



The Operations Triad Model and Youth Mental Health Assessments: Catalyzing a Paradigm Shift in Measurement Validation

Andres De Los Reyes, Mo Wang, Matthew D. Lerner, Bridget A. Makol, Olivia M. Fitzpatrick & John R. Weisz

To cite this article: Andres De Los Reyes, Mo Wang, Matthew D. Lerner, Bridget A. Makol, Olivia M. Fitzpatrick & John R. Weisz (2023) The Operations Triad Model and Youth Mental Health Assessments: Catalyzing a Paradigm Shift in Measurement Validation, *Journal of Clinical Child & Adolescent Psychology*, 52:1, 19-54, DOI: [10.1080/15374416.2022.2111684](https://doi.org/10.1080/15374416.2022.2111684)

To link to this article: <https://doi.org/10.1080/15374416.2022.2111684>



© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC.



[View supplementary material](#)



Published online: 30 Aug 2022.



[Submit your article to this journal](#)



Article views: 7151



[View related articles](#)



[View Crossmark data](#)



Citing articles: 62 [View citing articles](#)

The Operations Triad Model and Youth Mental Health Assessments: Catalyzing a Paradigm Shift in Measurement Validation

Andres De Los Reyes ^a, Mo Wang ^b, Matthew D. Lerner ^c, Bridget A. Makol ^a, Olivia M. Fitzpatrick ^d, and John R. Weisz ^d

^aComprehensive Assessment and Intervention Program, Department of Psychology, University of Maryland; ^bDepartment of Management, University of Florida; ^cDepartment of Psychology, Stony Brook University; ^dDepartment of Psychology, Harvard University

ABSTRACT

Researchers strategically assess youth mental health by soliciting reports from multiple informants. Typically, these informants (e.g., parents, teachers, youth themselves) vary in the social contexts where they observe youth. Decades of research reveal that the most common data conditions produced with this approach consist of discrepancies across informants' reports (i.e., *informant discrepancies*). Researchers should arguably treat these informant discrepancies as *domain-relevant information*: data relevant to understanding youth mental health domains (e.g., anxiety, depression, aggression). Yet, historically, in youth mental health research as in many other research areas, one set of paradigms has guided interpretations of informant discrepancies: *Converging Operations* and the *Multi-Trait Multi-Method Matrix (MTMM)*. These paradigms (a) emphasize shared or *common variance* observed in multivariate data, and (b) inspire research practices that treat *unique variance* (i.e., informant discrepancies) as *measurement confounds*, namely random error and/or rater biases. Several years ago, the Operations Triad Model emerged to address a conceptual problem that Converging Operations does not address: Some informant discrepancies might reflect measurement confounds, whereas others reflect domain-relevant information. However, addressing this problem requires more than a conceptual paradigm shift beyond Converging Operations. This problem necessitates a paradigm shift in *measurement validation*. We advance a paradigm (Classifying Observations Necessitates Theory, Epistemology, and Testing [CONTEXT]) that addresses problems with using the MTMM in youth mental health research. CONTEXT optimizes measurement validity by guiding researchers to leverage (a) informants that produce domain-relevant informant discrepancies, (b) analytic procedures that retain domain-relevant informant discrepancies, and (c) study designs that facilitate detecting domain-relevant informant discrepancies.

Scholars who conduct mental health research with children and adolescents (i.e., youth) perform studies that ultimately inform critical decisions regarding mental health services, including diagnosis, treatment selection, and evaluation of treatment outcomes. The “origin stories” of these services can be traced back through decades of research and theory (Bronfenbrenner, 1979; Cicchetti, 1984; Crick & Dodge, 1994; Darling & Steinberg, 1993). This work teaches us that youth vary in terms of their lived experiences and social worlds. Any one youth may experience mental health concerns in some contexts—such as home and school—to a greater degree than they do in other contexts—such as peer interactions. In fact, the intervention models informed by this work—such as classical conditioning, cognitive theory, radical behaviorism, and social learning theory (Bandura, 1977; Beck, 1993; Rescorla & Wagner, 1972;

Skinner, 1953)—hinge on tailoring techniques (e.g., cognitive restructuring, therapeutic exposures, token economies, homework activities) to the unique social contexts of youth clients (e.g., home, school, peer interactions; Alfano & Beidel, 2011; Sewart & Craske, 2020; Weisz, 2004; Weisz et al., 2004). In youth mental health research, sound theory building and testing, along with developing and testing conceptually grounded interventions and implementing these interventions in clinical work with youth clients, all depend on sound measurement approaches.

The most commonly used measurement approach in youth mental health research involves scholars implementing a process akin to what journalists do: collect reports from multiple *informants* (Achenbach, 2020; Hunsley & Mash, 2007; De Los Reyes, 2011; Weisz et al., 2005). Like journalists preparing a news story,

CONTACT Andres De Los Reyes  adlr@umd.edu  Comprehensive Assessment and Intervention Program, Department of Psychology, University of Maryland, Biology/Psychology Building, Room 3123H, College Park, MD 20742

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

 Supplemental material for this article can be accessed online at <https://doi.org/10.1080/15374416.2022.2111684>

© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

mental health researchers seek a holistic understanding of the youth undergoing evaluation. Researchers collect reports about youth mental health from informants who vary in the social contexts where they observe youth (e.g., parents, teachers, youth themselves). In fact, even when informants like parents and teachers make reports using identical instruments (e.g., surveys with parallel item content and response options), these informants provide unique data that, in part, reflect the underlying social environments within which they observe youth behavior (e.g., parents at home vs. teachers at school; see also, Kraemer et al., 2003). Thus, researchers solicit reports from *structurally different* informants: Informants distinguishable from one another by where they observe youth, namely by *social contexts* or structural features of the social environment (e.g., Eid et al., 2008; Geiser et al., 2012; Konold & Cornell, 2015; Konold & Sanders, 2020). In this paper, we focus on a fundamental problem in youth mental health research: The perennial disconnect between the conceptual models that most closely align with use of structurally different informants, and the validation paradigm that has governed researchers' approaches to using, interpreting, and examining multi-informant data. In turn, we advance a validation paradigm to resolve this disconnect.

Unsurprisingly, structurally different informants often tell researchers very different stories about the youth undergoing evaluation. For instance, a researcher collecting a report from a parent might learn that a youth displays oppositional behavior, but the youth's teacher does not corroborate the parent's impressions. At other times, a researcher observes the reverse pattern: a teacher reports that a youth in the classroom needs help with anxiety, but the youth's parent does not concur. Sometimes, a researcher collects a report from a youth, who reports struggling with relatively covert depressive symptoms that reports from adult authority figures fail to capture. Throughout this paper, we refer to these disparate patterns of reports as *informant discrepancies*.

We focus our paper on informant discrepancies. Indeed, across all known mental health domains we know the following to be true: Informant discrepancies define the data conditions that result from using structurally different informants' reports to assess youth mental health. Empirical observations of informant discrepancies date back to the 1950s (e.g., Lapouse & Monk, 1958), and have appeared in over 400 published studies (De Los Reyes & Makol, *in press*). These studies reveal robust findings about the nature of these discrepancies. For instance, a meta-analysis published in the 1980s of 119 studies yielded a relatively low mean

Pearson r of .28 between two informant's reports of the same youth's mental health (Achenbach et al., 1987). A meta-analysis conducted over 25 years later of 341 studies published between 1989 and 2014 yielded an identical estimate (i.e., .28; De Los Reyes et al., 2015). Considering the sheer number of changes in youth mental health research over the decades separating these meta-analyses—in the theories driving the research, measures and study designs researchers use to address research aims, and operational definitions of youth mental health domains—the stability of this effect is remarkable. Further, a meta-analysis of studies from over 30 countries reveals that informant discrepancies reliably appear in assessments all over the globe (De Los Reyes et al., 2019a).

That informant discrepancies define the data conditions resulting from use of structurally different informants' reports reveals a reality about this approach to multi-informant assessment. That is, researchers often use structurally different informants' reports because they each facilitate understanding youth mental health in their own way. Yet, such reports might nonetheless share data in common. Each informant's report has the ability to provide *unique variance* about youth mental health, as well as *common variance* that is shared with the reports of other informants. The breakdown between common and unique variance factors into how we use data from structurally different informants. In particular, we use these data to estimate relations between mental health domains and other clinically relevant domains (e.g., parenting, diagnostic status, treatment response). When estimating relations, how much of each source of variance might we expect to observe when using structurally different informants' reports? We have addressed this question using *regression commonality analyses*: tests of the overlapping variance from multiple predictors (e.g., parent, youth, and teacher reports) on a common criterion variable (Cohen et al., 2003). These tests reveal that parent, teacher, and youth reports display common variance in prediction of a criterion variable at levels of 3% for reports of youth externalizing concerns, 1.3% for youth internalizing concerns, and 0.8% for concerns overall (De Los Reyes et al., 2015). Taken together, we should expect relations between structurally different informants' reports of youth mental health and clinically relevant domains to consist of less than 5% of common variance and upwards of 95% of unique variance. Thus, understanding youth mental health involves detecting factors that underlie both common and unique variance.

The need to emphasize both common and unique variance in structurally different informants' reports clashes with a long-standing reality in youth mental

health. Historically, researchers' efforts to understand, interpret, and model multivariate data conditions of any kind—including multi-informant measures of youth mental health—have, ironically, emphasized common variance: The tiniest variance available to explain. The emphasis on common variance in youth mental health research is no accident. Scholars in Psychology generally emphasize common variance. In fact, one can trace the source of this emphasis back to an idea that is now over 60 years old. In the 1950s, Garner et al. (1956) advanced *Converging Operations*, a paradigm that they originally developed to improve the interpretability of experiments in psychophysics. In their view, the findings from one experiment often fail to hold a high degree of certainty in support of a hypothesis. One way to increase certainty might involve designing several methodologically distinct experiments to test the hypothesis. One could then use convergence of findings across experiments to gauge empirical support for the hypothesis. Here, Garner and colleagues equated converging outcomes with the truth. By extension, within Converging Operations, diverging outcomes would reflect inconclusive evidence, and in the case of informant discrepancies, *measurement confounds* like random error and/or rater biases.

In the decades since Garner et al. (1956), scholars have credited Converging Operations for improving the interpretability of research findings in areas as diverse as intelligence (Sternberg, 2005), face perception (Von Der Heide et al., 2018), memory (Vallar, 2006), personality (Brase, 2014), neuropsychology (Bruno & Paolo Battaglini, 2008), addiction science (Zeid et al., 2018), religious studies (Krause et al., 2010), and infant development (LoBue et al., 2020). In fact, Campbell and Fiske (1959) drew inspiration from Converging Operations to develop their seminal *Multi-Trait Multi-Method Matrix* (MTMM; Grace, 2001; Highhouse, 2009). This is the dominant measurement validation paradigm implemented in youth mental health research, and subdisciplines of Psychology (e.g., Clinical, Developmental, Personality) more broadly (see Watts et al., 2022). Yet, the MTMM paradigm instantiates in measurement a rather extreme view of common variance. This view guides users of the MTMM to assume that all informant discrepancies (i.e., low “monotrait-heteromethod” correlations) reflect measurement confounds, and thus threats to measurement validity. For users of this paradigm, this assumption is their “north star,” their guiding principle. In this paper, we articulate how this principle logically translates to users of the MTMM paradigm taking a specific approach to working with multi-informant data. This approach guides users' decisions as to which theories will inform their examinations of informant discrepancies, the research

practices they will implement when working with multi-informant data, and even the philosophical underpinnings of how they draw conclusions from research findings.

Although the Converging Operations and MTMM paradigms have proven useful to many areas of study, it would be unreasonable to expect their utility to extend to all areas of study. We contend that the data conditions that typify work in youth mental health violate the assumptions underlying use of these paradigms. As such, youth mental health requires a paradigm shift on these issues. This shift necessitates a two-step process, which we began a decade ago. We took the first step toward making this shift by developing the Operations Triad Model (De Los Reyes et al., 2013a). Here, the data conditions could reflect Converging Operations, insofar as each informant's report points to the same conclusion. For instance, reports from parents and teachers might each indicate that an intervention produced positive responses among the youth clients who received it. Clearly, Converging Operations scenarios reflect data conditions that ought to be retained. Yet, Converging Operations scenarios reflect relatively rare data conditions within youth mental health research. Thus, they provide an incomplete account of all data conditions observed in this area. This necessitates also conceptualizing informant discrepancies.

Within the Operations Triad Model, not all informant discrepancies are created equal. Structurally different informants often vary in their opportunities to observe youth behavior within and across clinically relevant contexts, such as parents at home and teachers at school. If youth also vary in where they display mental health concerns, then this raises the possibility that at least some informant discrepancies contain data germane to the specific contexts in which youth display these concerns. Those informant discrepancies that reflect variations in youth behavior across social contexts reflect what we refer to as *domain-relevant information*: data relevant to understanding youth mental health domains (e.g., anxiety, depression, aggression). In the Operations Triad Model, *Diverging Operations* reflect these exact kinds of informant discrepancies. Yet, to assume that all informant discrepancies are domain-relevant would be as mistaken an assumption as the MTMM paradigm's assumption that all discrepancies reflect measurement confounds. Thus, the third concept within the Operations Triad Model is *Compensating Operations* – that is, informant discrepancies that reflect measurement confounds like random error or rater biases. These confounds are irrelevant to understanding measured domains (see also, Millsap, 2011), hence their distinction from Diverging Operations.

Purpose and Aims

The purpose of this paper is to extend the analysis of informant discrepancies beyond the paradigms presented previously, with a specific focus on measurement validation.¹ In particular, the Converging Operations and MTMM paradigms optimize the use of multi-informant assessments when common variance accounts for the grand majority of all variance. They are less appropriate for scenarios where unique variance accounts for the grand majority of all variance, because they treat all of this variance as reflective of measurement confounds. We offer a path to addressing this problem with a new validation paradigm: Classifying Observations Necessitates Theory, Epistemology, and Testing (CONTEXT). In essence, if the Operations Triad Model catalyzed a paradigm shift in how we conceptualize informant discrepancies in youth mental health research, then CONTEXT catalyzes a paradigm shift in validation, in which we instantiate in measurement key principles underlying the Operations Triad Model. CONTEXT facilitates retaining common variance (i.e., Converging Operations) and, at the same time, detecting informant discrepancies that reflect Diverging Operations scenarios. In this paper we address three aims. First, we draw distinctions between CONTEXT and the MTMM paradigm, and cite evidence to demonstrate how these two paradigms inform fundamentally distinct approaches to collecting, using, and interpreting data from structurally different informants. We argue that the MTMM paradigm's core assumptions guide youth mental health researchers to leverage theory, interpret research findings, and engage in research practices that risk depressing measurement validity. In contrast, we argue, the features of CONTEXT guide youth mental health researchers to leverage theory, epistemological principles, and research practices that facilitate optimizing measurement validity. Second, we illustrate CONTEXT's features with a combination of published and new evidence. Third, we delineate several implications of CONTEXT and chart directions for future research.

The Three Features of CONTEXT and How They Differ from the MTMM Paradigm

The mathematical psychologist Philip Levy (1969) wrote, “nothing, not even real data, can contradict classical test

theory” (p. 276). This statement highlights a fact about paradigms like the MTMM: They are not inherently “wrong” or “right.” Paradigms all have assumptions. It is not the paradigm's task to align these assumptions with the data conditions to which it will be applied; that's the user's task. At the same time, a paradigm's assumptions chart a path, a guide for the user. In the case of multi-informant assessments, that path dictates the approach the user will take to working with data from these assessments. That path will influence the selection of theoretical models to guide hypothesis testing. That path will influence what or how many informants the user selects, the study designs they select, and the analytic procedures selected to carry out data analyses. If the user is correct to follow the path—such that the data conditions align with the paradigm's assumptions—then their decisions will likely optimize measurement validity. Yet, if the user incorrectly applies a paradigm—namely to data conditions that violate the paradigm's assumptions—then their decisions will likely depress measurement validity.

We designed CONTEXT to chart a path for youth mental health researchers that aligns with the data conditions typifying their use of multi-informant approaches to assessment: (a) use of structurally different informants that (b) often produce informant discrepancies that (c) contain domain-relevant information. In this section, we demonstrate how the core assumptions of CONTEXT and those of the MTMM paradigm logically result in distinct guidance on how to work with multi-informant data. In fact, the guidance CONTEXT provides users reflects a paradigm shift across multiple aspects of scholarly work, namely in theory, epistemology, and research practices when leveraging multi-informant approaches to assessing youth mental health.

Theory

The differences between CONTEXT and the MTMM paradigm become readily apparent when probing the guidance each provides users for selecting theories about what informant discrepancies reflect. To understand these differences, we have to go beyond the conceptual models that gave rise to each of them (i.e., Operations Triad Model and Converging Operations, respectively), and trace the “origin story” of Converging Operations. We can trace the conceptual origins of the MTMM paradigm back to Converging Operations (see Grace, 2001; Highhouse, 2009).

¹As recently reviewed elsewhere (De Los Reyes & Makol, 2022), prior work supports the notion that informant discrepancies frequently manifest when delivering services to clients. That said, a discussion of informant discrepancies observed in service delivery settings lies outside the scope of this paper. In fact, distinctions between the informant discrepancies that occur in research samples (i.e., the focus of this paper) versus those that occur when delivering services to individual clients (e.g., Fisher et al., 2017; Hawley & Weisz, 2003; Hoffman & Chu, 2015; Yeh & Weisz, 2001) necessitate varied considerations that encompass distinct frames of reference (i.e., measurement validation vs. implementation science; see also, De Los Reyes et al., 2022a).

However, we can trace the conceptual origins of Converging Operations back to the very origins of measurement in Psychology. It was Edgeworth (1888) who repurposed principles from Physics to posit that one might estimate mental states by gathering information from multiple raters of the same state. As Edgeworth saw it, no one rater provided a “perfect score,” but at the same time, none provided a rating with value unique to itself. Instead, each of their ratings only began to make sense when you assumed that they each reflected “error,” and only when you examined them collectively did they approximate what classical test theorists refer to as the *true score* of the measured mental state (see also, Borsboom, 2005). By extension, the “closer” these ratings were to each other—the more they converged with one another—the closer their assumed approximation to the true score of the measured state. For Edgeworth, convergence was the only path to the truth. The same could be said for Garner et al. (1956). Converging Operations equates converging findings with strong empirical support. Similarly, within the MTMM paradigm, high cross-informant agreement (i.e., high mono-trait-heteromethod correlations) signals strong support for measurement validity. This assumption logically results in also assuming that diverging findings like informant discrepancies can only signal findings that lack truth; they reflect measurement confounds and nothing more.

What implications does this core assumption of the MTMM paradigm have for informing the selection of theories for understanding what informant discrepancies reflect? We contend that this assumption compels users of the MTMM paradigm to select theories about informant discrepancies that fundamentally differ from those that CONTEXT guides users to select. In Table 1, we summarize our rationale for this contention. If users of the MTMM paradigm begin from the assumption that all informant discrepancies reflect measurement confounds, then it logically follows they will seek out theoretical explanations for why these confounds exist. For

instance, when users of the MTMM paradigm adopt a theory for informant discrepancies research, they tend to use the depression-distortion hypothesis, which holds that informant discrepancies stem from the fact that informants completing reports about youth (e.g., parents) often experience mental health concerns that impact their accuracy as raters. In particular, concerns such as depressive mood are quite common among parents of youth undergoing evaluation (Goodman et al., 2011). Consistent with the assumption that all informant discrepancies reflect measurement confounds, the depression-distortion hypothesis posits that informant discrepancies are fully accounted for by depressed informants attending to, encoding, recalling, and rating more negative behaviors (i.e., relative to neutral or positive behaviors) in the youth being evaluated, relative to the reports of non-depressed informants (Richters, 1992).

