ABSTRACT

Title of Dissertation: BIAS VERSUS CONTEXT MODELS FOR INTEGRATING MULTI-INFORMANT REPORTS OF YOUTH MENTAL HEALTH

Bridget Alexis Makol, Doctor of Philosophy, 2021

Dissertation directed by: Professor Andres De Los Reyes, Psychology

Best practices in youth mental health assessment entail collecting reports from multiple informants. However, multi-informant reports commonly yield different estimates of youth mental health (i.e., *informant discrepancies*), resulting in various clinical decision-making challenges and necessitating strategies for integrating them. Two leading theoretical models exist for interpreting informant discrepancies. Whereas one model posits that informant discrepancies reflect rater biases and thus depress measurement validity (i.e., *bias models*), the other posits that they reflect meaningful variations in behavior across social contexts (e.g., home, school) and thus enhance measurement validity (i.e., *context models*). Although greater empirical support exists for context models relative to bias models, measurement models extending from both bias (i.e., Trifactor Model [TFM]) and context (i.e., Trait Score Satellite Model [TSSM]) models have been developed. Across two studies, I rigorously compared the TFM and TSSM. In Study 1, a systematic review of TFM and TSSM research (*n* = 47) revealed that, relative to TFM studies, TSSM studies were more likely to include (a) informants who varied in where they observe behavior (e.g., parent [home] vs. teacher [school]) and (b) more informants. In Study 2, I subjected these models to validation testing using a sample (*n* = 134) that included three informants' reports of adolescent social anxiety and independent ratings of adolescent behavior within peer interactions. I found satisfactory fit for both models when integrating all three informants' reports. However, when predicting well-established, independent criterion variables (i.e., observed behavior, referral status), the primary score derived from the TSSM outperformed each individual informant's report, a composite of informants' reports, and the primary TFM-derived score. Relative to the TFM, the TSSM (a) more closely aligns with best practices in evidence-based assessment of youth mental health, and (b) more effectively integrates multi-informant reports in data conditions where informant discrepancies reflect valid information. When using measurement models designed to integrate multi-informant reports, users of these models must subject them to rigorous validation testing to discern their applicability to the data conditions in which they will be applied. In turn, integrating multi-informant reports requires explicitly linking theory, quantitative methodology, and empirical support observed within relevant data conditions.

BIAS VERSUS CONTEXT MODELS FOR INTEGRATING MULTI-INFORMANT REPORTS OF YOUTH MENTAL HEALTH

by

Bridget Alexis Makol

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Professor Andres De Los Reyes, Chair
Assistant Professor Jessica Magidson
Associate Professor Jonathan Mohr
Dr. Mo Wang
Professor Hedwig Teglasi, Dean’s Representative

Dedication

To my parents, who instilled in me a love for learning and always encouraged me to follow my passions. You modeled hard work and perseverance every day. To my amazing husband, who joined me on my PhD journey. I could not have done it without your support.

Acknowledgements

I am extremely grateful to my advisor, Dr. Andres De Los Reyes, a truly supportive and invested mentor. Thank you for always believing in me! I would also like to express my deepest appreciation to my committee, including Drs. Jessica Magidson, Jonathan Mohr, Mo Wang, and Hedwig Teglasi. You provided invaluable feedback, insight, and encouragement throughout the process of my dissertation. Further, I am very grateful to Dr. Wang for aiding me in understanding both the conceptual and technical aspects of the statistical models applied in my dissertation. Lastly, I would like to thank the graduate student colleagues and research assistants who made this research possible. In particular, I am grateful to Lauren Keeley and Noor Qasmieh. Thank you for your friendship and support throughout my PhD journey and for always helping me laugh and have fun. Thank you to Nicholas Bellamy and Lia Follet, who were instrumental in completing Study 1 of my dissertation. In addition, thank you to Hide Okuno, who served as a wonderful lab manager and colleague and was instrumental in collecting Study 2 data.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Background

Assessing child and adolescent (i.e., youth) mental health is a complex task. Many of the mental health concerns that youth commonly face involve maladaptive reactions to social contexts. Youth vary not only in the contexts that encompass their social worlds, but also *how* or *why* they react as they do. Thus, understanding youth's clinical presentations requires consideration of cross-contextual variation in behavior, development, and comorbidity. The complexity underlying youth's clinical presentations highlights the need for precise and psychometrically sound assessment tools that incorporate multi-dimensional, multi-method data. Indeed, by definition, no single gold standard measure of youth mental health can accurately capture all of this complexity (De Los Reyes, Augenstein, Aldao, 2017). Thus, best practices in youth mental health assessment entail collecting and interpreting reports from multiple informants (e.g., parents, teachers, youth, peers). These reports are assumed to provide unique and incrementally valid information when assessing youth mental health concerns and making clinical decisions related to these concerns (Hunsley & Mash, 2007). Given the relative efficiency of administering multi-informant reports and the rich information they provide, they represent one of the primary, and at times *the* primary, standardized data used in service settings to characterize and plan treatment for youth mental health concerns (Youngstrom & Van Meter, 2016). These reports are also often used as the primary "evidence" when identifying evidence-based treatments for youth (Southam-Gerow & Prinstein, 2014).

Collecting multi-informant reports is standard practice when assessing youth mental health. Yet, in light of the complexity of youth's clinical presentations, reports from multiple informants commonly yield unique estimates of mental health concerns (i.e., *informant discrepancies*; De Los Reyes, 2011). Specifically, multi-informant reports consistently display low-to-moderate levels of convergence (i.e., *r*s = .20s-.30s; Achenbach, McConaughy, & Howell, 1987; De Los Reyes et al., 2015). These informant discrepancies have been documented in large-scale epidemiological studies and meta-analyses across psychopathology domains, informants, measures, measurement methods, contexts (e.g., home, school, peer interactions), developmental periods (i.e., early childhood through adulthood), cultures, and countries (Achenbach et al., 1987; Achenbach, Krukowski, Dumenci, & Ivanova, 2005; De Los Reyes et al., 2015; De Los Reyes, Lerner, et al., 2019; Duhig, Renk, Epstein, Phares, 2000; Gresham et al., 2018; Hou et al., 2019; Jones et al., 2019; Korelitz & Garber, 2016; Narad et al., 2014; Rescorla et al., 2013, 2017; Romano, Weegar, Babchishin, Saini, 2018; Stratis & Lecavalier, 2015). Taken together, over 50 years of research indicate that in virtually no decision-making setting (e.g., laboratory, hospital, community mental health center) can an assessor avoid encountering discrepancies when collecting multi-informant reports. Thus, in this dissertation I examine existing strategies for *reconciling* informant discrepancies, or the means by which researchers interpret these discrepancies and take actions to resolve the uncertainties that they create. In particular, I focus on strategies for *integrating* multi-informant data: aggregating these data to arrive at a single estimate of the measured domain about which informants provide reports.

An important consideration is that service providers and researchers lack any empirically based consensus guidelines for reconciling informant discrepancies (Beidas et al., 2015; De Los Reyes, Cook, Gresham, Makol, & Wang, 2019; Hunsley & Mash, 2007; Offord et al., 1996). This is a surprising omission in the literature on evidence-based practices. Indeed, prior work indicates that the strategy one uses to reconcile these discrepancies can lead to vastly different conclusions when completing important tasks, such as assigning mental health diagnoses, estimating prognosis, determining prevalence rates of mental health disorders, planning treatment, and evaluating intervention effects (De Los Reyes & Kazdin, 2005; Hawley & Weisz, 2003). Further, although no consensus guidelines exist, many strategies have been proposed. The state of research on strategies for reconciling informant discrepancies is analogous to research on evidence-based treatments. Specifically, we have long known that although hundreds of interventions exist for youth mental health, few have undergone any kind of empirical testing (Weisz & Kazdin, 2017). Similarly, whereas some strategies for reconciling informant discrepancies have undergone rigorous validation testing, many others have not. Importantly, one cannot understand the effects of interventions without also understanding the qualities of the evidence gathered to test intervention effects (De Los Reyes et al., 2017).

A straightforward strategy for addressing the uncertainties created by informant discrepancies is to identify an “optimal” informant for the domain one seeks to assess. However, as previously mentioned, no criteria exist for selecting an optimal informant or gold standard measurement tool (De Los Reyes et al., 2017).

Further, youth behavior varies considerably across contexts, preventing a single informant from capturing all of the relevant information about the mental health domain being measured (Dirks et al., 2012). Consequently, assessors collect multi-informant reports and then leverage various integrative strategies for reconciling discrepant data. Conceptual and methodological issues with many of the existing strategies have been discussed at length, including the practice of averaging informants' reports (i.e., composite scores), combinational algorithms (i.e., AND/OR rules), and latent variable modeling (Barbot et al., 2016; Curran, Georgeson, Bauer, & Hussong, 2020; De Los Reyes, Kundey, & Wang, 2011; Makol, De Los Reyes, Ostrander, & Reynolds, 2019; Martel et al., 2021). Many of these strategies focus on *common variance*, or estimating the variance shared among informants' reports (for a review, see Eid et al., 2008). In using such a strategy, one assumes that informant-specific variance contained in informant discrepancies (i.e., *unique variance*) reflects *measurement confounds* (i.e., bias or measurement error contained within reports that do not relate to the domain being assessed; Millsap, 2011).

Importantly and in the absence of using an integrative strategy, low correspondence often results in assessors dismissing multi-informant approaches and instead making clinical decisions that rely on single informants' reports (De Los Reyes et al., 2015; Loeber Green, & Lahey, 1990; Loeber, Green, Lahey, & Stouthamer-Loeber, 1989; Marsh, De Los Reyes, & Lilienfeld, 2018). For instance, it is common for clinicians in treatment settings to make clinical decisions that are aligned with caregiver reports when youth and caregiver reports disagree (Brown-Jacobsen, Wallace, & Whiteside, 2011; Hawley & Weisz, 2003). Similarly, in treatment studies, researchers commonly select a *primary outcome measure*, which requires a decision about which informant (frequently the caregiver) is best able to capture change over time and the effectiveness of treatment, even though informants' reports consistently display unique patterns of treatment effects (De Los Reyes et al., 2011; Weisz et al., 2017). These decisions stem in part from the assumption that disagreement among informants' reports lacks clinical utility and instead reflects measurement confounds (De Los Reyes, 2011). Idiosyncrasies in use of strategies for reconciling discrepancies, as well as varied empirical support for the use of each strategy, represents a significant clinical decision-making problem for the youth mental health field.

Issues when integrating multi-informant reports are consistent with a broader issue in clinical practice: due to resource limitations and a lack of empirically based guidelines, clinicians often make clinical decisions based on their intuitive impressions about the information obtained in clinical assessments (i.e., *clinical prediction*; Youngstrom, Halverson, Youngstrom, Lindhiem, & Findling, 2018). However, a large body of research demonstrates that idiosyncrasies in clinical decision-making not only commonly occur in clinical work, but that statistical- or actuarial-based decision-making (i.e., *statistical prediction*) tends to result in improved accuracy in decision-making (Grove, Zald, Lebow, Snitz, & Nelson, 2000; Meehl, 1954; Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009; Youngstrom et al., 2018). Despite decades of research on informant discrepancies, integrative strategies that promote statistical- or actuarial-based decision-making and enhance the clinical utility of youth mental health assessments have yet to be developed. Without such strategies, researchers will continue to interpret multi-informant reports in idiosyncratic ways that lack a strong evidence-base. Thus, the pivotal next step for optimizing use of multi-informant reports involves developing evidence-based strategies for integrating these reports that facilitate sound, accurate clinical decision-making.

