used to identify evidence-based therapies largely focus on how effects converge across studies (see Chambless & Ollendick, 2001; Southam-Gerow & Prinstein, 2014; Weisz et al., 2023). What if these approaches to integrating evidence “leave out” valid unique variance that, if retained, could reveal new approaches to characterizing and understanding therapy effects?

Grounded in CONTEXT, an emerging body of research repurposes well-established validation techniques—tests of whether the properties of scores taken from assessment tools surpass thresholds for psychometric soundness—to test strategies for integrating the data produced by

Table 1
How investments in future directions in research about psychological assessment improve upon current assessment practices.

Current Assessment Practices	Future Directions in Research about Psychological Assessment	How Investing in Research about Psychological Assessment Will Improve Assessment Practices	Citation Support
Use of strategies for integrating multi-informant data that assume assessment results that converge (i.e., common variance) are the lone sources of valid data and discrepant results (i.e., unique variance) cannot contain valid data. When integrating data from a sample of youth, prior work indicates that the scores produced by these integration strategies retain roughly 10% (or less) of the variance produced by multi-informant assessments, and these integrated scores uniquely correlate with scores from non-informant or independent validity criteria (e.g., observed behavior) in the low-magnitude range (e.g., $r_s \approx .10s$).	Innovate evidence-based strategies for integrating data (EBSIDs)	The development of EBSIDs for integrating multi-informant data that extract valid data from both common and unique variance among informants' reports, resulting in integrated scores that correlate with independent validity criteria (e.g., observed behavior) in the large-magnitude range (e.g., $r_s \approx .70s$).	De Los Reyes et al., 2015 Dunning et al., 2004 Hendriks et al., 2018 Howe et al., 2019 Makol et al., 2020 Makol et al., 2025 Meyer et al., 2001 Weisz et al., 2017
When delivering therapy, therapists tend to make decisions about tailoring therapy without data, and largely based on a combination of intuition and experience.	Innovate tools that measure contingencies	The development of tools that facilitate therapists' use of evidence to make decisions about tailoring therapy.	Fine et al., 2025 Herman et al., 2025 Venturo-Conerly et al., 2023
Use of laboratory paradigms to validate links between survey content and actual behavior that are based on laboratory tasks developed decades ago and have not been updated to reflect contemporary social contexts.	Innovate laboratory paradigms tailored to social contexts	Leverage immersive VR and generative AI technologies to develop laboratory paradigms that allow researchers to tailor behavioral tasks to each person being evaluated, thus optimizing the development of clinically feasible tools (e.g., surveys) that accurately reflect behavior in contemporary and clinically relevant social contexts.	Cannon et al., 2020 De Los Reyes et al., 2023
Use of universal screening	Implement tools and EBSIDs in	The implementation of universal	De Los Reyes et al.,

Current Assessment Practices	Future Directions in Research about Psychological Assessment	How Investing in Research about Psychological Assessment Will Improve Assessment Practices	Citation Support
paradigms that largely rely on single-source assessment approaches, which are unlikely to capture the unique, context-specific ways that people display mental health concerns.	universal screening	screening protocols that not only detect those in need of services, but also the contexts in which their needs manifest.	2019 De Los Reyes et al., 2022b
Use of approaches to measurement-based care that are either based on single-informant methods of assessing client functioning or represent data using formats that require expertise in data science for interpretation and use (e.g., line graphs).	Implement tools and EBSIDs in measurement-based care	The implementation of protocols for representing data using formats (e.g., icons and symbols) that do not require backgrounds in data science to interpret, thus optimizing the use of evidence to inform decision-making.	Charamut et al., 2022 de Jong et al., 2021 Delgadillo et al., 2022 Lambert et al., 2003
Application of data integration strategies to samples whose data conditions violate the strategy's assumptions.	Implement EBSIDs using Monte Carlo simulations as guides	The implementation of data simulations to inform selecting data integration strategies, thus optimizing the likelihood that researchers select integration strategies that fit the data conditions to which they are applied.	Bauer et al., 2013 De Los Reyes et al., 2025 Kraemer et al., 2003 Makol et al., 2025

Directions in future research listed in this table are those depicted in [Fig. 2](#). VR = virtual reality; AI = artificial intelligence.

these tools ([De Los Reyes, 2024](#)). Researchers use these tests to detect *evidence-based assessments*—tools that produce scores that can be interpreted reliably and validly as markers of mental health domains ([Hunsley & Mash, 2018](#)). Researchers can use this same approach to detecting evidence-based assessments to construct falsifiable approaches to detecting what we define in this paper as *evidence-based strategies for integrating data* (EBSIDs). For an integration strategy to be deemed an EBSID, it too must pass a threshold: Does the strategy's *assumptions* fit the data conditions to which it is applied? By “assumptions” we mean the underlying elements of an integration strategy that a user must accept as true to apply the strategy to a given set of data conditions. Not only are these tests of fit crucial to execute, but also well-established validation techniques ground the execution of these tests.

When considering the concept of EBSID, it is important to note that not all integration strategies are created equal. In youth mental health, commonly used strategies for integrating data likely lack the ability to “pass the EBSID threshold.” This is because the most often used integration strategies assume that converging results (i.e., the common variance) are the lone sources of valid data and that discrepant results (i.e., the unique variance) cannot contain valid data. If users apply these strategies to discrepant results that contain valid data, then they violate these assumptions ([De Los Reyes, 2024](#)). Not only might these strategies not “count” as EBSIDs in youth mental health, but their use might have dire consequences when applied to data conditions that violate their

assumptions. Applying an integration strategy to data that violate the assumptions of that strategy means misclassifying valid data as measurement confounds or vice versa (Makol, Yang, Wang, & De Los Reyes, 2025). By leveraging validation techniques to test the fit between integration strategies and data conditions, researchers can (a) estimate the consequences of applying a strategy to data conditions that violate the strategy's assumptions, as well as (b) identify the benefits of applying a strategy to data conditions that fit the strategy's assumptions.

For example, consider recent work that tested the fit of strategies for integrating multiple informants' reports about adolescent social anxiety. This work involved use of the *Unfamiliar Peer Paradigm*—a set of controlled laboratory tasks that simulate social contexts relevant to understanding social anxiety among adolescents, namely interactions with unfamiliar same-age peers (Fig. 3; Cannon et al., 2020). Task scripts and training materials for the Unfamiliar Peer Paradigm exist on the Open Science Framework (<https://osf.io/zadbc/>). Researchers collected data from these tasks in two ways. First, they developed schemes for training independent observers to code anxiety behaviors via archived videos of adolescents participating in these tasks (Glenn et al., 2019). These coding schemes produced the validity criteria necessary for examining the fit between integration strategies and the data conditions produced when assessing adolescent social anxiety (i.e., via tests of criterion-related validity). Second, researchers solicited reports from the research personnel with whom the adolescents interacted in these tasks; essentially the people who simulated unfamiliar same-age peers (Deros et al., 2018). Unlike the independent observers, these personnel did not receive any training on how to rate adolescents' behavior. In fact, they provided their reports via the same surveys used by parents to rate their adolescent's social anxiety and by adolescents to rate their own social anxiety. In this respect, these personnel produced survey reports that were structurally different from the survey reports provided by parents. That is, they based their reports about adolescents on their observations of the social anxiety that adolescents displayed within controlled social interaction tasks. These tasks were designed to simulate social contexts that adolescents experience outside of the home. In line with this presumption, estimates of correspondence between parent reports and reports from these personnel were quite low in magnitude (i.e., Pearson r s as low as 0.01; Makol et al., 2025). Yet, these reports and the self-reports

of adolescents significantly related to the validity criteria—independent observers' ratings of adolescent anxiety exhibited in the tasks depicted in Fig. 3 (i.e., Pearson r s between 0.25 and 0.58; Glenn et al., 2019). Further, both the parent reports and reports from research personnel displayed moderate-to-large correlations with adolescent self-reports (i.e., Pearson r s as high as 0.51; Deros et al., 2018). These findings signal that the three informants' survey reports display some degree of valid common variance as well as a significant degree of valid unique variance.

Taken together, informants' reports about adolescent social anxiety reflect data conditions that contain valid common variance as well as valid unique variance. Do these data conditions have implications for the fit of strategies for integrating multi-informant data? To address this question, researchers tested the fit of several integration strategies to the data conditions we just described and depicted in Fig. 3 (Makol et al., 2025). Three of these strategies had one thing in common—they require users to assume that discrepant results reflect measurement confounds. They included the average of scores (i.e., *composite scores*), the best predictor among scores, and a sophisticated structural model designed to attain precise estimates of common variance (i.e., *Trifactor Model*; Bauer et al., 2013). Hereafter, we refer to these as *common variance integration strategies*. Alongside these three common variance integration strategies, researchers tested a fourth strategy. This strategy required users to hold different assumptions than those that underlie use of common variance integration strategies, namely that both converging and discrepant results may contain valid data. We depict this fourth integration strategy in Fig. 4. Briefly, the strategy requires users to strategically select informants whose reports are hypothesized to disagree and for valid reasons, using a schematic that metaphorically connects to the geographic positioning of satellites (i.e., *Satellite Model*; Kraemer et al., 2003).

A geographic scientist locks in on an object's location by positioning an array of satellites at distinct points in geographic space, using location estimates that vary by latitude and longitude. These satellites are positioned in such a way that they “form a triangle” above the object. As we depict in Fig. 4a, applying the Satellite Model to tools for assessing mental health involves the strategic selection of structurally different informants whose reports vary by the contexts in which they observe

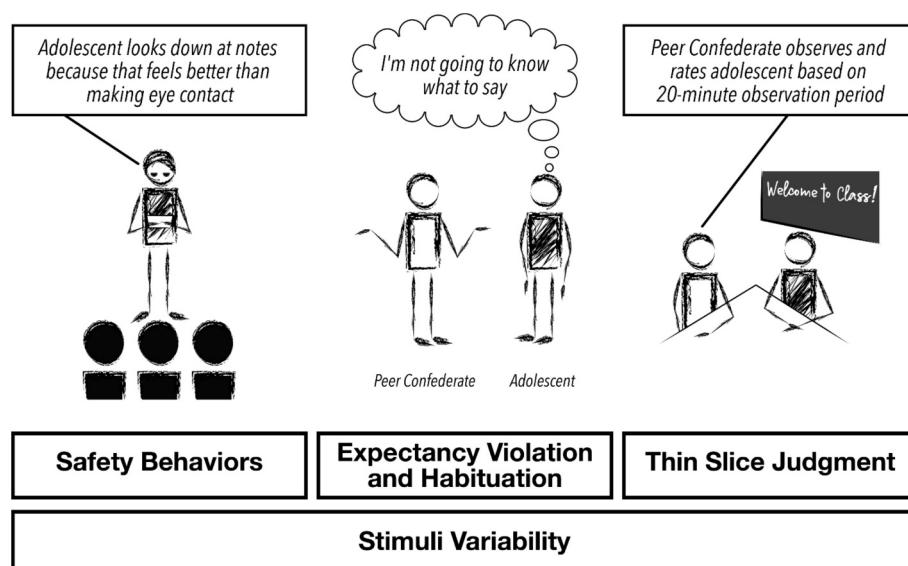


Fig. 3. The conceptual and empirical foundations of the Unfamiliar Peer Paradigm. *Note:* The Unfamiliar Peer Paradigm's origins lie in research and theory in clinical psychology on mechanisms of change in exposure-based therapies, namely work on clients' use of safety behaviors during therapeutic exposures, the need to consider stimuli variability or the degree to which exposures vary on key elements of anxiety-provoking situations, and the importance of expectancy violation and habituation processes in therapy. The paradigm's origins also lie in research and theory in personality and social psychology on thin slice judgments, or the ability to gather data about adolescent social anxiety based on the small “samples” of behavior displayed during tasks within the paradigm. Reproduced from the accepted manuscript version of Cannon et al. (2020).

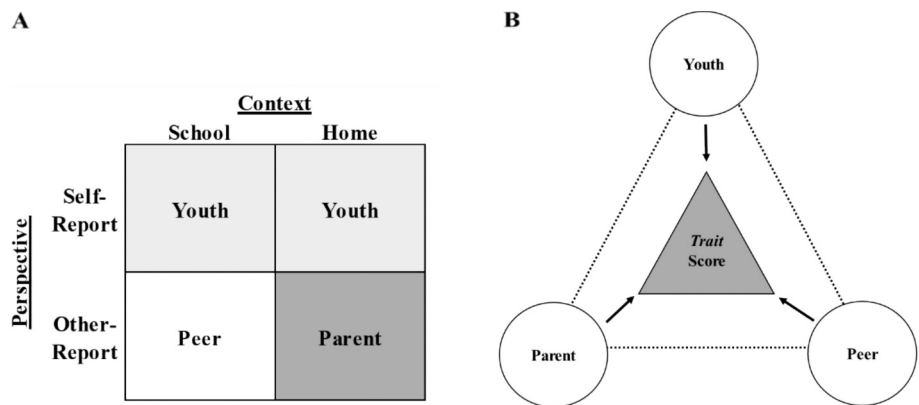


Fig. 4. Panel A depicts an example of the “mix-and-match” criterion to identify optimal informants to include in a multi-informant assessment. Informants systematically vary in the perspective and context from which they rate youth behavior, with the goal of effectively triangulating on a *Trait* score. Panel B provides a graphical depiction of multi-informant reports triangulating, much like the global positioning system, to identify the *Trait* score. Both peer- and parent-reports provide information from an other-perspective, with peers providing information about the school context and parents providing information about the home context. Youth reports provide the self-perspective and information about both the school and home contexts. Reproduced from the accepted manuscript version of Makol et al. (2020).

behavior (e.g., parent at home vs. research personnel outside the home) and the perspectives from which they rate behavior (e.g., parent and research personnel as observer-raters vs. adolescents as self-raters). Using principal components analysis (Dunteman, 1989), the Satellite Model synthesizes informants' reports into three interpretable components. When applied to the reports of informants depicted in Fig. 4, these analyses produce: (a) a component that estimates the social contexts in which adolescents experience social anxiety (i.e., the *context* score), (b) a component that estimates the perspectives from which social anxiety has been observed (i.e., the *perspective* score), and (c) a component that estimates the level of social anxiety rated across informants' contexts and perspectives (i.e., the *trait* score). As illustrated in Table 2, users of the Satellite Model interpret these scores based on how each informant's report loads onto the three components. For instance, a large and positive *trait* score indicates relatively high and stable levels of adolescent social anxiety across informants' contexts and perspectives, as indicated by each informant's report loading onto this component positively and with a loading above 0.40 (see also Kraemer et al., 2003). A large and positive *context* score indicates relatively high levels of adolescent social anxiety that is specific to reports from parents, and vice versa for reports from peers. A large and negative *perspective* score indicates relatively high levels of adolescent social anxiety that is specific to how adolescents perceive themselves, and vice versa for observers of adolescents (i.e., parents and peers). Taken together, scores produced by the Satellite Model facilitate extracting valid data from how informants' reports converge as well as from what makes each of these reports unique.