Importantly, users of the MTMM paradigm frequently cite the depression-distortion hypothesis as the theoretical foundation of their work (e.g., Bauer et al., 2013; Clark et al., 2016; Jungersen & Lonigan, 2021; T. M. Olinio et al., 2021; T.M. Olinio et al., 2018). Let us set aside the fact that this hypothesis rests on weak empirical evidence (for recent reviews, see De Los Reyes & Makol, 2022; De Los Reyes et al., 2022b). Instead, let us focus on the implications of this conceptual frame. If the depression-distortion hypothesis frames informant discrepancies as a “bug” or “artifact” of using multi-informant approaches to assess youth mental health, then work that flows from this frame cannot fairly evaluate informant discrepancies as a source of domain-relevant data.

With CONTEXT, we guide users to select theories that presume the structurally different informants from whom they solicited reports “did their job.” If scholars seek out informants who they expect will provide psychometrically sound reports, then it is more likely than not that the informant discrepancies produced by these reports will contain domain-relevant information. To

Table 1. Differences between CONTEXT and the MTMM paradigm in guidance on theory surrounding informant discrepancies.

Validation Paradigm	Core Assumption Regarding Informant Discrepancies	How Assumption Guides Conceptual Foundations of Paradigm-Informed Research	Examples of Theories Applied in Prior Research	Citations of Prior Work That Applied Paradigm-Informed Theories
MTMM	All informant discrepancies reflect measurement confounds	Guides users to apply theories about informant discrepancies that posit factors that produce rater biases	Depression-Distortion Hypothesis (Richters, 1992)	Bauer et al. (2013); Clark et al. (2016); Jungersen and Lonigan (2021); T.M. Olinio et al. (2018, 2021)
CONTEXT	At least some informant discrepancies contain domain-relevant information	Guides users to apply theories about informant discrepancies that posit factors that produce domain-relevant information	Situational Specificity (Achenbach et al., 1987)	De Los Reyes et al. (2022a, 2022b, 2022c); Deros et al. (2018); Lerner et al. (2017); Makol et al. (2020); Watts et al. (2022)

CONTEXT = Classifying Observations Necessitates Theory, Epistemology, and Testing. MTMM = Multi-Trait Multi-Method Matrix.

illustrate, prior work testing for the presence of domain-relevant information has historically relied on the Achenbach et al. (1987) notion of *situational specificity* to guide their work (e.g., see De Los Reyes et al., 2022a, 2022b, 2022c; Deros et al., 2018; Lerner et al., 2017; Makol et al., 2020; Watts et al., 2022). In direct contrast to the depression-distortion hypothesis, situational specificity holds that informant discrepancies might reflect key aspects of youth and their social contexts. In fact, contexts (e.g., home, school) often vary in their *contingencies*: Factors in the environment (e.g., inconsistent discipline, aversive peer interactions) that elicit specific kinds of behaviors indicative of mental health concerns (e.g., aggression, avoidance; see, Kazdin, 2013; Skinner, 1953). Within their reports, structurally different informants may very well be reflecting these variations in the social environments of the youth about whom they provide reports.

Facets of Situational Specificity Relevant to Understanding Informant Discrepancies

With CONTEXT, we extend the Achenbach et al. (1987) notion of situational specificity, by guiding users to theorize how situational specificity effects translate to the presence of domain-relevant information in informant discrepancies. In fact, with CONTEXT we guide users to develop tangible, testable explanations for informant discrepancies. It is these explanations—these theories—that inform the design of studies focused on detecting informant discrepancies that reflect domain-relevant information. In line with this goal, prior work indicates that situational specificity effects manifest as “facets” of domain-relevant information. These facets manifest in many assessments where informant discrepancies arise. Three such facets are particularly important to describe. The first facet we already discussed: The idea that sometimes informant discrepancies reflect contextual variations in youth mental health concerns.

A second facet of situational specificity effects stems from the notion that, as structurally different data sources, informants like parents and teachers are not detached bystanders. Not only do they observe youth behavior, they also shape the behavior of the youth about whom they provide reports. For instance, parents report about youth behavior and engage in parenting behaviors involved in the development of the mental health domains about which they provide reports (e.g., Darling & Steinberg, 1993). Similarly, teachers’ classroom management strategies impact the learning outcomes of the youth about whom they report, and these learning outcomes impact mental health (e.g., Atkins et al., 2017). Youth provide self-reports containing similarly unique information, such that they capture aspects

of their social environments for which adult informants (e.g., parents, teachers) often lack experiences with observing (e.g., peer relations outside of home and school contexts), and these environments are intimately related to the development and maintenance of youth mental health concerns (e.g., Crick & Dodge, 1994). Further, informants’ perspectives on youth mental health, or their thresholds for perceiving certain behaviors as clinically concerning, are intimately aligned with the behaviors about which they report. These perspectives and/or thresholds might often dictate how informants react to youth mental health concerns (e.g., disciplinary strategies); factors that contribute to the development and maintenance of these very concerns.

Informants’ roles in youth mental health are even evident in the nosological systems used to diagnose youth mental health conditions. In particular, these systems outline diagnostic criteria as well as *associated features* of mental health conditions (e.g., *Diagnostic and Statistical Manual of Mental Disorders* [DSM-5]; American Psychiatric Association, 2013). These associated features include characteristics that often co-occur with symptoms delineated in diagnostic criteria (e.g., maladaptive parenting behaviors; Hunsley & Lee, 2014). Associated features often promote and maintain youth mental health conditions. Further, they may reflect behavioral manifestations of informants’ perspectives on these conditions (e.g., parent’s use of harsh parenting in response to their child’s behavior). Thus, these features are often directly targeted in evidence-based interventions (e.g., parenting interventions for conduct disorder; Kazdin & Rotella, 2009; Weisz & Kazdin, 2017). In fact, knowledge about associated features often comes from research testing links between measures of these features (e.g., parenting) and informants’ reports of youth mental health. In turn, this research comprises some of the psychometric support for mental health instruments (see, Hunsley & Mash, 2018).

Taken together, informants’ reports about youth mental health are inextricably linked to features in the social environment. Yet, these associated features often display a situational specificity of their own. Associated features often manifest in one context (e.g., parenting, classroom management strategies) to a greater degree than in other contexts. In this respect, they often illuminate measured domains relevant to understanding variations among social contexts, particularly with regard to the contingencies that maintain youth mental health concerns. In support of this notion, consider that researchers often integrate structurally different informants’ reports using procedures focused on common variance (e.g., combinational algorithms); approaches

that often fail to identify known associated features of mental health concerns that manifest in situationally specific ways (e.g., Offord et al., 1996; Rubio-Stipec et al., 2003). By extension, the structural differences among informants sometimes work to the benefit of identifying situationally specific features, such that each informant's report facilitates detecting distinct aspects of mental health service delivery (e.g., domains of clinical concern vs. goals for mental health services; see, Follet et al., 2022). Thus, situational specificity effects might also produce domain-relevant informant discrepancies, insofar as they illuminate features of social environments linked to the development and maintenance of youth mental health concerns.

A third facet of situational specificity effects is related to the second facet. When informants who provide reports (e.g., parents) play key roles in youths' social environments, they typically inhabit these roles for extended periods. If these roles also reflect situationally specific characteristics (e.g., parenting behaviors) that are unique to the reports of a particular informant (e.g., parents but not teachers), then informant discrepancies may contain domain-relevant information pertinent to predicting youth mental health outcomes. As an example, consider research on parenting. Some research indicates that parents' own functioning (e.g., levels of stressors) impacts the development and maintenance of youth conduct problems (e.g., Kazdin & Rotella, 2009). Interestingly, when interventions successfully reduce levels of youth conduct problems, not only does youth functioning improve, but so does that of their parents (e.g., Kazdin & Wassell, 2000). In fact, therapeutically targeting parental stress in combination with these youth interventions improves youth and parental outcomes (Kazdin & Whitley, 2003). Thus, not only can features of youths' social environments vary across informants (e.g., parental stress at home), but these features may also change over time in ways that differ from those same features of informants in other social environments (e.g., teacher stress at school). If informants' reports vary in the social environments in which they observe youth behavior—and informants' mental health reports reflect, in part, features of these very environments—then these variations among informants are not static in nature. That is, they wax and wane over time or remain stable over time. These fluctuations have implications for using reports to predict outcomes (e.g., treatment response). CONTEXT guides users to interpret each informant's report relative to the reports of other informants, because they may each carry unique predictive capabilities, producing interactions or patterns that boost the overall predictive value of reports (see, Becker-Haimes et al., 2018; Makol et al., 2019,

2020). As such, situational specificity might produce informant discrepancies that reflect domain-relevant phenomena germane to predicting youth mental health outcomes (see also, Laird & Weems, 2011).

Situational Specificity and CONTEXT

We focused on facets of situational specificity that manifest as domain-relevant information in informant discrepancies. These facets may explain some, but certainly not all cases of domain-relevant information. That said, by conceptualizing informant discrepancies not merely as measurement confounds but rather as potential sources of domain-relevant information, CONTEXT guides users to detect cases of domain-relevant information, beyond those typified by common variance (i.e., informant agreement). In this respect, CONTEXT guides users to ground their work in conceptual models that optimize measurement validity. In turn, we expect CONTEXT to optimize the utility of structurally different informants' reports to address a host of research questions in youth mental health, including those focused on characterizing youth mental health and predicting outcomes. Of course, theory is the first step in the validation process. For CONTEXT to catalyze a true paradigm shift in validation, it must also represent a distinct departure from the MTMM paradigm in the epistemological guidance it provides researchers on how to draw conclusions from data provided by structurally different informants.

Epistemology

In his seminal treatise on psychological measurement—*Measuring the mind*—Borsboom (2005) wrote, “A blueprint may be invaluable in constructing a bar, but one cannot have a beer inside one.” The statement proves useful for understanding the links between validation paradigms and the measured realities they inform. Like the assumptions we have discussed previously regarding paradigms, blueprints carry assumptions regarding the physical world. Blueprints provide specifications to builders that ought to guide construction of shelves that properly adhere to walls, and walls that, unless tragedy strikes (e.g., really strong tornado), will remain upright. In this respect, blueprints are vitally important. Blueprints make the difference between a bar that stands the test of time and one that crumbles due to lousy planning. Yet, if the builder has agency over which blueprint to follow, then they own the consequences of what they build. Similarly, researchers must own the consequences of the paradigm that informs their approach to assessing domains of interest. Granted, the building blocks of multi-informant assessments are not of the

“brick and mortar” variety. Instead, the philosophical underpinnings of validation paradigms and the logical conclusions that result from their use dictate how users build multi-informant assessments and interpret their outcomes (i.e., the data conditions).

In this section, we leverage principles from Epistemology (i.e., branch of Philosophy focused on what is known and knowable) and Philosophy of Science in particular to demonstrate the philosophical differences between CONTEXT and the MTMM paradigm. We summarize in Table 2 that, down to an epistemological level, CONTEXT aligns with youth mental health assessments to a far greater extent than the MTMM paradigm ever could. In fact, the consequences of adhering to the epistemological process that the MTMM paradigm guides users to follow are so dire for the user to consider, that they inspire some fairly ill-informed—and baseless—decisions regarding what to do with multi-informant data. Work in Philosophy of Science highlights these errors in decision-making. In particular, Hempel (1966) delineates two processes germane to hypothesis testing in scholarly inquiry. In the first stage of this process, a scholar specifies a *test implication*: A specification of the data conditions arising from expected outcomes. The MTMM paradigm guides users toward adopting the test implication that converging findings are the only data that contain domain-relevant information. Of course, such a test implication would be a blueprint for success when interpreting findings in a scholarly area wherein the modal data conditions consist of converging findings. However, that same blueprint spells disaster when applied to data conditions that are typified by discrepant findings.

As we have previously noted, informant discrepancies characterize data conditions in youth mental health. What we have yet to articulate is that these discrepancies are not just linear in nature, as might be suggested by low-to-moderate levels of correspondence between informants’ reports. Rather, these discrepancies even

manifest when drawing discrete interpretations of the state of the science. Is exposure-based therapy efficacious for addressing youth anxiety concerns? Is corporal punishment a risk factor for conduct disorder? Are improvements in social skills a mechanism of action in interventions for attention-deficit/hyperactivity disorder (ADHD)? What is the 12-month prevalence rate for major depression among adolescents? For each of these kinds of questions, changing the informant who completes the measures results in qualitative differences in the conclusions drawn from research findings (for a recent review, see, De Los Reyes & Makol, 2022). In youth mental health, even discrete judgments about the state of the science fail to achieve convergence in research findings across multiple informants’ reports. Consequently, users of the MTMM paradigm are logically forced into accepting the fact that the data underlying all their research questions are inconclusive. Or are they?

When scholars encounter measured outcomes that conflict with the test implication outlined in a hypothesis they tested, Hempel (1966) describes a second step in the hypothesis-testing process. This step of the process allows users of the MTMM paradigm to nonetheless proceed in their scholarly work, even in the face of findings that would otherwise force them to interpret their work as inconclusive. Yet, it is also a step that is rife with commission of decision-making errors. Specifically, when the test implication fails to “match” observed outcomes, and a scholar is nonetheless motivated to accept their initial hypothesis, Hempel says they rely on a secondary or *ad hoc* hypothesis. An *ad hoc* hypothesis provides a researcher with a means of explaining away observed outcomes that contradict their initial hypothesis, in an effort to preserve its veracity. When confronted by informant discrepancies—and rather than accept the notion that they observed inconclusive findings—users of the MTMM paradigm must adopt the assumption that informant discrepancies

Table 2. Differences between CONTEXT and the MTMM paradigm in epistemology surrounding informant discrepancies.

Validation Paradigm	Test Implication	Logical Extension of Test Implication	Research Practices That Logically Extend from Test Implication	Relevant Citations
MTMM	Converging findings are the only data to contain domain-relevant information	Anything less than converging findings ought to be interpreted as inconclusive evidence	primary outcome measure methods; tests of measurement invariance; use of structural models that prioritize common variance; eschew estimating common or unique variance altogether and just examine informants’ reports individually	Bauer et al. (2013); Boutron et al. (2010); De Los Reyes et al. (2011); Howe et al. (2019); T.M. Olinio et al. (2018)
CONTEXT	At least some informant discrepancies reflect domain-relevant information	Informant discrepancies may optimize rather than hinder the interpretability of findings	strategic selection of informants that produce context-sensitive information; use of analytic procedures that facilitate identifying informant discrepancies and testing whether they contain domain-relevant information	Deros et al. (2018); Kraemer et al. (2003); Laird (2020); Lakind et al. (2022); Makol et al. (2020)

CONTEXT = Classifying Observations Necessitates Theory, Epistemology, and Testing. MTMM = Multi-Trait Multi-Method Matrix.

reflect only measurement confounds. Here, the *ad hoc* hypothesis reflects a kind of confirmation bias (Tversky & Kahneman, 1974): A tendency to focus on evidence supporting a hypothesis and downplay contradictory evidence. In fact, the tangible outcomes of these *ad hoc* hypotheses reveal themselves when considering common, MTMM-informed responses to observing informant discrepancies. We cite these examples in Table 2.

For instance, in developing an MTMM-informed structural model for which the depression-distortion hypothesis formed its conceptual foundations, the stated goal of Bauer et al. (2013) was to develop an analytic tool focused on creating, “integrative scores that are purged of the subjective biases of single informants” (p. 475). As currently applied (e.g., Curran et al., 2021; Shin et al., 2019; Soland & Kuhfeld, 2022), this model removes unique variance from reports to focus on attaining as “pure” an estimate of common variance as possible, and thus renders informant discrepancies as measurement confounds. The epistemological underpinnings of the MTMM paradigm also inform applications of measurement invariance techniques to model informant discrepancies. Here, users are required to treat the detection of informant discrepancies as synonymous with detecting measurement confounds. As evidence, look no further than on page 1 of Millsap’s (2011) authoritative text, in which he states, “Measurement invariance is built on the notion that a measuring device should function the same way across varied conditions, so long as those varied conditions are *irrelevant* [emphasis added] to the attribute being measured.” As with the model developed by Bauer et al. (2013), users who apply these techniques assume that informant discrepancies reflect *de facto* measurement confounds. That is, within the *ad hoc* step of the hypothesis-testing process that informs use of these models, the user does not test whether the informant discrepancies they seek to detect and/or model indeed reflect measurement confounds (e.g., Alvarez et al., 2021; Florean et al., 2022; Murray et al., 2021; Russell et al., 2016; Schiltz et al., 2021; T.M. Olino et al., 2018).

Sometimes, the *ad hoc* hypothesis process leads the user to simply “give up” on modeling or understanding common and/or unique variance altogether; we also cite examples of this in Table 2. Indeed, consider the common practice of selecting a *primary outcome measure* to test the effects of interventions within randomized controlled trials designs (for reviews, see, Boutron et al., 2010; De Los Reyes et al., 2011). When engaging in this practice, one selects a primary measure *a priori* and deems other measures as “secondary,” even those

that assess effects on the same domain as the primary measure (e.g., reductions in symptoms of the domain targeted for intervention). This approach treats one instrument as a “gold standard” for gauging intervention effects. Why would one even consider applying this approach to interpreting intervention effects in a multivariate environment, unless they encounter uncertainty with interpreting the modal outcome, namely discrepant findings as to intervention effects? Of course, these issues are compounded when using this approach in youth mental health. Indeed, if scholars in this area agree on anything, it is that there exists no definitive means to declare an instrument designed to assess a mental health domain as the “gold standard,” relative to alternative instruments for assessing that same domain (e.g., Achenbach et al., 1987; Hunsley & Mash, 2007; Kazdin, 2017; Kraemer et al., 2003; De Los Reyes, 2011; Richters, 1992). Here too, the *ad hoc* step of the hypothesis-testing process informs use of a practice that has no evidentiary basis.

The epistemological approach adopted by users of CONTEXT markedly departs from approaches that users of MTMM-informed procedures must adopt. We designed CONTEXT to guide users to subject notions about what informant discrepancies reflect to empirical scrutiny. CONTEXT-informed research involves testing the following research hypothesis: *Informant discrepancies contain unique domain-relevant information*. Operationally, this hypothesis reflects an expectation that informant discrepancies reflect variations in one or more of the facets of situational specificity described previously. This hypothesis also serves as a counterfactual to the null hypothesis in CONTEXT-informed research: *Informant discrepancies reflect only measurement confounds*. Unlike the MTMM paradigm, CONTEXT guides users to implement hypothesis testing processes, to guard against merely accepting as true an *ad hoc* hypothesis about informant discrepancies. As we describe in further detail below, studies that align with CONTEXT’s features develop hypothesis-testing procedures whereby if the multi-informant approach being tested fails to produce domain-relevant informant discrepancies, then the logical outgrowth of such an event involves observing inconclusive findings (e.g., Deros et al., 2018; Makol et al., 2020). When this event arises, we designed CONTEXT to guide users not toward creating *ad hoc* explanations for the inconclusive evidence, but rather toward conducting follow-up studies that address the possibility that the “null hypothesis” is true: The discrepancies they observed may, in fact, reflect measurement confounds (see also, De Los Reyes et al., 2013a).

In sum, distinctions between CONTEXT and the MTMM paradigm readily appear in the epistemological

principles they each guide users to adopt. When discrepant findings reflect the modal data conditions in an area of scholarly inquiry (e.g., informant discrepancies in youth mental health assessments), the *ad hoc* hypothesis process that logically follows from use of the MTMM paradigm plagues decision-making. This process bars direct tests of whether the initial hypothesis (e.g., findings should converge) cannot explain the data conditions (e.g., informant discrepancies reflect domain-relevant information). In contrast, CONTEXT leverages hypothesis testing to improve decision-making. Yet, in line with Hempel (1966), CONTEXT requires not just a logical hypothesis testing process, but also precise empirical conditions for interpreting the test implications inherent in understanding and interpreting informant discrepancies.