Theoretical Models

Operations Triad Model (OTM)

A key premise underlying my dissertation lies in a limitation in prior work. Specifically, developers of available strategies rarely draw explicit links between two important components of any sound strategy: (a) the *theoretical model* used to characterize informant discrepancies and the psychological phenomena they reflect, and (b) the *measurement model* designed to instantiate these concepts in quantified estimates of psychological phenomena. Linking theory and measurement is necessary for addressing clinical decision-making issues when using multi-informant reports. Theoretical models inform tests of the phenomena that informant discrepancies reflect. As such, they are instrumental in building an evidence-base for understanding these discrepancies. Further, both theory and this evidence-base help guide selection of informants for use in youth mental health assessments. Only after building these sound conceptual and empirical foundations can one design measurement models that accurately inform strategies for integrating informants' reports. Stated another way, perhaps a key reason for the absence of consensus strategies for integrating informants' reports is that we have yet to understand the degree to which existing measurement models are grounded in research and theory that validly characterizes informant discrepancies.

As with any measurement model, the importance of basing model input in basic science can be characterized by the idea of “garbage-in-garbage-out” (GIGO). That is, if one does not use appropriate measurement conditions (e.g., a well-designed assessment that includes psychometrically sound informants’ reports), no measurement model can glean clinically meaningful data. Similarly, understanding the conceptual and empirical underpinnings of these models allows end-users to determine whether assumptions underlying the models “fit” their intended use or the research questions driving their implementation. Mismatches between measurement models and their conceptual underpinnings and empirical support, as well as the measurement conditions in which the measurement models are used, may translate to errors when interpreting study findings or making clinical decisions. Thus, advancing youth mental health assessments and addressing longstanding issues when reconciling informant discrepancies requires attending to these three “pillars”: (a) theoretical modeling, (b) measurement modeling, and (c) empirical support for these modeling strategies.

The Operations Triad Model (OTM; De Los Reyes et al., 2013) offers a framework for critically evaluating what informant discrepancies reflect (see Figure 1). Specifically, the OTM provides a framework for forming a priori hypotheses about whether informants’ reports will converge, and when not, reasons that reports will diverge. The OTM is agnostic as to whether any particular set of informants converge or diverge for meaningful or methodological reasons. Rather, the model provides tools for generating hypotheses and testing them, and thus supports building the basic science for understanding phenomena linked to patterns of multi-informant reports. Importantly, the OTM does not provide a measurement model. That is, it does not provide strategies for integrating multi-informant reports to improve clinical decision-making. Rather, the OTM offers a first step, much like a "gating procedure", for determining whether patterns of reports validly reflect psychological phenomena. In line with GIGO, this first step ensures that meaningful information is entered into measurement models. The second step then involves merging theoretical and measurement models that allow one to integrate informants' reports.

All theories about the meaning underlying multi-informant reports fall in line with one or more OTM measurement conditions (De Los Reyes et al., 2013). First, *Compensating Operations* reflects a set of measurement conditions for interpreting patterns of inconsistent reports that is based on methodological features of the measures or informants used (see Figure 1, Panel C). For example, two informants' reports may diverge if the informants complete different measures (e.g., two measures of social anxiety with unique items) or measures that vary in their psychometric properties (e.g., two measures that vary in their reliability). As other examples, some informants may purposefully misrepresent information when providing reports (e.g., malingering; Gould, Rappaport, & Flens, 2018) or hold unconscious biases that lead to differential reporting across groups being rated (e.g., racial bias; Fadus et al., 2020; Kang & Harvey, 2020). Key to Compensating Operations conditions is that from a measurement validity standpoint, informant discrepancies reflect measurement confounds. As such, these discrepancies fail to validly reflect psychological phenomena germane to understanding the domain(s) about which informants provide reports (see also Millsap, 2011). Thus, when using discrepant reports that reflect Compensating Operations, one would enhance the clinical utility of reports by using an integrative strategy that minimizes the impact of unique variance on integrative scores (e.g., composite scoring and latent variable modeling). In some instances, support for Compensating Operations conditions would suggest that an informant's report lacks sufficient reliability and validity to justify their use in a mental health assessment.

The other two OTM measurement conditions focus on instances in which patterns of informants' reports reflect valid, domain-relevant psychological phenomena (De Los Reyes et al., 2013). *Converging Operations* reflects a set of measurement conditions for interpreting patterns of consistent reports, or circumstances in which reports yield the same conclusion (see Figure 1, Panel A). For example, informants' reports may converge if the informants observe youth behavior in the same context (e.g., two school teachers rating youth hyperactivity) or are providing ratings about a target that has a more severe symptom presentation (e.g., high levels of pervasive and impairing hyperactivity). In contrast, *Diverging Operations* reflects a set of measurement conditions for interpreting patterns of inconsistent reports based on hypotheses about variations in the behavior being assessed (see Figure 1, Panel B). For example, two informants' reports may diverge if the informants observe youth behavior in unique contexts (e.g., caregiver and teacher ratings of youth hyperactivity) or if youth have a relatively covert symptom presentation (e.g., caregiver and youth ratings of depression). Thus, evidence for

Converging and Diverging Operations conditions supports using information about patterns of convergence and divergence to *enhance* the clinical utility of reports. Stated otherwise, such conditions suggest that minimizing or erasing differences among informants' reports would only *subtract* meaningful information obtained in an assessment.

For decades, research involving multi-informant data has focused almost exclusively on using these data to test hypotheses in line with Converging and Compensating Operations, specifically that convergence among informants' reports reflects "truth" or valid phenomena, and divergence among informants' reports reflects measurement confounds (De Los Reyes, 2011). This emphasis on convergence as truth grew out of Campbell and Fiske's (1959) seminal construct validation paradigm, the multi-trait, multi-method (MTMM) matrix, which is commonly applied to multi-informant data (for a review, see Eid et al., 2008). Users of this paradigm assume that a measure's or set of measures' construct validity is supported by convergence among informants' reports of the same domain (i.e., high monotrait-heteromethod correlations), and conversely, is threatened by low correlations among informants' reports' of the same domain (i.e., low monotrait-heteromethod correlations). Thus, this paradigm emphasizes common variance, and this emphasis promotes the interpretation of informant discrepancies as a threat to validity. In fact, the prevailing strategies used to integrate multi-informant data stem from the assumptions underlying the MTMM matrix (e.g., Eid et al., 2008; Barbot et al., 2016; Howe et al., 2019; Martel et al., 2021; Piacentini et al., 1992). However, in the past decade, a burgeoning area of research has called into question the notion that multi-informant data and the discrepancies they produce should always be assumed to reflect Converging and/or Compensating Operations. Many times, these data might reflect Diverging Operations.

Bias versus Context Models in Informant Discrepancies Research

As previously mentioned, theoretical models offer a starting point for selecting informants and understanding discrepancies among informants in youth mental health assessments. Below, I review two theoretical models commonly used to interpret informant discrepancies in youth mental health assessments. I also review the degree to which evidence supports key contentions made in these models, namely the psychological phenomena (or lack thereof) reflected by informant discrepancies.

Bias Models: Theory and Empirical Support. Bias models focus on how informant-specific factors (i.e., subjective bias) compromise the validity of reports (De Los Reyes, 2011). Most often, these models focus on the degree to which an informant's psychological state "colors" their ability to provide psychometrically sound ratings of a target's behavior. Bias models conceptualize an informant's psychological state in a way that is consistent with classical test theory interpretations of bias: systematic variance that is independent of variance reflected by individual differences in the behavior being assessed (Nunnally & Bernstein, 1994). Thus, bias models are in line with a Compensating Operations hypothesis. Specifically, these models posit that informant discrepancies reflect measurement confounds and thus hold no inherent phenomenological value. Consequently, these discrepancies depress the measurement validity of informants' reports. Given that convergence among informants is conceptualized as being free of bias and containing the true domain being assessed, bias models are also aligned with a Converging Operations hypothesis. As previously mentioned, the emphasis on informant discrepancies as a threat to validity grew out of the use of the MTMM matrix for evaluating construct validity. When using this paradigm, researchers are constrained to the assumption that convergence among informants' reports represents truth and divergence represents measurement confounds (Campbell & Fiske, 1959).

Perhaps the most studied bias model is the *depression-distortion hypothesis*, which holds that depressed informants harbor a negative bias when rating a target's behavior (Ritchers, 1992; Ritchers & Pellegrini, 1989). This theoretical model rests on the assumption that depressed individuals are more likely than non-depressed individuals to attend to, encode, and recall negative information germane to the behaviors being assessed (i.e., relative to positive or neutral information). This selective attention to and memory for negative behavior is thought to result in depressed informants providing more negative reports about the target's behavior (e.g., higher levels of problem behavior), relative to non-depressed informants. This theory makes intuitive sense. Indeed, its popularity is likely due in part to the high base rates of caregiver psychopathology, and particularly depression, observed in youth mental health settings. Specifically, prevalence estimates suggest that approximately 20 to 60% of caregivers of youth in mental health treatment experience clinically significant psychopathology (Cooper, Fearn, Willetts, Seabrook, & Parkinson, 2006; Kim-Cohen, Moffitt, Taylor, Pawlby, & Caspi, 2005; Lahey et al., 1988; Middeldorp et al., 2016). Thus, given the association between caregiver and youth mental health problems, caregivers providing reports in most clinic settings have a high likelihood of having mental health concerns themselves.

Despite decades of research on the depression-distortion hypothesis, support for the hypothesis is inconsistent and most research is limited by significant conceptual and methodological issues. Some studies find that depressed caregivers provide higher ratings of youth mental health concerns compared to non-depressed caregivers and other informants (e.g., Boyle & Pickles, 1997; Briggs-Gowan, Carter, & Schwab-Stone, 1996; Fergusson, Lynskey, & Horwood, 1993; Najman et al., 2000; Lohaus, Rueth, Vierhaus, 2020; Muller, Achtergarde, & Furniss, 2011; Youngstrom, Loeber, & Stouthamer-Loeber, 2000). In contrast, other studies find no support for the depression-distortion hypothesis (e.g., De Los Reyes, Goodman, Kliewer, & Reid-Quiñones, 2010; Conrad & Hammen, 1989; Hawley & Weisz, 2003; Lewis et al., 2012; Makol & Polo, 2018; Olino, Michelini, Mennies, Kotov, & Klein, 2021; Weissman et al., 1987) or varied support across behaviors being rated (e.g., Affrunti & Woodruff-Borden, 2015; Gartstein, Bridgett, Dishion, & Kaufman, 2009; Madsen, Rask, Olsen, Niclasen, & Obel, 2020) or youth age and gender (e.g., Boyle & Pickles, 1997b; Perez, Coo, Irarrazaval, 2018; Renouf & Kovacs, 1994). Thus, no clear pattern of findings has emerged supporting that informants' mood states consistently and significantly bias their reports.

The vast majority of research on bias models and the depression-distortion hypothesis is hampered by issues stemming from a key methodological confound, namely *criterion contamination* (Garb, 2003). That is, in nearly all cases, informants reporting on their own mental health are the same individuals reporting on the target's mental health (i.e., caregiver provides reports about their own depressive symptoms *and* their child's behavior; Affrunti & Woodruff-Borden, 2015; Boyle & Pickles, 1997; Briggs-Gowan et al., 1996; Conrad & Hammen, 1989; De Los Reyes et al., 2010; Fergusson et al., 1993; Lohaus et al., 2020; Gartstein et al., 2009; Hawley & Weisz, 2003; Lewis et al., 2012; Makol & Polo, 2018; Madsen et al., 2020; Muller, Achtergarde, & Furniss, 2011; Najman et al., 2000; Olino et al., 2021; Perez et al., 2018). Thus, in these studies, links between informants' depressive symptoms and their reports about youth behavior may be due to use of the same modality to assess each of these constructs. To avoid criterion contamination and clarify the role of mood on an informant's ratings, one must use a measure of informant mood state that is completely independent from informants' reports (Garb, 2003). The issue of criterion contamination creates challenges when determining whether "bias" adds non-meaningful variance to informants' reports or can be simply explained by shared method variance. This is not a trivial question given that evidence for non-meaningful "bias" would suggest a very different strategy for integrating multi-informant reports than would evidence demonstrating that discrepancies validly reflect psychological phenomena.