Consider two ways in which the Satellite Model may facilitate use of both common and unique variance. First, in research about EBSIDs, the

Table 2
Principal components analysis of parent, adolescent, and peer confederate reports about adolescent social anxiety.

Component:	Trait	Context	Perspective
Informant	Component Weight		
Parent	0.73	0.63	0.27
Adolescent	0.81	−0.06	−0.59
Peer Confederate	0.75	−0.55	0.37
Total Variance Explained			
Eigenvalue	1.75	0.70	0.56
Variance attributable to component	58.24%	23.25%	18.52%

Multi-informant data based on parent, adolescent, and peer confederate reports taken using parallel versions of the same survey instrument. Reproduced from the accepted manuscript version of Makol et al. (2025).

trait score allows researchers to test the fit between integration strategies and data conditions like those produced when assessing adolescent social anxiety. Recent work reveals that integrating informants' survey data using the Satellite Model produces *trait* scores that, in tests of criterion-related validity (i.e., when predicting independent observers' ratings of adolescents' behavior), outperform the scores produced by common variance integration strategies (i.e., composite scores, best predictor, Trifactor Model; Makol et al., 2025). When assessing adolescent social anxiety, common variance integration strategies do not pass the EBSID threshold, because they misclassify valid data as measurement confounds; they let valid data go to waste that the Satellite Model otherwise conserves.

Second, recent work supports the construct validity of the Satellite Model scores (Charamut, Racz, Wang, & De Los Reyes, 2022), such that it is valid to interpret the *context* score as an index of mental health concerns as they manifest within specific environments (e.g., home vs. school), and the *perspective* score as an index of concerns that manifest within specific vantage points (e.g., self vs. observer). Along these lines, an innovative direction in future research would involve testing whether the Satellite Model can be used to probe moderators observed in meta-analyses of therapy effects. The mean therapy effect size in the meta-analytic dataset used to produce the effect sizes depicted in Fig. 1 can be characterized as positive and moderate in magnitude, indicating significant, potent effects for youth therapies (Hedges' $g = 0.46$; Weisz et al., 2017). Yet, as mentioned previously, this mean effect size is qualified to a significant extent by a key moderator—the informants completing the outcome measures (i.e., parent vs. youth vs. teacher; see Table 1 of Weisz et al., 2017). The moderating effects are substantial. They range from large-magnitude and positive effects of youth therapies to evidence indicating iatrogenic or harmful effects of these therapies, depending on the informant completing the outcome measure (see Table 2 of Weisz et al., 2017). For example, in this study effect sizes for depression therapies ranged from moderate and negative effects based on teacher reports (Hedges' $g = -0.41$) to positive and small-to-moderate effects based on youth reports (Hedges' $g = 0.32$). Further, effect sizes for family-focused behavioral therapies ranged from near-zero effects based on teacher reports (Hedges' $g = -0.06$) to positive and moderate effects based on parent reports (Hedges' $g = 0.51$). As another example, consider the variability in effect sizes produced by studies that compared active therapies to usual care—these range from small and negative effects for active therapies based on teacher reports (Hedges' $g = -0.11$) to positive and low-to-moderate effects for active therapies based on parent reports (Hedges' $g = 0.32$).

Synthesizing therapy effect sizes using the Satellite Model could

facilitate classifying therapies based on factors that are relevant to how therapists tailor their delivery (e.g., the contexts in which youth experience benefits from therapies). In turn, we expect these classifications to facilitate developing strategies for tailoring mental health services. Specifically, researchers could apply the Satellite Model to meta-analytic effect sizes based on parent, teacher, and youth outcome data, producing a principal components analysis solution akin to what we reported in Table 2. The *trait*, *context*, and *perspective* components facilitate characterizing the effect sizes produced from informants used by a sample of therapy studies, and component scores inform interpreting effect sizes within each therapy study in a sample. These Satellite Model scores would constitute a form of post-hoc probing of informant-based moderator effects that aligns with the rationale for taking a multi-informant approach to estimating therapy effects (i.e., situational specificity). For instance, if effect sizes from studies of a given therapy produced Satellite Model *context* scores that loaded strongly to the home context, then these findings could facilitate detecting youth whose presenting concerns would be a good “fit” for receiving this therapy (e.g., youth whose conduct problems manifest specifically in home contexts). These *context* scores might also inform future work that tests the combined effects of therapies that complement one another (e.g., combining one therapy for which its *context* scores for therapy effects load strongly to the home context with another therapy for which its *context* scores load strongly to the school context). Taken together, the work of Makol et al. (2025) points to a candidate EBSID (i.e., the Satellite Model) that retains estimates of common variance akin to those currently used to estimate therapy effects (*trait* score) and complements these estimates with estimates of unique variance pertinent to understanding individual differences in therapy effects (e.g., the contexts in which youth experience therapeutic benefits).

3.1.1.1. Technology and transparency. Taken together, researchers use validation techniques to identify evidence-based assessments (Becker-Haimes et al., 2020; Hunsley & Mash, 2018). Makol et al. (2025) demonstrate that these techniques also facilitate detecting EBSIDs, by discerning the fit between an integration strategy's assumptions and the data conditions to which it is applied. That these techniques allow for detecting EBSIDs relates to an earlier discussion about contemporary technologies and whether the scores they produce can measure individual differences in therapy research and service delivery.

In psychological assessment, both the origins of validation techniques (Campbell & Fiske, 1959) and observations of discrepant results in youth mental health (Lapouse & Monk, 1958) predate contemporary technologies, and by decades. Yet, only recently have approaches emerged that allow researchers to distinguish valid discrepant results from those that reflect measurement confounds (De Los Reyes, 2024). These approaches are the precursors to detecting EBSIDs—researchers cannot know which data to integrate unless they have an evidence base for what these data mean. Detecting EBSIDs requires an infrastructure: (a) psychometrically sound assessment tools, (b) a literature on the converging versus discrepant results that scores taken from these tools produce, (c) a literature on which of these results reflect domain-relevant information, (d) integration strategies that vary on their assumptions, and (e) study designs that allow for testing integration strategies with varying assumptions. As of this writing, traditional assessment tools have this psychometric infrastructure, and contemporary technologies do not. For the time being, traditional assessment tools are what researchers have available to them to characterize discrepant results and thus use these characterizations to detect EBSIDs. Conducting this same kind of work focused on contemporary technologies like neurobiological and experience sampling technologies will involve building a psychometric infrastructure for scores taken from these technologies. Let us raise three crucial elements on the state of the science surrounding psychological assessments produced by contemporary technologies.

First, we previously noted the lack of a strong evidence base for the reliability of scores produced by contemporary technologies but let us get specific about these estimates. For task-based fMRI, recent meta-analytic work indicates that the mean test-retest reliabilities of the scores they produce (i.e., estimated using intraclass correlation coefficients [ICCs]) are below 0.40 (Elliott et al., 2020). This estimate is well below established thresholds for bare-minimum levels of reliability or what is considered “adequate” (i.e., ICC of 0.70; see Hunsley & Mash, 2018; Youngstrom et al., 2017). Reliability estimates worsen when accounting for a reality about many neurobiological phenomena: Measuring how individuals vary means estimating *differences* between estimates of an individual's neural activity when exposed to two or more tasks (see Hajcak et al., 2017; Luck, 2014). When estimating the reliability of individual differences measures, this signals the need to estimate the reliability not of each instance of neural activity, but rather the reliability of between-task “changes” in neural activity. In a sample of 80 children administered EEG/ERP within a widely used paradigm for assessing neural activity linked to mood and anxiety, test-retest reliabilities using the Pearson *r* metric for change scores reflecting relevant neural components ranged from −0.18 to 0.16 (see Table 2 of Szenczy et al., 2021). These test-retest estimates fall well below bare-minimum thresholds (i.e., $r \geq 0.70$; see Hunsley & Mash, 2018; Youngstrom et al., 2017). Similarly, a large task-based fMRI sample of 139 adolescents revealed internal consistency estimates for relevant contrasts of neural activity of −0.06 (i.e., via Spearman-Brown coefficient; see Table 2 of Infantolino et al., 2018). As individual differences measures, scores taken from neurobiological technologies must gain significant ground in reliability performance to function as psychometrically sound indicators of therapy effects in research and viable tools for making clinical decisions in applied settings.

Second, developing a psychometric infrastructure for scores taken from contemporary technologies requires estimating discrepant results between scores taken from these technologies. Let us consider the state of the science here as well. The literature on discrepant results using traditional assessment tools traces back to the 1950s and has been characterized by decades of meta-analyses about hundreds of studies of youth mental health (De Los Reyes, 2024). In fact, these meta-analyses and the moderating factors they detected inspired Kraemer et al.' (2003) Satellite Model (Fig. 4). If the evolution of a psychometric infrastructure for scores taken from neurobiological and experience sampling technologies were to follow a similar course, then it will require large numbers of studies focused on estimating discrepant results among scores taken from these technologies. Here too, the evidence base is nascent. As of this writing, multi-modal technologies for assessing neurobiology linked to youth mental health are currently best characterized by the Research Domain Criteria initiative (Beauchaine & Hinshaw, 2020; Insel et al., 2010). Both systematic reviews and individual studies seeking to estimate discrepant results among these modalities are, relatively speaking, still in their infancy (e.g., Clarkson et al., 2020; Kaurin et al., 2022; Pasion et al., 2023). As another example, consider a recent systematic review of 50 studies that leveraged ecological momentary assessment of youth mental health (Thunissen, aan het Rot, van den Hoofdakker, & Nauta, 2022). In this review, the authors could only identify 8 studies that incorporated data from parent and youth reports, and none focused on any other informant (e.g., teachers). Taken together, we see a limited literature on discrepant results produced by contemporary technologies. The state of this literature precludes distinguishing which of these results reflect domain-relevant information from those that reflect measurement confounds.

Third, others have noted that as a subdiscipline of Psychology and relative to its other subdisciplines (e.g., Cognitive, Social; see Nosek, Spies, & Motyl, 2012), Clinical Psychology has lagged in confronting issues surrounding scientific transparency (Tackett et al., 2019b; Tackett & Miller, 2019). These issues typically focus on *questionable research practices*—these are actions scholars take when conducting and disseminating research that decrease the ability of research products like

peer-reviewed articles to accurately characterize scientific phenomena (e.g., *p*-hacking; selective reporting of results; Howard et al., 2025). These practices are obviously relevant to research about psychological assessment (see Tackett et al., 2019a). In particular, the psychometric infrastructure limitations we raise here have implications for Clinical Psychology broadly. Scholars in Clinical Psychology are well-aware of the disparities in access to funding for research grounded in traditional assessment tools versus funding for research that incorporates contemporary technologies (see Teachman et al., 2019). Some of this funded work involves researchers using scores taken from contemporary technologies as outcome measures in studies designed to estimate therapy effects (for a review, see Hyde, Bezek, & Michael, 2024). The agencies that funded this work expect the researchers they fund to disseminate their findings: To publish estimates of therapy effects using scores taken from contemporary technologies. These actions necessitate transparency. Researchers must be transparent in conveying the psychometric properties of scores taken from contemporary technologies, particularly when used to estimate therapy effects. Future work on transparency in Clinical Psychology ought to focus on delineating questionable research practices linked to reporting therapy effects based on scores produced by technologies with uncertain psychometric properties. In articles that estimate therapy effects, how often do researchers selectively report the psychometric properties of scores taken from these technologies (e.g., in the Method sections of these articles), to make them appear sounder than the evidence indicates? What open science practices might be implemented to ensure that scores taken from contemporary technologies are not presented as more psychometrically sound than their measured realities? In sum, detecting EBSIDs that incorporate scores taken from contemporary technologies involves investing in an infrastructure for (a) gauging the psychometric soundness of these scores and (b) interpreting what discrepant results among these scores reflect. Absent these investments, we must be transparent about what scores taken from contemporary technologies can and cannot do when reporting findings from therapy studies, and by extension, when using data to tailor therapies for youth in applied settings.

3.1.1.2. Community engagement and representation. Emerging work on EBSIDs has focused on a limited number of mental health domains and periods of youth development, using criterion variables extracted from a limited set of behavioral tasks. Yet, this work may serve as the “origin story” for a larger literature—studies that test the fit between integration strategies and the data conditions to which they are applied. In these respects, a larger literature might emanate from the work of Makol et al. (2020, 2025), and ultimately support the Satellite Model as an EBSID when applied to the data conditions that typify youth mental health assessments. That said, this strategy requires further study. The work of Makol and colleagues focused on use of the Satellite Model when integrating data produced by tools for assessing adolescent social anxiety. Do the integrated data produced by use of the Satellite Model fit data conditions other than those produced when assessing this one domain and within periods of development other than adolescence? Addressing this question requires a large literature of multi-modal approaches to assessing various mental health domains and across periods of youth development. Along these lines, two issues warrant comment.