Testing

I am beginning to recognize the fact that nothing is true . . . it's all down to perception . . . I have fantastic memories, but everybody's memory is different, so they're just my memories . . . I know that Maurice and Robin would have had a different kind of memory. (Barry Gibb, *The Bee Gees: How Can You Mend a Broken Heart*, 2020)

The notion that convergence equals the truth appears so ingrained in thinking about behavior that its adopters traverse scholars with research training and lay individuals without such training. In this respect, this notion exemplifies features of *folk psychology* (Knobe & Mendlow, 2004), an area of study as old as Psychology itself (Haeberlin, 1916; Wundt, 1916). We have long known that folk psychological principles often exhibit limited explanatory power and fail to withstand empirical scrutiny (Goldenweiser, 1912). In this respect, we already demonstrated some of the epistemological limitations of the MTMM paradigm. We also articulated how CONTEXT's own epistemological process corrects for these limitations (Table 2).

In this section, we highlight fundamental flaws with use of the MTMM paradigm when assessing youth mental health. In particular, the MTMM paradigm's assumptions guide users to engage in research practices that are not only antithetical to use of structurally different informants, but they also bar users from subjecting these assumptions to empirical scrutiny. The MTMM paradigm's assumptions act as a "force field" between users and research practices that would facilitate *falsifiability* (Popper, 1962). Although validation paradigms are not scientific theories (see, Levy, 1969), a prospective user ought to nonetheless be able to

empirically scrutinize their fit to a given set of data conditions. In fact, we cite below studies that engaged in research practices that contributed to drawing incorrect conclusions from research findings, and we draw connections between these errors in decision-making and the MTMM paradigm's assumptions. Our focus is on three sets of practices: (a) selecting informants, (b) selecting analytic procedures, and (c) designing studies. Along the way, we describe how CONTEXT guides users toward engaging in fundamentally distinct research practices from those guided by the MTMM paradigm, and how CONTEXT's guidance represents a closer alignment with the use of structurally different informants and the data conditions that typify youth mental health.

Selecting Informants

The "origin story" of how CONTEXT and the MTMM paradigm each inform fundamentally distinct research practices regarding selecting informants—both in terms of how many informants as well as which informants to select—traces back to our earlier discussion about the origins of psychological measurement. We summarize these distinctions in Table 3. Let us take a second look at Edgeworth (1888), namely at the idea that no one rater provides a perfect score, and at the same time, none provides a rating with value unique to itself. We discussed how Garner et al. (1956) extended this idea to interpret findings from different experiments designed to address the same hypothesis, and that one can draw a throughline between these ideas and core assumptions underlying the MTMM paradigm. If the MTMM paradigm originates from these sources, how might it inform research practices surrounding selecting informants?

The answer to this question begins with a fact: The MTMM paradigm draws inspiration from conceptual models that ascribe no unique value to individual data points (e.g., findings from separate experiments, raters of a single assessment target). Edgeworth (1888), Garner et al. (1956), and Campbell and Fiske (1959) all focused on attaining precise estimates of common variance, and in fact, created paradigms that, ideally, would result in data conditions with as much common variance—and as little unique variance—as possible. A validation paradigm informed by such concepts—where common variance is what matters—must necessarily view individual data sources as either exclusively or primarily the means by which one models common variance. Thus, the guidance on selecting informants that the MTMM paradigm provides users logically leads them down one path: Select more than one informant, but which informants one selects does not matter. In fact, one can easily observe this guidance in both statements by adherents of the MTMM paradigm and

Table 3. Differences between CONTEXT and the MTMM paradigm in guidance on research practices surrounding selection of informants.

Validation Paradigm	Central Focus of Paradigm	Treatment of Individual Informants' Reports that Logically Results from Focus	Research Practice That Logically Extends from Focus	Relevant Citations
MTMM	Attain precise estimates of common variance	If the means for selecting informants is to achieve the end of estimating common variance, then no one informant's report has value unique to itself; each report simply reflects "error" around a measure of central tendency (i.e., common variance)	Select multiple informants for the sole purpose of maximizing precision in estimating common variance; the actual informants do not matter (e.g., could be structurally different, random raters, parents, teachers, youth, peers), so long as one attains precise estimates of common variance	Bauer et al. (2013); Edgeworth (1888); Fiske and Campbell (1992); Roberts and Caspi (2001)
CONTEXT	Strike a balance between attaining precise estimates of both common variance and domain-relevant unique variance (i.e., informant discrepancies)	If the means of the multi-informant approach involve collecting reports from structurally different informants, and the informant discrepancies resulting from this approach ought to contain domain-relevant information, then determining which specific informants will provide reports is as an important decision as determining the content of the measures the informants will complete	Strategically select informants who will produce both common variance and informant discrepancies that contain domain-relevant information; the actual informants are instrumental to optimizing measurement validity	De Los Reyes et al. (2009); Kraemer et al. (2003); Makol et al. (2020)

CONTEXT = Classifying Observations Necessitates Theory, Epistemology, and Testing. MTMM = Multi-Trait Multi-Method Matrix.

research practices leveraged by scholars who address their questions with MTMM-informed analytic procedures.

Consider two examples. The first example comes from the developers of the MTMM paradigm themselves (Fiske & Campbell, 1992), who articulated their perspective on the ideal matrix: "one that is based on fairly similar concepts and plausibly independent methods" (p. 393). The developers of the MTMM paradigm never made any other mention of the characteristics of the methods used, just that users of the paradigm (a) use more than one method and (b) that the methods are "plausibly independent." Again, the data sources reflect the means by which one models common variance.

A second example highlights another MTMM-informed strategy for selecting informants. Arguably, this method focuses on maximizing common variance, even at the expense of who the informants are, or even in the absence of a sound rationale for which informants a user inputs into the model. Earlier, we discussed the MTMM-informed structural model developed by Bauer et al. (2013) designed to "purge" integrative scores of the effects of informant discrepancies. What we did not mention was how they demonstrated use of the model. As with all models, the viability of their model is dictated by its fit with the data conditions to which it is applied. Thus, to demonstrate the fit of their model, Bauer and colleagues leveraged mother and father reports of youth emotions, recruited from a large sample of parents who provided reports for youth between the ages of 2 and

18. Importantly, the authors provided little rationale for which specific informants they selected, other than to note that they required a set of informants who could report about youth emotions across a large age range of youth (see p. 484 of Bauer et al., 2013). However, because the authors' modeling approach was informed by the MTMM paradigm, we can surmise that at least one other factor was at play.

In the case of the informants selected by Bauer et al. (2013), it is perhaps no coincidence that, relative to other pairs of informants (e.g., parent-teacher, teacher-youth, parent-youth), none display larger levels of correspondence than mothers and fathers (see, Achenbach et al., 1987; Duhig et al., 2000; De Los Reyes et al., 2015, 2019a). Thus, if someone wanted to maximize the amount of common variance estimated in a structural model, no other pair of informants beyond mothers and fathers would encounter greater success in doing so. Importantly, no other rationale linked to selecting informants would justify selecting only mothers and fathers to assess an internal process like youth emotions. In fact, every set of recommendations available to the authors at the time (i.e., pre-2013) would have made the same recommendations that researchers have available to them now. Specifically, it is established practice (and perhaps ethically incumbent) in this particular sub-field that the informants used to assess youths' emotions should include youth self-reports (for a pre-2013 review,

see, Hunsley & Mash, 2007). Further, two decades ago Kraemer et al. (2003) successfully integrated youth, parent, and teacher reports when assessing youth internalizing concerns, even in samples of youth as young as four years of age. Here too, it appears that the MTMM paradigm guides users to see the use of multiple informants as a means of modeling common variance, arguably at the expense of violating prevailing recommendations for which informants to use.

As we summarize in Table 3, the guidance that CONTEXT provides users on research practices in selecting informants could not be more distinct from that of the MTMM paradigm. With CONTEXT, not only do the actual informants matter, but the theory and evidence that informed the development of CONTEXT indicates that the informants matter as much—and sometimes, arguably more than (e.g., see, Makol et al., 2019)—the domains about which researchers solicit their reports. Indeed, taken to its logical conclusion, the notion of situational specificity dictates that a researcher leverages structurally different informants who have the expertise to provide them with a holistic view of the youth the researcher seeks to assess. One cannot achieve that aim unless they strategically select informants; the informants have to matter. Each of the studies we summarize in our illustration below of CONTEXT's features make this guidance abundantly clear (e.g., Becker-Haimes et al., 2018; Lerner et al., 2017; De Los Reyes et al., 2009; Makol et al., 2020). Taken together, CONTEXT and the MTMM paradigm each provide diametrically opposed guidance on research practices surrounding selecting informants.

Selecting Analytic Procedures

The core assumptions of CONTEXT and the MTMM paradigm impact at least two other research practices. We previously clarified differences between their assumptions: The latter emphasizes common variance, whereas the former emphasizes both common variance and domain-relevant unique variance. What are the implications of these assumptions for selecting analytic procedures that account for domain-relevant unique variance? Put simply, a researcher cannot possibly rely on the MTMM paradigm for guidance on such procedures, particularly when modeling data from structurally different informants. Indeed, if use of structurally different informants produces domain-relevant unique variance, how could the MTMM—a paradigm that assumes such variance does not exist—provide users with guidance on modeling it?

In Table 4, we summarize the distinctions between CONTEXT and the MTMM paradigm as to the guidance they provide on practices relevant to selecting analytic procedures. Understanding these distinctions involves stipulating the prerequisites to considering common variance (i.e., informant agreement) and domain-relevant unique variance (i.e., informant discrepancies) as unique predictors. When using structurally different informants to predict a criterion variable (e.g., treatment response, observed behavior on a laboratory task), modeling both common variance and domain-relevant unique variance means that you allow both to explain variance in the criterion. To start, it is reasonable to assume that Converging Operations reflect a reality: Portions of reports from structurally different informants display commonalities in how they predict the criterion. By construction, this assumption holds that informant agreement reflects domain-relevant information. Beyond this variance, we previously highlighted

Table 4. Differences between CONTEXT and the MTMM paradigm in guidance on research practices surrounding selection of analytic procedures.

Validation Paradigm	Central Focus of Paradigm	Logical Result of Focus	Research Practices That Logically Extend from Focus	Relevant Citations
MTMM	Attain precise estimates of common variance, because informant discrepancies reflect measurement confounds	If informant discrepancies reflect measurement confounds, then there is no legitimate reason for modeling them, other than creating estimates of error for modeling purposes	Use of analytic procedures that cannot estimate the incremental value of informant discrepancies	Alvarez et al. (2021); Murray et al. (2021); T.M. Olino et al. (2018); Russell et al. (2016); van der Ende et al. (2020)
CONTEXT	Strike a balance between attaining precise estimates of both common variance and unique variance, because informant discrepancies may reflect domain-relevant information	If informant discrepancies may reflect domain-relevant information, and common variance estimates are likely domain-relevant as well, then one must leverage analytic procedures that allow one to determine the relative value of both common and unique variance for the purpose of scholarly inquiry (e.g., characterizing behavior, predicting outcomes, theory testing)	Use of analytic procedures that facilitate testing for the incremental value of informant discrepancies	Becker-Haimes et al. (2018); De Los Reyes et al. (2009); Laird and De Los Reyes (2013); Makol et al. (2019, 2020)

CONTEXT = Classifying Observations Necessitates Theory, Epistemology, and Testing. MTMM = Multi-Trait Multi-Method Matrix.

that in a given sample of structurally different informants' reports of youth mental health, one could leverage regression commonality analyses to estimate overlapping variance from these reports in prediction of a common criterion variable (see, Cohen et al., 2003), and that these analyses facilitate decomposing common and unique variance in such prediction (see, De Los Reyes et al., 2015). By definition, such analyses reveal that in any one sample of multi-informant predictors, one will likely discover a non-zero estimate of common variance and a non-zero estimate of domain-relevant unique variance. Thus, it becomes crucial when modeling data from structurally different informants to leverage analytic procedures that allow for not only demonstrating the utility of common variance in predicting a criterion, but also the incremental value of domain-relevant unique variance for predicting that same criterion.

As with guidance on selecting informants, examples in the published literature highlight distinctions between CONTEXT and the MTMM paradigm on research practices regarding selection of analytic procedures. For instance, van der Ende et al. (2020) recently tested an MTMM-informed structural model to determine the predictive utility of multi-informant assessments of youth mental health in predicting mental health outcomes in adulthood. To examine whether informant discrepancies contribute to prediction, the authors incorporated estimates of the measured differences between informants' reports (i.e., *difference scores*; Laird, 2020). However, it has been long-known that use of difference scores precludes the ability to demonstrate the incremental predictive value of informant discrepancies (see, Edwards, 1994; Laird & Weems, 2011). To address this concern, van der Ende and colleagues used a procedure that, unlike difference scores, accounts for incremental value (i.e., polynomial regression; Laird & De Los Reyes, 2013). Importantly, whereas analyses relying on difference scores produced evidence supporting the predictive value of informant discrepancies, polynomial regression analyses all indicated null effects. As the authors relied on a relatively large sample of nearly 600 youth, statistical power issues were not likely to have contributed to differences in the findings of these two procedures. Although the polynomial regression analyses indicated null effects, the authors nonetheless concluded that informant discrepancies "contributed to the prediction of adult internalizing and externalizing DSM disorders" (p. 343). One could argue that drawing such a conclusion stemmed from overreliance on the MTMM paradigm to select analytic procedures, even at the expense of drawing erroneous conclusions from research findings.

As another example of an MTMM-informed analytic procedure that precludes testing for the incremental value of domain-relevant unique variance, consider the measurement invariance techniques described previously. We previously mentioned that assumptions underlying use of the technique require users to treat the detection of informant discrepancies as synonymous with detecting measurement confounds. What we have yet to articulate is exactly what users of these techniques do with the evidence gleaned from these techniques. Users of these techniques leverage the evidence gleaned from tests of measurement invariance to identify items that function differentially across measurement conditions (see, Meredith, 1993; Millsap, 2011; Osterlind & Everson, 2009). In the case of informant discrepancies, these would consist of items that function differently between informants' reports on parallel instruments. In essence, evidence from tests of measurement invariance inform revising item content, such that the only items that remain on the instrument are those that demonstrate convergence between informants on item functioning. Consider the implications if these techniques were misapplied. What if they were applied to detect measurement conditions that the user wrongfully assumed reflect measurement confounds? By construction, using this technique under these circumstances precludes the ability of testing for domain-relevant unique variance. Yet, this currently occurs in youth mental health research, where researchers have applied these techniques to detect informant discrepancies (e.g., Alvarez et al., 2021; Florean et al., 2022; Murray et al., 2021; Russell et al., 2016; Schiltz et al., 2021; T.M. Olino et al., 2018). The logical result of using these techniques is to remove items that solicit domain-relevant responses from informants. Within data conditions in which informant discrepancies contain domain-relevant information, application of these techniques cannot possibly optimize measurement validity.

Taken together, the core assumptions underlying the MTMM paradigm prevent it from accurately guiding users to select analytic procedures that facilitate understanding and modeling the most frequently observed data conditions when using structurally different informants. In contrast, CONTEXT guides users to accurately select analytic procedures that fit the data conditions that typically result from use of structurally different informants. Indeed, unlike the MTMM paradigm, a core feature of CONTEXT involves assuming that structurally different informants' reports may produce both common variance and domain-relevant unique variance. Consequently, CONTEXT guides users to consider analytic procedures that fit that assumption, or procedures that facilitate characterizing both of these kinds of variance. With regard to informant discrepancies research, one issue with articulating exactly which analytic procedures allow users to

characterize both common variance and domain-relevant unique variance has to do with the history of this research as it has manifested in youth mental health. Historically, much of the work in this area has been hindered by use of procedures that cannot achieve these aims (e.g., difference scores; see, Laird & De Los Reyes, 2013; Laird & Weems, 2011; De Los Reyes & Ohannessian, 2016). Only recently have researchers developed and/or implemented procedures that correct for this limitation of prior work.

Here, we briefly describe three of these procedures. The first procedure was one we mentioned previously, polynomial regression (Edwards, 1994). Using this procedure, researchers have examined structurally different informants' reports as predictors of youth outcomes (for reviews, see, Laird, 2020; De Los Reyes & Ohannessian, 2016). Relevant to this discussion, the procedure facilitates detecting variance explained by each individual informant's report. Over-and-above the variance explained by these reports, statistical interactions calculated between informants' reports allow researchers to characterize specific reporting patterns. These patterns include instances in which informants agree in their reports (i.e., common variance) and disagree in their reports (i.e., unique variance; see, De Los Reyes et al., 2019b). When applied to multi-informant data (e.g., Lerner et al., 2017; Makol et al., 2019), person-centered models such as latent class analysis reflect a second analytic procedure (Bartholomew et al., 2002). Similar to polynomial regression, latent class analysis allows researchers to characterize subgroups of informants' reports about a sample of youth. Here as well, this procedure facilitates detecting instances in which informants' reports agree (i.e., yield similar findings) as well as disagree (i.e., yield discrepant findings). A third example can be found in factor-analytic work by Kraemer et al. (2003), who developed a form of principal components analysis that facilitates characterizing multi-informant data in terms of a common variance component and multiple unique variance components. Below, we illustrate studies that have recently demonstrated how each of the analytic procedures we described facilitate characterizing common variance and domain-relevant unique variance germane to using structurally different informants. In sum, CONTEXT and the MTMM paradigm each provide fundamentally distinct guidance on research practices surrounding selection of analytic procedures.

Designing Studies

We previously described CONTEXT-informed analytic procedures as those that allow users to "characterize" multi-informant data in terms of the common variance (i.e., informant agreement) and unique variance (i.e., informant discrepancies) they produce. To merely assume

these characterizations reflect something valid and domain-relevant would reflect a well-known decision-making error in the psychometric literature known as the *naming fallacy*: "Just because a factor is named does not mean that the hypothetical construct is understood or even correctly labeled" (Kline, 2016, p. 300). If users of CONTEXT relied only on these characterizations to, for instance, deem an instance of informant discrepancies as domain-relevant, then they would be engaging in the same kinds of confirmation biases that users of the MTMM paradigm engage in when assuming that all informant discrepancies reflect measurement confounds (Table 2). Making claims about what informant discrepancies represent requires carefully conducted studies that link characterizations of informant discrepancies—or informant agreement in the case of common variance—to domain-relevant criterion variables. The phenomena these criterion variables were constructed to reflect, along with their methodological characteristics, dictate the claims researchers are justified in making. In this section, we demonstrate how CONTEXT and the MTMM paradigm produce differential guidance for users on at least one more set of research practices, namely on designing studies.

In Table 5, we summarize the distinctions between CONTEXT and the MTMM paradigm as to the guidance they provide on practices relevant to designing studies. As before, these distinctions stem from the "origin stories" of the two paradigms (see, Tables 1 and 2). If the MTMM paradigm assumes that informant discrepancies signal measurement confounds, then the logical extension of this assumption is that measurement confounds also account for discrepancies between informants' reports and data from non-informant modalities, such as trained observers' behavioral ratings or indices drawn from a performance-based task. Not surprisingly, discrepancies between informants' reports and non-informant modalities robustly manifest across multi-modal assessments of youth mental health (e.g., Clarkson et al., 2020; De Los Reyes et al., 2020; Meyer et al., 2001). When viewed through the lens of the MTMM paradigm, it is also not surprising to see a plethora of instances in which researchers have used these discrepancies to cast doubt on the veracity of informants' reports, non-informant methods, or both (for reviews, see, Dunning et al., 2004; De Los Reyes et al., 2019a). This perspective has dire consequences for the decisions that researchers make when designing studies. In particular, this perspective may guide researchers toward thinking that either the informants or modalities chosen do not matter, or that they each inhibit one's ability to observe converging findings.

Consider some examples. In youth mental health, researchers typically construct MTMM-informed

Table 5. Differences between CONTEXT and the MTMM paradigm in guidance on research practices surrounding study design.