As a result of issues with criterion contamination, many alternative hypotheses might explain the association between informant psychopathology and their ratings of a target. First, when interpreting depression-distortion research, it is important to consider that children of depressed caregivers are at heightened risk for psychopathology (Cheung & Theule, 2018; Goodman & Gotlib, 1999). Parent psychopathology may translate to increased risk for child psychopathology via many mechanisms (e.g., genetics, parenting styles, environmental stress, social learning, gene-environment interactions; Carlone & Milan, 2021; Caspi et al., 2004; Goodman & Gotlib, 1999; Moffitt, 2005; Monroe & Harkness, 2005). For instance, compared to a non-depressed caregiver rating youth behavior, a depressed caregiver has a higher likelihood of providing reports about behavior occurring in a home environment that is characterized by stressors (e.g., marital discord) and parent-child interaction patterns associated with the development of depression (e.g., higher levels of expressed criticism, fewer positive interactions; Goodman & Gotlib, 1999; Lindhiem et al., 2020). For these reasons, many evidence-based prevention and intervention programs for young children specifically target maternal depression and the parent-child relationship (e.g., Child-Parent Psychotherapy; Lieberman, Ghosh Ippen, & Van Horn, 2015). Thus, heightened youth mental health reports provided by a depressed caregiver may *meaningfully* reflect heightened mental health problems on the part of the child. As stated by Goodman and Gotlib (1999), higher ratings of mental health concerns among depressed mothers "may not reflect 'accuracy' on the part of the mothers, but instead, may be inadvertently capitalizing on a 'match' of children's more negative behaviors" and characteristics of the environment provided by depressed caregivers (p. 467). The inverse is also true. Caregivers of youth with mental health concerns experience higher levels of stress. For example, having a child diagnosed with attention deficit hyperactivity disorder (ADHD) is associated with elevated parenting stress (Cheung, Aberdeen, Ward, & Theule, 2018; Cheung & Theule, 2019). Thus, a caregiver reporting elevated youth mental health concerns is more likely to also self-report elevated levels of depression and stress. In this way, elevated reports provided by a depressed caregiver may *meaningfully* reflect heightened parenting difficulties.

Given these issues, the fairest and most rigorous evaluation of the depression-distortion hypothesis requires an experimental design in which informants' moods are experimentally manipulated, and I know of two peer-reviewed studies that have done so. One study found that depressed mood had a non-significant effect on informants' reports (Jouriles & Thompson, 1993). Another study that treated caregiver mood as an individual differences variable within an experimental design found that caregiver mood had a statistically significant and moderate effect on the caregiver's reports (i.e., explaining roughly 10% of incremental variance in the informants' reports; Youngstrom, Izard, & Ackerman, 1999). Overall, support for the depression-distortion hypothesis is inconsistent, with only a single experimental study supporting that caregiver mood is associated with over-reporting of youth mental health concerns. Despite these empirical findings and methodological issues, research on the depression-distortion hypothesis continues to be published (e.g., Lohaus et al., 2020; Madsen et al., 2020; Olino et al., 2021).

Context Models: Theory and Empirical Support. In contrast to bias models, context models focus on the impact of where informants observe the youth about whom they provide mental health reports. As mentioned previously, context is crucial to consider when assessing and conceptualizing youth mental health and is a key rationale for why multiple informants' reports are collected. One can trace this theoretical model back to what Achenbach and colleagues (1987) referred to as *situational specificity*. Essentially, youth may vary considerably in how they behave and why they behave as they do, depending on the social context. Further, the informants tasked with reporting on youth behavior typically vary in the social contexts within which they observe youth. As such, multi-informant reports are thought to provide incrementally valid information that is unique to their observational context (De Los Reyes et al., 2013). Home and school are developmentally appropriate contexts that are commonly examined in research that leverages informants' reports (Achenbach et al., 1987; De Los Reyes et al., 2015). However, the contexts navigated by youth vary in numerous other ways that exert a significant influence on their behavior (Kraemer et al., 2003). For instance, contexts can vary in the extent to which they include familiar or unfamiliar peers, include an evaluative component or are free of evaluation, or are structured or unstructured. Youth's broader contexts also vary substantially in ways that shape their behavior. As examples, families vary in communication and parenting practices, classrooms vary in the level of organization provided by teachers, and neighborhoods vary in the likelihood of exposure to adverse childhood experiences (Gonzales et al., 2011; Rimm-Kaufman, Curby, Grimm, Nathanson, & Brock, 2009). Indeed, youth navigate multifaceted contexts in their daily lives and the role of context on behavior is key to understanding their lived experiences.

Consequently, a key premise of various evidence-based youth mental health interventions (e.g., cognitive-behavioral, behavior modification, exposure-based, family systems) is that treatment techniques should be tailored to fit clients' contexts (Weisz & Kazdin, 2017). In fact, those tasked with researching and delivering youth mental health services receive extensive practice with assessing and understanding contextual variations in youth behavior. As evidence of this, consider that work in developmental and child psychology demonstrates that context influences youth in both reciprocal and transactional ways, and serves as protective, risk, and maintaining factors for mental health concerns (Drabick & Kendall, 2010). Relatedly, behavioral norms differ across youth's contexts and different coping strategies are needed to adapt to the requirements of each context (Teglasi, Ritzau, Sanders, Kim, & Scott, 2017). As a second example, consider that several psychiatric diagnoses require assessment of whether symptoms are present across contexts (e.g., ADHD, autism spectrum disorder [ASD], selective mutism) or include specifiers for whether symptoms are present across contexts (e.g., oppositional defiant disorder; social anxiety disorder [SAD]; American Psychiatric Association [APA], 2013).

Given that context is conceptualized as closely linked to the behaviors about which informants provide reports, context theories about the meaning underlying informants' reports are in line with Converging and Diverging Operations conditions. That is, when informants provide reports that converge on a common estimate of a domain (e.g., parent and teacher reports of elevated youth ADHD symptoms), this reflects a cross-contextual consistency in manifestations of that domain (e.g., youth displays concerns across home and school contexts). Conversely, when informants provide reports that diverge or yield distinct estimates of a domain (e.g., elevated parent but not teacher report of youth ADHD symptoms), this reflects a context-specific manifestation of that domain (e.g., youth displays concerns at home to a greater degree than school).

Recent work on informant discrepancies consistently demonstrates that these discrepancies reflect contextual variations in youth behavior. First, context is an established moderator of informants' reports. In Achenbach and colleagues' (1987) seminal meta-analysis, informants providing reports about youth behavior within the same context (e.g., two caregivers or teachers; *r*s = .54-.64) exhibited overall higher levels of agreement than two informants providing reports about unique contexts (e.g., caregivers and teachers; *r*s = .24-.34). In addition, intervention effects for youth mental health treatments are often moderated by informant (Weisz et al., 2017). This finding suggests that interventions likely exert an influence on youth behavior in some contexts (and not others), giving informants unique opportunities to observe these changes in youth behavior. As evidence of this, parent training interventions for conduct problems demonstrate their most robust effects with outcomes completed by the parent, or the informant with the best access to observations of the primary context targeted in treatment (i.e., home; De Los Reyes & Kazdin, 2009). The impact of context on youth behavior speaks to the need to move away from categorical conceptualizations of mental health (i.e., disorder is present or not present) to multifaceted conceptualizations rooted in the social environment's influence on mental health (Achenbach, 2005; Beauchaine & Hinshaw, 2020; Markon, Chmielewski, & Miller, 2011).

Despite overall low-to-moderate levels of convergence in multi-informant reports, informants consistently display individual differences in their reporting patterns. That is, within any one sample that contains two or more informants' reports (e.g., parent and teacher), not all informants' reports disagree with one another (De

Los Reyes, Lerner et al., 2019; Makol, De Los Reyes, Garrido, Harlaar, & Taussig, 2021). Sometimes, informants converge in their reports, and other times, they diverge in their reports. Emerging work suggests that these patterns of convergence and divergence relate to contextual variations in youth behavior. Most of this research focuses on caregiver and teacher reports of externalizing problems and finds that these informants' reports relate to psychosocial functioning and observed behavior across home and school contexts (e.g., caregiver and teacher reports of disruptive behavior track with observed disruptive behavior across home and school context; De Los Reyes, Henry, Tolan, & Wakschlag, 2009; Drabick, Gadow, & Loney, 2007; Wakschlag et al., 2007).

Informants' reports also yield information about youth social functioning across contexts. Specifically, teacher and caregiver reports of aggression and social withdrawal relate to social events encountered by youth across contexts (Hartley, Zakriski, & Wright, 2011), teacher and peer reports of social skills each provide incremental validity in predicting context-relevant social functioning (Kwon, Kim, & Sheridan, 2012), and adolescent-unfamiliar peer reports (but not parent reports) of adolescent social anxiety relate to adolescents' perceived arousal when interacting with unfamiliar peers (Deros et al., 2018). These findings demonstrate that when assessing youth social functioning across contexts, it is important to consider the relevance of the behavior being rated by informants to the child's social demands in that context (Teglasi et al., 2017). Specifically, behaviors are more likely to be expressed in the contexts in which they are relevant and observers are more likely to attend to behaviors when they are relevant to functioning in that context (e.g., teachers focus on youth hyperactivity in the classroom, caregivers focus on their child's relationship with siblings). Overall, this work demonstrates that variations among informants' reports reflect, at least in part, contextual variations in the behavior being assessed.

Patterns among informants' reports across contexts also reflect other psychological phenomena germane to understanding youth mental health. First, patterns among informants' reports yield meaningful information about the severity and impairment of youth mental health concerns. For example, when informant dyads (e.g., caregivers, teachers, or youth) report elevated mental health concerns, youth are more likely to exhibit a more severe symptom presentation or greater impairment (e.g., psychiatric medication use, clinician diagnosis, clinician-rated severity; Azad, Reisinger, Xie, & Mandell, 2016; De Los Reyes, Alfano, Lau, Augenstein, & Borelli; Lerner, De Los Reyes, Drabick, Gerber, & Gadow, 2017; Makol et al., 2019; Wakschlag et al., 2007; Wall, Ahmed, & Sharp, 2018). As one example, Lerner et al. (2017) found that convergence in teacher and caregiver reports of high levels of ASD symptoms at school and home was associated with a higher likelihood that youth received an ASD diagnosis, psychiatric medication, and special education services. Patterns of informants' reports may also signal information about informants' engagement in treatment. Specifically, in outpatient and psychiatric inpatient settings, youth self-reporting lower levels of internalizing problems than their caregiver at intake are at increased risk for poor treatment outcomes (Becker-Haimes et al., 2018) and the provision of higher levels of intervention (Makol et al., 2019). Finally, patterns of informants' reports may signal information about youth risk over time

(Lippold, Greenberg, & Collins, 2013, 2014; Fergusson, Boden, & Horwood, 2009). For instance, Lippold and colleagues (2013, 2014) found that higher parent than youth reports of parental knowledge of youth activities was associated with increased risk for developing substance use problems over time.

Overall, a growing evidence-base supports that informant discrepancies may serve as markers of how youth mental health varies as a function of context. These findings are consistent with developmental psychopathology research supporting that context plays a key role in the onset and maintenance of youth mental health problems, and are aligned with the rationale for collecting multi-informant reports. As discussed previously, the strongest evidence for what informant discrepancies reflect comes from studies that include independent criterion variables (e.g., observed behavior across contexts, academic records) so as to avoid criterion contamination (Garb, 2003). Importantly, context alone does not explain all variations in informants' reports and particularly when they diverge despite observing youth in the same context (e.g., two caregivers; Duhig et al., 2000). Even so, it is important to consider that informants in the same setting each provide a unique interaction context that can differentially impact youth behavior (e.g., two caregivers who vary in their caretaking roles of the youth about whom they provide reports). Further, the salient qualities of contexts and their influence on behavior can vary within and across youth (e.g., individual differences in response to teacher scaffolding or peer influence).