First, the beginning of this multi-modal literature exists. Consider the wide availability of multi-informant assessment tools across a host of domains and for use in assessing youth across development (see Becker-Haimes et al., 2020; Hunsley & Mash, 2018). Further, in Supplementary Table 1 we list laboratory tasks and naturalistic observation approaches that allow researchers to construct validity criteria along the lines of the tasks depicted in Fig. 3. These approaches allow researchers to construct indices of how youth respond to social contexts relevant to their mental health. These are indices that researchers can use to construct validity criteria to test integration strategies and their fit to data conditions (e.g., via trained observers' ratings), and in ways that are relevant to assessing

youth across various periods of development.

Second, although the beginnings of the literature on multi-modal tasks and approaches needed to test EBSIDs exists, advancing the literature on testing EBSIDs requires much more than what we describe in Supplementary Table 1. Let us reconsider the data in Fig. 1. A few years after learning about the lack of improvements in therapy, scholars probing the same meta-analytic dataset used by Weisz et al. (2019) discovered another key moderator of the effects of youth therapies. Specifically, for youth living in communities with elevated levels of anti-Black racism, the effects of therapies “flatlined” or dropped to few therapeutic benefits for Black youth, but not White youth in these communities (Price et al., 2022). Similarly, psychotherapy for girls was most beneficial for those living in communities with low levels of cultural sexism (Price et al., 2021). Therefore, when youth with marginalized identities receive therapies within structural environments that are culturally affirming, the effects of well-established therapies are likely to be more powerful (i.e., produce medium- to large-magnitude effects; see also Price & Hollinsaid, 2022). What these effects may reveal is the fact that although existing therapies can be tailored to youth and the social contexts that typify day-to-day life, the contexts for these therapies are largely limited to immediate, interpersonal relations (e.g., social interactions with individuals and peers; see also De Los Reyes, 2024). The next generation of therapies must consider adaptations or tailoring among youth who experience broader, community-level contextual factors beyond those that are typically considered within therapies (e.g., structural stigma, systemic discrimination or racism; Price & Hollinsaid, 2022). Grounding this tailoring in evidence will require advancements in assessment technologies—tools for gathering evidence that are sensitive to broader, community-level contextual factors (i.e., contexts beyond those interpersonal contexts described in Supplementary Table 1).

Approaches to assessing structural contexts (environments) may take the form of controlled tasks that reference the laws, policies, and community norms that typify youth and their social environments, and the degree to which these structural contexts affirm or stigmatize youth with specific identities (e.g., see also Anderson, Heard-Garris, & DeLapp, 2022; Fish, 2020; Hatzenbuehler, 2017). They may also take the form of characterizing the structural contexts that youth experience in their natural environment (e.g., ratings of neighborhood factors; see Odgers, Caspi, Bates, Sampson, & Moffitt, 2012). Developing and testing these approaches will require tailoring tools to structural contexts that are relevant to the communities involved in the research. As “proof of concept” for this form of tailoring, consider previous efforts in tailoring the content of survey tools (e.g., Hausman et al., 2013). In this work, researchers took a participatory action approach to seeking feedback from community partners (e.g., parents, youth) on the relevance of item content from traditional assessment tools. They then modified this item content to align with the behaviors and experiences that partners deemed relevant to assessing mental health domains. Along these lines, in future work researchers might apply participatory action approaches to understanding how to infuse assessment of structural contexts into tasks and approaches that have historically only focused on interpersonal contexts. Using these modified tasks as validity indicators will ultimately result in EBSIDs that are sensitive to structural contexts and not just interpersonal contexts. Beyond this specific direction in future research, EBSIDs play a crucial role in several of the research directions described below.

3.1.2. Future direction #2: Innovate tools for assessing contingencies experienced by youth

Let us return to the anxiety example that opened this paper. Pretend that now you are delivering therapy to two children. Whereas one child's anxiety leads them to avoid making new friends with children their own age, the other child's anxiety leads them to avoid group project assignments at school or asking questions in class. How might you deliver services to address their needs? Both children might receive a service

that gives them “practice” with interacting with people, like an exposure-based therapy (e.g., [Sewart & Craske, 2020](#)). However, would these practices look the same? The anxieties these two children display occur in different social contexts. The social contexts where anxiety impacts their lives vary in their contingencies—those “if-then” scenarios that we described earlier. We previously mentioned that social contexts may vary in the contingencies within them. These variations between contexts and their contingencies play key roles in how youth behave, how others respond to their behaviors, and thus how informants report about these behaviors. When delivering therapies, this second direction in future research poses the question, *Can we improve how therapists use data to make decisions, knowing that youth vary in the social contexts related to their needs for services?*

We ground our second direction in future research in contemporary intervention science, namely in the tailoring of therapies for individual youth. Contemporary work about this tailoring pertains to a defining feature of many evidence-based therapies: They were designed as standardized programs with core components that were intended to be delivered in specific ways to all youth who receive them (for a review, see [Albers et al., 2024](#)). This feature of evidence-based therapies necessitates attending to the degree to which therapists deliver its components to youth as intended or with *fidelity*. Yet, a second feature typifies these therapies, namely that they were often initially developed and tested within (a) controlled environments via (b) highly trained and supervised therapists, and (c) within samples of youth who might not reflect those who receive therapies outside of the settings in which they were initially developed and tested (see also [Wiltsey Stirman & Beidas, 2020](#)). These two (seemingly) competing features of evidence-based therapies inspired scholarship designed to guide decisions about tailoring therapies. This scholarship places a particular emphasis on when to instill the *flexibility* needed for tailoring therapies, while maintaining fidelity to their core components (i.e., *flexibility within fidelity*; see [Kendall et al., 2023](#)). For example, [Georgiadis, Peris, and Comer \(2020\)](#) designed their Strategic Flexibility Model to guide flexibly tailoring decisions based on youth characteristics (e.g., cultural backgrounds, comorbid diagnoses), therapy components (e.g., metaphors therapists use to explain core components to youth), timing of tailoring decisions (i.e., when in therapy), setting (e.g., is tailoring warranted based on where therapy is delivered), and the underlying rationale for the tailoring (e.g., to maintain rapport or therapeutic engagement with the youth).

Our second direction in future research deals more with the fidelity of delivering evidence-based therapies rather than flexibility in their delivery. To understand why, let us return to the question posed at the introduction of this future direction. If therapists ground their decisions in evidence, then this practice reveals a third feature of evidence-based therapies—the delivery of knowledge and skills to youth to address their needs for services, and as these needs manifest in real life (i.e., outside of therapy). Recall the two children we described earlier. If these children each receive exposures to address their anxiety, then delivering these exposures with fidelity means using evidence to tailor them to youth and their everyday experiences with anxiety. The child who experiences anxiety when meeting same-age peers receives an exposure tailored to these meetings, whereas the child who has difficulty with social interactions connected to school performance receives an exposure tailored to school-based activities. To do otherwise—failing to use evidence to tailor exposures to real-world challenges—would conflict with a therapy's fidelity. Tailoring exposures reduces risks of a *return of fear*—when therapy gains are lost because the exposures embedded in therapy failed to prepare the youth to confront the social contexts connected to their needs for services (see [Sewart & Craske, 2020](#)).

3.1.2.1. Technology and transparency. The notion of “flexibility within fidelity”—a notion that typifies much contemporary implementation science (e.g., [Albers et al., 2024](#))—requires transparency. That is,

scholars in Clinical Psychology should be transparent about the fact that, for at least some therapeutic components, delivery with fidelity means *evidence-based tailoring*. Along these lines, let us consider the key tailoring decisions linked to ensuring that the therapies youth receive reflect real life—the “practice activities” therapists create when working with youth. For these activities to produce meaningful changes for youth receiving therapy, they must simulate the contingencies that youth experience in day-to-day life. No two youth should be assumed to experience the exact same contingencies. Thus, therapists construct practice activities with an assumption—they must tailor these activities to each youth and their individual contingencies. Yet, as mentioned previously, decisions about tailoring therapy tend to be based on intuition and experience (see [Venturo-Conerly et al., 2023](#)). These decisions have immense implications for therapeutic progress, particularly if they fail to consider the views of youth, their parents or caregivers, or other partners in service delivery (see also [De Los Reyes et al., 2022a](#); [Fitzpatrick, Rusch, Chiffer, & Weisz, 2025](#)). Would it not improve how therapists make decisions, if they had tools that produce data to complement their intuition and experience? These tools would produce data to inform decisions about tailoring practice activities to a youth's unique contingencies.

Decisions regarding tailoring services are not unique to anxiety. Mental health is intimately tied to the contexts and contingencies that typify day-to-day life. The oppositional behavior exhibited by a child might make it hard for them to get along with their parents. The inattention a child experiences might interfere with completing their homework. The depression a child feels might decrease their motivation to pursue hobbies or make friends. The tasks described in Supplementary Table 1 reflect some of these contexts and contingencies. That said, we also require clinically feasible tools that reflect the needs of youth and partners in service delivery, to ensure that their views inform therapeutic activities (see also [Fitzpatrick et al., 2025](#)). If decades of research support the ability of researchers and therapists to solicit reports from informants like parents, teachers, and youth to measure youth mental health, then might they also rely on these same informants to measure contingencies linked to youth mental health?

Let us consider the possibility that informants can capably provide psychometrically sound reports about contingencies linked to service delivery. How might researchers design such a tool? If a tool for assessing contingencies might facilitate tailoring service delivery, then it logically follows that such a tool ought to be tailored to measure contingencies for the specific needs a youth has for services, as well as the active ingredients of the therapy (i.e., *therapeutic goals*). Further, this tool must be amenable to completion by structurally different informants (e.g., teachers), and not just the typical participants in youth services (i.e., parents and youth). Recently, our team built a tailored tool for collecting data about contingencies that informants perceive as relevant to addressing a youth's mental health needs and the goals of the therapies they receive. The tool tailors item content to the informants completing them, using computer-adapted technologies now commonplace in educational testing (see [AERA, 2014](#)). The Kids' Behavior in Context Scales (KICS; [Fine et al., 2025](#); [Herman et al., 2025](#)), consists of item content that is calibrated so that informants can rate any youth, regardless of their current need for mental health services. Informants use the KICS to report both qualitative data about a youth's contingencies, as well as quantitative data about how stable across social contexts their contingencies might be (e.g., multiple contingencies at school vs. relatively few at home). On the KICS, the contingencies that informants rate are tailored to the domains for which they believe the youth they are rating would benefit from services. For instance, an informant who identifies anxiety and managing emotions as areas for improvement in the youth's functioning receives a tailored set of items to rate about contingencies relevant to anxiety (e.g., “When a situation doesn't go as planned”) and managing emotions (e.g., “When excited about doing something”).

3.1.2.2. Community engagement and representation. The discrepant results produced by informants' KICS reports reflect the contextual stability in a youth's contingencies. That said, our team designed the KICS to detect a particular set of contingencies relevant to delivering services to youth, namely those that occur in interpersonal social contexts. A crucial next step in this area of work involves designing tools like the KICS that are tailored to contingencies relevant to service delivery for youth and as they manifest in other kinds of social contexts.

When considering these other contexts, let us return to the structural (environmental) contexts we highlighted previously in our discussion of EBSIDs. When detecting EBSIDs, we previously discussed the value of developing laboratory-controlled tasks and approaches that simulate social contexts relevant to understanding youth mental health (see Fig. 3; Supplementary Table 1). We also raised the prospect of leveraging participatory action approaches to research that facilitate tailoring the content of these tasks so that they accurately reflect structural contexts relevant to those living in communities where researchers recruit their participants. Just as has been demonstrated with tailoring the item content of surveys (see Hausman et al., 2013), research on the KICS provides proof-of-concept support for the idea that informants can rate tailored sets of item content in psychometrically sound ways. For instance, in a sample of 173 youth, parent, and teacher ratings on the KICS, internal consistency estimates of informants' 10-item ratings of contingencies ranged from 0.81 to 0.93, and their ratings significantly related to scores taken from well-established measures of youth psychosocial functioning (Fine et al., 2025; Herman et al., 2025). As such, the KICS, which already functions as a tailored assessment tool, may be well-positioned for further tailoring. For example, applying a participatory action approach to research might reveal new approaches to developing sets of contingencies rated on the KICS that are tailored not just to specific domains, but also to the specific contingencies that parents and youth find particularly relevant to day-to-day life in their local community. Further, the contingencies contained in these "community tailored" KICS items can reference contingencies produced by interpersonal interactions, but also signs of contingencies produced by the structural environment (e.g., closure of a local park in the community, implementation of exclusionary discipline at a local school that disproportionately affects Black youth). The KICS represents an exciting starting point for future research that leverages participatory action approaches to developing and tailoring assessment tools.

3.1.3. Future direction #3: innovate laboratory tasks that are tailored to social contexts

We previously described validation studies focused on detecting EBSIDs. This research involves designing laboratory-controlled tasks that simulate social contexts germane to understanding youth mental health. Arguably, the value of this work hinges on evidence that the data produced by these tasks accurately simulate the social contexts that youth experience outside of the laboratory. Along these lines, consider a key observation made by Makol et al. (2020, 2025) when testing levels of correspondence between multi-informant data integrated with the Satellite Model as depicted in Fig. 4 and behavioral ratings from independent observers who viewed adolescents' behavior in the tasks depicted in Fig. 3. The *trait* score produced by the Satellite Model corresponds (i.e., using the coefficient beta [β] metric) with behavioral data about anxiety at magnitudes as high as 0.70. To properly frame these effects, they are in the range of the effect size conventions in the social sciences that have historically been considered large in magnitude (see Cohen, 1988). In contrast, decades of work across assessments of youth and adults have revealed a stark reality about survey data, namely that scores taken from any one survey tend to correlate at low magnitudes with data from behavioral tasks (Pearson r s in the .10s and .20s; see De Los Reyes, 2024; De Los Reyes, et al., 2019a; Dunning, Heath, & Suls, 2004; Hendriks, Van der Giessen, Stams, & Overbeek, 2018; Meyer et al., 2001).