Validation Paradigm	Core Assumption Regarding Informant Discrepancies	Logical Result of Focus	Research Practices That Logically Extend from Focus	Relevant Citations
MTMM	All informant discrepancies reflect measurement confounds	If informant discrepancies reflect measurement confounds, then it is logical to assume that discrepancies between informants' reports and other non-informant modalities (e.g., observed behavior) also reflect measurement confounds	Use of study designs that cannot distinguish informant discrepancies that reflect measurement confounds from those that reflect domain-relevant information	Bauer et al. (2013); De Los Reyes et al. (2019a); Dunning et al. (2004); van der Ende et al. (2020); Watts et al. (2022)
CONTEXT	At least some informant discrepancies contain unique domain-relevant information	If the means of the multi-informant approach involve collecting reports from structurally different informants, and the informant discrepancies resulting from this approach ought to contain domain-relevant information, then detecting this information requires careful attention to selecting non-informant modalities to serve as domain-relevant validity indicators; selecting validity indicators is as an important decision as selecting the informants	Use of study designs that facilitate distinguishing informant discrepancies that reflect measurement confounds from those that reflect domain-relevant information; these designs involve "matching" the structural characteristics of the informants providing reports with the structural characteristics of the non-informant modalities used to construct validity indicators	Al Ghriwati et al. (2018); De Los Reyes et al. (2009, 2013c, 2022c); Makol et al. (2019, 2020)

CONTEXT = Classifying Observations Necessitates Theory, Epistemology, and Testing. MTMM = Multi-Trait Multi-Method Matrix.

structural models focused on estimating common variance among indicators, and the indicators chosen most often consist of multi-informant reports completed on parallel instruments (for a review, see, Eid et al., 2008). Recent work indicates that these multi-rater models not only produce informant discrepancies, but also that these same models cannot distinguish informant discrepancies that reflect measurement confounds from those that reflect domain-relevant information (Watts et al., 2022). This is largely due to the fact that these models often only consist of multiple informants' reports, not only to construct estimates of common variance on the same domains (e.g., internalizing, externalizing) but also in the construction of covariates or criterion variables. This reliance on the same modality to assess all domains means researchers who rely on this study design cannot rule out the possibility that their findings arose simply because of *criterion contamination* or rater-specific variance (see Garb, 2003). This was true of the van der Ende et al. (2020) study described previously, which relied on self-reports to estimate informant discrepancies (i.e., among parents, teachers, and youth) as well as to assess the outcomes used to estimate the predictive utility of these discrepancies.

The implications of these issues are perhaps more serious for researchers conducting MTMM-informed studies of factors that they hypothesize might account for informant discrepancies. With the Bauer et al. (2013) study, we previously mentioned that their MTMM-informed structural model carried the assumption that all informant discrepancies reflect measurement confounds. What we have yet to mention is that the model also included a feature that allows users to input as covariates factors that are hypothesized to account for rater biases. Because the model was

informed by the depression-distortion hypothesis, Bauer and colleagues demonstrated this feature by inputting various mental health characteristics of the mothers and fathers to "stand in" as these biasing factors. The authors assessed these characteristics via self-report assessments of substance use history and history of mood or personality disorders, which mothers and fathers provided on structured diagnostic instruments administered by lay interviewers. Because assessments of informant discrepancies and these covariates involved reports of domains of interest from mothers and fathers, issues surrounding criterion contamination apply as much with this study as we previously described for the study by van der Ende et al. (2020). Yet, one other study design decision by Bauer and colleagues compounds the issues raised, namely how they treated the biasing factors. Within the design of this study, the authors assumed that any variance in informant discrepancies explained by factors such as parental substance use and mood disorder history reflected a rater bias. Yet, we also know from decades of research that these same factors are implicated in the development and maintenance of youth mental health concerns, including the emotional functioning domains about which the mothers and fathers in the sample provided reports (see also, Goodman et al., 2011; De Los Reyes & Makol, 2022). The authors made no attempt to decompose variance in these factors in terms of domain-relevant associations with youth emotions (i.e., the domain rated by mothers and fathers) and rater biases. Thus, in this study, any links observed between informant discrepancies and parental mental health could simply reflect a combination of rater-specific variance and long-known effects in the youth mental health literature (i.e., that parental functioning impacts youth functioning).

The features of CONTEXT guide users to make fundamentally distinct decisions regarding study design, relative to the guidance the MTMM paradigm provides its users. We previously mentioned that CONTEXT guides users to assume that (a) which informants they select is as important as the domains about which they provide reports and (b) the analytic procedures chosen to characterize informant discrepancies must facilitate understanding common variance as well as domain-relevant unique variance. The guidance CONTEXT provides users regarding study design allows users to capitalize on these research practices and draw valid inferences from study findings. This guidance starts with the composition of criterion variables used to validate characterizations of informant discrepancies and/or informant agreement.

To detect domain-relevant information in informant discrepancies and/or informant agreement, users of CONTEXT pay special attention to the key characteristics of criterion variables, namely their underlying methodology and the connections between this methodology and the structural differences among informants. Indeed, these aspects of criterion variables dictate the inferences a researcher could justifiably draw from studies seeking to validate characterizations of informant discrepancies and/or informant agreement. Consider criterion contamination. CONTEXT guides users to leverage criterion variables that facilitate ruling out the confound that significant links between characterizations of discrepancies and/or agreement and criterion variables resulted from using the same measurement source for all variables. In fact, in Table 5 we cite studies that leveraged a range of criterion variables to test the validity of characterizations of informant discrepancies and/or informant agreement, including independent observers' behavioral ratings, physiological indices, and scores from performance-based tasks.

Addressing the issue of criterion contamination is essential in drawing valid inferences from studies purporting to demonstrate links between characterizations of informant discrepancies and/or informant agreement and criterion variables. However, this element of guidance that CONTEXT provides its users does not fully address the naming fallacy issue described previously. Beyond addressing issues with criterion contamination, CONTEXT provides users with two more explicit pieces of guidance relevant to addressing the naming fallacy. One piece of guidance directly relates to core assumptions surrounding use of structurally different informants.

Specifically, we previously mentioned that when informants structurally differ from each other, these structural differences ought to reflect domain-relevant phenomena, such as parents observing youth at home and teachers observing youth at school. What we have yet to discuss is the implications this assumption has for constructing criterion variables: The criterion variables must demonstrate domain-relevant, structural characteristics of their own. Consequently, in Table 5 the studies cited in relation to CONTEXT not only used criterion variables that were independent in modality relative to the informants' reports, but they also reflected indices that displayed domain-relevant structural characteristics. Examples of these criterion variables include school observations, laboratory tasks designed to reflect home-based family interactions, tasks designed to index clinical impairments linked to social communication, performance-based tasks designed to index emotion recognition, tasks designed to index physiological functioning, and treatment-related decisions made by trained clinical personnel.

A second piece of guidance derived from CONTEXT's features concerns the explicit linkage of the structural characteristics of criterion variables with the structural differences among the informants providing reports. It is this second piece of guidance that directly addresses issues germane to the naming fallacy described previously. That is, do characterizations of informant discrepancies and/or informant agreement relate to or predict variations on these criterion variables in domain-relevant ways? For instance, consider a task that was designed to index how consistently a youth undergoing evaluation displayed a specific behavior across home and school contexts (e.g., oppositional behavior). In a study examining parent and teacher reports in relation to such a criterion variable, do discrepancies between reports signal context-specific behavior on the criterion variable, namely behavior that manifests in the home-based portion of the task but not the school-based portion, or vice-versa? Conversely, does agreement between reports signal cross-contextual behavior on the criterion variable, namely behavior that manifests across both the home- and school-based portions of the task? By testing links between characterizations of informant discrepancies and/or informant agreement and domain-relevant criterion variables, users of CONTEXT can directly test the validity of these characterizations and thus avoid decision-making errors linked to criterion contamination and the naming fallacy.²

²CONTEXT allows for this process to occur in both directions, such that multi-informant data could be used to validate discrepancies observed on non-informant modalities. Consider a researcher who designed a battery of context-specific tasks to index domain-relevant unique variance (i.e., discrepancies between indices reflecting behavior in different social contexts). The researcher would be incorrect to assume, absent validation testing, that between-task discrepancies reflect domain-relevant information. To conduct the validation tests that would justify such an assumption, the researcher could use as a criterion measure indices of informant discrepancies for which prior CONTEXT-informed research had already demonstrated reflect domain-relevant information.

Summary Comments

We provided an overview of the key features of CONTEXT, namely the guidance it provides users on theory, epistemology, and testing as it pertains to measurement validation using structurally different informants' reports of youth mental health. On the surface, it would be reasonable to argue that CONTEXT simply guides users to engage in sound research practices. To address this argument, we carefully articulated the core assumptions of the MTMM paradigm and how these assumptions guide users to make decisions regarding use of structurally different informants' reports. We also cited actual decisions regarding research practices users of the MTMM paradigm made, in line with this guidance. MTMM-informed guidance and practices fundamentally differ from the guidance and practices that stem from CONTEXT's features (Tables 1–5). Thus, we revealed a “ground truth” about use of structurally different informants: What one considers a “sound research practice” is dictated by the validation paradigm that informed the practice. Scholars who leveraged the MTMM paradigm to address their aims made decisions they clearly thought reflected sound research practices. As did the peers who reviewed their work. Importantly, the MTMM-informed work as cited in Tables 1–5 appeared in some of the most influential peer review journals in Psychology, including *Psychological Methods*, *Journal of Abnormal Psychology*, and *Psychological Assessment*. Yet, we also articulated how these practices do not align with the data conditions that typify use of multi-informant approaches to assessing youth mental health. In turn, we described how CONTEXT guides users to engage in research practices that do align with these same data conditions. This raises the question: What evidence exists to indicate that researchers can carry out studies that align with CONTEXT's features and the guidance it provides users? We address this question in the section that follows.

Illustrations of Research that Align with the Features of CONTEXT

A widely read and widely cited *Psychological Bulletin* article is one that researchers can use. Is our article being used as a crutch? Matrices published today continue to be about as unsatisfactory as those published more than 33 years ago. Editors and readers are accepting matrices showing limited convergence or discrimination, or both . . . The only published ones that are fairly good are those using rather similar methods or quite disparate traits or other attributes or both. We have yet to see a really good matrix . . . (Fiske & Campbell, 1992; p. 393)

In their article, *Citations do not solve problems*, Fiske and Campbell (1992) had the luxury of time and data—over 30 years of research—to form a perspective on the MTMM paradigm's impact on measurement validation. As the above-referenced quote states, they submitted a fairly pessimistic evaluation, as indicated by the quality of matrices developed over the course of several decades of researchers' use of the MTMM paradigm. In a very real sense, Fiske and Campbell “tipped their hand,” so to speak. Their statements connoted their view that matrices informed by the MTMM paradigm had never lived up to the paradigm's assumptions. We also previously noted that these issues exist to this day (see, Watts et al., 2022).

Similar to the MTMM paradigm, with CONTEXT we had the luxury of both prior theory and evidence to guide development of its features. However, unlike the MTMM paradigm, we can already see studies that not only align with CONTEXT's features, but are also making significant inroads into optimizing measurement validity. In turn, these studies appear poised to inform the next generation of research on key areas of work in youth mental health. Along these lines, in this section we illustrate research in three areas: (a) characterizing youth mental health (Table 6), (b) predicting youth mental health outcomes (Table 7), and (c) testing theories about links between informant discrepancies and youth mental health (Table 8).

Characterizing Youth Mental Health

Researchers frequently leverage multi-informant approaches to assessment to understand a youth's mental health functioning at a given moment in time, and for use in research on such issues as diagnosis, classification, and treatment planning. If the informant discrepancies produced when using structurally different informants reveal domain-relevant information, then they ought to improve the validity of scores taken from multi-informant data. Under these circumstances, use of structurally different informants ought to boost the ability of researchers to draw valid inferences from findings on studies that leverage multi-informant data. Use of these multi-informant data ought to facilitate characterizing domain-relevant links between youth mental health and clinical phenomena. Examples of these phenomena include the contexts in which mental health concerns manifest. In line with these notions, prior studies that aligned themselves with CONTEXT's features illustrate the ability of multi-informant data—namely patterns of informant discrepancies and/or informant agreement—to facilitate characterizing links between youth mental health concerns and the contexts in which these concerns manifest.

Table 6. Examples of how features of CONTEXT facilitate using informant discrepancies to optimize characterizations of youth mental health concerns.

Authors	Informants	Domains Informants Reported	Analytic Procedure Used to Characterize Informant Discrepancies	Criterion Variables
De Los Reyes et al. (2009)	Parent, Teacher	Disruptive behavior	Latent class analysis	Behavior rated by independent observers, based on interactions between the child and adult authority figures on a controlled laboratory task
De Los Reyes et al. (2022c)	Parent, Teacher, Child	Psychosocial impairments	Kraemer et al. (2003) Satellite Model	Behavior rated by independent observers, based on a combination of school observations and parent-child interaction tasks
Lerner et al. (2017)	Parent, Teacher	Autism spectrum	Latent class analysis	Trained raters' evaluations of clinical severity, based on observations of client during a standardized diagnostic observation task
Makol et al. (2020)	Adolescent, Parent, Peer confederate	Social anxiety	Kraemer et al. (2003) Satellite Model	Behavior rated by independent observers, based on interactions between the adolescent and personnel trained to simulate a same-age, unfamiliar peer

CONTEXT = Classifying Observations Necessitates Theory, Epistemology, and Testing. MTMM = Multi-Trait Multi-Method Matrix.

Leveraging Multi-Informant Data to Detect Contextual Variations in Youth Mental Health

In a demographically diverse sample of 327 preschoolers, De Los Reyes et al. (2009) examined patterns of informant discrepancies and informant agreement among parents' and teachers' reports about preschoolers' disruptive behavior. Researchers identified preschoolers for whom reports varied as to whether they displayed relatively high levels of disruptive behavior. Consistent with diagnostic procedures in the DSM-IV (American Psychiatric Association, 2000), a report was deemed "relatively high" if the informant endorsed three or more symptoms of a disruptive behavior disorder (i.e., conduct disorder, oppositional defiant disorder). Researchers then created four groups of preschoolers to characterize patterns of informant discrepancies and informant agreement on reports of high levels of disruptive behavior based on: (a) both parent and teacher report, (b) parent but not teacher report; (c) teacher but not parent report, and (d) neither parent nor teacher report. These characterizations operate in ways akin to the latent class analysis procedures described previously. In fact, the groups created by De Los Reyes et al. (2009) have been replicated by a number of investigative teams, who conducted formal tests of their presence using latent class analyses of parent and teacher data about youth externalizing concerns (e.g., Fergusson et al., 2009; Makol et al., 2021; Sulik et al., 2017).

To validate characterizing patterns of informant discrepancies and informant agreement, De Los Reyes et al. (2009) administered an independent, cross-contextual behavioral task, the Disruptive Behavior Diagnostic Observation Schedule (DB-DOS; Wakschlag et al., 2010). This task yielded indices of the degree to which preschool children displayed observable signs of disruptive behavior with parental and non-parental adult

authority figures. The DB-DOS included a series of activities between the preschooler and an adult authority figure (i.e., parent, unfamiliar examiner), designed to "press" for the elicitation of disruptive behavior (e.g., frustration and compliance tasks). These activities were standardized across interaction partners. Thus, measured indices of observed disruptive behavior across these tasks could serve as proxies for how the child displayed such behaviors when interacting with their parent and a non-parental adult. Using independent observers' ratings of preschoolers' displays of maladaptive emotional (i.e., anger modulation) and behavioral (i.e., behavioral regulation) signs of disruptive behavior, researchers used latent class analysis to detect contextual variations in disruptive behavior.

De Los Reyes et al. (2009) made two crucial observations relevant to leveraging CONTEXT's features to facilitate characterizing youth mental health concerns. First, latent class analyses of DB-DOS data revealed that, like characterizations of the parent and teacher data, preschoolers varied in observed disruptive behavior. That is, each preschooler displayed one of four forms of disruptive behavior: (a) with both parental and non-parental adults, (b) with the parental but not non-parental adult; (c) with the non-parental but not parental adult, and (d) with neither parental nor non-parental adults. Second, characterizations of informant discrepancies and informant agreement related to these DB-DOS classes. Specifically, preschoolers who displayed observed disruptive behavior across parental and non-parental adults tended to have parents and teachers who agreed in reports of high levels of disruptive behavior. Conversely, preschoolers who displayed observed disruptive behavior when interacting with parental but not non-parental adults tended to have parents who reported disruptive behaviors that teachers did not

report, and vice versa for teachers' reports. In sum, the study by De Los Reyes and colleagues shows how researchers who align their work with CONTEXT's features can leverage data from structurally different informants' reports to detect both context-specific (i.e., informant discrepancies) and cross-contextual (i.e., informant agreement) displays of youth mental health concerns.

Leveraging Multi-Informant Data to Estimate Youth Mental Health in a Specific Context

Prior work illustrates the ability of CONTEXT's features to facilitate leveraging data from structurally different informants in a way that optimizes prediction of youth mental concerns that manifest in a specific context. This work may address a long-standing issue in assessment. Indeed, consider that researchers select informants based on structurally different aspects of their expertise in observing youth in specific contexts, signaling that youth may vary in where they display mental health concerns. However, many researchers believe that one "optimal" informant exists for assessing specific domains (e.g., teachers for ADHD and parents for anxiety; see, Loeber et al., 1990). These inconsistencies in logic trace back to the MTMM paradigm, in that its use has forced researchers to surmise that, if informant discrepancies signal measurement confounds, then that means that only some informants contribute valid reports (see also, De Los Reyes et al., 2013a). If this notion were true, then it would be logical, perhaps intuitive, to believe that a specific informant provides the "most valid" report about a specific youth mental health domain (e.g., teacher > parent and youth when assessing ADHD). Studies that align with CONTEXT's features facilitate challenging this notion. In fact, CONTEXT reveals the question about who the "optimal informant" is for assessing domains like ADHD or anxiety for the "trick question" that it is. If a researcher strategically selects the multiple informants who will contribute reports to assess a domain, then there is no "optimal informant" to assess that domain.

Recent work leveraged procedures developed by Kraemer et al. (2003) that capitalize on the incremental value of integrating multi-informant data from structurally different informants, over-and-above an "optimal informant" for predicting domain-relevant criterion variables. The procedures (hereafter referred to as the *Satellite Model*) assume that an individual informant's data function like the data contributed by a single satellite within a global positioning system. That is, each informant within a "satellite array" of informants provides incrementally valuable data about a target youth's

mental health, insofar as that informant contributes data that cannot be obtained from the other informants in the array. To ensure that each informant provides incrementally valuable data, Kraemer and colleagues call for an *a priori* selection of domains that one predicts will logically produce discrepant estimates of mental health, in ways akin to latitude and longitude location estimates in a global positioning system. In such a system, one optimizes the accuracy of location estimates (i.e., for a target building or person) by strategically placing satellites at disparate, coordinated locations, such that the satellites collectively triangulate on the target's location. Based on meta-analytic data available to them at the time (i.e., Achenbach et al., 1987), Kraemer and colleagues surmised that the metaphorical "latitudes and longitudes" for assessing youth mental health were the contexts in which informants observed youth (i.e., home, non-home) and the perspectives from which they observed youth (self, other). By strategically selecting informants who fell into disparate points on these context and perspective domains, Kraemer and colleagues posited that researchers could leverage factor analytic techniques historically focused on synthesizing variability of items on survey instruments (i.e., principal components analysis [PCA]; Duntzman, 1989; Nunnally & Bernstein, 1994) to instead synthesize summary score data from structurally different informants' reports. In essence, accurately triangulating on youth mental health estimates should translate into a PCA solution yielding three components reflecting variability among informants in their contexts and perspectives of observation, along with a *trait* component that reflects youth mental health concerns that manifest consistently across informants' contexts and perspectives. Integrated scores from the *trait* component of the Kraemer et al. (2003) *Satellite Model* should optimize the validity of scores taken from multi-informant assessments. This is because such an estimate capitalizes on the collective expertise of each of the informants in the "satellite array." If so, then a conservative test of this notion would involve leveraging this approach to test links between structurally different informants' reports and data from a domain-relevant criterion variable designed to reflect behavior as it manifests in a specific context. That is, a context for which only a subset of informants could directly observe. Makol et al. (2020) conducted just such a conservative test, in a sample of 127 adolescents. Makol and colleagues collected social anxiety survey reports from adolescents and parents, as is typical in multi-informant assessments of social anxiety (see, Hunsley & Mash, 2007). Adolescents and parents provided their reports within a battery of mental health surveys. Following completion of this survey battery,

adolescents participated in the Unfamiliar Peer Paradigm, a set of social interaction tasks designed to simulate how adolescents interact with same-age, unfamiliar peers (for a review, see, Cannon et al., 2020). The research personnel who “stood in” as these unfamiliar peers (i.e., *peer confederates*) also served as the third informant who, along with parents and adolescents, provided survey reports on a parallel set of instruments (see, Deros et al., 2018). In essence, these peer confederates used the expertise they developed from interacting with adolescents within the Unfamiliar Peer Paradigm to complete parallel versions of the social anxiety surveys completed by parents and adolescents. To construct the criterion variable (i.e., anxiety displayed within the Unfamiliar Peer Paradigm), trained independent observers made ratings of adolescent social anxiety based on archived videos of these tasks, using a well-established coding scheme (Glenn et al., 2019). Importantly, whereas independent observers received extensive training to make their behavioral ratings, peer confederates received no training and as such, made their reports using the same instructions parents and adolescents received when completing their own reports on the parallel survey instruments.