Summary: Bias versus Context Models. Given the ubiquity of informant discrepancies, several theoretical models have been developed that posit what these discrepancies reflect. In the youth mental health assessment field, two theoretical models largely dominate the discussion. These models vary in whether they assume that informant discrepancies validly reflect psychological phenomena, or alternatively, measurement confounds. They also vary in their empirical support. Bias models are rooted in the assumption that convergence among informants' reports captures the "truth" and, consequently, that informant biases drive discrepancies and reduce measurement validity. Empirical work evaluating the most commonly used bias model (i.e., depression-distortion hypothesis) exhibits significant methodological flaws, and at best provides modest support for the impact of depressed mood on informants' ratings. In contrast, context models are in line with leading developmental psychopathology, evidence-based assessment, and evidence-based treatment research. In turn, empirical work evaluating context models demonstrates that variations among informants' contexts of observation (e.g., parent [home] vs. teacher [school) contribute meaningful variance in informants' ratings. These two theoretical models for understanding informant discrepancies suggest qualitatively distinct strategies for integrating informants' reports (i.e., measurement models). Rigorously comparing these distinct theoretical models of informant discrepancies and their empirical support facilitates an equally rigorous comparison of the distinct measurement models they have inspired.

Exemplar Measurement Models

Measurement models follow directly from theoretical models and their empirical evidence. This is a key principle underlying quantitative methodology in Psychology (see Borsboom, 2005). Specifically, through the application of statistical techniques, measurement models allow one to integrate informants' reports in a manner consistent with theory and the content area informing the collection of multi-informant data (e.g., youth mental health). Below, I review two exemplar measurement models derived from bias and context models for understanding what informant discrepancies reflect. Although these two measurement models differ in the underlying theory and empirical support from which they are derived, both are used to arrive at quantifiable, integrated multi-informant indices. As previously mentioned, clinicians and researchers lack empirically based consensus guidelines for reconciling informant discrepancies (Beidas et al., 2015; De Los Reyes et al., 2019; Hunsley & Mash, 2007; Offord et al., 1996). Similarly, although one can identify a number of studies that leverage these models to quantify youth mental health, I know of no prior work that critically evaluates these models, compares and contrasts indices derived from them, and tests the incremental value of these indices, relative to each other.

Bias Model: Bauer et al.'s (2013) Trifactor Model (TFM)

Bauer and colleagues (2013) used the MTMM matrix to inform the development of their TFM for integrating informants' reports. Specifically, Bauer and colleagues assume that by removing variance unique to each informant's report and isolating common variance, one obtains the most accurate index of the domain being assessed. Thus, the TFM adheres to Converging Operations in that it views common variance among informants as signaling the "true" level of the domain being assessed. Yet, in terms of unique variance, the TFM adheres to a Compensating Operations hypothesis about discrepancies among multi-informant reports.

The TFM includes three levels of latent factors that represent sources of variability contributing to individual informants' reports of a target's behavior. First, the *Common Factor* is defined as the consensus among informants in the target's behavior being rated. Bauer and colleagues (2013) state that this factor captures both the construct being assessed as well as shared sources of variability including informants' shared contexts, similar roles of the informant relative to the target (e.g., two parents), or direct information sharing among informants. Second, *Perspective Factors* are defined as the unique views or biases of individual informants (e.g., caregiver depression) as well as other independent sources of variation contributing to informants' ratings including informants' unique observation contexts (e.g., home vs. school) and roles relative to the target (e.g., parent vs. teacher). Thus, the TFM's Perspective Factors are theorized to capture information about unique views, biases, observation contexts, and roles. This suggests that they capture *both* meaningful and non-meaningful variation in informants' reports, and importantly, without distinguishing one from the other. Thus, there is potential for the Perspective Factors to capture meaningful variations between informants' reports. Yet, Bauer and colleagues state that the goal of the TFM is to "generate integrative scores that are purged of the subjective biases of single informants" (p. 475). In this respect, they focus exclusively on informant bias when interpreting the Perspective Factor, even though there remains the possibility that this factor might also contain meaningful information. Third, the TFM includes *Specific Factors*, which are defined as item-level variations in informants' ratings. According to Bauer and colleagues, including Specific Factors allows one to separate out informant-specific variance contributing to ratings at the item-level, and thus provides a more refined assessment of the construct being measured.

Bauer et al. (2013) developed two versions of the TFM (see Figure 2). The *Unconditional TFM* includes the three latent variables that are conceptually defined and mathematically modelled by imposing constraints on factor loadings and factor correlations. First, all informant ratings are loaded onto the Common Factor, given that this factor represents the consensus view or shared variability in item responses across informants (i.e., common variance). Second, unique Perspective Factors are loaded onto each informants' ratings, which captures unique variance for each informant that is unshared with the other informants. Restrictions on model parameters can be imposed that allow researchers to model informants as *interchangeable* (i.e., randomly drawn from one set of raters) or *structurally different* (i.e., selected for the unique information provided; Eid et al., 2008). However, there is a lack of clarity on criteria for determining how to model informants (e.g., based on statistical tests of invariance or the extent to which informants are conceptually similar) across applications of the TFM (e.g., two caregivers are at times modeled as interchangeable and at other times are modeled as structurally different; Bauer et al., 2013; Haeny, Littlefield, Wood, & Sher, 2018). Finally, Specific Factors are modeled to capture unique variance attributable to individual items in the model. Each factor in the TFM is assumed to be orthogonal to all other factors.

Bauer et al. (2013) also describe a *Conditional TFM*, which extends the Unconditional TFM by including predictors of the factors in the model, including factors hypothesized to contribute to biases in informants' ratings. Specifically, one might include predictors of informants' ratings at the target-level (e.g., youth gender) or informant-level (e.g., caregiver depression). Bauer and colleagues argue that including predictors results in increased precision of score estimates, as this allows model users to test hypotheses about sources of variability contributing to informants' reports and remove informants' subjective biases. However, the researchers do not provide guidance on how to use theory or empirical work to select predictors to include in the model.

Original Application of the TFM. In their original study, Bauer and colleagues (2013) applied their TFM to mother and father reports of youth negative affect using 13 parallel items from the Child Behavior Checklist (CBCL; Achenbach, 2001). The sample consisted of youth aged two-to-18 years, with the sample split between youth who had a parent with substance use concerns and matched controls. The researchers included a *calibration sample* to fit, evaluate, and refine the model, and *cross-validation sample* to evaluate the stability of the model. Overall, Bauer and colleagues found good model fit for both the Unconditional and Conditional TFMs. Informant-level predictors included mother and father lifetime history of an alcohol use disorder, depression or dysthymia, and antisocial personality disorder. The researchers did not include any measure of *current* caregiver mental health functioning, an important omission in light of the model's foundation in the depression-distortion hypothesis. Caregivers with a lifetime history of dysthymia or depression and antisocial personality disorder were more likely to rate higher levels of child negative affect compared to caregivers without this lifetime history. The researchers concluded that these caregivers perceived their child's affect to be "greater than it is commonly perceived to be" (p. 486). Thus, within the TFM, unique variance attributed to these caregivers is deemed a measurement confound.

Bauer and colleagues (2013) found that standardized factor loadings for the Common Factor (*Range* = .08-.65, *M* = .43) were often lower than for the Perspective Factors (*Range* = .33-.63, *M* = .50) and Specific Factors (*Range* = .00-.69, *M* = .39). The researchers interpreted this pattern of factor loadings as indicating that the “true” level of the domain being rated contributes less variance to informants’ ratings than do informant- or item-level factors. The researchers conducted sensitivity analyses in which the final model was applied to the cross-validation sample and found overall excellent fit when comparing intercepts, loadings, and factor regression parameter estimates.

Bauer and colleagues (2013) also correlated the Common Factor with other integrative strategies including: (a) average proportion of items endorsed by caregivers (much like the “AND rule”), (b) average proportion of items endorsed by either mothers or fathers (much like the “OR rule”), (c) average factor score estimates obtained in a two-parameter logistic item response theory model, and (d) average factor score estimates obtained in a moderated nonlinear factor analysis model. The researchers found that the Common Factor correlated at large magnitudes with all other integrated scores (*r*s = .79-.84) but that the other integrated scores correlated with each other at even larger magnitudes (*r*s = .93-.99). Bauer and colleagues concluded that although all integrative strategies assessed negative affect, the TFM produced a more interpretable score due to separating out unique sources of variance and removing caregivers’ subjective biases. However, the researchers did not evaluate the ability of the Common Factor, or other integrative scores, to predict independent criterion variables. Thus, the *validity* of data derived from the TFM (i.e., incremental, criterion-related), and in particular the Common Factor, awaits proper testing.

Summary of the TFM. Bauer et al.'s (2013) TFM focuses on common variance as reflecting validity and is rooted in theories about the role of bias on informants' ratings. Using a confirmatory factor analytic approach offers many important strengths, and in particular the ability to model factors and external predictors often hypothesized to impact informants' ratings. Doing so allows for hypothesis-testing, given that factors underlying informants' reports are determined a priori and then tested statistically using global fit statistics. If poor statistical fit is obtained, the theorized factors and their relations to observed variables can be rejected or refined. In addition, the TFM is unique in its modeling of item-level data. The emphasis on items encourages researchers to select parallel measures across informants, thus reducing measurement confounds.

There are several limitations to the TFM. First and perhaps most notably, the TFM's focus on depressive bias is rooted in an inconsistent and methodologically flawed literature. As mentioned previously, a key principle underlying quantitative methods in Psychology involves applying measurement models to data in ways that adhere to how researchers collect, use, and interpret those data in specific content areas (Borsboom, 2005). In this respect, the TFM is misaligned with the science on informant discrepancies in youth mental health assessments. Thus, even when using methodologically rigorous strategies, placing emphasis on informant bias, and in particular a depressive bias, are unlikely to result in valid indicators of measured domains if the data conditions fail to support such an emphasis.

Second, the TFM does not emphasize theory when determining model input including selection of informants and predictors to include in the model. This lies in stark contrast to long-held “best practices” in factor-analytic modeling, which hold that these models only yield meaningful results when users select psychometrically sound and empirically based indictors (Fabrigar & Wegener, 2011; Kline, 2016; Nunnally & Bernstein, 1994; Pett, Lackey, & Sullivan, 2003; Tabachnick & Fidell, 2013). Interestingly, the TFM appears to presume that informants are inherently biased but that applying a factor analytic approach transforms their reports to be free of bias (i.e., “garbage in” does not lead to “garbage out”). Third, when implementing the TFM, assumptions are imposed at multiple steps and may need to be modified to be consistent with the data. For example, informants in the TFM can be modeled as structurally similar by imposing equal item intercepts and factor loadings for each informant. Thus, applying the TFM requires users to make numerous sample-specific modifications in an effort to yield acceptable model fit. In this way, as a general rule use of the TFM may result in over-specified models. This over-specification logically results in (a) decreased clarity when interpreting factors in the TFM for any one study, (b) increased uncertainty when comparing findings across TFM studies, and thus (c) barriers when seeking to understand the generalizability of findings for studies that use the TFM.

Exemplar Context Model: Kraemer et al.’s (2003) Trait Score Satellite Model (TSSM)

Kraemer and colleagues’ (2003) *Trait Score Satellite Model (TSSM)* provides an exemplar measurement model that is rooted in evidence supporting that context meaningfully impacts informants' ratings of a target's behavior. Key to Kraemer and colleagues' TSSM is the strategic selection of informants who systematically vary in the contexts (e.g., home vs. school) and perspectives (e.g., self vs. other) from which they rate youth mental health (see Figure 3). The concepts underlying this measurement model are rooted in Achenbach and colleagues' (1987) theory of situational specificity and share metaphorical links with the global positioning systems (GPS) used to track objects or people in geographic space. Specifically, GPS systems acquire accurate location information, insofar as the positioning of satellites focuses on systematically placing them at strategically identified locations. In essence, three satellites placed at the same latitude and longitude in geographic space make for a rather imprecise system for locating that object. In contrast, three satellites placed at varying latitudes and longitudes—in an effort to form a *triangulated* position relative to the object's location—will, on average, result in accurate location information.