Imagine a world that invested more resources into validation studies.

Investments could focus on improving the ability of laboratory tasks to accurately simulate social contexts. In turn, researchers could calibrate informants' reports on clinically feasible tools like surveys, based on how well the EBSIDs used to integrate their reports correspond to how youth behave in simulated social contexts. Imagine the possibilities produced with this calibration process. This process would allow survey developers to sensitively, and strategically, select informants (and perhaps even the item content of surveys) based on the reports they produce and the connections between these reports and how youth behave when exposed to simulated social contexts. EBSIDs that are calibrated to how youth behave within these simulations may help researchers significantly improve upon the effects observed by Makol et al. (2020, 2025). With the right investments, this future is within reach.

Given the magnitudes of correspondence between EBSIDs and behavior observed by Makol et al. (2020, 2025), you might ask, *How might additional validation studies improve upon these magnitudes of correspondence?* Indeed, it is reasonable to ask if these magnitudes—which go as high as a β of 0.70—represent a ceiling to the connections that researchers can make between EBSIDs and validity criteria. To give you a sense of the potential for future research to improve upon these effects, let us reconsider the tasks depicted in Fig. 3. These are tasks embedded in a laboratory paradigm that was first described in a paper published in 2020. Yet, the origins of these tasks trace back decades. The tasks in Fig. 3 are developmental adaptations of tasks developed in the 1990s (Turner, Beidel, Cooley, Woody, & Messer, 1994), 1980s (Beidel, Turner, Jacob, & Cooley, 1989; Curran, 1982), and 1970s (Richardson & Tasto, 1976). These tasks have drawbacks that rival those of the survey and interview tools most often used to estimate therapy effects (Fig. 1). They are simply outdated technologies and are long overdue for improvement.

3.1.3.1. Technology and transparency. Using technologies available as of this writing, let us consider what additional investments in advancing the tasks in Fig. 3 might entail. These tasks reflect the "practice" elements of therapeutic exposures, or what a social interaction might look like without the aid of a therapist. They simulate social environments experienced by youth, such as those social contexts that elicit fear and avoidance behaviors among youth receiving exposure-based therapies. The laboratory tasks depicted in Fig. 3 include multiple social contexts, because youth vary in the social contexts where they experience social anxiety. Yet, these tasks are standardized—all youth are exposed to the same social contexts. What if we could tailor these tasks for each individual youth's social contexts?

A few of this paper's authors have an affinity for science fiction and perhaps you might as well. Like us, you might have seen a 2018 film directed by Steven Spielberg, *Ready Player One* (De Line, Krieger, Spielberg, & Farah, 2018). The film takes place in the year 2045 and much of the story centers around its central character's immersion in a virtual world called *OASIS* (Ontologically Anthropocentric Sensory Immersive Simulation). Immersion in this world takes place within a virtual reality device in which the user not only wears a virtual reality headset but is also tethered to a platform that rests on a treadmill governed by gyro sensors. These sensors allow users to move in a 360-degree environment, creating the potential for complete immersion not just of a user's visual system, but also of their sensorimotor system. We wrote this paper two decades before the fictional world of *Ready Player One*. Yet, as of this writing the immersive virtual reality technology we just described is commercially available (<https://vrituix.com/omni-one/>).

Coupled with the generative artificial intelligence platforms now available for constructing video media (e.g., <https://sora.com/>), technology exists to make tailored versions of the tasks depicted in Fig. 3. For example, generative artificial intelligence platforms allow users to produce detailed media based on short, text-based prompts (Torous et al., 2025). Using these platforms, researchers could gather some information about a youth study participant's social contexts (e.g.,

neighborhood, school, living room), and build video-based scenes to simulate these contexts. Consider a youth study participant for whom research personnel generate social contexts that reflect tailored simulations of the contexts the youth experiences in day-to-day life. By immersing that youth in social contexts that are uniquely salient to their own experiences, researchers could leverage the power of laboratory-controlled procedures for manipulating contingencies of the same social contexts that contribute to the youth's functioning (e.g., interactions with peers vs. adults, interactions with individuals vs. groups, interactions at school vs. parties). Researchers could then calibrate EBSIDs so that they reference behavioral reactions to these simulations. Further, generative artificial intelligence platforms keep records of the text-based prompts used to construct stimuli, making all this work amenable to posting on the Open Science Framework in ways akin to prior recommendations for posting other forms of materials relevant to psychological assessment (Tackett et al., 2019). We expect this work to optimize the ability of EBSIDs to inform decision-making in applied settings.

In sum, if funding bodies were to invest in validation studies that leverage just those technologies available as of this writing, then we expect the social scientists of the future to require new conventions for what constitutes a small-magnitude effect. For decades, we have thought of a small-magnitude correlation between variables as quite low, a Pearson r of 0.10, for instance (Cohen, 1988). We as scholars have learned to live with this as our “floor” in terms of the explanatory power of our work, and as mentioned previously, this has long been the standard set by observed correlations between survey data and data from behavioral tasks. What if new technologies were to allow us to reach a new floor of explanatory power? What if this new floor were to become what we currently consider a large-magnitude effect (i.e., Pearson r of 0.50)? Addressing this question requires significantly greater investments. We would say the opportune time to invest is now, except that this time came and went decades ago. Boosting investments in validation research now will give scientists a chance to make up for lost time.

3.1.3.2. Community engagement and representation. A theme described in this first set of future directions may be the key to optimizing the value of validation research. Specifically, the next generation of research must embrace the use of participatory action and community engagement approaches to infuse input from such partners as youth, families, teachers, and therapists. Use of these approaches will facilitate the calibration process we described previously, so that the actions of researchers align with the social contexts and experiences that are salient to those who participate in research and receive mental health services. When considering recent advances in technology, this tailoring process takes on new meaning. Imagine a version of the immersive virtual reality approach we just described, for which the prompts used to create simulated social contexts are co-constructed between researchers and participants in these simulations (e.g., youth, parents). How much more accurate and relevant will these situations become when co-constructed between researchers and participants? The technology exists for building these tailored social contexts in real-time, and we are overdue to capitalize on the potential for innovations in psychological assessment that these technologies afford.

3.2. Future directions: implementation studies

The first three research directions focused on validation studies. The validation studies we described inform implementation studies, namely studies that focus on infusing assessment tools and EBSIDs into service delivery. But what characteristics distinguish implementation studies from validation studies? In contrast to validation studies, implementation studies sacrifice the certainty in assessments observed in validation studies to optimize the feasibility of using assessment tools in applied settings (De Los Reyes, 2024). For instance, implementation studies are

less likely than validation studies to include modalities beyond surveys (although see De Los Reyes et al., 2022b). In this sense, validation studies leverage time-intensive assessment tools like behavioral tasks to optimize what we can learn from scores taken from clinically feasible tools like surveys. In turn, these validation studies produce knowledge that implementation studies can capitalize on when devising strategies for infusing assessment tools in settings that typify service delivery (e.g., schools, hospitals, clinics). That said, the future directions we describe below share two features with the validation studies we previously described. First, these future directions in implementation research include studies that researchers can pursue with technologies that exist as of this writing. Second, as with validation studies, implementation studies must involve close collaboration with partners in service delivery—youth, their families, teachers, therapists, and administrators, for example—to ground assessment tools and EBSIDs in social environments that are relevant to youth mental health.

3.2.1. Future direction #4: implement assessment tools and EBSIDs in universal screening

We previously described directions in future research about EBSIDs—evidence-based strategies for integrating data. These research directions involved testing EBSIDs within controlled laboratory settings. One candidate EBSID, the Satellite Model, has been tested in a setting focused on improving evidence-based assessments of adolescent social anxiety. Along these lines, let us return to our youth anxiety therapy example. When delivering a therapy to address a youth's anxiety, what if you had access to data that, from the beginning of therapy, informed you about where and under what conditions the youth experiences their concerns? What if you could acquire these data from informants and in less than two minutes' time? An emerging line of implementation research applies an innovative lens to *universal screening* (Romer et al., 2020)—the use of brief, free, and accessible tools to identify youth who experience mental health domains and their associated features. In terms of feasibility, a particularly exciting portion of this work involves screening for *psychosocial impairments*—evidence that how youth think, feel, or behave interferes with their ability to function in everyday social contexts. This evidence includes grade performance and making friends at school, and relationships with family at home (Fabiano & Pelham, 2016).

As with mental health domains such as anxiety, depression, and hyperactivity, the psychosocial impairments connected to these domains manifest within social contexts. They also manifest across diverse contexts that include decreases in school performance, the presence of family dysfunction at home, the inability to meet demands at home (e.g., chores, homework), and impaired relationships with peers at school (e.g., initiating and maintaining peer relationships). In line with situational specificity and what we know about discrepant results when assessing mental health (e.g., Achenbach et al., 1987; De Los Reyes, 2024), universal screening should involve assessing psychosocial impairments based on reports from multiple informants. Yet, historically tools for assessing impairments have been limited in their psychometric rigor (e.g., single-item measures), item content (e.g., items that can only be administered to diagnosed youth; items for which responses could also be explained by physical health), and informants (e.g., single-source raters such as therapists; see Blake, Cangelosi, Johnson-Brooks, & Belcher, 2007; De Los Reyes, et al., 2022b; Gadow, Kaat, & Lecavalier, 2013; Rapee, Bögels, van der Sluis, Craske, & Ollendick, 2012; Tabachnick & Fidell, 2001). These are important limitations in research about assessing impairments. These impairments dictate detecting who ought to receive mental health services and whose mental health reflects diagnosable conditions (American Psychiatric Association, 2013; Goldstein & Naglieri, 2016). For universal screening to meaningfully inform decision-making about mental health services, we must address limitations in not only the assessment of psychosocial impairments, but also approaches to integrating data about these impairments. Along these lines, what if universal screenings could incorporate EBSIDs in ways that

facilitate context-sensitive decision-making?

3.2.1.1. Technology and transparency. Recent developments in the literature on assessing psychosocial impairments pave the way for multi-informant approaches to universal screening and incorporating EBSIDs in the screening process. Consider the Work and Social Adjustment Scale for Youth (WSASY). This 5-item, open access tool measures psychosocial impairments and is available on the Open Science Framework (<https://osf.io/xfqym/>). An innovative feature of the WSASY is that it measures any youth's impairments, regardless of their current need for mental health services. That is, its items assess these impairments without mention of mental health concerns or status (e.g., "Because of the ways I think, feel or behave, my ability to do well in school is impaired."). This ensures that the WSASY can be used to rate youth impairments regardless of the presence of diagnosable mental health concerns. This also allows the WSASY to inform the delivery of prevention-focused services, not just therapies to treat youth who already have mental health diagnoses. Severity of impairment is indicated using a Likert scale from "0" to "8" with higher scores indicating greater levels of impairment. WSASY reports from teachers, caregivers, and youth demonstrate high levels of internal consistency ($\alpha > 0.80$). The criterion-related validity of scores taken from the WSASY has been demonstrated via significant links to scores taken from measures of diverse internalizing and externalizing domains, independent measures of impairment domains (e.g., peer functioning, social skills, caregivers' functioning, family conflict), and criterion measures of diverse modalities (e.g., surveys, home and school observations, structured peer interactions; Charamut et al., 2022; De Los Reyes et al., 2022b; De Los Reyes et al., 2019b; Rapee et al., 2024). Further, recent work indicates that the WSASY can be feasibly administered in such widely varied settings as schools and inpatient hospitals (e.g., Rapee et al., 2024; Thomas et al., 2024; Vaillant-Coindard, Briet, Lespiau, Gisclard, & Charbonnier, 2024). As such, the WSASY appears poised for implementation in a variety of screening contexts.

Yet, psychometrically sound measures like the WSASY must also facilitate decision-making about screening. As such, future research must involve testing whether the WSASY (a) produces universal screening data that can be integrated using an EBSID like the Satellite Model and (b) can be implemented within universal screening settings. Recent work supports this key direction in future research. Specifically, our team implemented the WSASY to collect parent, teacher, and youth data in a universal screening conducted among a set of middle schools in Minnesota (De Los Reyes et al., 2022b). Further, in this study we integrated these informants' reports using the Satellite Model, and in a way that facilitated detecting links between these reports and data from tasks designed to index behavior in both the school context (i.e., naturalistic behavioral observations) and home context (i.e., laboratory-controlled discussions between parents and youth). This study provides key proof-of-concept support for the WSASY's implementation in universal screenings and use of an EBSID to integrate WSASY scores.

In sum, the Satellite Model appears to facilitate integrating WSASY data within universal screening settings. Beyond implementation research, we also require additional validation research focused on attaining certainty that users of the WSASY can interpret the scores it produces precisely and accurately across diverse samples of youth not only in terms of clinical status (e.g., screening vs. outpatient care vs. inpatient care) but also in terms of cultural backgrounds, identities, and the social environments in which they live. An additional consideration is that the WSASY is but one tool for assessing psychosocial functioning among youth. Future work might involve testing the ability of data produced by other open access tools to be not only feasibly administered in universal screening settings (e.g., schools, primary care), but also integrated using EBSIDs like the Satellite Model, and prior work provides resources for identifying these kinds of tools (e.g., Becker-Haimes et al., 2020).

3.2.1.2. Community engagement and representation. Importantly, the infusion of participatory action approaches into this future direction may provide researchers with a unique opportunity to scrutinize long-held assumptions about the breadth, depth, and content of universal screening procedures. Current best practices for implementing universal screening call for (a) a finite list of short screening measures completed by (b) a limited number of informants or data sources and (c) a completion time of just a few minutes (see Romer et al., 2020). However, consider if researchers took a participatory action approach to understanding what the communities they serve want to learn from screening. In recent work, Dryburgh et al. (2025) developed a tool for assessing the organizational context and its readiness for implementing school-based mental health programming. The authors followed an approach that involved tailoring the item content of their tool to the needs of the organization as well as the domains assessed on the tool. If such an approach were also applied to universal screening, then the result might be an increased motivation among research participants to complete screening batteries. What if this increased motivation also meant that screening batteries could expand beyond current estimates? Consider if it were a norm of scholarly work for researchers to invest time in learning from prospective participants what they would want to achieve from screening, and incorporated participants' preferences in the screening process. Could such an approach boost the ability of researchers to develop batteries that extend their aims, scope, and content beyond what are seen as the current limits of batteries, which are arguably constructed largely based on researcher-driven approaches? Testing competing approaches to developing and administering universal screenings (e.g., researcher-driven screening batteries vs. participatory action-driven screening batteries) represents an exciting direction in future research in this key area of implementation research.