An additional element of the study design in Makol et al. (2020) further enhances the interpretability of their tests. Specifically, the three informants each completed two separate sets of social anxiety survey instruments. This element of the study design allowed the authors to use one set of informants’ reports to integrate data using the Satellite Model approach, and the other set for use as control variables for their tests. That is, Makol and colleagues could use this second set of reports in tests of the integrated *trait* score based on the Satellite Model, relative to the individual informant’s reports. This second set of reports could also be used for tests of this integrated *trait* score, relative to an integrated score where the user assumes that informant discrepancies reflect measurement confounds (e.g., composite scoring). Thus, the implication of the null hypothesis (i.e., informant discrepancies reflect measurement confounds) is that strategically selecting structurally different informants’ reports results in no improvements to prediction. Concretely, data consistent with the null hypothesis would involve the *trait* score and composite score yielding similar findings (i.e., no incremental value of the *trait* score).

Makol et al. (2020) observed three components consistent with the Satellite Model. Specifically, analyses revealed that the *trait* component accounted for most of the variance (58.5%), followed by *context* (24.1%), and *perspective* (17.4%). Makol and colleagues tested the incremental value of the *trait* score in predicting

independent observers’ ratings of adolescent social anxiety relative to (a) individual informant’s reports and (b) a composite score of these reports. The *trait* score demonstrated incremental value, relative to both of these approaches. In this respect, we highlight three aspects of their findings, in light of their relevance to CONTEXT. First, the *trait* score incrementally predicted independent observers’ ratings of adolescent social anxiety, over-and-above the variance accounted for by the individual parent, adolescent, and peer confederate reports, with β s for the *trait* score ranging from .47–.67. This is an important finding, in light of prior work in this same sample indicating that among the three informants, peer confederates’ reports best predicted observed anxiety (Cannon et al., 2020; Glenn et al., 2019). If peer confederates truly operated as “optimal informants,” then integrating multiple informants’ reports to predict adolescent anxiety within peer interactions should not improve prediction. Indeed, in this study the informants integrated with peer confederates consisted of informants who, unlike peer confederates, did not base their reports on the Unfamiliar Peer Paradigm, as parents and adolescents completed their reports before administration of the peer interaction tasks. This first aspect of the findings suggests that integrating informants’ reports improves prediction, even when the criterion variable consists of context-specific behavior that only one informant who—at the time when they made their report—had the capacity to directly observe.

A second aspect of the findings demonstrates the large amount of domain-relevant unique variance present in the sample. Consistent with the Satellite Model variance components reported previously, cross-informant correlations among the adolescent, parent, and peer confederate reports ranged from low-to-moderate in magnitude (i.e., r s = .08–.49). This raises a question: Is there at least a portion of this unique variance that is domain-relevant? In fact, supplementary analyses reported by Makol et al. (2020) indicated that the *context* and *perspective* scores, at times, demonstrated incremental value in prediction, relative to the individual informants’ reports. This second aspect of the findings not only highlights the large amount of informant discrepancies present in the sample, but also indicates that the *context* and *perspective* scores reflected, in part, domain-relevant unique variance.

Related to the second aspect of the findings, the third aspect of findings is that the *trait* score demonstrated incremental prediction of observed anxiety, relative to the composite score, β = .60. In fact, when Makol et al. (2020) reversed the analysis to conduct tests of the composite score, over-and-above the variance

accounted for by the *trait* score, the composite score failed to demonstrate significant incremental prediction, $\beta = .01$; $p = .94$. This last aspect of the findings—the differences in prediction between the *trait* score and composite score—is crucial in detecting the presence of domain-relevant information. That is, if the components used to arrive at the *trait* score (i.e., context and perspective) were irrelevant to understanding informant discrepancies (i.e., they did not reflect domain-relevant information), then the *trait* score and composite score would yield roughly similar findings. This is because in both cases, the only variance accounting for prediction would be common variance. However, the *trait* score outperformed the composite score in predicting observed anxiety, and as mentioned previously, the *context* and *perspective* scores also predicted observed anxiety. In sum, the Makol and colleagues study shows that both common variance and domain-relevant unique variance are crucial to optimizing measurement validity. Further, the study shows how researchers who align their work with CONTEXT's features could leverage data from structurally different informants' reports to estimate displays of youth mental health concerns, even for criterion variables constructed to purposefully reflect how youth behave within a specific context.

Predicting Youth Mental Health Outcomes

Beyond optimizing research on characterizing youth mental health, studies that align with CONTEXT's features optimize the ability of researchers to predict outcomes in longitudinal studies. Indeed, CONTEXT guides researchers to construct multi-informant assessments such that they capitalize on the predictive utility of both informant discrepancies and informant agreement. Thus, CONTEXT's features ought to translate to prediction of outcomes based on instances in which informants disagree in their reports but also instances in which they agree. By construction, if the nature and extent of these predictions differs between patterns of informant discrepancies and informant

agreement, the implication is clear: When using multi-informant data to predict outcomes, optimal prediction comes not by examining each informant's report in isolation of the other informants' reports, but rather by interpreting each informant's report relative to reports completed by the other informant(s). In this respect, two recent studies highlight the value of CONTEXT for optimizing prediction of youth mental health outcomes.

In the first study, Becker-Haimes et al. (2018) examined a treatment sample of 488 youth participating in a large, multi-site trial of various treatments for youth anxiety disorders, in which youth were randomly assigned to one of four conditions (i.e., cognitive behavioral treatment, sertraline, their combination, and pill placebo). The enrollment process for the study began with parents contacting the investigative team via an initial telephone screen, after which time the team determined eligibility for participation. This process—where parents served as the key stakeholder tasked with initiating clinical services—is typical of both treatment outcome studies in youth mental health and service delivery processes generally (Hunsley & Lee, 2014). This recruitment process also figures prominently in interpreting the findings of this study. Becker-Haimes et al. (2018) examined pre-treatment anxiety symptom data, derived from parent and youth reports on a parallel set of survey measures about youth anxiety. To characterize patterns of informant discrepancies and informant agreement observed between parent and youth reports, the authors constructed polynomial regression models as described previously. Consistent with CONTEXT's features, the criterion variable within this model consisted of diagnostic remission ratings completed by trained independent evaluators who were both masked to the youth's treatment condition and did not serve as the youth's treating clinician. Within this model, and specifically among youth assigned to the cognitive behavioral treatment condition, when youth self-reported relatively low levels of symptoms, higher parent-reported symptoms predicted increased likelihood of youth not achieving diagnostic remission at post-treatment. This specific pattern of informant discrepancies is important because the study's inclusion criteria required

Table 7. Examples of how features of CONTEXT facilitate using informant discrepancies to optimize predictions of youth mental health outcomes.

Authors	Informants	Domains Informants Reported	Analytic Procedure Used to Characterize Informant Discrepancies	Criterion Variables
Becker-Haimes et al. (2018)	Parent, Child/Adolescent	Anxiety	Polynomial regression	Diagnostic remission
Makol et al. (2019)	Parent, Adolescent	Internalizing	Latent class analysis	Characteristics of care during adolescent's stay on inpatient unit

CONTEXT = Classifying Observations Necessitates Theory, Epistemology, and Testing. MTMM = Multi-Trait Multi-Method Matrix.

youth to meet diagnostic criteria for an anxiety disorder, as rated by an independent evaluator. Thus, for treatment contexts typified by the parent—but not the youth—endorsing the need for treatment at baseline, youth experience relatively low treatment responses. We can speculate that the mechanisms underlying this effect may consist of youth within this treatment context experiencing disengagement from receipt of treatment, perhaps because youth who do not perceive a problem are not motivated for treatment designed to address the problem. If true, then this pattern of informant discrepancies might demonstrate utility in future research on implementing services for improving therapeutic engagement, among youth at risk for experiencing low treatment responses.

The second study we illustrate here supports the predictive value of the specific pattern of informant discrepancies observed by Becker-Haimes et al. (2018), but also highlights the predictive value of informant agreement. Makol et al. (2019) examined a large sample of 765 adolescents receiving services in an acute care and hospitalization unit. Similar to the study by Becker-Haimes and colleagues, the adolescents involved in this study needed to have been voluntarily admitted to the unit via guardian consent. Within this treatment context, the authors examined internalizing symptom data collected at the intake admission assessment, consisting of parent and adolescent reports on parallel survey instruments. To characterize patterns of informant discrepancies and informant agreement observed between parent and adolescent reports, the authors examined these reports using latent class analyses as described previously. Consistent with CONTEXT's features, the criterion variables the authors examined consisted of independently assessed treatment characteristics, namely such indices as the number of days the adolescent spent in the acute care unit, as well as the use of intensive treatment regimens during care (e.g., use of standing antipsychotic medications, locked door seclusion). Germane to the findings described below, intake assessments also included diagnostic indices for conditions other than those on the internalizing spectrum (e.g., externalizing disorders). Latent class analyses characterized four kinds of reporting patterns, including parents who reported greater internalizing symptoms relative to adolescent self-reports, and dyads for whom both parent and adolescent agreed on reports of high internalizing symptoms. Each of these reporting patterns predicted important treatment characteristics. Consistent with Becker-Haimes and colleagues, the “parent > adolescent” latent class predicted increased likelihood during care of adolescents receiving both locked door seclusion and standing antipsychotic medication. Importantly, the informant agreement latent class either did not predict these same outcomes or predicted them at a much lower

magnitude than that observed for the “parent > adolescent” latent class. However, the informant agreement class had a predictive value of its own, namely as a marker of the number of days an adolescent would ultimately stay on the unit. Importantly, all of these findings were robust to controlling for number of externalizing disorders diagnosed at intake, and neither of these discrepancies/agreement classes predicted administration of medication for aggression, indicating that the findings we summarize here were unique to the predictive value of reports about internalizing symptoms. Taken together, these studies illustrate how researchers who align their work with CONTEXT's features could leverage data from structurally different informants' reports to predict youth mental health outcomes, and in a way that informs future research on service delivery (e.g., improving therapeutic engagement).

Testing Theories about Links between Informant Discrepancies and Youth Mental Health

We previously noted how the Achenbach et al. (1987) notion of situational specificity represents a key foundation for theory underlying CONTEXT-informed research. What we have yet to articulate is the historical context in which Achenbach and colleagues developed this notion. Situational specificity emerged roughly 30 years after Campbell and Fiske (1959), at a time when the MTMM paradigm's core assumptions were already firmly rooted in common assessment practices and notions about informant discrepancies. Situational specificity represents a sound conceptual rationale for studying informant discrepancies. Yet, users of the MTMM paradigm have been conditioned to think of these discrepancies as measurement confounds, and thus they have grounded their work on theories that align with the view that informant discrepancies reflect measurement confounds (i.e., depression-distortion hypothesis; Table 1).

The consequence of this history of informant discrepancies research is that scholars in youth mental health have yet to subject situational specificity to experimentation. Importantly, the features of CONTEXT guide researchers seeking to conduct this exact kind of theory testing. In particular, CONTEXT calls for direct tests of theories underlying use of structurally different informants' reports. If situational specificity holds true, then we would expect that experimental manipulations of informants' access to observations of youth behavior—essentially, an experimental manipulation of the structural differences between informants—would cause discrepancies between informants' behavioral reports. We tested

this expectation with a CONTEXT-informed, controlled experiment. In online supplementary material, we provide information relevant to our rationale for focusing this experiment on informant discrepancies in reports about youth social skills, as well as procedures for stimuli development and pre-testing.

For this experiment, we recruited an online sample of 182 undergraduates. Broadly, the experiment involved randomly assigning participants to observing videotaped renditions of youth displaying behaviors indicative of specific social skills. Following their observations of these videotaped stimuli, we then prompted participants to make responses on a behavioral checklist consisting of a mix of behaviors they observed and did not observe.

We recruited participants at the institution of the first author, which is in a different geographic region from the site of stimuli development (see online supplementary material). This element of our study design reduced the likelihood that study participants bore any relation to the youth involved in the videotaped segments. Further, all participants observed stimuli within the same study conditions, and thus we held constant both the relationship status of participants to the youth they observed and the manner in which they observed behavior. In this way, we preserved the integrity of our experimental manipulation of structural differences between informants. We recruited these participants through the SONA experiment management system (SONA). Within SONA, undergraduates participate in studies and receive course credit in exchange for doing so. Once a participant signed up for a timeslot on the SONA system, we directed them to an online consent form explaining all study procedures.

Following consent, we directed participants to an online platform (i.e., Qualtrics) where we executed the experiment. We then randomly assigned participants to two groups. These groups varied in their opportunities to observe a subset of the videotaped segments described previously. Thus, we created a matrix of observations designed to manipulate situational specificity. Cell 1 of this matrix consisted of videos of social skills behaviors observed by Group 1 and not Group 2. Cell 2 consisted of videos of social skills behaviors observed by Group 2 and not Group 1. Cell 3 consisted of videos of neutral behaviors (e.g., sitting in chair) observed by both Groups 1 and 2. We created an “empty” Cell 4, which included item descriptions that we presented to both Groups 1 and 2 during the testing phase, but for which they had not observed a videotaped segment. Taken together, in this experiment we utilized a 2 (between-person factor: Group 1 vs. Group 2) X 4 (within-person factors: social skills behaviors observed only by Group 1, social skills

behaviors observed only by Group 2, “neutral items” videos observed by both Groups 1 and 2, non-video items observed by neither Group 1 nor 2) mixed study design. This design induced measured outcomes reflecting situational specificity. In fact, we exposed participants to assessment scenarios consistent with the Operations Triad Model. Specifically, we induced measured outcomes reflecting Diverging Operations via the situational specificity effects brought about by Cells 1 and 2. Further, using Cell 3 we expected to induce measured outcomes reflecting Converging Operations, given that participants in both groups would observe the same videotaped behaviors in Cell 3. Lastly, we expected Cell 4 to induce responses reflecting measurement error. That is, if a participant endorsed behaviors in the item descriptions in Cell 4, they would be making a “false positive” endorsement – i.e., an indication that they observed stimuli to which they were not exposed. If our experiment revealed informant discrepancies in Cell 4, then this would signal a form of discrepancies reflecting Compensating Operations.

Following exposure to these videos, participants entered the testing phase of the experiment, in which they completed a “behavior checklist,” consisting of a series of items reflecting social skills behaviors and neutral behaviors. We instructed participants to rate whether the items described behaviors on the videos that they just observed. Both groups were administered the same instrument, which included a set of 120 items, designed to simulate the number of items on a behavior checklist (e.g., Achenbach & Rescorla, 2001). When an informant provides a report on such a checklist, they tend to endorse some, but not all items. Creating more items than an informant typically endorses allows a researcher to develop an instrument that avoids *ceiling effects*. By *ceiling effects*, we mean the possibility that an informant endorsed the highest possible score, not because of the assessed youth’s “true level” on the measured domain, but because the instrument included an insufficiently small sampling of behaviors that reflect that domain (see also, Nunnally & Bernstein, 1994). Thus, the 120 items consisted of: (a) social skills behaviors observed only by Group 1 (Cell 1); (b) social skills behaviors observed only by Group 2 (Cell 2); (c) neutral behaviors in videos observed by both groups (Cell 3); and (d) behaviors not displayed in any videos and thus not observed by either group (Cell 4).

We report the findings of our experiment in Table 8. Means and standard deviations reported in Table 8 reflect participants’ mean probabilities of item endorsement (possible range: 0–100), which we report separately for Groups 1 and 2. By *item endorsement*, we

mean probabilities reflecting endorsement by a participant that they observed the behavior described in a given item. To test group differences, we first constructed a repeated-measures analysis of variance model. This model consisted of a 4-level within-person Item factor (i.e., social skills behaviors observed only by Group 1, social skills behaviors observed only by Group 2, “neutral items” videos observed by both Groups 1 and 2, non-video items observed by neither Groups 1 or 2), and a 2-level between-person Group factor (i.e., Group 1 vs. Group 2). Our experiment produced variations in situational specificity. First, we observed significant omnibus effects (i.e., via the Roy’s Largest Root F test), namely a significant Item effect, $F(3, 178) = 565.57$; $p < .001$; $\eta^2 = .90$, and a significant Item X Group interaction effect, $F(3, 178) = 196.45$; $p < .001$; $\eta^2 = .77$.

Second, we probed the Item X Group interaction effect. Group 1 was significantly more likely than Group 2 to positively endorse an item describing a behavior displayed in a video that Group 1 participants observed and Group 2 participants did not, $t(180) = 16.82$, $p < .001$. Similarly, Group 2 was significantly more likely than Group 1 to positively endorse an item describing a behavior displayed in a video that Group 2 participants observed and Group 1 participants did not, $t(180) = -19.02$, $p < .001$. Importantly, effect sizes denoting between-group differences for these item endorsements (Table 8) were at magnitudes over three times higher than effect size conventions for “large” effects on the Cohen’s d metric (i.e., 0.80; Cohen, 1988). Findings reported in Table 8 revealed that we also “removed” situational specificity effects: Groups 1 and 2 had nearly identical levels of positive item endorsement for videotaped behaviors that both groups observed, $t(180) = -0.35$, $p = .73$. The effect size for this finding was near zero and well below effect size conventions for “small” effects on the Cohen’s d metric (i.e., 0.20). Further, item endorsements for behaviors that neither group observed were below chance levels and yielded non-significant group differences as well, $t(180) = 0.90$, $p = .37$. Importantly, these null differences

reflected non-zero levels of magnitude. Thus, at times, participants endorsed the appearance of behaviors that, by construction, they had not seen (i.e., “false positives”). Yet, the between-group effect was non-significant, indicating that our study produced informant discrepancies for which little of the variance reflected measurement confounds (Table 8).

In causally manipulating situational specificity, we produced variations in measured informant discrepancies. In these respects, imagine that our two experimental groups represented types of structurally different informants (e.g., Group 1 = parent; Group 2 = teacher). Assume that the two “informants” provided reports about the same youth, and the youth displayed considerable variations in displays of social skills across the two separate observational contexts available to them (e.g., Group 1 = home; Group 2 = school). In an analogue sense, this is a valid reflection of situational specificity effects, not only across clinical populations, but also across measured domains and development (i.e., early childhood through adulthood; Glenn et al., 2019; Lerner et al., 2017; De Los Reyes et al., 2013b; Wakschlag et al., 2010). The “informants” in our study varied in their opportunities to observe behaviors displayed by the youth, and made reports consistent with situational specificity as advanced by Achenbach et al. (1987).