Similarly, Kraemer et al. (2003) argue for the importance of identifying factors that allow assessors to systematically "mix and match" the three informants used in the measurement model (see Figure 3, Panel A). As with latitudes and longitudes in GPS, the informants should vary on their "positioning" relative to factors that predict or explain variance in discrepancies among informants' reports (see Figure 3, Panel B). If, like GPS, the goal is to triangulate on estimating the target youth's level of the behaviors being assessed, then by definition a key premise of the approach involves selecting informants for whom past research indicates their reports will disagree, and for meaningful reasons. As stated by Kraemer and colleagues: "The lack of correlation (orthogonality) between informants, to date considered problematic, becomes precisely the phenomenon that facilitates a more valid measure" (p. 1568). In this way, Kraemer and colleagues argue for an approach that markedly departs from that of Bauer and colleagues (2013), namely that the unique variance reflected by informant discrepancies should be used to enhance the precision, accuracy, and clinical utility of multi-informant assessments. Their approach adheres to Converging and Diverging Operations hypotheses, and the need to capture both common *and* unique variance when assessing youth mental health and integrating multi-informant reports.

To implement the TSSM, Kraemer et al. (2003) applied principal components analysis (PCA) to multi-informant reports. This statistical approach linearly transforms a set of variables into a smaller set of uncorrelated variables (Dunteman, 1989). Further, it is an exploratory technique that has no underlying statistical model; it is most commonly used to reduce the dimensionality of a set of item responses or variables by transforming them into a smaller set of variables that are easy to interpret and use in subsequent analyses. Kraemer and colleagues adapted this approach to aggregate multiple informants' reports, under the assumption that the informants most often used in youth mental health research are of the structurally different variety as defined by Eid and colleagues (2008). This assumption of structurally different informants is in line with PCA modeling parameters and assumptions, as it yields linear composites of a set of latent variables, with each composite being orthogonal to all other composites. Further, when using approaches such as PCA, the "key to success" is carefully selecting theoretically meaningful items that are correlated but not redundant (Fabrigar & Wegener, 2011; Nunnally & Bernstein, 1994; Tabachnick & Fidell, 2013). In line with this aim of PCA, the TSSM requires a user to select structurally different informants who systematically vary in ways that result in low-to-moderate convergence among reports, and with each informant contributing incrementally valid information when assessing youth mental health.

When strategically selecting informants who vary in their contexts and perspectives in the TSSM, Kraemer and colleagues (2003) argue that three components can be accurately identified through examination of components weights. Specifically, one component (i.e., *Trait* score) reflects variability for which all three informants' reports load strongly and in the same direction. As the first component obtained in PCA analyses, the Trait score explains the most variance in informants' reports and is the key integrative score to be used in subsequent analyses. A second component (i.e., *Context* score) reflects informants' contexts such that informants from different contexts load in opposite directions. A third component (i.e., *Perspective* score) reflects informants' perspectives such that self-reports load in the opposite direction of observer informants' reports. Although informants are selected who vary in their context and perspective, PCA provides the crucial next step of removing extraneous variance due to these components. Kraemer and colleagues state that their model allows for random factors related to the method by which the informants' reports are collected (i.e., error), but conceptualize error as explaining only a small and insignificant amount of variance in informants' reports.

Original Application of the TSSM. In their original study, Kraemer et al. (2003) applied the TSSM to mother, teacher, and child reports of internalizing problems, externalizing problems, and academic functioning. Mothers and teachers completed parallel forms of the MacArthur Health and Behavior Questionnaire (Essex, Klein, Miech & Smider, 2001) and children completed the Berkeley Puppet Interview (Ablow et al., 1999). Kraemer and colleagues observed that for each area of functioning assessed, these reports, which met their "mix-and-match criterion", yielded component weights consistent with Trait (i.e., all informants' reports loaded strongly and in the same direction), Context (i.e., mother reports loaded in the opposite direction of teacher reports, with child reports loading between parent and teacher reports), and Perspective Scores (i.e., self-reports loaded in the opposite direction of mother and teacher reports). Interestingly, Kraemer and colleagues also applied the TSSM to three informants who did not meet their mix-and-match criterion. Specifically, using the same sample, the researchers applied the TSSM to mother, father, and teacher reports of internalizing problems. When doing so, they found that PCA did not yield component weights consistent with Trait, Context, or Perspective scores. In essence, effectively implementing the TSSM requires strategic selection of informants whose reports "fit" assumptions underlying its use. That is, entering informants into the model whose reports do not systematically vary along key dimensions results in component scores that fail to reflect hypothesized patterns. In this respect, the TSSM results in data that offers clinical utility, but only insofar as a user adheres to the mix-and-match criterion.

Kraemer et al.'s (2003) original TSSM study did not conduct validation testing for scores derived from the model. Recently, Makol et al. (2020) applied the TSSM to caregiver, adolescent, and unfamiliar peer confederate reports of social anxiety in a mixed clinical/community sample of adolescents. These informants systematically varied in the context (i.e., home vs. non-home context) and perspective (i.e., self vs. other) from which they observe adolescent social anxiety. Consistent with the development of the TSSM, Makol et al. found that the Trait score provided incremental validity over-and-above individual informants' reports in predicting adolescent referral status (i.e., community control vs. clinical-referred group; odds ratios [ORs] = 2.66-6.53) and observed social anxiety in interactions with unfamiliar peer confederates (i.e., βs = .47-.67).

Additionally, Makol et al. (2020) evaluated whether the Trait score outperformed a commonly used integrative strategy, namely taking a composite or simple average of informants' ratings. Interestingly, the composite score strategy shares similarities with Bauer et al.'s (2013) TFM. In particular, the composite score treats variations among informants' reports as measurement confounds, consistent with a Compensating Operations hypothesis. This idea has its origins in classical test theory (Borsboom, 2005), and in particular Edgeworth's (1888) assertion that variations among individual raters represents "error" around a "true score" of the construct being rated. When comparing the TSSM and composite score, Makol et al. found that the Trait score provided incremental validity over-and-above the composite score when predicting adolescent referral status and observed social anxiety. In contrast, the composite score did not provide incremental validity over-and-above the Trait score when predicting these independent criterion variables. Overall, this study demonstrated that integrating strategically selected informants using the TSSM may optimize informants' reports when used as predictors in youth mental health research.

Summary of the TSSM. Kraemer and colleagues' (2003) TSSM is rooted in research demonstrating that context influences informants' ratings of youth mental health. The researchers leveraged PCA, a factor analytic approach typically used to aggregate multiple item responses on a measure, to obtain orthogonal composites of factors contributing to informants' ratings. However, being an exploratory approach, statistically testing the "fit" of the data with these components is not possible. Although I know of one study that has conducted validation testing of scores derived from the TSSM (Makol et al., 2020), more research is needed to understand how and when this integrative strategy can be used to optimize multi-informant assessments and improve clinical decision-making when using multi-informant reports.

Kraemer and colleagues' TSSM has some limitations. First, it is designed to be used with three or more informants and may not be suitable for assessments leveraging only two informants. Further, determining whether one has appropriately identified the components is a qualitative process (i.e., examining component weights for "Trait", "Context", "Perspective") for which Kraemer and colleagues do not offer thresholds. Further, although the components described make conceptual sense, little work has been done to validate the meaning of scores derived from the TSSM (e.g., does the "Context" score reflect context?). Relatedly, although the Trait score is the primary score obtained, it is unclear if the Context and Perspective scores could be used to understand how youth behavior varies across key contexts (e.g., is elevated in the home, and not school, setting) and perspectives (e.g., intrapersonal vs.

interpersonal aspects of mental health). Finally, because PCA is an exploratory technique, hypothesis testing of the derived components is not appropriate (Tabachnick & Fidell, 2013). As such, and as with any use of PCA, when using the TSSM it is *imperative* to leverage validation testing strategies to determine whether a user is appropriately interpreting the observed components.

The Need for Rigorous Research

I reviewed two key theoretical and exemplar measurement models for reconciling informant discrepancies. Both measurement models are based in distinct theoretical models that seek to explain patterns among informants' reports. Each of these theoretical models have been featured prominently in the past several decades as foundations for the integrative strategies researchers use to reconcile informant discrepancies. Both measurement models use factor analytic approaches for integrating informants' reports of a single domain and for a target individual. In these measurement models, common and unique sources of variance are modeled across informants to obtain an integrated score that can be used in subsequent analyses. Further, both measurement models focus on how the observed variables collected in multi-informant assessments relate to underlying latent factors or components.

The two measurement models also differ in key ways. First, while the TSSM offers an explicit theoretical model for how to select informants, the TFM does not (Bauer et al., 2013; Kraemer et al., 2003). The two models also differ in how sources of variation among informants' reports are interpreted. The TFM focuses on common variance and how informants' subjective biases contribute non-meaningful, unique variance to their ratings, which is in line with Converging and Compensating

Operations hypotheses. In contrast, the TSSM focuses primarily on how informants' contexts contribute meaningful variance to their ratings, and the idea that youth might behave in ways that manifest consistently across contexts or alternatively, manifest uniquely in specific contexts. As such, the TSSM is in line with Converging and Diverging Operations hypotheses. Perhaps most importantly, the two measurement models rest on two theoretical models that differ in their evidentiary foundations. The TFM's focus on informant bias lacks strong empirical support, or demonstrations that bias robustly, and at large magnitudes, explains informant discrepancies. In contrast, the TSSM's focus on informant context is supported by a great deal of empirical support for the role of context in explaining informant discrepancies.

Second, the two measurement models differ in whether they are exploratory or confirmatory. The TSSM offers an exploratory approach for deriving components that synthesize informants' reports in ways that are consistent with the literature on what informant discrepancies reflect (e.g., contextual variations in youth behavior). This prevents users of the TSSM from evaluating model fit, an important step in determining whether a measurement model reflects the data. Regardless of conceptual and empirical issues, a strength of the TFM's confirmatory approach is that factors hypothesized to explain discrepancies between informants' reports can be represented by a series of structural equations and then empirically tested through evaluation of model fit. In contrast, exploratory approaches are best suited for instances when the links between the observed and latent variables are unknown (Byrne, 2013). Based on decades of research on youth mental health assessment, there is strong evidence that context factors prominently into informant discrepancies, whereas bias does not.

In sum, these issues suggest that the degree to which measurement models differ on their theoretical foundations may result in not only differences among measured variables derived from these models, but also their incremental value, relative to each other. Thus, an important next step in research on integrating multi-informant reports of youth mental health involves directly testing which measurement model is best supported by the evidence. That is, when implementing each model, which performs best in predicting key indices relevant to understanding youth mental health? To accomplish this, I will independently evaluate and then compare scores derived from the TFM and TSSM (Bauer et al., 2013; Kraemer et al., 2003). Very little validation testing has been conducted for the two integrative strategies separately, and no prior study has tested them against each other.

Key Study Characteristics of Rigorous Validation Testing

I identified four key study characteristics for rigorous validation testing of these integrative strategies. First, validation testing should include use of parallel instruments across informants (i.e., parallel item content, scaling, and response options). Doing so decreases the likelihood that variations among informants' reports are the result of measurement artifacts associated with the instruments used to collect informants' reports (De Los Reyes et al., 2013). Stated another way, using parallel instruments decreases the likelihood that any *Compensating Operations* hypothesis other than rater biases explains differences among informants' reports. Second, validation testing should examine various dyad pairs (e.g., parent-child, teacher-child, parent-teacher) to evaluate the role of informant selection when applying each theoretical model and associated measurement models. In line with GIGO, informant selection is integral when drawing on empirical and theoretical literatures.