3.2.2. Future direction #5: implement assessment tools and EBSIDs in measurement-based care

We opened this paper with evidence indicating that estimates of the effects of youth psychotherapies have not improved in decades (Fig. 1). Throughout this paper, we have highlighted directions in future research that involve developing new assessment tools and testing EBSIDs that optimize the accuracy of data produced by assessment tools. These directions in future research hinge on the ability of scholars to improve upon long-held assessment practices (Table 1), using approaches that can be implemented in applied settings. As such, the directions in future research we propose include lines of research focused on infusing assessment tools into decision-making throughout service delivery, not just at specific points of decision-making (e.g., universal screening). Specifically, scholars must devote time to improving technologies for measurement-based care, which we previously described as an area of implementation science devoted to infusing data into decision-making surrounding service delivery. To facilitate describing directions in future research about measurement-based care, we will continue with our previous examples of strategies for implementing the WSASY and using EBSIDs to integrate the multi-informant data produced by its use.

Our previous example focused on implementing the WSASY and Satellite Model into universal screening procedures, with the goal of detecting the specific contexts in which youth experience psychosocial impairments. Extending these efforts into measurement-based care will require new validation studies. This is because measurement-based care requires repeated administrations of assessment tools and within a finite period such as a course of therapy (e.g., before every weekly session of an intervention program). Tools that can be implemented in this way must harbor a series of properties beyond their brevity and low cost. In the case of the WSASY, these properties include the ability to detect changes in psychosocial impairments, as well as the specific social contexts in which these impairments manifest (e.g., home, school). This calls for studies that include both validation and implementation techniques. For example, studies that probe use of the WSASY and Satellite

Model when administered session-by-session to estimate changes in a youth's functioning in therapy, and in ways that are sensitive to change within specific social contexts (e.g., home vs. school).

When we first described the Satellite Model (Fig. 4), we focused specifically on its value to research about validation and interpreting the outcomes of meta-analyses of therapy effects. Yet, the Satellite Model also provides scores relevant to implementation. When using the Satellite Model to integrate WSASY data, researchers generate scores to characterize the multi-informant impairment data for all youth in the sample. Previously, we described a hypothetical application of the Satellite Model with meta-analytic data, in which the effect sizes produced by multi-informant outcome measures in each therapy study could be synthesized to reveal variability in the contexts and perspectives of therapy effects. Similarly, when the Satellite Model is applied to characterize multi-informant data in a single sample, each youth in that sample receives the set of three integrated Satellite Model scores described previously (i.e., *trait*, *context*, *perspective*). The *context* score characterizes youth functioning in terms of where concerns with functioning manifest (e.g., home vs. school). If there were a way to characterize session-by-session change in this score, then it would move the WSASY and the Satellite Model one step closer to advancing technologies in measurement-based care, namely by implementing tools for tailoring therapy activities session-by-session and to specific contexts. For instance, implementing such a procedure within measurement-based care would allow for the repeated use of context-sensitive data to inform the selection of role-play or exposure activities within therapeutic sessions. For instance, a *context* score at session 4 indicating greater impairments at home versus school might guide a therapist to plan home-based activities for that session. In contrast, a *context* score at session 8 indicating greater impairments at school versus home might guide a therapist to shift in focus and plan school-based activities for that session.

3.2.3. Technology and transparency

The work we just described comprises an important first step in implementation research. Yet it is insufficient to advance technologies in measurement-based care. This is because measurement-based care requires advancements in not only the psychometric properties of data

produced by assessment tools, but also advancements in how the data produced by these tools is represented or utilized by those involved in service delivery (De Los Reyes, 2024; McLeod et al., 2022). When tools are integrated into measurement-based care, it is perhaps prudent to assume that partners in service delivery—therapists, clinic administrators, youth and their parents, for example—will not have advanced training in score interpretation and data-analysis. To implement a tool like the WSASY or an integration strategy like the Satellite Model in measurement-based care requires new approaches to using evidence when delivering services. It requires approaches akin to how marketing executives approach *product placement*—marketing a product in recognizable ways but in non-advertising settings (Gurevitch, 2009).

Consider Fig. 5, which provides an example of one way to represent the *context* score produced when using the Satellite Model to integrate WSASY data from parent, youth, and teacher reports. Rather than present the *context* score numerically, researchers could develop an icon for representing this score. The icon could reference an aspect of day-to-day life that many people could understand or interpret, even without training in data science. Along these lines, as of this writing researchers estimate that 96% of the world is now covered by mobile broadband networks, and nearly 60% has access to a mobile device connected to these networks (Shanahan & Bahia, 2024). Consequently, most of the world gets regular “practice” interfacing with icons used to index network connectivity (e.g., “signal bars” on a mobile phone). By representing the *context* score as bars of network connectivity, researchers could provide feedback to partners in service delivery that articulate the specific contexts in which youth experience psychosocial impairments, and in a way that could document changes in these contexts session-by-session. The proof-of-concept for these possibilities lies in a now extensive literature of using icons to give therapists feedback about their clients' functioning (Lambert, 2007; Lambert et al., 2003). In this work, feedback about clients' functioning is communicated with specific colors, corresponding to the signals on a traffic light (e.g., red = at risk of deterioration in functioning; yellow = attend to functioning as progress could be improved; green = client's functioning is progressing as expected). Meta-analytic reviews support use of this approach to feedback. Specifically, relative to assessment as usual in which therapists only get access to data in chart form, representing these same data using icons



Fig. 5. Graphical depiction of a measurement-based care approach to representing *context* score data derived from a Satellite Model used to integrate reports from parent, youth, and teacher about the youth's psychosocial impairments (i.e., using the Work and Social Adjustment Scale for Youth; De Los Reyes et al., 2019; De Los Reyes et al., 2022b). In this representation, the icon used to represent reception on a mobile phone is placed above icons representing home and school (i.e., greater darkened bars = greater phone reception). For youth with a *context* score that indicates greater impairments at home relative to school, the mobile phone icon would be calibrated to indicate greater reception at home than at school (i.e., number of darkened bars for home icon > number of darkened bars for school icon). For youth with a *context* score that indicates greater impairments at school relative to home, the mobile phone icon would be calibrated to indicate greater reception at school than at home (i.e., number of darkened bars for school icon > number of darkened bars for home icon). Images produced using the DALL-E generative artificial intelligence platform.

like stoplights reduces rates of client deterioration (e.g., de Jong et al., 2021; Shimokawa, Lambert, & Smart, 2010), as well as disparities among more versus less effective therapists as to client outcomes (see Delgadillo et al., 2022).

3.2.3.1. Community engagement and representation. Leveraging product placement techniques that typify marketing represent exciting new paths for the next generation of implementation research. Yet, these lines of work must account for the reality that no one product placement strategy will be applicable to all partners in service delivery. Stated another way, representing data using symbols derived from mobile technology (e.g., bars of network connectivity) or transportation (e.g., stoplights) have a finite scalability or reach, and not all participants in measurement-based care will understand data represented in this way. The participatory action examples we described earlier facilitated the development of tailored approaches to developing standardized measures (e.g., Hausman et al., 2013). We can think of this approach as akin to an *ideographic* or individualized approach to culturally informed assessment, and these ideographic assessment approaches have the potential to boost how people engage with mental health services (see also Sanchez et al., 2022). Consistent with prior work on measurement-based care, future work that infuses product placement techniques into data representations ought to involve working with partners in service delivery to identify a suite of symbols and strategies that serve as the basis of tailored approaches to data representation, amenable to use with youth from a variety of backgrounds, experiences, cultures, and values.

3.2.4. Future direction #6: implement EBSIDs in service delivery settings using Monte Carlo simulations as guides

The thesis of this paper is that researchers must modify the balance of their efforts in studying technologies for therapies versus technologies for assessment tools. No one paper can be expected to rebalance scientific efforts. Thus, our final direction in future research focuses on use of technologies that might be implemented without having to collect any new data.

For centuries, researchers have leveraged Monte Carlo simulations to study phenomena that were infeasible to study with real data (Harrison, 2010; Raychaudhuri, 2008). We expect Monte Carlo simulations to set the stage for making monumental progress on implementation. To understand why, let us reconsider research described previously. The data conditions that typify youth mental health assessments are thoroughly well-characterized. In this sense, the lack of progress in technologies for assessment tools has facilitated these characterizations. As mentioned previously, the methodologies that underlie assessment tools (i.e., surveys and interviews) have remained relatively stable over decades, and that has allowed researchers to document remarkably stable data conditions for youth mental health research. This stability paves the way for lines of future research that leverage Monte Carlo simulation techniques.

3.2.4.1. Technology and transparency. The data conditions described previously facilitate producing datasets that simulate how data “behave” in real life. In fact, Makol et al. (2025) recently leveraged these data conditions to construct Monte Carlo simulations that tested the criterion-related validity performance of the Satellite Model and Trifactor Model. These simulations complement the “real world” validation tests described previously. These validation tests focused on a single sample of youth assessed on one mental health domain (social anxiety), and as that domain manifests within a single developmental period (adolescence). It would be unreasonable to attest that findings from such a study would generalize to all other data conditions. It would also be unreasonable to expect researchers to conduct studies that replicate and extend the time-intensive validation work of Makol and colleagues. As such, Monte Carlo simulations involve constructing a series of simulated datasets to represent the variability among data conditions that typifies a scientific literature.

Based on the deep literature that has characterized assessment of youth mental health and the converging and discrepant results produced within this literature (De Los Reyes, 2024), Makol et al. (2025) constructed 48 Monte Carlo scenarios (i.e., with 1000 simulated datasets per scenario) to characterize a range of data conditions. These scenarios included what would be considered (a) modal or *base case* scenarios (e.g., informants disagree a fair amount, their reports each contain a fair amount of measurement error, and they each correlate to some degree with validity criteria), (b) *best case* scenarios (e.g., informants disagree a great deal, their reports each contain little measurement error, and they each correlate well with validity criteria), and (c) *unlikely case* scenarios (e.g., informants agree a great deal, their reports each contain little measurement error, and they each correlate well with validity criteria). Although Makol and colleagues varied these scenarios considerably, not one of the 48 scenarios resulted in the Trifactor Model outperforming the Satellite Model; even the scenario in which the data conditions fit the Trifactor Model's assumptions (i.e., unlikely case). The Satellite Model produced more favorable estimates of criterion-related validity relative to the Trifactor Model; the only variation was in the magnitudes of performance differences between these strategies.

The Monte Carlo simulation findings we just described have important implications for other technological applications relevant to integration strategies. For instance, consider the implications of these findings for implementing generative artificial intelligence platforms focused on how researchers implement integration strategies. Both the first and last directions in future research described in this paper illustrated how applications of well-established methodological and statistical techniques facilitate detecting those data conditions that violate the user assumptions of integration strategies, as well as estimating the consequences of these violations (Makol et al., 2020, 2025). These violations are relatively easy to detect. They reveal themselves through observed relations among predictors and validity criteria. For example, when predictors consist of multi-informant surveys and the criterion variable consists of independent observers' ratings of performance during behavioral tasks, these patterns consist of (a) cross-informant correlations among survey reports and (b) correlations between each of these survey reports and independent observers' ratings. Stated another way, when researchers take a multi-informant, multi-modal approach to psychological assessment, they tend to report these relations in their published work (e.g., correlation matrices reported as preliminary analyses). These are the exact kinds of data that can be probed by generative artificial intelligence platforms to understand the degree to which researchers use integration strategies with data that violate the assumptions underlying use of the strategies. A key direction in future research lies at the intersection of technology and scientific transparency, namely on use of technology to reduce biased decision-making in how researchers select integration strategies.

Recent work highlights the potential for applications of various forms of artificial intelligence (e.g., machine learning, generative artificial intelligence; Dwyer, Falkai, & Koutsouleris, 2018; Torous et al., 2025) to perpetuate biases in a host of areas of mental health research and service delivery, including racial and ethnic disparities in diagnostic decision-making errors and systemic inequalities in access to mental health services (see Timmons et al., 2023). Of course, these concerns are well-founded. As of this writing, forms of artificial intelligence consist of algorithms that are constructed or supervised by humans, or function generatively in unsupervised learning environments. These algorithms are grounded in work built by humans. This means that any biases present in human decision-making have the potential to be factored into the algorithms built with machine learning, or text-based prompts built within generative artificial intelligence platforms. Arguably, if these biases were “scaled up” to levels of biased decision-making made possible by artificial intelligence, then they could exacerbate the consequences of biases. That is, biases could occur at scales that far exceed the biased decision-making made possible by even large groups of humans and over prolonged periods.

Notwithstanding the risks of some artificial intelligence algorithms, in other cases algorithms might facilitate improving scientific transparency in mental health research. Consider one implication of Makol et al. (2020, 2025). In recent work, we have considered the misuse of statistical techniques in mental health research, namely when researchers implement these techniques within data conditions that violate their assumptions (De Los Reyes et al., 2025). In this work, we advocate for framing this misuse as a questionable research practice, akin to *p*-hacking. Misuse of statistical techniques might occur for the same reasons as other questionable research practices. Use of a sophisticated statistical technique might make the findings of a study appear more innovative and thus publishable than would be the case if the study involved use of less sophisticated techniques. As such, a technique's sophistication might increase the likelihood of its use, even when the data conditions do not support its use. In this sense, a questionable research practice like misuse of a statistical technique instills a bias in research findings just as any other methodological confound (e.g., drawing causal inferences from the findings of studies built with nonexperimental designs; Kazdin, 2024). Logically, this misuse compels researchers to draw inferences from their work that do not conform to reality.