Two key limitations of our experiment warrant comment. First, our experiment confirms that situational specificity effects can cause informant discrepancies. Yet, our experiment involved recruiting undergraduate participants to serve as “informants” who observed youth behaviors and reported about these behaviors. These informants shared few characteristics with the structurally different informants used in youth mental health research. Second, the youth in our video stimuli were simulating social skills behaviors. This element of our experiment allowed us to manipulate a key aspect of youth behavior, namely that it varies considerably within and across contexts. However, in doing so we created an assessment setting that does not reflect the conditions that typify how researchers assess youth

Table 8. Between-group differences in identification probabilities of videotaped social skills behavior, as a function of variations in situational specificity.

Behaviors Observed by Experimental Groups	Group N	Mean Probabilities of Video Identification	Standard Deviations of Video Identification	Effect Sizes of Mean Differences Reflecting Situational Specificity
Videotaped behaviors observed by Group 1 and not Group 2 ($n = 10$ items)	Group 1 N: 87 Group 2 N: 95	Group 1 Mean: 74.71 Group 2 Mean: 22.00	Group 1 SD: 20.79 Group 2 SD: 21.42	Cohen’s d : +2.50
Videotaped behaviors observed by Group 2 and not Group 1 ($n = 10$ items)	Group 1 N: 87 Group 2 N: 95	Group 1 Mean: 29.08 Group 2 Mean: 79.16	Group 1 SD: 18.40 Group 2 SD: 17.11	Cohen’s d : -2.82
Videotaped behaviors observed by both groups ($n = 15$ items)	Group 1 N: 87 Group 2 N: 95	Group 1 Mean: 89.43 Group 2 Mean: 90.11	Group 1 SD: 13.08 Group 2 SD: 13.11	Cohen’s d : -0.05
Behaviors observed by neither group ($n = 85$ items)	Group 1 N: 87 Group 2 N: 95	Group 1 Mean: 20.07 Group 2 Mean: 18.25	Group 1 SD: 12.85 Group 2 SD: 14.24	Cohen’s d : +0.13

mental health. Thus, our controlled experiment should be considered an analogue to how youth mental health assessments occur in research settings. In this respect, our experiment informs additional theory tests of situational specificity, preferably experiments conducted with the structurally different informants and clinical populations that typify assessments conducted in youth mental health research.

Summary Comments on the Need for a Paradigm Shift in Validation

We described CONTEXT, its distinct set of features relative to the MTMM paradigm, how each of these paradigms inform fundamentally distinct research practices, and how CONTEXT-informed practices more closely align with how youth mental health researchers leverage multi-informant approaches to assessment (Tables 1–5). We also illustrated CONTEXT “in action” with a set of studies that demonstrate how the paradigm’s features translate to optimizing the ability of multi-informant assessments to facilitate scholarly inquiry across a diverse array of research topics germane to youth mental health (Tables 6–8). Nonetheless, we imagine that users of the MTMM paradigm might contend that the issues we raise could be accomplished in ways that do not necessitate CONTEXT. For instance, users of the MTMM paradigm could reform some of the practices that reflect the issues we raised, such as incorporating non-informant indices in structural models that typically only include informants (see also, Watts et al., 2022).

We contend that reforming research practices, while still retaining use of the MTMM paradigm, would merely address the “symptoms” of the issues we raised. The “disorder” is the paradigm that informs the practices. Consider that the first observations of informant discrepancies in youth mental health assessments (e.g., Lapouse & Monk, 1958) predate the MTMM paradigm itself (Campbell & Fiske, 1959). Additionally, consider that the theoretical rationale for addressing the issues raised in this paper (e.g., situational specificity) has been around for nearly 40 years (Achenbach et al., 1987). Users of the MTMM paradigm have had decades’ worth of both rationales and time to address the issues raised in this paper.

Scholars in youth mental health have reached an inflection point. Should we catalyze a paradigm shift that empirical work already demonstrates will translate to both immediate and long-term improvements in measurement validity? Alternatively, should we give users of the MTMM paradigm another few decades to

enact a “course correction” that they have yet to enact in six decades’ worth of work? Which of these seems like the more reasonable approach?

We close this summary with one more argument in favor of a paradigm shift. We demonstrated that when users of the MTMM paradigm apply it to data conditions that violate its core assumptions (i.e., informant discrepancies contain domain-relevant information), they engage in research practices that have the untoward effect of depressing measurement validity. In contrast, within these exact data conditions, CONTEXT guides users to engage in practices that optimize measurement validity. We have long-known that measurement validity has important, downstream effects on all aspects of scholarly inquiry and study design, including observed magnitudes of effect sizes (e.g., between predictor and outcome variables) and in turn, the statistical power to detect hypothesized effects (see, Markon et al., 2011; Nunnally & Bernstein, 1994). We also know that some of the best predictors of the replicability of study findings are exactly these two factors: effect size and statistical power (e.g., Open Science Collaboration, 2015; Tackett et al., 2017). To the degree that scholars in youth mental health seek to address “replication crises” in Psychology, CONTEXT is not a panacea for addressing these issues. That said, the guidance it provides in terms of research practices reflects “low-hanging fruit” for addressing issues surrounding the replicability of study findings (see also, De Los Reyes, 2021).

Research Implications of CONTEXT

Applying CONTEXT to Instrument Development and Evaluation

As a validation paradigm, CONTEXT carries with it several key research implications. One implication pertains to instrument development and evaluation. For those illustrations of CONTEXT’s principles “in action” that involved informants completing standardized instruments (Tables 6–7), the instruments informants completed were not designed to conduct CONTEXT-informed research. Rather, they were optimized to conduct MTMM-informed research. Having been developed in a “Converging Operations world,” these instruments were largely optimized to model common variance, namely by including item content that functioned similarly across informants’ reports (see also, Charamut et al., 2022; De Los Reyes et al., 2019c, 2022b). This element of the work we reviewed provides three further insights into the nature of CONTEXT-informed research. First, we demonstrated the potential

for work informed by CONTEXT's principles to detect domain-relevant informant discrepancies. It is perhaps the most impressive aspect of this work that researchers could detect these kinds of discrepancies, using multi-informant instruments that were ill-equipped to do so. This aspect of prior work speaks to the robust nature of informant discrepancies observed in youth mental health research (see, Achenbach et al., 1987; De Los Reyes et al., 2015, 2019a). Equally important, the findings we described previously (Tables 6–7), though promising, likely provide a conservative estimate of the clinical significance of detecting domain-relevant informant discrepancies.

Second, because current instrumentation emphasizes common variance, the multi-informant data produced using these instruments are optimized for analytic procedures that emphasize common variance. Thus, we suspect it is possible to improve the sensitivity of instruments to detect domain-relevant informant discrepancies. Optimizing measurement validity in youth mental health research and its ability to inform clinical decisions in youth mental health services may require refining existing instruments (or developing new instruments). We previously mentioned that CONTEXT guides users to leverage analytic procedures that retain both common variance and domain-relevant unique variance (Table 4). We suspect that optimizing the utility of these analytic procedures may involve infusing in multi-informant instruments item content that reflects a mix of items that function invariantly across informants' reports (i.e., common variance) as well as items that display distinct functions across informants' reports (see also, De Los Reyes et al., 2019a). This idea merits further study.

Third, as we recently argued elsewhere (see, De Los Reyes et al., 2022b), informant discrepancies may not be the only forms of data conditions which have been treated as measurement confounds in prior work, but may also contain domain-relevant information. Consider the data conditions underlying multivariate estimates of longitudinal change in youth mental health concerns. Here too, researchers frequently leverage measurement invariance techniques to determine whether instruments function similarly across data conditions, only this time the techniques typically involve comparing data derived from the same informant's reports across multiple assessment points (e.g., youth completes self-report at baseline, 1-year, and 2-year follow-up assessments; T.M. Olino et al., 2018). Granted, measurement invariance techniques obviously serve useful purposes under some of these circumstances, namely when the presence of measurement confounds provides a sound rationale for observations of time-variant response patterns. An example of such a circumstance

might be longitudinal studies across developmental periods, wherein the developmental changes estimated are so prolonged that changes in instrumentation occur across assessment points (e.g., middle childhood through emerging adulthood; see, Tyrell et al., 2019). Yet, when efforts have been made to rule out such confounds (i.e., informants complete the same instrument across assessment periods), one has to wonder: Might some instances in which researchers have used measurement invariance techniques to detect time-variant response patterns also be detecting domain-relevant response patterns? That is, just as with domain-relevant informant discrepancies, might researchers also be depressing measurement validity when they try to construct instruments that only include items which function invariantly across time? Here too, this question merits further study.

Applying CONTEXT to Meta-Analytic Reviews of Intervention Outcome Studies

CONTEXT has important implications for interpreting the findings of meta-analytic reviews of intervention outcome studies. We have long-known that informant discrepancies frequently characterize intervention outcomes in controlled trials of youth interventions, with findings ranging from small-to-large (e.g., Cohen's, 1988 *d* ranging from 0.3 to 0.8+), depending on the informant completing the outcome measure (for a recent review, see, De Los Reyes & Makol, 2022). In these respects, CONTEXT can facilitate generating hypotheses about what these meta-analytic findings reflect, in an effort to guide future research on whether discrepant intervention outcomes contain domain-relevant information.

As one example, consider the literature on psychosocial treatments for youth depression. Meta-analyses consistently reveal large informant discrepancies in intervention response, with effect sizes from outcomes based on youth self-report anywhere from three times higher than parent-reported outcomes (effect size: 0.72 vs. 0.24; Weisz et al., 2006) to far larger discrepancies between these informants' outcome reports (effect size: 0.39 vs. –0.06; Eckshtain et al., 2020), and with effects based on teacher reports much more modest than those based on parent reports (effect size 0.27 vs. 0.48; Weisz et al., 2017). Consistent with the Converging Operations and MTMM paradigms, researchers' characterizations of these effects have focused on estimating the average intervention response (i.e., common variance). This approach has led to characterizations of interventions for youth depression as weaker in their ability to bring

about positive treatment responses, relative to use of these same interventions when administered to clients from later developmental periods (e.g., emerging adulthood, middle and older adulthood; Cuijpers et al., 2020). Further, by focusing on the consistency of effects across studies (and between informants), we argue that this has led researchers down a path of questioning whether existing interventions for youth clients are “capable of producing enduring positive outcomes” (see, Weersing et al., 2017, p. 38). Yet, what if the discrepancies between parent- and youth-reported outcomes contain domain-relevant information?

CONTEXT has the potential to promote an evolution in how researchers interpret meta-analytic outcomes, and it begins within instilling a balanced emphasis on both common variance and domain-relevant unique variance. Let us consider the possibility that the informant discrepancies observed within intervention studies for at least one subgroup of youth clients, namely adolescents, might contain domain-relevant information. Consider that adolescents, relative to youth from earlier developmental periods, tend to spend considerably greater amounts of time navigating social environments outside of the home context (Smetana, 2008). In fact, among adolescents many of the stressful life circumstances that factor into the maintenance of their internalizing concerns tend to manifest outside of the home (e.g., interactions with peers at school and other non-home contexts; for a review, see Prinstein et al., 2018). In these respects, there exists the potential that, within intervention outcome studies, parents and youth systematically vary in their opportunities for observing treatment responses in contexts where they often manifest. As evidence of this, consider that in recent work, adolescent self-reports of youth depressive symptoms (but not parent reports) predict adolescents’ arousal in reaction to laboratory tasks designed to simulate interactions with same-age peers (Rausch et al., 2017).

Researchers may approach interpreting the meta-analyses of interventions for youth depression in ways that are informed by CONTEXT. Granted, it will always be of value to characterize the average effects observed across intervention studies, the common variance. Yet, if these “main effects” are qualified by domain-relevant, unique effects reflected by informant discrepancies, then CONTEXT may inform a deeper level of interpretation beyond estimates of average intervention effects. In fact, with the wealth of information generated from examining multiple studies, meta-analysis might serve as a key tool for characterizing informant discrepancies. In this respect, we can reimagine meta-analytic tools to function as techniques for characterizing informant discrepancies, in ways akin to how we illustrated use of such techniques

as polynomial regression and latent class analysis (see, Tables 6 and 7). As with the techniques described previously, we envision applying meta-analytic tools this way to follow a two-step process. First, we can leverage meta-analytic procedures to discover patterns of informant discrepancies in intervention effects. Discovering patterns of discrepancies may facilitate generating hypotheses to guide future research. Consider the case of interpreting effects of interventions for adolescent depression. We previously mentioned that these meta-analyses revealed far larger effects for youth report relative to parent report. These findings might reflect domain-relevant information. Perhaps the interventions may be yielding potent effects, and in specific contexts which youth have considerable expertise to observe, such as within peer interactions outside of the home (see also, Talbott et al., 2021).

The second step of this process involves direct tests of the hypotheses generated from interpreting patterns of informant discrepancies discovered via meta-analyses. In this step, we envision CONTEXT informing the design of future studies specifically devoted to testing hypotheses surrounding informant discrepancies in intervention outcomes. In the case of interventions for adolescent depression, such a study might bear similarities to the Makol et al. (2020) study we described previously. Based on the meta-analytic findings, researchers might design an intervention study that includes pre- and post-treatment administration of a battery of adolescent and parent survey measures of depressive symptoms, as well as an independent criterion measure designed to index adolescents’ internalizing concerns when interacting within contexts that typically occur outside of the home (e.g., reports of youth depression by peers, trained observers’ ratings of youth within peer interactions; see, Cannon et al., 2020). Using this battery of outcome measures, researchers could test whether estimates of treatment responses observed in adolescent-reported outcomes “match” the treatment response patterns observed on the independent criterion measure, to a greater degree than the parent-reported outcomes. Studies of this kind directly test an interpretation of the meta-analytic outcomes described previously: Discrepancies in the outcomes of youth and parent reports may reflect, in part, domain-relevant information.

Taken together, meta-analyses may become potent tools for characterizing informant discrepancies in intervention outcomes, and CONTEXT can inform the design of follow-up studies dedicated to testing hypotheses generated from this reimagined approach to meta-analyses. In fact, our call for revising research practices germane to meta-analysis also illustrates the practical implications of using CONTEXT. Indeed, we clearly

centered the traditional controlled trials approach used in studies included in meta-analyses, namely trials within which the only outcome data consist of multi-informant assessments (see, Weisz et al., 2005). These kinds of studies are instrumental to characterizing patterns of informant discrepancies in intervention outcomes. By extension, they are also crucial for guiding the design of the multi-modal, CONTEXT-informed studies needed to determine if it is valid to interpret the informant discrepancies observed in intervention outcomes as reflecting domain-relevant information. In this respect, we suspect that this same kind of two-step process would also be of value to other areas of meta-analytic work, including studies of the associated features of youth mental health.

Applicability of CONTEXT to Recent Innovations in Intervention Science

Related to our discussion of interpreting meta-analyses of intervention outcome studies, let us anticipate that in some intervention literatures, CONTEXT-informed studies will demonstrate that informant discrepancies on outcome measures reflect domain-relevant information. Within these literatures, recent work on methods for tailoring intervention sequences to individual clients might also benefit from CONTEXT (for a review, see, Ng & Weisz, 2016). For instance, the Sequential Multiple Assignment Randomized Trial (SMART) design involves taking a data-driven approach to testing different intervention sequences (e.g., medication first vs. psychosocial intervention first or vice versa), and then using the resulting data to determine what to do next if the first step is successful, and if it is not (for a review, see, Lei et al., 2012). Researchers embed discrete decision points throughout intervention testing (e.g., end of a course of treatment). The data collected at these points can take the form of such aspects of care as symptom reduction, client preferences, and/or the degree to which the client adhered to the intervention. Researchers then use these data to determine whether a course of an intervention should continue or be modified to consider new intervention approaches.

Importantly, in SMART studies researchers often operationally define intervention non-response as displaying suboptimal outcomes on any of several outcomes (e.g., August et al., 2016), or on a single or “primary outcome” (see, Lei et al., 2012). In essence, these definitions of non-response do not account for the possibility of a client experiencing non-response in one social context but positive responses in another context. Yet, if informant discrepancies in reports of intervention outcomes contain domain-relevant information, then CONTEXT might help improve the

precision of individualized intervention sequences. Consider a client who receives a first-line, comprehensive intervention designed to address their concerns across home and school contexts. If the client experiences positive responses based on parent report but non-response based on teacher report, this may signal a true, context-specific intervention response. If true, then it may benefit the client to continue with a version of the intervention tailored to the home context but adapt the sequencing to an alternative intervention tailored to the school context. Thus, future research should test whether applying CONTEXT to SMART studies facilitates determining whether a client is, in fact, displaying a relatively low intervention response or, alternatively, displaying positive responses within specific contexts.

Applicability of CONTEXT to Assessing Mental Health across the Lifespan

We expect CONTEXT to inform multi-informant approaches to assessing mental health beyond those developed to assess youth. For instance, as with multi-informant assessments of youth mental health, meta-analytic reviews of assessments of adult mental health also indicate the ubiquitous presence of informant discrepancies (e.g., Achenbach et al., 2005). In fact, the presence of these discrepancies not only replicates across studies, but also within longitudinal studies over the course of adult development (e.g., van der Ende et al., 2012). That is, if you collect multi-informant data to assess adults, then you will likely encounter informant discrepancies, regardless of the developmental period of the adults undergoing evaluation.

The presence of informant discrepancies when assessing adults begs the question: Do discrepancies observed in these assessments contain domain-relevant information? CONTEXT may provide researchers with guidance on addressing this question when assessing adults. Yet, key characteristics of research on multi-informant assessments of adult mental health may pose barriers to addressing this question. As mentioned previously, multi-informant approaches to assessing youth trace back to the 1950s (e.g., Lapouse & Monk, 1958). In contrast, use of this approach to assess adults is a relatively new phenomenon (Achenbach, 2020). In fact, in the first meta-analysis of cross-informant correspondence of adult mental health, Achenbach et al. (2005) identified 108 articles from which to estimate correspondence, but this number represented just 0.2% of all of the articles that came up in their literature searches (i.e., 51,000). By comparison, the first meta-analysis of cross-informant correspondence of youth mental health identified relatively more articles (i.e.,

119; Achenbach et al., 1987), and did so nearly 20 years before the adult-focused meta-analysis. Thus, one barrier to applying CONTEXT to multi-informant assessments of adults is that scholars in adult mental health have decades less worth of data on this phenomenon than scholars in youth mental health.

In light of the relatively nascent literature regarding cross-informant correspondence when assessing adults, it should not come as a surprise that use of non-informant methods (e.g., observed behavior, performance-based tasks) is also relatively rare (see, De Los Reyes & Makol, *in press*). Applying CONTEXT to multi-informant assessments of adult mental health will prove challenging if studies on these issues encounter difficulty with avoiding criterion contamination. Yet, perhaps the most challenging barrier to overcome involves common research practices regarding use of structurally different informants when assessing adults.

The modal informant discrepancies study in adult mental health seeks to characterize discrepancies between adult self-reports and observer or *collateral informants*. In these studies, collateral informants have taken on a variety of forms, including spouses, coworkers, close friends, caregivers in the case of elderly adults, and parents in the case of emerging adults. This was true for the studies meta-analyzed by Achenbach et al. (2005), as well as for more recent work (e.g., Horwitz et al., 2016; L.A. Rescorla et al., 2016; L. A. Rescorla et al., 2022; van der Ende et al., 2012). Now, the diversity of collateral informants is, on the surface, not an issue. In fact, the option for researchers to leverage many different kinds of collateral informants, coupled with the ability of these informants to provide psychometrically sound data, speaks to the feasibility of the approach (see also, Vazire, 2006). That said, one problem lies in the fact that, although there are exceptions (see, Liu et al., 2021; Nelson & Lovett, 2019; Szkody et al., 2022), studies tend to “bundle” these collateral informants together. That is, researchers often compare self-reports to a single group of collateral informants that includes a “mix” of all kinds of informants (e.g., self vs. parent but also self vs. spouse and self vs. coworker). If one treats these comparisons as a study of “structurally different informants” in an omnibus sense—i.e., comparing self versus many other kinds of collateral informants, and with no distinction as to the characteristics of the collaterals—then doing so carries with it the assumption that all of the collateral informants share the same kinds of structural characteristics. For instance, is it safe to assume that the structural environment of spouses is similar enough to that of coworkers, such that a researcher can both

aggregate spouses and coworkers into a single group of collateral informants, and meaningfully compare this aggregated group to the self-reports of adult clients?