Third, in psychometrics and factor analysis, identifying the factor structure of items is a prelude to the crucial step of evaluating how the structure relates to external criterion variables (DiStefano & Hess, 2005; Kline, 2016; Nunnally & Bernstein, 1994; Pett et al., 2003). Stated otherwise, determining the validity of scores taken from these models factors prominently in determining their utility in research and applied settings. This issue is relevant to both the TFM and TSSM given that researchers subjectively interpret factor and component scores without directly testing these interpretations (e.g., youth negative affect with caregiver depressed bias removed or contextual variations in youth behavior; Bauer et al., 2013; Kraemer et al., 2003). Doing so has been termed the "naming fallacy" or false belief that naming a factor indicates that what the factor measures is understood (Kline, 2016, p. 466). In research on both measurement models, there is a lack of well-constructed validation studies focused on testing the "truth" underlying interpretations of factors and components. In particular, we require rigorous validation testing to determine whether the scores that users extract from the models capture what they purport to. In particular, validation testing should determine how well each measurement model performs when estimating criterion variables that are independent from individual informants' reports (Garb, 2003). Commonly used independent criterion variables may include diagnosis, observed behavior, treatment response as determined by an independent rater, referral status, and psychiatric medication use (Azad et al., 2016; Becker-Haimes et al., 2018; Lerner et al., 2017; Makol et al., 2019, 2020). Using independent criterion variables allows one to avoid *criterion contamination*, which occurs when informants' reports provide differential prediction of criterion variables, given that the criterion variable has an overlapping information source with one of the predictors (e.g., caregiver rates their own mood as well as their child's mood; De Los Reyes et al., 2015; Garb, 2003).

Fourth, rigorous research should conduct tests of the incremental validity of each measurement model to determine the incremental value of each model over-and-above (a) alternative information sources (e.g., individual informants) and (b) alternative integrative strategies (e.g., composite scores; Hunsley & Meyer, 2003; Smith, Fischer, & Fister, 2003). These tests ensure that science maximally informs clinical decision-making and usability in research and applied settings. A key goal of incremental validation strategies is determining the maximal complexity needed to make accurate, valid, and efficient clinical decisions. In essence, simpler analytic strategies requiring fewer data points are preferred if more complex strategies requiring greater data points do not offer incremental prediction of key clinical indices (Youngstrom et al., 2018). In this respect, we require incremental validity tests to directly examine whether the more complex modeling techniques used to derive scores taken from the TFM produce incrementally valuable data, relative to TSSM-derived scores, which require relatively less complex techniques.

More broadly, both the TFM and TSSM require use of a comprehensive multi-informant assessment approach. This approach is relatively more resource intensive compared to collecting single informants' reports (i.e., administering, scoring, and interpreting each informant's reports). However, it is unclear whether informants’ reports can be used in such a way as to enhance the information obtained in assessments. Thus, incremental validity tests are important for determining if an integrative tool increases the predictive power of multi-informant assessments, relative to examining individual informants’ reports or an alternative integrative strategy. Rigorous research meeting these four criteria will provide an empirically based approach to determining which measurement model, and associated theoretical model, best meets the underlying assumptions and rationale for collecting, interpreting, and integrating multiple informants’ reports of youth mental health.

Chapter 2: Dissertation Studies

This dissertation addresses major gaps in theory, methodology, and research on youth mental health assessments. In my Introduction, I critically examined the two leading theoretical models of informant discrepancies and the two measurement models they inspired. My dissertation studies described below examine the methodology and empirical support for these measurement models. In Study 1, I characterize and compare the methods of published studies using the TFM and TSSM. In Study 2, I empirically compare data derived from the measurement models, namely evidence of their criterion and incremental validity when predicting independent criterion variables. These studies have important implications for developing consensus strategies for integrating multi-informant assessments of youth mental health.

Study 1: Systematic Review of Informant Selection across Measurement Models

Ideally, theoretical models directly inform measurement models. Theoretical models provide conceptual frameworks for understanding what informant discrepancies reflect. Empirical work evaluating these models facilitates determining the validity of multi-informant assessment approaches and appropriate informants to select in assessments. The two leading measurement models evaluated in this dissertation differ in their conceptual foundations, namely in the degree to which their usage assumes that informant discrepancies reflect measurement confounds (i.e., bias models such as the TFM) or contextual variations in behavior (i.e., context models such as the TSSM). The overarching approach I took to this dissertation involved linking the three “pillars” of integrative strategies for informant discrepancies described previously (i.e., theoretical modeling, measurement modeling, empirical support for each modeling approach). In line with this approach, in my first study I examined the degree to which researchers leveraging the TFM and TSSM varied in which and how many informants they included in their model.

Aim 1

I conducted a systematic review of published studies in which researchers leveraged the TFM and TSSM, to elucidate how users of these measurement models select informants.

Hypothesis 1

I hypothesize that, relative to TFM studies, TSSM studies will be more likely to (a) include a greater number of informants in the measurement model and (b) select informants who vary in the contexts in which they observe behavior. This hypothesis is based on the theory underlying the two measurement models. Following from situational specificity (Achenbach et al., 1987), the TSSM offers explicit guidance on how to select informants. That is, informants should systematically vary in *where* they observe behavior (i.e., context) and *how* they rate behavior (i.e., perspective; Kraemer et al., 2003). Consequently, I posit that users of the TSSM focus on which informants they select, and the degree to which these informants vary in their opportunities to observe youth mental health domains within and across multiple contexts. In contrast, following from research on the depression-distortion hypothesis (Richters, 1992), the TFM provides no guidance on selecting informants (Bauer et al., 2013). I posit that, in light of these theoretical foundations, those who leverage the TFM assume that the informants they select have little bearing on the validity of estimates of youth mental health domains. Indeed, this logically follows from a key premise underlying use of the TFM, namely that all unique variance from informants' reports ought to be "purged" to attain valid, integrated estimates of youth mental health (i.e., presumably from Common Factor scores).

Study 2: Application and Comparison of the TFM and TSSM

As previously described, we require validation testing that critically evaluates leading measurement models for integrating multi-informant data in youth mental health assessments. Specifically, rigorous research should (a) use parallel measures across informants to reduce methodological artifacts, (b) examine model fit and performance when using varied informant dyads, (c) evaluate the ability of factor scores derived from measurement models in predicting clinically relevant independent criterion variables, and (d) conduct incremental validity tests. In Study 2 of this dissertation, I implemented the TFM and TSSM using a study design that meets these methodological criteria. Specifically, I applied the TFM and TSSM to multi-informant reports of adolescent social anxiety collected in a controlled laboratory study that includes a well-characterized, mixed clinical/community sample of adolescents (for an overview, see Cannon et al., 2020). This sample is ideal for initial validation testing of the two measurement models, given the high degree of experimental control, use of parallel measures across three informants, and inclusion of clinically relevant independent criterion variables. In addition, the multi-informant and multi-modal assessments collected in this sample demonstrate strong psychometric properties and have been thoroughly tested across several investigations (e.g., Cannon et al., 2020; Deros et al., 2018; Glenn et al., 2019; Makol et al., 2020; Qasmieh et al., 2018; Rausch et al., 2017).

Aim 1

Implement the TFM using various informant dyads' reports (i.e., caregiver-adolescent, caregiver-peer, adolescent-peer) as well as three informants' reports (i.e., caregiver-adolescent-peer) of adolescent social anxiety.

This aim is exploratory in nature and I make no specific hypotheses regarding model fit for the TFMs tested. The exploratory nature of this aim follows from prior work. Specifically, the original TFM study (Bauer et al., 2013) found good model fit across all fit indices. In contrast, subsequent work either (a) observed variability across model fit indices when implementing the TFM (e.g., Clark, Durbin, Donnellan, & Neppl, 2017; Martel, Nigg, & Schimmack, 2017b) or (b) did not report model fit indices (e.g., Curran et al., 2020). As such, I made no specific hypotheses regarding observations of model fit for the TFM.

Aim 2

Implement the TSSM consistent with its use in past work (Makol et al., 2020), namely using caregiver, adolescent, and peer confederate reports of adolescent social anxiety. I only implemented one version of the TSSM using three informants' reports, given that the model, by design, "triangulates" on the domain being assessed through use of three informants' reports who are assumed to vary systematically in their perspectives and contexts of observation (Kraemer et al., 2003).

Hypothesis 2

Consistent with prior work on an earlier version of this sample (i.e., before data collection was complete; Makol et al., 2020) and the original TSSM study (Kraemer et al., 2003), I hypothesize that the TSSM will meet factor analytic criteria and yield components consistent with Trait (i.e., all informants' reports load strongly and in the same direction), Context (i.e., caregiver reports load in the opposite direction of peer confederate reports, with adolescent reports loading between caregiver and peer confederate reports), and Perspective scores (i.e., self-reports load in the opposite direction of caregiver and peer confederate reports).

Aim 3

Evaluate whether the TFM Common Factor and TSSM Trait scores display incremental validity in distinguishing adolescents on clinical referral status and predicting their observed social anxiety in controlled laboratory tasks, relative to well-established measures of social anxiety taken from individual informants' reports as well an average of informants' report (i.e., composite score). As previously mentioned, a common strategy for reconciling informant discrepancies in research and clinical practice is to identify the "optimal" informant or analyze multiple informants' reports separately (De Los Reyes et al., 2011; Howe et al., 2019). Another common strategy, the composite score, assumes that unique variance reflects measurement confounds and needs to be removed to accurately capture the domain being measured (Borsboom, 2005; Edgeworth, 1888). I selected these two competing strategies for integrating informants' reports given that they are commonly used and share assumptions with one of the measurement models (i.e., using individual informants' reports assumes each informant provides useful information, averaging informants' reports assumes unique variance represents measurement confounds). Thus, this aim will facilitate determining whether the integrated, multi-informant reports derived from TSM and TSSM models provide incremental prediction of independent criterion variables, beyond other widely used strategies for extracting data from multi-informant assessments of youth mental health.

Hypothesis 3

Given the lack of prior research on the criterion and incremental validity of the Common Factor score, I make no specific hypotheses about the ability of the Common Factor scores to predict independent criterion variables, over-and-above individual informants' reports and the composite score. One possibility is that by capitalizing on the common variance among informants' reports and removing informants' "subjective biases" (Bauer et al., 2013), the Common Factor scores will display criterion-related and incremental validity, relative to individual informants' reports. Following this same logic, the Common Factor scores should display criterion-related and incremental validity, relative to the composite score, an integrative strategy that emphasizes common variance but includes no procedures for isolating unique variance or treating such variance as "bias". Alternatively, bias models (a) lack robust empirical support and (b) remove unique variance from informants' reports and treat such variance as "bias". Consequently, Common Factor scores may be unable to contribute unique variance in criterion-related and incremental validity analyses. In contrast, I hypothesize that the Trait score will provide criterion and incremental validity over-and-above individual informants and the composite score when predicting independent criterion variables. This hypothesis is supported by prior work in the same sample, albeit with a smaller sample size, finding that the Trait score provides criterion-related and incremental validity relative to these alternative strategies for integrating informants' reports (Makol et al., 2020).

Aim 4

Evaluate whether the TFM Common Factor score derived using caregiver, adolescent, and peer confederate reports provides incremental validity in distinguishing youth on clinical referral status and predicting observed adolescent social anxiety in controlled laboratory tasks, relative to the Trait score, and vice versa. As previously mentioned, it is important to empirically evaluate the incremental validity of measurement models over-and-above alternative integrative strategies. Thus, this aim will aid in determining whether integrating multi-informant reports using each of these two measurement models increases the predictive power of multi-informant assessments, relative to use of the other measurement model.