In this paper, we described integration strategies whose assumptions vary in whether they violate the data conditions that typify youth mental health assessments. If these violations can be readily detected in prior work, then future research can leverage generative artificial intelligence platforms to detect them. These platforms might be used to generate automated, regular probes of published work for these violations, using well-established data science approaches (i.e., web scraping; see S. Hu et al., 2024; Landers, Brusso, Cavanaugh, & Collmus, 2016). Beyond detecting these violations, these platforms could include features that guard against future misuse. For instance, researchers could build into the platform an automated feature that would send a correspondence (e.g., via email) to the author(s) of the detected work, letting them know of the violation, along with a list of recommended EBSIDs for which its assumptions align with the data conditions used in the detected work. If such an automated platform were to become a regular element of the science transparency infrastructure, it might contribute to a research culture that prioritizes not the sophistication of integration strategies (i.e., for their ability to facilitate publication), but rather their fit to the data conditions to which they are applied.

3.2.4.2. Community engagement and representation. The Monte Carlo simulations produced by Makol et al. (2025) focused on validation. Future implementation research requires thinking of simulations focused on use of assessment tools in service delivery. Consider that when delivering services, we have long known that therapists hold beliefs about the informants they consider “best” for assessing youth mental health (e.g., Loeber, Green, & Lahey, 1990). More recent experimental work indicates that, like researchers, therapists who deliver services to youth appear to integrate data using strategies that fail to capitalize on the valid data that discrepant results might contain (Marsh, Zeveney, & De Los Reyes, 2020). What if Monte Carlo simulations focused on implementation science had the capacity to motivate therapists to not only value the perspectives of multiple informants' reports, but also to access tools that facilitate making decisions with diverse data sources? These Monte Carlo simulations might estimate how an implementation-ready version of the Satellite Model would perform in service delivery. By “perform” we refer to a version of the Satellite Model that could facilitate decision-making based on multi-informant data and for such service-related decisions as making diagnoses, planning therapies, and monitoring therapeutic responses. How would such a model perform in terms of facilitating accurate decision-making, relative to how therapists tend to use data from informants to guide their decision-making? How would the Satellite Model perform when compared to decision-making based on data from one

informant, two informants, or several informants whose data meaningfully differ (e.g., parent or teacher as “best informant” for assessing hyperactivity)? These questions merit further inquiry. If the Satellite Model outperforms the usual ways in which therapists leverage data to make decisions about service delivery, then it justifies investing resources to develop versions of the Satellite Model that can be feasibly implemented in service delivery. As with the directions in future research we previously described, researchers might boost the impact of Monte Carlo simulations by working in partnership with communities to develop the questions that inform use of these simulations and the interpretations of the results with regards to their social environments and experiences.

4. Concluding comments

Scholars in youth mental health find themselves in a moment of reckoning. Investments in time, thinking, and resources to improve therapies have not produced evidence of a “payoff” in improved therapeutic effects. If these therapies were engines, then the evidence is telling us that the engines researchers built and tested in the 1970s perform about as well as the ones that researchers build and test today. We are skeptical that these investments have truly been made in vain. Perhaps you are skeptical as well. Yet, we are talking about “evidence-based therapies.” This moment should prompt us to rethink how we gather and integrate evidence. This moment should compel us to confront long-standing assessment practices (Table 1) and consider the possibility that a lack of investments in research about assessments allowed these practices to outlive their utility. We must invest in research on assessment tools and integration strategies that improve our ability to accurately characterize therapy effects, and facilitate the use of evidence when tailoring therapies to individual youth and their unique needs.

We opened this paper by highlighting that therapies and the tools used to estimate their effects are complementary variables. If your investments focus too much on improving the quality of one of these variables, then that means you have fewer resources available to invest in improving the quality of the other variable. Today's approaches to gathering and integrating evidence look a lot like those used decades ago (Table 1). If we continue taking these approaches to evidence, then we will find ourselves ill-equipped to test whether the next generation of investments made to improve therapies pay off.

This paper presented a new approach to investing in research about psychological assessment. We described directions in future research to improve technologies for assessment tools, develop strategies for integrating the data they produce, and test approaches to implementing these tools and strategies in applied settings. Importantly, simply rebalancing investments between evidence and therapies is not enough. To ensure success, researchers and the communities they study must work in close partnership. Thus, we highlighted ways to infuse participatory action approaches that set up for success the directions in future research that we described. If we empower the assessment researchers of today to improve how we approach evidence, then we equip the therapy researchers of tomorrow to demonstrate that investing in better engines pays off.

Declaration of competing interest

The authors declare that there are no competing interests to report.

Acknowledgements

The first author thanks Christine Cha, Anastasia Dimitropoulos, Alan E. Kazdin, James C. Overholser, and Wendy K. Silverman for a series of insightful conversations in October 2023 and January 2024 that inspired this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.cpr.2026.102705>.

Data availability

No data was used for the research described in the article.

References

- Achenbach, T. M. (2020). Bottom-up and top-down paradigms for psychopathology: A half century odyssey. *Annual Review of Clinical Psychology*, 16, 1–24. <https://doi.org/10.1146/annurev-clinpsy-071119-115831>
- Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, 101(2), 213–232. <https://doi.org/10.1037/0033-2909.101.2.213>
- Albers, B., Verweij, L., Blum, K., Oesch, S., Schultes, M. T., Clack, L., & Naef, R. (2024). Firm, yet flexible: A fidelity debate paper with two case examples. *Implementation Science*, 19, Article 79. <https://doi.org/10.1186/s13012-024-01406-3>
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-V)* (5th ed.). Author.
- Anderson, R. E., Heard-Garris, N., & DeLapp, R. C. (2022). Future directions for vaccinating children against the American endemic: Treating racism as a virus. *Journal of Clinical Child and Adolescent Psychology*, 51(1), 127–142. <https://doi.org/10.1080/15374416.2021.1969940>
- Bauer, D. J., Howard, A. L., Baldasaro, R. E., Curran, P. J., Hussong, A. M., Chassin, L., & Zucker, R. A. (2013). A trifactor model for integrating ratings across multiple informants. *Psychological Methods*, 18(4), 475–493. <https://doi.org/10.1037/a0032475>
- Beauchaine, T. P., & Hinshaw, S. P. (2020). RDoC and psychopathology among youth: Misplaced assumptions and an agenda for future research. *Journal of Clinical Child and Adolescent Psychology*, 49(3), 322–340. <https://doi.org/10.1080/15374416.2020.1750022>
- Becker-Haimes, E. M., Tabachnick, A. R., Last, B. S., Stewart, R. E., Hasan-Granier, A., & Beidas, R. S. (2020). Evidence base update for brief, free, and accessible youth mental health measures. *Journal of Clinical Child and Adolescent Psychology*, 49(1), 1–17. <https://doi.org/10.1080/15374416.2019.1689824>
- Beg, M. J., Verma, M., & Verma, M. K. (2024). Artificial intelligence for psychotherapy: A review of the current state and future directions. *Indian Journal of Psychological Medicine*. <https://doi.org/10.1177/02537176241260819>, 02537176241260819.
- Beidas, R. S., Stewart, R. E., Walsh, L., Lucas, S., Downey, M. M., Jackson, K., Fernandez, T., & Mandell, D. S. (2015). Free, brief, and validated: Standardized instruments for low-resource mental health settings. *Cognitive and Behavioral Practice*, 22(1), 5–19. <https://doi.org/10.1016/j.cbpra.2014.02.002>
- Beidel, D. C., Turner, S. M., Jacob, R. G., & Cooley, M. R. (1989). Assessment of social phobia: Reliability of an impromptu speech task. *Journal of Anxiety Disorders*, 3(3), 149–158. [https://doi.org/10.1016/0887-6185\(89\)90009-1](https://doi.org/10.1016/0887-6185(89)90009-1)
- Bettis, A. H., Vaughn-Coaxum, R. A., Lawrence, H. R., Hamilton, J. L., Fox, K. R., Augsberger, A., & REACH Youth Advisory Board. (2024). Key challenges and potential strategies for engaging youth with lived experience in clinical science. *Journal of Clinical Child and Adolescent Psychology*, 53(5), 733–746. <https://doi.org/10.1080/15374416.2023.2264389>
- Blake, K., Cangelosi, S., Johnson-Brooks, S., & Belcher, H. M. (2007). Reliability of the GAF and CGAS with children exposed to trauma. *Child Abuse and Neglect*, 31(8), 909–915. <https://doi.org/10.1016/j.chiabu.2006.12.017>
- Blanco, D., Roberts, R. M., Gannoni, A., & Cook, S. (2024). Assessment and treatment of mental health conditions in children and adolescents: A systematic scoping review of how virtual reality environments have been used. *Clinical Child Psychology and Psychiatry*, 29(3), 1070–1086. <https://doi.org/10.1177/13591045231204082>
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(1), 81–105. <https://doi.org/10.1037/h0046016>
- Cannon, C. J., Makol, B. A., Keeley, L. M., Qasmieh, N., Okuno, H., Racz, S. J., & De Los Reyes, A. (2020). A paradigm for understanding adolescent social anxiety with unfamiliar peers: Conceptual foundations and directions for future research. *Clinical Child and Family Psychology Review*, 23(3), 338–364. <https://doi.org/10.1007/s10567-020-00314-4>
- Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology*, 52, 685–716. <https://doi.org/10.1146/annurev.psych.52.1.685>
- Charamut, N. R., Racz, S. J., Wang, M., & De Los Reyes, A. (2022). Integrating multi-informant reports of youth mental health: A construct validation test of Kraemer and colleagues' (2003) satellite model. *Frontiers in Psychology*, 13, Article 911629. <https://doi.org/10.3389/fpsyg.2022.911629>
- Chen, T., Ou, J., Li, G., & Luo, H. (2024). Promoting mental health in children and adolescents through digital technology: A systematic review and meta-analysis. *Frontiers in Psychology*, 15, Article 1356554. <https://doi.org/10.3389/fpsyg.2024.1356554>
- Clarkson, T., Kang, E., Capriola-Hall, N., Lerner, M. D., Jarcho, J., & Prinstein, M. J. (2020). Meta-analysis of the RDoC social processing domain across units of analysis in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, 49(3), 297–321. <https://doi.org/10.1080/15374416.2019.1678167>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.
- Cortese, S., Purper-Ouakil, D., Apter, A., Arango, C., Baeza, I., Banaschewski, T., ... Moreno, C. (2024). Psychopharmacology in children and adolescents: Unmet needs and opportunities. *The Lancet Psychiatry*, 11(2), 143–154. [https://doi.org/10.1016/S2215-0366\(23\)00345-0](https://doi.org/10.1016/S2215-0366(23)00345-0)
- COVID-19 Working Group. (2024). Successes and lessons learned in responding to the needs of pediatricians, children, and families during the COVID-19 pandemic. *Pediatrics*, 153(6), Article e2024066634. <https://doi.org/10.1542/peds.2024-066634>
- Curran, J. P. (1982). A procedure for the assessment of social skills: The simulated social interaction test. In J. P. Curran, & P. M. Monti (Eds.), *Social skills training: A practical handbook for assessment and treatment* (pp. 348–373). New York, NY: Guilford.
- Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth*. Free Press.
- De Line, D., Krieger, K. M., Spielberg, S., & Farah, D. (2018). *Ready player one [Motion picture]*. Warner Brothers.
- De Los Reyes, A. (2024). *Discrepant results in mental health research: What they mean, why they matter, and how they inform scientific practices*. Oxford.
- De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin*, 141(4), 858–900. <https://doi.org/10.1037/a0038498>
- De Los Reyes, A., Cook, C. R., Sullivan, M., Morrell, N., Hamlin, C., Wang, M., ... Qasmieh, N. (2022). The work and social adjustment scale for Youth: Psychometric properties of the teacher version and evidence of contextual variability in psychosocial impairments. *Psychological Assessment*, 34(8), 777–790. <https://doi.org/10.1037/pas0001139>
- De Los Reyes, A., & Epkins, C. C. (2023). Introduction to the special issue. A dozen years of demonstrating that informant discrepancies are more than measurement error: Toward guidelines for integrating data from multi-informant assessments of youth mental health. *Journal of Clinical Child and Adolescent Psychology*, 52(1), 1–18. <https://doi.org/10.1080/15374416.2022.2158843>
- De Los Reyes, A., Lerner, M. D., Keeley, L. M., Weber, R., Drabick, D. A. G., Rabinowitz, J., & Goodman, K. L. (2019). Improving interpretability of subjective assessments about psychological phenomena: A review and cross-cultural meta-analysis. *Review of General Psychology*, 23(3), 293–319. <https://doi.org/10.1177/108926801983764>
- De Los Reyes, A., Makol, B. A., Racz, S. J., Youngstrom, E. A., Lerner, M. D., & Keeley, L. M. (2019). The work and social adjustment scale for Youth: A measure for assessing youth psychosocial impairment regardless of mental health status. *Journal of Child and Family Studies*, 28(1), 1–16. <https://doi.org/10.1007/s10826-018-1238-6>
- De Los Reyes, A., Oswald, F. L., Racz, S. J., Piña, A., McLeod, B. D., Wang, M., & Charamut, N. (2025). Editorial: Fairness, validity, and transparency in what researchers assume when testing for measurement invariance. *Journal of Clinical Child and Adolescent Psychology*, 54(3), 299–317. <https://doi.org/10.1080/15374416.2025.2484813>
- De Los Reyes, A., Talbott, E., Power, T., Michel, J., Cook, C. R., Racz, S. J., & Fitzpatrick, O. (2022). The needs-to-goals gap: How informant discrepancies in youth mental health assessments impact service delivery. *Clinical Psychology Review*, 92, Article 102114. <https://doi.org/10.1016/j.cpr.2021.102114>
- De Los Reyes, A., Thomas, S. A., Goodman, K. L., & Kunder, S. M. A. (2013). Principles underlying the use of multiple informants' reports. *Annual Review of Clinical Psychology*, 9, 123–149. <https://doi.org/10.1146/annurev-clinpsy-050212-185617>
- De Los Reyes, A., Wang, M., Lerner, M. D., Makol, B. A., Fitzpatrick, O., & Weisz, J. R. (2023). The operations triad model and youth mental health assessments: Catalyzing a paradigm shift in measurement validation. *Journal of Clinical Child and Adolescent Psychology*, 52(1), 19–54. <https://doi.org/10.1080/15374416.2022.2111684>
- Delgado, J., Deisenhofer, A.-K., Probst, T., Shimokawa, K., Lambert, M. J., & Kleinstäuber, M. (2022). Progress feedback narrows the gap between more and less effective therapists: A therapist effects meta-analysis of clinical trials. *Journal of Consulting and Clinical Psychology*, 90(7), 559–567. <https://doi.org/10.1037/ccp0000747>
- Deros, D. E., Racz, S. J., Lipton, M. F., Augenstein, T. M., Karp, J. N., Keeley, L. M., & De Los Reyes, A. (2018). Multi-informant assessments of adolescent social anxiety: Adding clarity by leveraging reports from unfamiliar peer confederates. *Behavior Therapy*, 49(1), 84–98. <https://doi.org/10.1016/j.beth.2017.05.001>
- Dryburgh, N. S., Wang, L., Repchuck, R., Matte, A. R., Runions, K., & Georgiades, K. (2025). *Psychometric evidence for the school organizational conditions for mental health programming measure: Assessing the organizational context for implementing evidence-informed programming in Ontario schools*. School Mental Health. <https://doi.org/10.1007/s12310-025-09742-5>. Advance online publication.
- Dunlap, G., & Kern, L. (2018). Perspectives on (functional) behavioral assessment. *Behavioral Disorders*, 43(2), 316–321. <https://doi.org/10.1177/0198742917746633>
- Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, 5(3), 69–106. <https://doi.org/10.1111/j.1529-1006.2004.00018.x>
- Duntman, G. H. (1989). *Principal components analysis* (No. 69). Sage.

- Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, 14(1), 91–118. <https://doi.org/10.1146/annurev-clinpsy-032816-045037>
- Eid, M., Geiser, C., & Koch, T. (2024). *Structural equation modeling of multiple rater data*. Guilford.
- Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods*, 13(3), 230–253. <https://doi.org/10.1037/a0013219>
- Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., & Hariri, A. R. (2020). What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. *Psychological Science*, 31(7), 792–806. <https://doi.org/10.1177/0956797620916786>
- Fabiano, G. A., & Pelham, W. E. (2016). Impairment in children. In S. Goldstein, & J. A. Naglieri (Eds.), *Assessing impairment* (pp. 105–119). NY: Springer.
- Fine, E., Herman, T., Makol, B. M., Wang, M., Talbott, E., McLeod, B., ... De Los Reyes, A. (2025). The kids' behavior in context scales (KICS) I. Teacher, caregiver, and youth views on needs and goals for intervention. *Behavioral Disorders*. <https://doi.org/10.1177/01987429251371076>
- Fish, J. N. (2020). Future directions in understanding and addressing mental health among LGBTQ youth. *Journal of Clinical Child and Adolescent Psychology*, 49(6), 943–956. <https://doi.org/10.1080/15374416.2020.1815207>
- Fiske, D. W., & Campbell, D. T. (1992). Citations do not solve problems. *Psychological Bulletin*, 112(3), 393–395. <https://doi.org/10.1037/0033-2909.112.3.393>
- Fitzpatrick, O. M., Rusch, T., Chiffer, M., & Weisz, J. R. (2025). Alignment between clinician treatment choices and client data as a predictor of youth clinical outcomes. *Journal of Clinical Child and Adolescent Psychology*. <https://doi.org/10.1080/15374416.2025.2484802>
- Gadow, K. D., Kaat, A. J., & Lecavalier, L. (2013). Relation of symptom-induced impairment with other illness parameters in clinic-referred youth. *Journal of Child Psychology and Psychiatry*, 54(11), 1198–1207. <https://doi.org/10.1111/jcpp.12077>
- Gallop, L., Westwood, S. J., Hemmings, A., Lewis, Y., Campbell, I. C., & Schmidt, U. (2024). Effects of repetitive transcranial magnetic stimulation in children and young people with psychiatric disorders: A systematic review. *European Child and Adolescent Psychiatry*. <https://doi.org/10.1007/s00787-024-02475-x>
- Gensler, G., Johnson, S., Panizza, U., & di Mauro, B. W. (2025). *The economic consequences of the second trump administration: A preliminary assessment*. Centre for Economic Policy Research.
- Georgiadis, C., Peris, T. S., & Comer, J. S. (2020). Implementing strategic flexibility in the delivery of youth mental health care: A tailoring framework for thoughtful clinical practice. *Evidence-Based Practice in Child and Adolescent Mental Health*, 5(3), 215–232. <https://doi.org/10.1080/23794925.2020.1796550>
- Giusto, A., Triplett, N. S., Foster, J. C., & Gee, D. G. (2024). Future directions for community-engaged research in clinical psychological science with youth. *Journal of Clinical Child and Adolescent Psychology*, 53(3), 503–522. <https://doi.org/10.1080/15374416.2024.2359650>
- Glenn, L. E., Keeley, L. M., Szollos, S., Okuno, H., Wang, X., Rausch, E., ... De Los Reyes, A. (2019). Trained observers' ratings of adolescents' social anxiety and social skills within controlled, cross-contextual social interactions with unfamiliar peer confederates. *Journal of Psychopathology and Behavioral Assessment*, 41(1), 1–15. <https://doi.org/10.1007/s10862-018-9676-4>
- Goldstein, S., & Naglieri, J. A. (2016). *Assessing impairment: From theory to practice*. Springer.
- Groth-Marnat, G., & Wright, A. J. (2016). *Handbook of psychological assessment* (6th ed.). Wiley.
- Gurevitch, L. (2009). Problematic dichotomies: Narrative and spectacle in advertising and media scholarship. *Popular Narrative Media*, 2(2), 143–159. <https://doi.org/10.3828/pnm.2009.3>
- Haefl, G. J., Jeronimus, B. F., Kaiser, B. N., Weaver, L. J., Soyter, P. D., Fisher, A. J., ... Lu, W. (2022). Folk classification and factor rotations: Whales, sharks, and the problems with the hierarchical taxonomy of psychopathology (HiTOP). *Clinical Psychological Science*, 10(2), 259–278. <https://doi.org/10.1177/21677026211002500>
- Hajcak, G., Meyer, A., & Kotov, R. (2017). Psychometrics and the neuroscience of individual differences: Internal consistency limits between-subjects effects. *Journal of Abnormal Psychology*, 126(6), 823–834. <https://doi.org/10.1037/abn0000274>
- Harrison, R. L. (2010). Introduction to Monte Carlo simulation. In , Vol. 1204. *AIP Conference Proceedings* (pp. 17–21). <https://doi.org/10.1063/1.3295638>
- Hatzenbuehler, M. L. (2017). Advancing research on structural stigma and sexual orientation disparities in mental health among youth. *Journal of Clinical Child and Adolescent Psychology*, 46(3), 463–475. <https://doi.org/10.1080/15374416.2016.1247360>
- Hausman, A. J., Baker, C. N., Komaroff, E., Thomas, N., Guerra, T., Hohl, B. C., & Leff, S. S. (2013). Developing measures of community-relevant outcomes for violence prevention programs: A community-based participatory research approach to measurement. *American Journal of Community Psychology*, 52(3), 249–262. <https://doi.org/10.1007/s10464-013-9590-6>
- Heisenberg, W. (1927). Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik [the actual content of quantum theoretical kinematics and mechanics]. *Zeitschrift für Physik*, 43(3–4), 172–198. <https://ntrs.nasa.gov/citation/19840008978>
- Hendriks, A. M., Van der Giessen, D., Stams, G. J. J. M., & Overbeek, G. (2018). The association between parent-reported and observed parenting: A multi-level meta-analysis. *Psychological Assessment*, 30(5), 621–633. <https://doi.org/10.1037/pas0000500>
- Herman, T., Fine, E., Makol, B. M., Wang, M., Talbott, E., McLeod, B., ... De Los Reyes, A. (2025). The kids' behavior in context scales (KICS) II. An instrument for detecting contexts relevant to youth interventions. *Behavioral Disorders*. <https://doi.org/10.1177/01987429251371074>
- Howard, A. L., Neal, T. M., Kirtley, O. J., Urry, H. L., & Tackett, J. L. (2025). Open science at clinical psychological science: Reflections on progress, lessons learned, and suggestions for continued improvement. *Clinical Psychological Science*, 13(1), 195–204. <https://doi.org/10.1177/21677026241255882>
- Howe, G. W., Dagne, G. A., Brown, C. H., Brincks, A. M., Beardslee, W., Perrino, T., & Pantin, H. (2019). Evaluating construct equivalence of youth depression measures across multiple measures and multiple studies. *Psychological Assessment*, 31(9), 1154–1167. <https://doi.org/10.1037/pas0000737>
- Hu, S., Liu, J., Cornacchi, S. D., Klassen, A. F., Pusic, A. L., & Kaur, M. N. (2024). Extracting big data from the internet to support the development of a new patient-reported outcome measure for breast implant illness: A proof of concept study. *Quality of Life Research*, 33(7), 1975–1983. <https://doi.org/10.1007/s1136-024-03672-6>
- Hu, X., Sgherza, T. R., Nothrup, J. B., Fresco, D. M., Naragon-Gainey, K., & Bylsma, L. M. (2024). From lab to life: Evaluating the reliability and validity of psychophysiological data from wearable devices in laboratory and ambulatory settings. *Behavior Research Methods*, 56(7), 1–20. <https://doi.org/10.3758/s13428-024-02387-3>
- Hunsley, J., & Mash, E. J. (Eds.). (2018). *A guide to assessments that work* (2nd ed.) Oxford.
- Hyde, L. W., Bezek, J. L., & Michael, C. (2024). The future of neuroscience in developmental psychopathology. *Development and Psychopathology*, 36(5), 2149–2164. <https://doi.org/10.1017/S0954579424000233>
- Infantolino, Z. P., Luking, K. R., Sauder, C. L., Curtin, J. J., & Hajcak, G. (2018). Robust is not necessarily reliable: From within-subjects fMRI contrasts to between-subjects comparisons. *NeuroImage*, 173, 146–152. <https://doi.org/10.1016/j.neuroimage.2018.02.024>
- Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., ... Wang, P. (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, 167(7), 748–751. <https://doi.org/10.1176/appi.ajp.2010.09091379>
- de Jong, K., Conijn, J. M., Gallagher, R. A. V., Reshetnikova, A. S., Heij, M., & Lutz, M. C. (2021). Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: A multilevel meta-analysis. *Clinical Psychology Review*, 85, article 102002. <https://doi.org/10.1016/j.cpr.2021.102002>
- Kaurin, A., Sequeira, S. L., Ladouceur, C. D., McKone, K. M. P., Rosen, D., Jones, N., ... Silk, J. S. (2022). Modeling sensitivity to social threat in adolescent girls: A psychoneurometric approach. *Journal of Psychopathology and Clinical Science*, 131(6), 641–652. <https://doi.org/10.1037/abn0000532>
- Kazdin, A. E. (2024). Randomized controlled trials: Characteristics, options, and challenges. In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (5th ed., pp. 281–308). American Psychological Association. <https://doi.org/10.1037/0000409-013>
- Kendall, P. C., Ney, J. S., Maxwell, C. A., Lehrbach, K. R., Jakubovic, R. J., McKnight, D. S., & Friedman, A. L. (2023). Adapting CBT for youth anxiety: Flexibility, within fidelity, in different settings. *Frontiers in Psychiatry*, 14, Article 1067047. <https://doi.org/10.3389/fpsy.2023.1067047>
- Kiukas, J., Lahti, P., Pellonpää, J. P., & Ylinen, K. (2019). Complementary observables in quantum mechanics. *Foundations of Physics*, 49(6), 506–531. <https://doi.org/10.1007/s10701-019-00261-3>
- Kleiman, E. M., Glenn, C. R., & Liu, R. T. (2019). Real-time monitoring of suicide risk among adolescents: Potential barriers, possible solutions, and future directions. *Journal of Clinical Child and Adolescent Psychology*, 48(6), 934–946. <https://doi.org/10.1080/15374416.2019.1666400>
- Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *The American Journal of Psychiatry*, 160(9), 1566–1577. <https://doi.org/10.1176/appi.ajp.160.9.1566>
- Lambert, M. (2007). Presidential address: What we have learned from a decade of research aimed at improving psychotherapy outcome in routine care. *Psychotherapy Research*, 17(1), 1–14. <https://doi.org/10.1080/10503300601032506>
- Lambert, M. J., Whipple, J. L., Hawkins, E. J., Vermeersch, D. A., Nielsen, S. L., & Smart, D. W. (2003). Is it time for clinicians to routinely track patient outcome? A meta-analysis. *Clinical Psychology: Science and Practice*, 10(3), 288–301. <https://doi.org/10.1093/clipsy.bpg025>
- Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the internet for use in psychological research. *Psychological Methods*, 21(4), 475–492. <https://doi.org/10.1037/met0000081>
- Lapouse, R., & Monk, M. A. (1958). An epidemiologic study of behavior characteristics in children. *American Journal of Public Health*, 48(9), 1134–1144. <https://doi.org/10.2105/AJPH.48.9.1134>
- Leifer, A. M., Liu, A. J., & Nagel, S. R. (2025). US researchers must stand up to protect freedoms, not just funding. *Nature*, 641, 592–593. <https://doi.org/10.1038/d41586-025-01466-5>
- Loeber, R., Green, S. M., & Lahey, B. B. (1990). Mental health professionals' perception of the utility of children, mothers, and teachers as informants on childhood psychopathology. *Journal of Clinical Child Psychology*, 19(2), 136–143. https://doi.org/10.1207/s15374424jccp1902_5
- Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT Press.