Needless to say, often these decisions are made out of convenience. Researchers often have no choice but to examine collateral informants as one big group, because in their samples, no one specific pair of informants (e.g., self and coworker) exists in a large enough supply to estimate informant discrepancies with much precision (for an exception, see van der Ende et al., 2012). That said, these considerations often pose few issues for youth mental health researchers. Researchers who assess youth routinely implement inclusion criteria, such that they require youth participants to live in a stable household and attend school. These criteria result in recruiting the same structurally different informants for the whole sample (i.e., parents, teachers, and youth). This is a privilege of conducting research on informant discrepancies in youth mental health. In any one sample of youth, the “inner workings” of the structural differences among informants might vary from youth-to-youth, but the same overall structures are apparent for all youth in the sample (home, school). Conversely, in samples of adults for whom collateral informants display wide variations in structural characteristics (e.g., parents/spouses at home, coworkers in employment settings, caregivers in hospital settings, friends in neighborhood settings), might researchers nonetheless be able to detect domain-relevant information, particularly in reference to validity criteria as illustrated previously (Tables 6 and 7)? These issues merit further study.

Concluding Comments

A decade ago, we observed a profound disconnect between the rationale researchers used to collect data from structurally different informants, and how they treated the informant discrepancies that invariably arise from taking this approach to multi-informant assessment (De Los Reyes et al., 2013a). Back then, we termed this disconnect *The Grand Discrepancy*, and we developed the Operations Triad Model to resolve this disconnect. Ten years later, and in the process of writing this paper, we made a discovery: The Operations Triad Model represented a paradigm shift in how we conceptualize informant discrepancies, but our work was incomplete. Fully resolving *The Grand Discrepancy* that we designed the Operations Triad Model to address requires a paradigm shift in how we approach measurement validation when leveraging data from structurally different informants. In line with *The Grand Discrepancy*, in this paper we revealed a disconnect

between the conceptual models that most closely align with use of structurally different informants, and the validation paradigm that has governed researchers' practices regarding use, interpretation, and examination of multi-informant data (i.e., the MTMM). To address this disconnect, we advanced a validation paradigm that instantiates the Operations Triad Model's principles in measurement (CONTEXT).

In this paper, we described how the features of CONTEXT provide users with explicit guidance on theory, epistemological principles, and research practices germane to multi-informant assessments (e.g., selecting informants, analytic procedures, study design, interpretation of findings). We drew distinctions between CONTEXT and the MTMM paradigm, and cited evidence to demonstrate how these two paradigms inform fundamentally distinct practices regarding use and interpretation of data from structurally different informants. In focusing on common variance and assuming that the unique variance typified by informant discrepancies reflects measurement confounds, we highlighted how the MTMM paradigm guides users down a path of research practices that risks depressing measurement validity, namely when informant discrepancies, in fact, reflect domain-relevant information. In contrast, we described how CONTEXT guides users down a very different path; a path of research practices that, by attending to both common variance and domain-relevant unique variance, facilitates optimizing measurement validity. To demonstrate these distinctions in concrete terms, we cited MTMM-informed studies and the fundamental limitations inherent in the informants that researchers selected, the analytic procedures they implemented, and key features of the designs of their studies. We also illustrated CONTEXT's features with a combination of published and new evidence, and in doing so highlighted how CONTEXT-informed research addresses the limitations of the MTMM-informed research cited throughout the paper. Lastly, we highlighted directions for future research. In particular, we expect CONTEXT to (a) facilitate innovations in measurement development and evaluation, (b) improve the interpretability of the findings of meta-analytic reviews, (c) optimize the accuracy of treatment response estimates, and (d) inform our understanding of multi-informant approaches to assessing mental health across the lifespan.

Acknowledgment

We thank Thomas M. Achenbach for his extremely helpful comments on a previous version of this paper. We also thank Catherine C. Epkins for serving as Action Editor and

overseeing the review process for this paper, as well as three anonymous reviewers for their own commentary on multiple versions of this paper.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

Efforts by the first author were supported by a grant from the Fulbright U.S. Scholars Program (Fulbright Canada Research Chair in Mental Health). Efforts by the first, second, and fourth authors were supported by a grant from the Institute of Education Sciences [R324A180032].

ORCID

Andres De Los Reyes  <http://orcid.org/0000-0003-0214-983X>

Mo Wang  <http://orcid.org/0000-0001-7004-3549>

Matthew D. Lerner  <http://orcid.org/0000-0002-7373-6663>

Bridget A. Makol  <http://orcid.org/0000-0002-7627-2614>

Olivia M. Fitzpatrick  <http://orcid.org/0000-0001-8868-6702>

John R. Weisz  <http://orcid.org/0000-0002-5560-6814>

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., Krukowski, R. A., Dumenci, L., & Ivanova, M. Y. (2005). Assessment of adult psychopathology: Meta-analyses and implications of cross-informant correlations. *Psychological Bulletin*, 131(3), 361–382. <https://doi.org/10.1037/0033-2909.131.3.361>
- 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>
- Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles*. University of Vermont Research Center for Children, Youth, & Families.
- Al Ghriwati, N., Winter, M. A., Greenlee, J. L., & Thompson, E. L. (2018). Discrepancies between parent and self-reports of adolescent psychosocial symptoms: Associations with family conflict and asthma outcomes. *Journal of Family Psychology*, 32(7), 992–997. <https://doi.org/10.1037/fam0000459>
- Alfano, C. A., & Beidel, D. C. (2011). *Social anxiety in adolescents and young adults: Translating developmental science into practice*. American Psychological Association.
- Alvarez, I., Herrero, M., Martínez-Pampliega, A., & Escudero, V. (2021). Measuring perceptions of the therapeutic alliance in individual, family, and group therapy from a systemic perspective: Structural validity of the

- SOFTA-s. *Family Process*, 60(2), 302–315. <https://doi.org/10.1111/famp.12565>
- American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed.).
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.).
- Atkins, M. S., Cappella, E., Shernoff, E. S., Mehta, T. G., & Gustafson, E. L. (2017). Schooling and children's mental health: Realigning resources to reduce disparities and advance public health. *Annual Review of Clinical Psychology*, 13, 123–147. <https://doi.org/10.1146/annurev-clinpsy-032816-045234>
- August, G. J., Piehler, T. F., & Bloomquist, M. L. (2016). Being "SMART" about adolescent conduct problems prevention: Executing a SMART pilot study in a juvenile diversion agency. *Journal of Clinical Child and Adolescent Psychology*, 45(4), 495–509. <https://doi.org/10.1080/15374416.2014.945212>
- Bandura, A. (1977). *Social learning theory*. Prentice-Hall.
- Bartholomew, D. J., Steele, F., Moustaki, I., & Galbraith, J. I. (2002). *The analysis and interpretation of multivariate data for social scientists*. Chapman & Hall/CRC.
- 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>
- Beck, A. T. (1993). Cognitive therapy: Past, present, and future. *Journal of Consulting and Clinical Psychology*, 61(2), 194–198. <https://doi.org/10.1037/0022-006X.61.2.194>
- Becker-Haimes, E. M., Jensen-Doss, A., Birmaher, B., Kendall, P. C., & Ginsburg, G. S. (2018). Parent–youth informant disagreement: Implications for youth anxiety treatment. *Clinical Child Psychology and Psychiatry*, 23(1), 42–56. <https://doi.org/10.1177/1359104516689586>
- Borsboom, D. (2005). *Measuring the mind*. Cambridge University Press.
- Boutron, I., Dutton, S., Ravaud, P., & Altman, D. G. (2010). Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*, 303(20), 2058–2064. <https://doi.org/10.1001/jama.2010.651>
- Brase, G. L. (2014). Behavioral science integration: A practical framework of multi-level converging evidence for behavioral science theories. *New Ideas in Psychology*, 33, 8–20. <https://doi.org/10.1016/j.newideapsych.2013.11.001>
- Bronfenbrenner, U. (1979). *The ecology of human development*. Harvard University Press.
- Bruno, N., & Paolo Battaglini, P. (2008). Integrating perception and action through cognitive neuropsychology (broadly conceived). *Cognitive Neuropsychology*, 25(7–8), 879–890. <https://doi.org/10.1080/02643290802519591>
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 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>
- 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, 911629. <https://doi.org/10.3389/fpsyg.2022.911629>
- Cicchetti, D. (1984). The emergence of developmental psychopathology. *Child Development*, 55(1), 1–7. <https://doi.org/10.2307/1129830>
- Clark, D. A., Listro, C. J., Lo, S. L., Durbin, C. E., Donnellan, M. B., & Neppl, T. K. (2016). Measurement invariance and child temperament: An evaluation of sex and informant differences on the child behavior questionnaire. *Psychological Assessment*, 28(12), 1646–1662. <https://doi.org/10.1037/pas0000299>
- 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.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Routledge.
- Crick, N. R., & Dodge, K. A. (1994). A review and reformulation of social information-processing mechanisms in children's social adjustment. *Psychological Bulletin*, 115(1), 74–101. <https://doi.org/10.1037/0033-2909.115.1.74>
- Cuijpers, P., Karyotaki, E., Eckshtain, D., Ng, M. Y., Corteselli, K. A., Noma, H., Quero, S., & Weisz, J. R. (2020). Psychotherapy for depression across different age groups: A systematic review and meta-analysis. *JAMA Psychiatry*, 77(7), 694–702. <http://doi.org/10.1001/jamapsychiatry.2020.0164>
- Curran, P. J., Georgeson, A. R., Bauer, D. J., & Hussong, A. M. (2021). Psychometric models for scoring multiple reporter assessments: Applications to integrative data analysis in prevention science and beyond. *International Journal of Behavioral Development*, 45(1), 40–50. <https://doi.org/10.1177/0165025419896620>
- Darling, N., & Steinberg, L. (1993). Parenting style as context: An integrative model. *Psychological Bulletin*, 113(3), 487–496. <https://doi.org/10.1037/0033-2909.113.3.487>
- De Los Reyes, A. (2011). More than measurement error: Discovering meaning behind informant discrepancies in clinical assessments of children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, 40(1), 1–9. <https://doi.org/10.1080/15374416.2011.533405>
- De Los Reyes, A. (2021). (Second) inaugural editorial: How the journal of clinical child and adolescent psychology can nurture team science approaches to addressing burning questions about mental health. *Journal of Clinical Child and Adolescent Psychology*, 50(1), 1–11. <https://doi.org/10.1080/15374416.2020.1858839>
- 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., Bunnell, B. E., & Beidel, D. C. (2013b). Informant discrepancies in adult social anxiety disorder

- assessments: Links with contextual variations in observed behavior. *Journal of Abnormal Psychology*, 122(2), 376–386. <https://doi.org/10.1037/a0031150>
- De Los Reyes, A., Cook, C. R., Gresham, F. M., Makol, B. A., & Wang, M. (2019c). Informant discrepancies in assessments of psychosocial functioning in school-based services and research: Review and directions for future research. *Journal of School Psychology*, 74, 74–89. <https://doi.org/10.1016/j.jsp.2019.05.005>
- De Los Reyes, A., Cook, C. R., Sullivan, M., Morrell, N., Gresham, F. M., Wang, M., Gresham, F. M., Makol, B. M., Keeley, L. M., & Qasmieh, N. (2022c). 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., Drabick, D. A. G., Makol, B. A., & Jakubovic, R. (2020). Introduction to the special section: The research domain criteria's units of analysis and cross-unit correspondence in youth mental health research. *Journal of Clinical Child and Adolescent Psychology*, 49(3), 279–296. <https://doi.org/10.1080/15374416.2020.1738238>
- De Los Reyes, A., Henry, D. B., Tolan, P. H., & Wakschlag, L. S. (2009). Linking informant discrepancies to observed variations in young children's disruptive behavior. *Journal of Abnormal Child Psychology*, 37(5), 637–652. <https://doi.org/10.1007/s10802-009-9307-3>
- De Los Reyes, A., Kundey, S. M. A., & Wang, M. (2011). The end of the primary outcome measure: A research agenda for constructing its replacement. *Clinical Psychology Review*, 31(5), 829–838. <https://doi.org/10.1016/j.cpr.2011.03.011>
- De Los Reyes, A., Lerner, M. D., Keeley, L. M., Weber, R., Drabick, D. A. G., Rabinowitz, J., & Goodman, K. L. (2019a). 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., Lerner, M. D., Thomas, S. A., Daruwala, S. E., & Goepel, K. A. (2013c). Discrepancies between parent and adolescent beliefs about daily life topics and performance on an emotion recognition task. *Journal of Abnormal Child Psychology*, 41(6), 971–982. <https://doi.org/10.1007/s10802-013-9733-0>
- De Los Reyes, A., & Makol, B. A. (2022). Informant reports in clinical assessment. In G. Asmundson (Ed.), *Comprehensive clinical psychology* (2nd ed., Vol. 4, pp. 105–122). Elsevier. <https://doi.org/10.1016/B978-0-12-818697-8.000113-8>
- De Los Reyes, A., & Makol, B. A. (in press). Interpreting convergences and divergences in multi-informant, multi-method assessment. In J. Mihura (Ed.), *The Oxford handbook of personality and psychopathology assessment* (2nd ed.). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780190092689.013.33>
- De Los Reyes, A., & Ohannessian, C. M. (2016). Introduction to the special issue: Discrepancies in adolescent-parent perceptions of the family and adolescent adjustment. *Journal of Youth and Adolescence*, 45(10), 1957–1972. <https://doi.org/10.1007/s10964-016-0533-z>
- De Los Reyes, A., Ohannessian, C. M., & Racz, S. J. (2019b). Discrepancies between adolescent and parent reports about family relationships. *Child Development Perspectives*, 13(1), 53–58. <https://doi.org/10.1111/cdep.12306>
- De Los Reyes, A., Talbott, E., Power, T., Michel, J., Cook, C. R., Racz, S. J., & Fitzpatrick, O. (2022a). The needs-to-goals gap: How informant discrepancies in youth mental health assessments impact service delivery. *Clinical Psychology Review*, 92, 102114. <https://doi.org/10.1016/j.cpr.2021.102114>
- De Los Reyes, A., Thomas, S. A., Goodman, K. L., & Kundey, S. M. A. (2013a). 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., Tyrell, F. A., Watts, A. L., & Asmundson, G. J. G. (2022b). Conceptual, methodological, and measurement factors that disqualify use of measurement invariance techniques to detect informant discrepancies in youth mental health assessments. *Frontiers in Psychology*, 13, 931296. <https://doi.org/10.3389/fpsyg.2022.931296>
- Deros, D. E., Racz, S. J., Lipton, M. F., Augenstein, T. M., Karp, J. N., Keeley, L. M., Qasmieh, N., Grewe, B. I., Aldao, A., & 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>
- Duhig, A. M., Renk, K., Epstein, M. K., & Phares, V. (2000). Interparental agreement on internalizing, externalizing, and total behavior problems: A meta-analysis. *Clinical Psychology: Science and Practice*, 7(4), 435–453. <https://doi.org/10.1093/clipsy.7.4.435>
- 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>
- Dunteman, G. H. (1989). *Principal components analysis* (No. 69). Sage.
- Eckstain, D., Kuppens, S., Ugueto, A., Ng, M. Y., Vaughn-Coaxum, R., Corteselli, K., & Weisz, J. R. (2020). Meta-analysis: 13-year follow-up of psychotherapy effects on youth depression. *Journal of the American Academy of Child and Adolescent Psychiatry*, 59(1), 45–63. <https://doi.org/10.1016/j.jaac.2019.04.002>
- Edgeworth, F. Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society*, 51(3), 598–635. <https://www.jstor.org/stable/2339898>
- Edwards, J. R. (1994). The study of congruence in organizational behavior research: Critique and a proposed alternative. *Organizational Behavior and Human Decision Processes*, 58(1), 51–100. <https://doi.org/10.1006/obhd.1994.1029>
- 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>
- Fergusson, D. M., Boden, J. M., & Horwood, L. J. (2009). Situational and generalized conduct problems and later life outcomes: Evidence from a New Zealand birth cohort. *Journal of Child Psychology and Psychiatry*, 50(9), 1084–1092. <https://doi.org/10.1111/j.1469-7610.2009.02070.x>
- Festa, J. E., Producer, Sinclair, N., Producer, Monroe, M., Producer, & Marshall, F., Director. (2020). *The Bee Gees: How can you mend a broken heart*. Home Box Office. Documentary.

- Fisher, E., Bromberg, M. H., Tai, G., & Palermo, T. M. (2017). Adolescent and parent treatment goals in an internet-delivered chronic pain self-management program: Does agreement of treatment goals matter? *Journal of Pediatric Psychology*, 42(6), 657–666. <https://doi.org/10.1093/jpepsy/jsw098>
- 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>
- Flores, I. S., Dobrea, A., Balazsi, R., Roşan, A., Păsăreanu, C. R., Predescu, E., & Rad, F. (2022). Measurement invariance of Alabama parenting questionnaire across age, gender, clinical status, and informant. *Assessment*, 107319112110681. <https://doi.org/10.1177/10731911211068178>
- Follet, L., Okuno, H., & De Los Reyes, A. (2022). Assessing peer-related impairments linked to adolescent social anxiety: Strategic selection of informants optimizes prediction of clinically relevant domains. *Behavior Therapy*. Advance online publication. <https://doi.org/10.1016/j.beth.2022.06.010>
- Garb, H. N. (2003). Incremental validity and the assessment of psychopathology in adults. *Psychological Assessment*, 15(4), 508–520. <http://dx.doi.org/10.1037/1040-3590.15.4.508>
- Garner, W. R., Hake, H. W., & Eriksen, C. W. (1956). Operationism and the concept of perception. *Psychological Review*, 63(3), 149–159. <https://doi.org/10.1037/h0042992>
- Geiser, C., Eid, M., West, S. G., Lischetzke, T., & Nussbeck, F. W. (2012). A comparison of method effects in two confirmatory factor models for structurally different methods. *Structural Equation Modeling*, 19(3), 409–436. <https://doi.org/10.1080/10705511.2012.687658>
- Glenn, L. E., Keeley, L. M., Szollos, S., Okuno, H., Wang, X., Rausch, E., Deros, D. E., Karp, J. N., Qasmieh, N., Makol, B. A., Augenstein, T. M., Lipton, M. F., Racz, S. J., Scharfstein, L., Beidel, D. C., & 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>
- Goldenweiser, A. A. (1912). Folk-psychology. *Psychological Bulletin*, 9(10), 373–380. <https://doi.org/10.1037/h0074365>
- Goodman, S. H., Rouse, M. H., Connell, A. M., Broth, M. R., Hall, C. M., & Heyward, D. (2011). Maternal depression and child psychopathology: A meta-analytic review. *Clinical Child and Family Psychology Review*, 14(1), 1–27. <https://doi.org/10.1007/s10567-010-0080-1>
- Grace, R. C. (2001). On the failure of operationism. *Theory and Psychology*, 11(1), 5–33. <https://doi.org/10.1177/0959354301111001>
- Haerlein, H. K. (1916). The theoretical foundations of Wundt's folk-psychology. *Psychological Review*, 23(4), 279–302. <https://doi.org/10.1037/h0075449>
- Hawley, K. M., & Weisz, J. R. (2003). Child, parent, and therapist (dis)agreement on target problems in outpatient therapy: The therapist's dilemma and its implications. *Journal of Consulting and Clinical Psychology*, 71(1), 62–70. <https://doi.org/10.1037/0022-006X.71.1.62>
- Hempel, C. G. (1966). *Philosophy of natural science*. Prentice-Hall.
- Highhouse, S. (2009). Designing experiments that generalize. *Organizational Research Methods*, 12(3), 554–566. <https://doi.org/10.1177/1094428107300396>
- Hoffman, L. J., & Chu, B. C. (2015). Target problem (mis) matching: Predictors and consequences of parent–youth agreement in a sample of anxious youth. *Journal of Anxiety Disorders*, 31, 11–19. <https://doi.org/10.1016/j.janxdis.2014.12.015>
- Horwitz, E. H., Schoevers, R. A., Ketelaars, C. E. J., Kan, C. C., Van Lammeren, A. M. D. N., Meesters, Y., Spek, A. A., Wouters, S., Teunisse, J. P., Cuppen, L., Bartels, A. A. J., Schuringa, E., Moorlag, H., Raven, D., Wiersma, D., Minderaa, R. B., & Hartman, C. A. (2016). Clinical assessment of ASD in adults using self- and other-report: Psychometric properties and validity of the Adult Social Behavior Questionnaire (ASBQ). *Research in Autism Spectrum Disorders*, 24, 17–28. <https://doi.org/10.1016/j.rasd.2016.01.003>
- 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>
- Hunsley, J., & Lee, C. M. (2014). *Introduction to clinical psychology* (2nd ed.). Wiley.
- Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology*, 3, 29–51. <https://doi.org/10.1146/annurev.clinpsy.3.022806.091419>
- Hunsley, J., & Mash, E. J., (Eds.). (2018). *A guide to assessments that work* (2nd ed.). Oxford University Press.
- Jungersen, C. M., & Lonigan, C. J. (2021). Do parent and teacher ratings of ADHD reflect the same constructs? A measurement invariance analysis. *Journal of Psychopathology and Behavioral Assessment*, 43(4), 778–792. <https://doi.org/10.1007/s10862-021-09874-3>
- Kazdin, A. E. (2013). *Behavior modification in applied settings* (7th ed.). Waveland Press.
- Kazdin, A. E. (2017). *Research design in clinical psychology* (5th ed.). Pearson.
- Kazdin, A. E., & Rotella, C. (2009). *The Kazdin method for parenting the defiant child: With no pills, no therapy, no contest of wills*. Houghton Mifflin Harcourt.
- Kazdin, A. E., & Wassell, G. (2000). Therapeutic changes in children, parents, and families resulting from treatment of children with conduct problems. *Journal of the American Academy of Child and Adolescent Psychiatry*, 39(4), 414–420. <https://doi.org/10.1097/00004583-200004000-00009>
- Kazdin, A. E., & Whitley, M. K. (2003). Treatment of parental stress to enhance therapeutic change among children referred for aggressive and antisocial behavior. *Journal of Consulting and Clinical Psychology*, 71(3), 504–515. <https://doi.org/10.1037/0022-006X.71.3.504>
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.
- Knobe, J., & Mendlow, G. S. (2004). The good, the bad and the blameworthy: Understanding the role of evaluative reasoning in folk psychology. *Journal of Theoretical and Philosophical Psychology*, 24(2), 252–258. <https://doi.org/10.1037/h0091246>
- Konold, T., & Cornell, D. (2015). Multilevel multitrait–multimethod latent analysis of structurally different and interchangeable raters of school climate. *Psychological*