Hypothesis 4

When comparing the two integrative strategies, I hypothesize that the Trait score will display incremental prediction of independent criterion variables, over-and-above the Common Factor score, whereas the reverse will not be found. This hypothesis is based on the greater empirical support for the meaningful impact of context on patterns among informants' reports, compared to the inconsistent and methodologically weak empirical base supporting the impact of bias on informants' reports. Further, this hypothesis is supported by prior work demonstrating that the Trait score provides incremental validity, over-and-above the composite score (Makol et al., 2020). As stated above, the composite score, much like the TFM, assumes that unique variance reflects measurement confounds (Borsboom, 2005; Edgeworth, 1888; Martel et al., 2021).

Chapter 3: Systematic Review of Informant Selection across Measurement Models (Study 1)

Study 1 Method

Search Strategy and Inclusion Criteria

I conducted a search using PsycINFO and PubMed to identify peer-reviewed articles applying the TFM and TSSM. For TFM studies, the search first entailed obtaining PsycINFO and PubMed references citing the original TFM study (i.e., Bauer et al., 2013) as well as a recent article from the same research team applying the TFM (i.e., Curran et al., 2020). Next, I used PsycINFO and PubMed to search for any additional articles published between September 29, 2013 (i.e., date of publication) and August 31, 2020 by querying the following term: “Trifactor Model”. Given that Bauer et al.’s (2013) TFM is the only model with this name, I used no other search terms. For TSSM studies, the search first entailed obtaining PsycINFO and PubMed references citing the original TSSM study (i.e., Kraemer et al., 2003). Next, I used PsycINFO and PubMed to search for any additional articles published between September 1, 2003 (i.e., date of publication) and August 31, 2020 by querying the following terms: “Trait score” AND “principal components analysis”. I used both search terms given that the term “trait score” is used in other research areas (e.g., personality traits, trait anxiety). I removed any duplicate articles obtained through these searches. To maximize the number of studies included in the review, my final study list was inclusive of all studies using these models, regardless of: (a) the domain being assessed using the measurement model, (b) developmental stage of the target being rated by informants, and (c) country in which the study was conducted.

Coding Procedure and Codebook

In the first stage of coding (*Screening and Eligibility*), the coders screened for eligibility using the inclusion criteria described above. In the second stage of coding, the coders identified the methodological features of included articles using the variable codebook (*Methods Coding*, see Table 1).

Coding was completed by two masters-level research assistants who were masked to study hypotheses. Prior to beginning any coding, the coders read the original TFM (Bauer et al., 2013) and TSSM (Kraemer et al., 2003) studies and participated in training meetings to learn about each measurement model. In the first stage of coding (*Screening and Eligibility*), the coders evaluated studies using three key inclusion criteria: (1) published in a peer-reviewed journal (*Peer Review*), (2) written in English (*Language*), and (3) applied the TFM or TSSM as described by the original research team (*Measurement Model*). If a study was deemed ineligible based on an exclusion criterion, I instructed the coders to stop after completing coding for that variable. The two coders completed a trial of ten sample studies and discussed and reconciled any coding errors with me in a group meeting. For this trial coding, one coder completed coding with 100% accuracy and the other with 80% accuracy. For this reason, the second coder completed additional training with me and coded an additional 10 sample articles which were coded with 100% accuracy. Coders completed a second round of training and trial coding for five sample articles to prepare for the second stage of coding (*Methods Coding*). The coders completed this trial with 84-86% accuracy and received individualized feedback from me on any errors made. In a team meeting, we discussed coding discrepancies at length to ensure team consensus on the appropriate approach to coding each variable, making minor modifications as needed to the coding instructions to reflect the complexity of the studies included (e.g., studies with a longitudinal design or employing non-traditional "informants" like official records). While reviewing studies, I instructed coders to note any additional applications of the TFM and TSSM that they found in the text of the article. This resulted in an additional three applications of the TSSM being identified that were not found through the PsycINFO and PubMed search results.

Search Results and Reliability

I identified a total of 646 articles in the initial PsycINFO and PubMed search and removed 175 duplicate articles. The two independent coders each screened 50% of the identified articles for eligibility. For reliability purposes, the two coders also double coded an additional 30% of the articles (471 total articles per coder). Of note, the total number of double coded articles varied across the three exclusion criteria given that coding ceased with the variable at which the article was deemed to be ineligible. To estimate inter-rater reliability for coding of continuous variables, I calculated the Intraclass Correlation Coefficient (ICC) statistic and interpreted it using thresholds described by Cicchetti (1994). To estimate inter-rater reliability for coding of categorical variables, I calculated the kappa statistic and interpreted it using thresholds described by Viera and Garrett (2005). Coders had perfect or almost perfect agreement across the Screening and Eligibility variables: Peer Review (1 discrepancy; $\kappa = .89$, $p <.001$), Language (0 discrepancies; $\kappa = 1.00$, $p <.001$), and Measurement Model (7 discrepancies; $\kappa = .86$, $p <.001$). In a team meeting, we reconciled these discrepancies to ensure consensus among coders. In the Methods Coding phase, all included studies ($n = 47$) were coded by both coders. For categorical variables, the coders achieved moderate agreement in their coding of Age (19 discrepancies; $\kappa = .51$, $p <.001$) and moderate to substantial agreement in their coding of Informant Type (average discrepancies = 7; average $\kappa = .67$; *p*'s <.001). For continuous variables, the coders achieved moderate agreement in their coding of Number of Informants (*ICC* = .74, $p < .001$) and excellent agreement in their coding of Sample Size (*ICC* = .99, $p < .001$). Coder discrepancies were most common for articles with multiple applications of the measurement model, wide age ranges, multiple time points, or a lack of clarity between coders on the target being rated when the domain measured captured aspects of the target's environment (e.g., marital discord, parenting style). We reconciled these discrepancies in team meetings, and came to final codes only after consensus was reached between myself and the two coders.

Study 1 Data Analytic Plan

I categorized informants included in TFM and TSSM studies into one of the following informant context groups, consistent with key social contexts examined in the evidence-based assessment literature focused on youth (e.g., De Los Reyes et al., 2015; Hunsley & Mash, 2007): *Home Context* (i.e., mothers, fathers, caregivers), *School Context* (i.e., teachers), *Peer Context* (i.e., peers), *Mixed Contexts* (i.e., self-reporters, clinicians), or *Various Informants* (i.e., multiple informant types combined, official records). I conducted analyses to statistically compare coding results across TFM and TSSM studies. First, I conducted a chi-square analysis to statistically compare the likelihood that TFM and TSSM studies included informants who represent one context (e.g., home only) or two or more contexts (e.g., home and school contexts). Second, I conducted an independent samples *t*-test to compare the mean number of informants included in TFM and TSSM studies.

I interpreted statistical significance for all analyses using a *p*-value threshold of < .05. I inferred magnitudes of effect sizes based on Cohen's (1988) effect size conventions for the effect size *r* (low: .10; moderate: .30; large: .50) and *d* (low: .30; moderate: .50; large: .80), and based on Gravetter and Wallnau (2013) for the Cramer's *V* statistic with two degrees of freedom (low: .10; moderate: .30; large: .50).

Study 1 Results

Search Results

As reported in Figure 6, the coding team completed Methods Coding for 47 articles (TFM *n* = 8, TSSM *n* = 39). As reported in Tables 2 and 3, the measurement models were used to integrate informants' reports on a wide range of domains that included broadband measures of mental health (e.g., internalizing and externalizing problems), symptom measures (e.g., depression, ADHD), temperament (e.g., behavioral inhibition), and associated features of mental health concerns (e.g., parenting style). The majority of the applications of the measurement models integrated multi-informant reports of youth mental health and behavior (62.5% of TFM applications, 93% of TSSM applications). However, compared to TSSM studies, nearly twice as many TFM studies integrated multi-informant reports of adult mental health and behavior (37.5% of TFM studies, 62.5% of TSSM studies).[1]

Differences in Informants and Contexts across Measurement Models

I conducted a chi-square analysis to statistically compare the likelihood that TFM and TSSM studies included informants representing more than one context versus two or greater contexts. These analyses revealed a significant and moderate effect for context across applications of the two measurement models, $\chi^2(1) = 8.45$, $p < .01$, Cramer's $V = 0.41$. Specifically, TFM studies included informants representing two or greater contexts 50.0% of the time, whereas the other 50% of the time these studies included only mothers and fathers as the informants in the assessment (i.e., representing the home context only). In contrast, TSSM studies included informants representing two or greater contexts 90.7% of the time, with the most common application being consistent with Kraemer et al.'s (2003) original application which included caregiver, youth, and teacher reports (60.5% of unique TSSM applications). Using an independent samples *t*-test, I conducted additional analyses to compare the mean number of informants included in TFM and TSSM studies. Representing a large effect, these analyses revealed that TSSM studies ($M = 2.77$, $SD = .48$) included a greater number of informants on average when applying the measurement model, relative to the number of informants used in TFM studies ($M = 2.00$, $SD = 0.53$), $t(49) = 4.08$, $p < .001$, $d = 1.51$. Overall, as hypothesized, TSSM were more likely to include informants representing a greater number of contexts and total number of informants, relative to the TFM.

[1] Totals are greater than 100% for TSSM studies given that some included samples with both youth and adults.

Study 1 Discussion

Study 1 examines *how* researchers use the two measurement models, and in particular *who* is included when the models are applied to data. Advancing use and integration of multi-informant assessments rests not only on the validity of the measurement models themselves, but also the validity of the information entered into the models. In line with GIGO, it is important for researchers to base model input on the basic science on informants' reports and informant discrepancies. Specifically, researchers should base model input on which informants, both individually and collectively, provide psychometrically sound and clinically meaningful reports for the domain being measured. As a general principle, best practices in youth mental health assessment entail collecting reports from more than one informant, and selecting informants who provide unique and incrementally valuable information (Hunsley & Mash, 2007). In these respects, "best practices" typically involve (a) using two or more informants and (b) sampling behavior across more than one context (e.g., home and school). Are these best practices reflected in how users of the TFM and TSSM select and use informants? Study 1 directly addressed this larger question.

I found that the TFM and TSSM have been applied to a diverse range of informants, problem types, and research aims, which speaks to the broad utility of these measurement models in research leveraging multi-informant reports. Interestingly, nearly five times as many published studies implemented the TSSM compared to the TFM. This may be due to the fact that the TSSM was developed prior to the TFM, but also due to the relative simplicity of implementing PCA in the TSSM compared to the structural equation modeling approach necessitated by the

TFM. When examining how these measurement models have been used in published research, I found that applications of the two measurement models systematically varied in *who* was included in the measurement model. Consistent with hypotheses, I found that TSSM studies were more likely than TFM studies to include a greater number of informants representing a greater number of contexts. In general, applications of the two measurement models closely aligned with Kraemer and colleagues' (2003) and Bauer and colleagues' (2013) original applications of their respective measurement models. That is, a majority of TSSM studies included caregiver, teacher, and youth reports and the most common application of the TFM included mother and father reports. Thus, by design, TSSM studies were more likely to include informants representing more than one context (i.e., home, school, mix) and perspective (i.e., self, other) while TFM studies were more likely to include informants representing one context (i.e., home) and one perspective (i.e., other). These systematic differences are unsurprising when considering that Kraemer and colleagues provide an explicit theoretical model for selecting informants while Bauer and colleagues do not. In fact, it would be antithetical to the TFM for Bauer and colleagues to provide such guidance. That is, consistent with the assumptions underlying the MTMM matrix, if multi-informant models ought to prize common variance and treat unique variance as data that should be "purged" from further analysis, then it logically follows that the specific informants selected should be immaterial to the eventual scores. Thus, the unique goal of this integrative strategy is to "protect" scores derived from the TFM by including variables in the model that reflect factors hypothesized to “cause” biases in informants’ reports (e.g., caregivers’ depressive symptoms).