- Makol, B. A., Yang, J., Wang, M., & De Los Reyes, A. (2025). *Youth development in context: Integrating multiple informants to assess behavior*. Springer.
- Makol, B. A., Youngstrom, E. A., Racz, S. J., Qasimieh, N., Glenn, L. E., & De Los Reyes, A. (2020). Integrating multiple informants' reports: How conceptual and measurement models may address long-standing problems in clinical decision-making. *Clinical Psychological Science*, 8(6), 953–970. <https://doi.org/10.1177/2167702620924439>
- Marsh, J. K., Zeveney, A. S., & De Los Reyes, A. (2020). Informant discrepancies in judgments about change during mental health treatments. *Clinical Psychological Science*, 8(2), 318–332. <https://doi.org/10.1177/2167702619894905>
- McLeod, B. D., Jensen-Doss, A., Lyon, A. R., Douglas, S., & Beidas, R. S. (2022). To utility and beyond! Specifying and advancing the utility of measurement-based care for youth. *Journal of Clinical Child and Adolescent Psychology*, 51(4), 375–388. <https://doi.org/10.1080/15374416.2022.2042698>
- Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, 56(2), 128–165. <https://doi.org/10.1037/0003-066X.56.2.128>
- Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102(2), 246–268. <https://doi.org/10.1037/0033-295X.102.2.246>
- Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6), 615–631. <https://doi.org/10.1177/1745691612459058>
- Ogders, C. L., Caspi, A., Bates, C. J., Sampson, R. J., & Moffitt, T. E. (2012). Systematic social observation of children's neighborhoods using Google street view: A reliable and cost-effective method. *Journal of Child Psychology and Psychiatry*, 53(10), 1009–1017. <https://doi.org/10.1111/j.1469-7610.2012.02565.x>
- Pasion, R., Macedo, I., Paiva, T. O., Patrick, C. J., Krueger, R. F., & Barbosa, F. (2023). Neurobehavioral mechanisms of comorbidity in internalizing and externalizing psychopathology: An RDoC multimethod assessment. *Journal of Psychopathology and Behavioral Assessment*, 45(3), 793–808. <https://doi.org/10.1007/s10862-023-10073-5>
- Price, M. A., & Hollinsaid, N. L. (2022). Future directions in mental health treatment with stigmatized youth. *Journal of Clinical Child and Adolescent Psychology*, 51(5), 810–825. <https://doi.org/10.1080/15374416.2022.2109652>
- Price, M. A., McKetta, S., Weisz, J. R., Ford, J. V., Lattanner, M. R., Skov, H., Wolock, E., & Hatzenbuehler, M. L. (2021). Cultural sexism moderates efficacy of psychotherapy: Results from a spatial meta-analysis. *Clinical Psychology: Science and Practice*, 28(3), 299–312. <https://doi.org/10.1037/cps0000031>
- Price, M. A., Weisz, J. R., McKetta, S., Hollinsaid, N. L., Lattanner, M. R., Reid, A. E., & Hatzenbuehler, M. L. (2022). Meta-analysis: Are psychotherapies less effective for black youth in communities with higher levels of anti-black racism? *Journal of the American Academy of Child and Adolescent Psychiatry*, 61(6), 754–763. <https://doi.org/10.1016/j.jaac.2021.07.808>
- Rapee, R. M., Bögels, S. M., van der Sluis, C. M., Craske, M. G., & Ollendick, T. H. (2012). Annual research review: Conceptualising functional impairment in children and adolescence. *Journal of Child Psychology and Psychiatry*, 53(5), 454–468. <https://doi.org/10.1111/j.1469-7610.2011.02479.x>
- Rapee, R. M., Kuhnert, R., Spence, S. H., Bowsher, I., Burns, J., Coen, J., & Wuthrich, V. (2024). The brief evaluation of adolescents and children online (BEACON): Psychometric development of a mental health screening measure for school students. *Journal of Clinical Psychology*, 80(6), 1420–1447. <https://doi.org/10.1002/jclp.23673>
- Rapee, R. M., Kuhnert, R., Spence, S. H., Bowsher, I., Burns, J., Coen, J., ... Wuthrich, V. (2024). The Brief Evaluation of Adolescents and Children Online (BEACON): Psychometric development of a mental health screening measure for school students. *Journal of Clinical Psychology*, 80(6), 1420–1447. <https://doi.org/10.1002/jclp.23673>
- Raychaudhuri, S. (2008). Introduction to Monte Carlo simulation. In *2008 winter simulation conference* (pp. 91–100). IEEE. <https://doi.org/10.1109/WSC.2008.4736059>
- Richardson, F. C., & Tasto, D. L. (1976). Development and factor analysis of a social anxiety inventory. *Behavior Therapy*, 7(4), 453–462. [https://doi.org/10.1016/S0005-7894\(76\)80164-5](https://doi.org/10.1016/S0005-7894(76)80164-5)
- Richters, J. E. (1992). Depressed mothers as informants about their children: A critical review of the evidence for distortion. *Psychological Bulletin*, 112(3), 485–499. <https://doi.org/10.1037/0033-2909.112.3.485>
- Romer, N., von der Embse, N., Eklund, K., Kilgus, S., Perales, K., Splett, J. W., Sudlo, S., & Wheeler, D. (2020). Best practices in social, emotional, and behavioral screening: An implementation guide. Version 2.0. <https://smhcollaborative.org/resources/#>
- Sanchez, A. L., Jent, J., Aggarwal, N. K., Chavira, D., Cox, S., Garcia, D., La Roche, M., & Comer, J. S. (2022). Person-centered cultural assessment can improve child mental health service engagement and outcomes. *Journal of Clinical Child and Adolescent Psychology*, 51(1), 1–22. <https://doi.org/10.1080/15374416.2021.1981340>
- Sewart, A. R., & Craske, M. G. (2020). Inhibitory learning. In J. S. Abramowitz, & S. M. Blakey (Eds.), *Clinical handbook of fear and anxiety: Maintenance processes and treatment mechanisms* (pp. 265–285). American Psychological Association.
- Shanahan, M., & Bahia, K. (2024). The state of mobile internet connectivity 2024 [Technical report]. *GSMA Intelligence*.
- Shimokawa, K., Lambert, M. J., & Smart, D. W. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology*, 78(3), 298–311. <https://doi.org/10.1037/a0019247>
- Southam-Gerow, M. A., & Prinstein, M. J. (2014). Evidence base updates: The evolution of the evaluation of psychological treatments for children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, 43(1), 1–6. <https://doi.org/10.1080/15374416.2013.855128>
- Steen, S. (2025). Science is at a crossroads—And physiology with it: A statement from the APS chief executive officer. *Function*, 6(3), Article zqaf025. <https://doi.org/10.1093/function/zqaf025>
- Stone, A. A., Schneider, S., & Smyth, J. M. (2023). Evaluation of pressing issues in ecological momentary assessment. *Annual Review of Clinical Psychology*, 19(1), 107–131. <https://doi.org/10.1146/annurev-clinpsy-080921-083128>
- Szenczy, A. K., Levinson, A. R., Hajcak, G., Bernard, K., & Nelson, B. D. (2021). Reliability of reward- and error-related brain activity in early childhood. *Developmental Psychobiology*, 63(6), Article e22175. <https://doi.org/10.1002/dev.22175>
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Allyn and Bacon.
- Tackett, J. L., Brandes, C. M., King, K. M., & Markon, K. E. (2019). Psychology's replication crisis and clinical psychological science. *Annual Review of Clinical Psychology*, 15, 579–604. <https://doi.org/10.1146/annurev-clinpsy-050718-095710>
- Tackett, J. L., Brandes, C. M., & Reardon, K. W. (2019). Leveraging the Open Science framework in clinical psychological assessment research. *Psychological Assessment*, 31(12), 1386–1394. <https://doi.org/10.1037/pas0000583>
- Tackett, J. L., Brandes, C. M., & Reardon, K. W. (2019). Leveraging the Open Science Framework in clinical psychological assessment research. *Psychological Assessment*, 31(12), 1386–1394. <https://doi.org/10.1037/pas0000583>
- Tackett, J. L., & Miller, J. D. (2019). Introduction to the special section on increasing replicability, transparency, and openness in clinical psychology. *Journal of Abnormal Psychology*, 128(6), 487–492. <https://doi.org/10.1037/abn0000455>
- Teachman, B. A., McKay, D., Barch, D. M., Prinstein, M. J., Hollon, S. D., & Chambless, D. L. (2019). How psychosocial research can help the National Institute of Mental Health achieve its grand challenge to reduce the burden of mental illnesses and psychological disorders. *American Psychologist*, 74(4), 415–431. <https://doi.org/10.1037/amp0000361>
- Thomas, S. A., Thompson, E. C., Peters, J. R., Micalizzi, L., Meisel, S. N., Maron, M., & Wolff, J. C. (2024). Investigating substance use as a coping strategy among adolescent psychiatric inpatients: A comparative analysis before and during the COVID-19 pandemic. *Child Psychiatry and Human Development*. <https://doi.org/10.1007/s10578-024-01731-0>
- Thunissen, M. R., aan het Rot, M., van den Hoofdakker, B. J., & Nauta, M. H. (2022). Youth psychopathology in daily life: Systematically reviewed characteristics and potentials of ecological momentary assessment applications. *Child Psychiatry and Human Development*, 53(6), 1129–1147. <https://doi.org/10.1007/s10578-021-01177-8>
- Timmons, A. C., Duong, J. B., Simo Fiallo, N., Lee, T., Vo, H. P. Q., Ahle, M. W., & Chaspari, T. (2023). A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. *Perspectives on Psychological Science*, 18(5), 1062–1096. <https://doi.org/10.1177/17456916221134490>
- Torous, J., Linaaron, J., Goldberg, S. B., Sun, S., Bell, I., Nicholas, J., & Firth, J. (2025). The evolving field of digital mental health: Current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. *World Psychiatry*, 24(2), 156–174. <https://doi.org/10.1002/wps.21299>
- Turner, S. M., Beidel, D. C., Cooley, M. R., Woody, S. R., & Messer, S. C. (1994). A multicomponent behavioral treatment for social phobia: Social effectiveness therapy. *Behaviour Research and Therapy*, 32(4), 381–390. [https://doi.org/10.1016/0005-7967\(94\)90001-9](https://doi.org/10.1016/0005-7967(94)90001-9)
- Vaillant-Coindard, E., Briet, G., Lespiau, F., Gisclard, B., & Charbonnier, E. (2024). Effects of three prophylactic interventions on French middle-schoolers' mental health: Protocol for a randomized controlled trial. *BMC Psychology*, 12(1), 204. <https://doi.org/10.1186/s40359-024-01723-8>
- Venturo-Conerly, K. E., Reynolds, R., Clark, M., Fitzpatrick, O. M., & Weisz, J. R. (2023). Personalizing youth psychotherapy: A scoping review of decision-making in modular treatments. *Clinical Psychology: Science and Practice*, 30(1), 45–62. <https://doi.org/10.1037/cps0000130>
- Watts, A. L., Greene, A. L., Bonifay, W., & Fried, E. I. (2024). A critical evaluation of the p-factor literature. *Nature Reviews Psychology*, 3(2), 108–122. <https://doi.org/10.1038/s44159-023-00260-2>
- Weisz, J. R., Doss, A. J., & Hawley, K. M. (2005). Youth psychotherapy outcome research: A review and critique of the evidence base. *Annual Review of Psychology*, 56, 337–363. <https://doi.org/10.1146/annurev.psych.55.090902.141449>
- Weisz, J. R., Kuppens, S., Ng, M. Y., Eckshtain, D., Ugueto, A. M., Vaughn-Coaxum, R., Jensen-Doss, A., Hawley, K. M., Krumholz Marchette, L. S., Chu, B. C., Weersing, V. R., & Fordwood, S. R. (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. *American Psychologist*, 72(2), 79–117. <https://doi.org/10.1037/a0040360>
- Weisz, J. R., Kuppens, S., Ng, M. Y., Vaughn-Coaxum, R. A., Ugueto, A. M., Eckshtain, D., & Corteselli, K. A. (2019). Are psychotherapies for young people growing stronger? Tracking trends over time for youth anxiety, depression, ADHD, and conduct problems. *Perspectives on Psychological Science*, 14(2), 216–237. <https://doi.org/10.1177/1745691618805436>
- Weisz, J. R., Venturo-Conerly, K. E., Fitzpatrick, O. M., Frederick, J. A., & Ng, M. Y. (2023). What four decades of meta-analysis have taught us about youth psychotherapy and the science of research synthesis. *Annual Review of Clinical Psychology*, 19(1), 79–105. <https://doi.org/10.1146/annurev-clinpsy-080921-082920>

- Wiltsey Stirman, S. W., & Beidas, R. S. (2020). Expanding the reach of psychological science through implementation science: Introduction to the special issue. *The American Psychologist*, 75(8), 1033–1037. <https://doi.org/10.1037/amp0000774>
- Youngstrom, E. A., Van Meter, A., Frazier, T. W., Hunsley, J., Prinstein, M. J., Ong, M. L., & Youngstrom, J. K. (2017). Evidence-based assessment as an integrative model for applying psychological science to guide the voyage of treatment. *Clinical Psychology: Science and Practice*, 24(4), 331–363. <https://doi.org/10.1111/cpsp.12207>
- Zhang, M., Fan, C., Ma, L., Wang, H., Zu, Z., Yang, L., & Li, X. (2024). Assessing the effectiveness of internet-based interventions for mental health outcomes: An umbrella review. *General Psychiatry*, 37(4), Article e101355. <https://doi.org/10.1136/gpsych-2023-101355>