- Assessment*, 27(3), 1097–1109. <https://doi.org/10.1037/pas0000098>
- Konold, T., & Sanders, E. A. (2020). The nature of rater effects and differences in multilevel MTMM latent variable models. *Measurement: Interdisciplinary Research and Perspectives*, 18(4), 177–195. <https://doi.org/10.1080/15366367.2020.1746897>
- 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. *American Journal of Psychiatry*, 160(9), 1566–1577. <https://doi.org/10.1176/appi.ajp.160.9.1566>
- Krause, N., Liang, J., Bennett, J., Kobayashi, E., Akiyama, H., & Fukaya, T. (2010). A descriptive analysis of religious involvement among older adults in Japan. *Aging and Society*, 30(4), 671–696. <https://doi.org/10.1017/S0144686X09990766>
- Laird, R. D. (2020). Analytical challenges of testing hypotheses of agreement and discrepancy: Comment on Campione-Barr, Lindell, and Giron (2020). *Developmental Psychology*, 56(5), 970–977. <https://doi.org/10.1037/dev0000763>
- Laird, R. D., & De Los Reyes, A. (2013). Testing informant discrepancies as predictors of early adolescent psychopathology: Why difference scores cannot tell you what you want to know and how polynomial regression may. *Journal of Abnormal Child Psychology*, 41(1), 1–14. <https://doi.org/10.1007/s10802-012-9659-y>
- Laird, R. D., & Weems, C. F. (2011). The equivalence of regression models using difference scores and models using separate scores for each informant: Implications for the study of informant discrepancies. *Psychological Assessment*, 23(2), 388–397. <https://doi.org/10.1037/a0021926>
- Lakind, D., Bradley, W. J., Patel, A., Chorpita, B. F., & Becker, K. D. (2022). A multidimensional examination of the measurement of treatment engagement: Implications for children's mental health services and research. *Journal of Clinical Child and Adolescent Psychology*, 51(4), 453–468. <https://doi.org/10.1080/15374416.2021.1941057>
- 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>
- Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D., & Murphy, S. A. (2012). A “SMART” design for building individualized treatment sequences. *Annual Review of Clinical Psychology*, 8, 21–48. <https://doi.org/10.1146/annurev-clinpsy-032511-143152>
- Lerner, M. D., De Los Reyes, A., Drabick, D. G., Gerber, A. H., & Gadow, K. D. (2017). Informant discrepancy defines discrete, clinically useful autism spectrum disorder subgroups. *Journal of Child Psychology and Psychiatry*, 58(7), 829–839. <https://doi.org/10.1111/jcpp.12730>
- Levy, P. (1969). Platonic true scores and rating scales: A case of uncorrelated definitions. *Psychological Bulletin*, 71(4), 276–277. <https://doi.org/10.1037/h0026855>
- Liu, J., Dong, F., Lee, C. M., Reyes, J., & Ivanova, M. (2021). The application of the adult self-report and the adult behavior checklist form to Chinese adults: Syndrome structure, inter-informant agreement, and cultural comparison. *International Journal of Environmental Research and Public Health*, 18(12), 6352. <https://doi.org/10.3390/ijerph18126352>
- LoBue, V., Reider, L. B., Kim, E., Burris, J. L., Oleas, D. S., Buss, K. A., Pérez-Edgar, K., & Field, A. P. (2020). The importance of using multiple outcome measures in infant research. *Infancy*, 25(4), 420–437. <https://doi.org/10.1111/inf.12339>
- 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
- Makol, B. A., De Los Reyes, A., Garrido, E., Harlaar, N., & Taussig, H. (2021). Assessing the mental health of maltreated youth with child welfare involvement using multi-informant reports. *Child Psychiatry and Human Development*, 52(1), 49–62. <https://doi.org/10.1007/s10578-020-00985-8>
- Makol, B. A., De Los Reyes, A., Ostrander, R., & Reynolds, E. K. (2019). Parent-youth divergence (and convergence) in reports of youth internalizing problems in psychiatric inpatient care. *Journal of Abnormal Child Psychology*, 47(10), 1677–1689. <https://doi.org/10.1007/s10802-019-00540-7>
- Makol, B. A., Youngstrom, E. A., Racz, S. J., Qasmieh, 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>
- Markon, K. E., Chmielewski, M., & Miller, C. J. (2011). The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review. *Psychological Bulletin*, 137(5), 856–879. <https://doi.org/10.1037/a0023678>
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525–543. <https://doi.org/10.1007/BF02294825>
- 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>
- Millsap, E. (2011). *Statistical methods for studying measurement invariance*. Taylor & Francis.
- Murray, A. L., Speyer, L. G., Hall, H. A., Valdebenito, S., & Hughes, C. (2021). Teacher versus parent informant measurement invariance of the strengths and difficulties questionnaire. *Journal of Pediatric Psychology*, 46(10), 1249–1257. <https://doi.org/10.1093/jpepsy/jsab062>
- Nelson, J. M., & Lovett, B. J. (2019). Assessing ADHD in college students: Integrating multiple evidence sources with symptom and performance validity data. *Psychological Assessment*, 31(6), 793–804. <https://doi.org/10.1037/pas0000702>
- Ng, M. Y., & Weisz, J. R. (2016). Building a science of personalized intervention for youth mental health. *Journal of Child Psychology and Psychiatry*, 57(3), 216–236. <http://dx.doi.org/10.1111/jcpp.12470>
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
- Offord, D. R., Boyle, M. H., Racine, Y., Szatmari, P., Fleming, J. E., Sanford, M., & Lipman, E. L. (1996). Integrating assessment data from multiple informants. *Journal of the American Academy of Child and Adolescent*

- Psychiatry*, 35(8), 1078–1085. <https://doi.org/10.1097/00004583-199608000-00019>
- Olino, T. M., Finsaas, M., Dougherty, L. R., & Klein, D. N. (2018). Is parent–child disagreement on child anxiety explained by differences in measurement properties? An examination of measurement invariance across informants and time. *Frontiers in Psychology*, 9, 1295. <https://doi.org/10.3389/fpsyg.2018.01295>
- Olino, T. M., Michelini, G., Mennies, R. J., Kotov, R., & Klein, D. N. (2021). Does maternal psychopathology bias reports of offspring symptoms? A study using moderated non-linear factor analysis. *Journal of Child Psychology and Psychiatry*, 62(10), 1195–1201. <https://doi.org/10.1111/jcpp.13394>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. <https://doi.org/10.1126/science.aac4716>
- Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning*. Sage.
- Popper, K. R. (1962). *Conjectures and refutations: The growth of scientific knowledge*. Basic Books.
- Prinstein, M. J., Rancourt, D., Adelman, C. B., Ahlich, E., Smith, J., & Guerry, J. D. (2018). Peer status and psychopathology. In W. Bukowski, B. Laursen, & K. H. Rubin (Eds.), *Handbook of peer interactions, relationships, and groups* (2nd ed., pp. 617–636). Guilford Press.
- Rausch, E., Racz, S. J., Augenstein, T. M., Keeley, L., Lipton, M. F., Szollos, S., Riffle, J., Moriarity, D., Kromash, R., & De Los Reyes, A. (2017). A multi-informant approach to measuring depressive symptoms in clinical assessments of adolescent social anxiety using the Beck Depression Inventory-II: Convergent, incremental, and criterion-related validity. *Child and Youth Care Forum*, 46(5), 661–683. <https://doi.org/10.1007/s10566-017-9403-4>
- Rescorla, L. A., Achenbach, T. M., Ivanova, M. Y., Turner, L. V., Árnadóttir, H., Au, A., Caldas, J. C., Chen, Y.-C., Decoster, J., Fontaine, J., Funabiki, Y., Guðmundsson, H. S., Leung, P., Liu, J., Maraš, J. S., Marković, J., Oh, K. J., da Rocha, M. M., Samaniego, V. C., ... Zasepa, E. (2016). Collateral reports and cross-informant agreement about adult psychopathology in 14 societies. *Journal of Psychopathology and Behavioral Assessment*, 38(3), 381–397. <https://doi.org/10.1007/s10862-016-9541-2>
- Rescorla, L. A., Ivanova, M. Y., Achenbach, T. M., Almeida, V., Anafarta-Sendag, M., Bite, I., Caldas, J. C., Capps, J. W., Chen, Y. C., Colombo, P., da Silva Oliveira, M., Dobrean, A., Erol, A., Frigerio, A., Funabiki, Y., Gedutienė, R., Guðmundsson, H. S., Heo, M. Q., Kim, Y. A., ... Zasepa, E. (2022). Older adult psychopathology: International comparisons of self-reports, collateral reports, and cross-informant agreement. *International Psychogeriatrics*, 34(5), 467–478. <https://doi.org/10.1017/S1041610220001532>
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Prokasy (Ed.), *Classical conditioning II: Current research and theory* (pp. 64–99). Appleton-Century-Croft.
- 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>
- Roberts, B. W., & Caspi, A. (2001). Personality development and the person-situation debate: It's déjà vu all over again. *Psychological Inquiry*, 12(2), 104–109. https://doi.org/10.1207/S15327965PLI1202_04
- Rubio-Stipec, M., Fitzmaurice, G., Murphy, J., & Walker, A. (2003). The use of multiple informants in identifying the risk factors of depressive and disruptive disorders: Are they interchangeable? *Social Psychiatry and Psychiatric Epidemiology*, 38(2), 51–58. <https://doi.org/10.1007/s00127-003-0600-0>
- Russell, J. D., Graham, R. A., Neill, E. L., & Weems, C. F. (2016). Agreement in youth-parent perceptions of parenting behaviors: A case for testing measurement invariance in reporter discrepancy research. *Journal of Youth and Adolescence*, 45(10), 2094–2107. <https://doi.org/10.1007/s10964-016-0495-1>
- Schiltz, H. K., Magnus, B. E., McVey, A. J., Haendel, A. D., Dolan, B. K., Stanley, R. E., Willar, K. A., Pleiss, S. J., Carson, A. M., Carlson, M., Murphy, C., Vogt, E. M., Yund, B. D., & Van Hecke, A. V. (2021). A psychometric analysis of the social anxiety scale for adolescents among youth with autism spectrum disorder: Caregiver–adolescent agreement, factor structure, and validity. *Assessment*, 28(1), 100–115. <https://doi.org/10.1177/1073191119851563>
- 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.
- Shin, H. J., Rabe-Hesketh, S., & Wilson, M. (2019). Trifactor models for multiple-ratings data. *Multivariate Behavioral Research*, 54(3), 360–381. <https://doi.org/10.1080/00273171.2018.1530091>
- Skinner, B. F. (1953). *Science and human behavior*. MacMillan.
- Smetana, J. G. (2008). “It’s 10 o’clock: Do you know where your children are?” Recent advances in understanding parental monitoring and adolescents’ information management. *Child Development Perspectives*, 2(1), 19–25. <https://doi.org/10.1111/j.1750-8606.2008.00036.x>
- Soland, J., & Kuhfeld, M. (2022). Examining the performance of the trifactor model for multiple raters. *Applied Psychological Measurement*, 46(1), 53–67. <https://doi.org/10.1177/01466216211051728>
- Sternberg, R. J. (2005). The importance of converging operations in the study of human intelligence. *Cortex*, 41(2), 243–244. [https://doi.org/10.1016/S0010-9452\(08\)70908-0](https://doi.org/10.1016/S0010-9452(08)70908-0)
- Sulik, M. J., Blair, C., & Greenberg, M., & Family Life Project Investigators. (2017). Child conduct problems across home and school contexts: A person-centered approach. *Journal of Psychopathology and Behavioral Assessment*, 39(1), 46–57. <https://doi.org/10.1007/s10862-016-9564-8>
- Szkody, E., Rogers, M. M., & McKinney, C. (2022). Discrepancy analysis of emerging adult and parental report of psychological problems and relationship quality. *Journal of Psychopathology and Behavioral Assessment*, 44(2), 444–455. <https://doi.org/10.1007/s10862-021-09949-1>
- Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., Oltmanns, T. F., & Shrout, P. E.

- (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science*, 12(5), 742–756. <https://doi.org/10.1177/1745691617690042>
- Talbott, E., De Los Reyes, A., Power, T., Michel, J., & Racz, S. J. (2021). A team-based collaborative care model for youth with attention deficit hyperactivity disorder in education and pediatric health care settings. *Journal of Emotional and Behavioral Disorders*, 29(1), 24–33. <https://doi.org/10.1177/1063426620949987>
- Tversky, A., & Kahneman, D. (1974, September 27). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>
- Tyrell, F. A., Yates, T. M., Widaman, K. F., Reynolds, C. A., & Fabricius, W. V. (2019). Data harmonization: Establishing measurement invariance across different assessments of the same construct across adolescence. *Journal of Clinical Child and Adolescent Psychology*, 48(4), 555–567. <https://doi.org/10.1080/15374416.2019.1622124>
- Vallar, G. (2006). Memory systems: The case of phonological short-term memory. A festschrift for cognitive neuropsychology. *Cognitive Neuropsychology*, 23(1), 135–155. <https://doi.org/10.1080/02643290542000012>
- van der Ende, J., Verhulst, F. C., & Tiemeier, H. (2012). Agreement of informants on emotional and behavioral problems from childhood to adulthood. *Psychological Assessment*, 24(2), 293–300. <https://doi.org/10.1037/a0025500>
- van der Ende, J., Verhulst, F. C., & Tiemeier, H. (2020). Multitrait-multimethod analyses of change of internalizing and externalizing problems in adolescence: Predicting internalizing and externalizing DSM disorders in adulthood. *Journal of Abnormal Psychology*, 129(4), 343–354. <https://doi.org/10.1037/abn0000510>
- Vazire, S. (2006). Informant reports: A cheap, fast, and easy method for personality assessment. *Journal of Research in Personality*, 40(5), 472–481. <https://doi.org/10.1016/j.jrp.2005.03.003>
- Von Der Heide, R. J., Wenger, M. J., Bittner, J. L., & Fitousi, D. (2018). Converging operations and the role of perceptual and decisional influences on the perception of faces: Neural and behavioral evidence. *Brain and Cognition*, 122, 59–75. <https://doi.org/10.1016/j.bandc.2018.01.007>
- Wakschlag, L. S., Tolan, P. H., & Leventhal, B. L. (2010). Research review: “Ain’t misbehavin’” Toward a developmentally-specified nosology for preschool disruptive behavior. *Journal of Child Psychology and Psychiatry*, 51(1), 3–22. <http://dx.doi.org/10.1111/j.1469-7610.2009.02184.x>
- Watts, A. L., Makol, B. A., Palumbo, I. M., De Los Reyes, A., Olino, T. M., Latzman, R. D., DeYoung, C. G., Wood, P. K., & Sher, K. J. (2022). How robust is the p-factor? Using multitrait-multimethod modeling to inform the meaning of general factors of psychopathology in youth. *Clinical Psychological Science*, 10(4), 640–661. <https://doi.org/10.1177/21677026211055170>
- Weersing, V. R., Jeffreys, M., Do, M. C. T., Schwartz, K. T., & Bolano, C. (2017). Evidence base update of psychosocial treatments for child and adolescent depression. *Journal of Clinical Child and Adolescent Psychology*, 46(1), 11–43. <https://doi.org/10.1080/15374416.2016.1220310>
- Weisz, J. R. (2004). *Psychotherapy for children and adolescents: Evidence-based treatments and case examples*. Cambridge University Press.
- Weisz, J. R., Hawley, K. M., & Doss, A. J. (2004). Empirically tested psychotherapies for youth internalizing and externalizing problems and disorders. *Child and Adolescent Psychiatric Clinics of North America*, 13(4), 729–815. <https://doi.org/10.1016/j.chc.2004.05.006>
- Weisz, J. R., Jensen Doss, A., & 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., & Kazdin, A. E., (Eds.). (2017). *Evidence-based psychotherapies for children and adolescents* (3rd ed.). Guilford Press.
- Weisz, J. R., Kuppens, S., Ng, M. Y., Eckshtain, D., Ugueto, A. M., Vaughn-Coaxum, R., Jensen-Doss, A., & 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., McCarty, C. A., & Valeri, S. M. (2006). Effects of psychotherapy for depression in children and adolescents: A meta-analysis. *Psychological Bulletin*, 132(1), 132–149. <https://doi.org/10.1037/0033-2909.132.1.132>
- Wundt, W. (1916). *Elements of folk psychology: Outlines of a psychological history of the development of mankind* E. L. Schaub, (Trans.). George Allen & Unwin. <https://doi.org/10.1037/13042-000>
- Yeh, M., & Weisz, J. R. (2001). Why are we here at the clinic? Parent-child (dis)agreement on referral problems at outpatient treatment entry. *Journal of Consulting and Clinical Psychology*, 69(6), 1018–1025. <https://doi.org/10.1037/0022-006X.69.6.1018>
- Zeid, D., Carter, J., & Lindberg, M. A. (2018). Comparisons of alcohol and drug dependence in terms of attachments and clinical issues. *Substance Use and Misuse*, 53(1), 1–8. <https://doi.org/10.1080/10826084.2017.1319865>