Perhaps most importantly, the strategy required when using the TSSM aligns closely with best practices in youth mental health assessment, which emphasizes selecting informants for the unique information they provide, so as to maximize the utility of information obtained in the assessment. In contrast, the original and most common application of the TFM uses a set of informants (i.e., two caregivers) who are rarely sought out as the *only* informants when assessing youth mental health (De Los Reyes et al., 2015; Hunsley & Mash, 2007). In this way, the TFM in both construction and use is aligned with the MTMM paradigm’s emphasis on common variance. Both measurement models aim to address long standing issues when integrating informants’ reports and in particular discrepant reports, which are more commonly observed among informants representing unique contexts and perspectives (i.e., *r*s between caregiver, teacher, and youth reports = .20-.32) than informants representing one context and perspective (i.e., *r*s between two caregivers or two teachers = .48-.64; Achenbach et al., 1987; De Los Reyes et al., 2015; Duhig et al., 2000). In this way, the TSSM by design and in use appears better suited to addressing issues when integrating multi-informant assessments that produce large informant discrepancies. In contrast, the TFM’s emphasis on common variance appears better suited to integrating reports from informants whose reports tend to correspond to a significant extent. At least based on use of the two measurement models in published research, these findings may speak to the “fit” between assumptions underlying each model and the data conditions that fit these assumptions.

Importantly, this systematic review only answers questions about the use of leading bias and context models for integrating informants' reports, but it does not speak to their validity. Unfortunately, research on the two measurement models largely ignores validity issues for the scores obtained from the TFM and TSSM (e.g., Common Factor and Trait scores). Further, the methodological design of these studies lacks key criteria for rigorous validation testing of derived scores (i.e., use of parallel measures across informants, ability to determine how integrated scores perform when predicting independent criterion variables and over-and-above alternative strategies). Thus, a key question regarding use of these two measurement models involves the degree to which scores derived from these models reflect valid characterizations of youth mental health. Study 2 addresses this question.

Chapter 4: Application and Comparison of the TFM and TSSM (Study 2)

Study 2 Method

Participants

As part of a larger study, investigators at the study site (i.e., Comprehensive Assessment and Intervention Program) recruited adolescents and caregivers from the Washington, DC, Maryland, and Northern Virginia areas using online advertisements (e.g., Craigslist) and flyers posted in local businesses (e.g., libraries, pediatricians' office). Participants responded to one of two posted advertisements: (1) study providing a no-cost clinical social anxiety evaluation for adolescents (i.e., *clinic-referred adolescents*) and (2) nonclinical study on family relationships (i.e., *community control adolescents*). Following the study, we provided feedback to families in the clinic-referred sample on their adolescent's mental health concerns and referrals for mental health services. We held all study procedures consistent across participants, regardless of referral status.

To be eligible for the study, we required dyads to: (a) be fluent in English, (b) understand the consenting and interview process, (c) have a 14-15-year-old adolescent currently living in the home with the caregiver completing the study, and (d) have an adolescent who the caregiver did not report as having a history of learning or developmental disabilities. The total sample included 134 caregiver-adolescent dyads (45 clinic-referred, 89 community control). Adolescents were 14 or 15 years old (M_{age} = 14.49, *SD* = 0.50); 89 adolescents identified as female and 45 identified as male. Based on parent report, adolescents' race/ethnicity included African American/Black (53.0%), Caucasian/European American/White (33.6%), Hispanic/Spanish/Latino/a (10.4%), Asian American/Asian (5.2%), American Indian (0.7%) or Other (e.g., Caribbean; 7.5%). Race/ethnicity values total above 100% given that caregivers could select multiple racial/ethnic categories. Caregivers included the adolescent's biological mother or father (95.5%), adoptive mother or father (2.2%), or another caregiver (2.2%). Caregivers reported household income using a scale with 10 categories that varied by $100 increments (i.e., less than $100 per week through $901 or more per week). Based on this scale, 26.1% of families had a weekly household income of $500 or less, 22.4% had a weekly household income between $501 and $900, and 51.5% had a weekly household income of $901 or more per week.

In all study analyses, I pooled the clinic-referred and community control samples. By design, we recruited the community control sample in larger numbers than the clinic-referred sample. We selected this approach given that it mimics displays of dimensionally varying social anxiety in the general population and is consistent with dimensional models of psychopathology (Casey, Oliveri, & Insel, 2014). We also used this recruitment strategy given that dimensional approaches offer greater reliability and validity relative to categorical approaches (Markon et al., 2011). I conducted analyses to determine whether the two referral groups differed on key demographic characteristics including adolescent age, adolescent gender, adolescent racial/ethnic background, family income, caregiver relation to the adolescent, and caregiver marital status. Given the exploratory nature of these tests, I applied a Bonferroni correction (i.e., 11 tests and thus a corrected *p*-value cutoff of .0045). These analyses revealed no significant demographic differences between the clinic-referred and community control groups, thus justifying the pooled sample approach.

Measures

Caregivers completed a demographic questionnaire. Caregivers, adolescents, and peer confederates completed measures to assess the adolescent's social anxiety, and caregivers completed a measure to assess their own depressive symptoms. For measures completed by more than one informant, parallel survey items were used, with only minor modifications made to fit each informant's perspective (i.e., "My child" for caregiver measures, "I" for adolescent measures, "The participant" for peer confederates). Caregivers and adolescents completed measures prior to the adolescent's participation in social interaction tasks. Peer confederates completed survey measures following their interactions with adolescents in social interaction tasks.

Social Phobia Scale (SPS; Mattick & Clarke, 1998). Caregivers, adolescents, and peer confederates completed the 20-item SPS. To reduce the number of indicators when using the SPS within the TFMs, I used a 6-item short form (i.e., SPS-6; Peters, Sunderland, Rapee & Mattick, 2012). The SPS assesses adolescents' fears of being scrutinized by others during routine activities (example item: "I/My child/The participant is/am worried about shaking or trembling when watched by other people.”). Informants rated items on a Likert scale ranging from 0 (*Not at all characteristic or true of me/my child/the participant*) to 4 (*Extremely characteristic or true of me/my child/the participant*). The 20-item SPS exhibits strong validity, differentiates individuals on diagnostic status, and is sensitive to treatment response (Mattick & Clarke, 1998; Deros et al., 2018). Similarly, the SPS-6 demonstrates adequate internal consistency, convergent validity, differentiates individuals with and without SAD, and is sensitive to treatment response (Carleton et al., 2014; Fergus, Valentiner, Kim, & McGrath, 2014; Le Blanc et al., 2014; Peters et al., 2012). Adolescent (SPS α = .92; SPS-6 α = .87), caregiver (SPS α = .94; SPS-6 α = .89), and peer confederates’ reports (SPS α = .96; SPS-6 α = .90) on the SPS and SPS-6 exhibited good-to-excellent internal consistency and were highly correlated (*r*s = .95-.96, *p*s < .001). I used informants’ SPS-6 reports in the TFMs. In addition, I used informants’ SPS reports in incremental validity analyses comparing the performance of the TSSM Trait score relative to individual informants’ reports and the composite score when predicting independent criterion variables.

Social Interaction Anxiety Scale (SIAS; Mattick & Clarke, 1998). Informants completed the 20-item SIAS, which is a measure for assessing social anxiety concerns that adolescents may experience during social interactions (example item: “I/my child/ the participant has difficulty talking with other people.”). Informants rated items on a Likert scale ranging from 0 (*Not at all characteristic or true of me/my child/the participant*) to 4 (*Extremely characteristic or true of me/my child/the participant*). Informants’ reports on the SIAS display convergent validity and distinguish adolescents on referral status (Deros et al., 2018; Glenn et al., 2019).

Adolescent (α = .94), caregiver (α = .95), and peer confederate reports (α = .96) exhibited excellent internal consistency. I used informants' SIAS reports in the TSSM. In addition, I used informants' SIAS reports in incremental validity analyses comparing the performance of the TFM Common Factor score relative to individual informants' reports and the composite score when predicting independent criterion variables.

Social Phobia and Anxiety Inventory for Children (SPAIC; Beidel, Turner, & Morris, 1995). Caregivers and adolescents completed the 26-item SPAIC, which is one of the most widely used youth social anxiety measures. The SPAIC requires informants to endorse how often the adolescent feels nervous or scared when in various social scenarios (e.g., meeting new peers). Several items include "sub-items" that require informants to rate the adolescent's social anxiety with different interaction partners (e.g., "boys or girls his/her age that he/she knows" vs. "boys or girls his/her age that he/she doesn't know"). These sub-items are averaged at the item level to create a composite score. Caregivers and adolescents rated items using a Likert scale ranging from 0 (*Never*) to 2 (*Always*). Caregiver and adolescent reports on the SPAIC display strong construct and convergent validity, relate to independent observers' ratings of anxiety and social skills, differentiate adolescents on referral status, and are sensitive to treatment response (Beidel et al., 1995; Beidel et al., 2000a; Beidel, Turner & Morris, 2000b). Adolescent (α = .95) and caregiver reports (α = .96) exhibited excellent internal consistency. I used adolescents' and caregivers' SPAIC reports in incremental validity analyses comparing the performance of the

TFM Common Factor and TSSM Trait scores to individual informants' reports when predicting independent criterion variables.

Beck Depression Inventory-II (BDI-II; Beck, Steer, & Brown, 1996). Caregivers completed a modified version of the BDI-II, a 21-item measure commonly used to screen for depression. Caregivers rated their own depressive symptoms using a 3-point scale on which higher scores indicated higher levels of depressive symptoms. Based on the study's protocol, we did not administer items 9 and 21 of the BDI-II (see Rausch et al., 2017). For this reason, we prorated items 9 and 21 by calculating a mean item score for each participant and including a prorated item score for items 9 and 21. The BDI-II demonstrates strong psychometric properties across studies, including strong reliability, strong criterion-related validity, and good sensitivity and specificity when predicting depression diagnosis (Wang & Gorenstein, 2013). Caregiver reports on the BDI-II exhibited excellent internal consistency (α = .92). I entered caregiver BDI-II reports as an informant-level predictor in TFMs that included caregiver reports.

Behavioral Tasks

Task descriptions. Following completion of self-report measures, adolescents participated in a series of counterbalanced social interaction tasks with peer confederates. These tasks were developed in prior work with children, adolescents, and adults (e.g., Anderson & Hope, 2009; Beidel et al., 2000a; Beidel, Rao, Scharfstein, Wong, & Alfano, 2010) and took approximately 20 minutes to complete. Tasks included the Unstructured Conversation Task (UCT; adapted from Beidel et al., 2010), Simulated Social Interaction Test (SSIT; adapted from Curran, 1982; Beidel et al., 2000a), and Impromptu Speech Task (IST; adapted from Beidel et al., 2010). In each task, adolescents interacted with research assistants trained to pose as adolescents (i.e., peer confederates). We masked peer confederates to adolescents' referral status and clinical information, and ensured they had no prior interaction with adolescents or their caregivers prior to beginning the task. In all one-on-one social interactions across tasks, we paired adolescents with a gender-matched peer confederate to reduce any potentially confounding factors related to anxiety when interacting with individuals of the opposite sex.

In the UCT, adolescents participated in an unstructured three-minute role-play with a peer confederate. We provided the following instructions to adolescents: "*Pretend that you are at a new school and don't know anyone.*" We trained peer confederates to respond neutrally and allow the adolescent to drive the conversation. In the SSIT, adolescents participated in a series of five one-to-three-minute role-playing scenes (e.g., offering/accepting assistance, responding to inappropriate behavior) with a peer confederate. In each role play, adolescents were provided two opportunities to speak and peer confederates were trained to provide two scripted responses. In the IST, adolescents participated in a speech task in which they delivered a speech to a small audience about topics infrequently discussed by adolescents (e.g., politics, public health). The audience included a task administrator and two trained peer confederates with whom the adolescent had no prior contact. We provided adolescents three minutes to prepare their speech and ten minutes to complete their speech. Following a minimum period of three-minutes into the speech, we permitted adolescents to terminate the speech if they wished to do so. We videorecorded social interaction tasks from two angles to allow for coding of adolescent behavior.

Behavioral ratings. To obtain independent observers' ratings of adolescent social anxiety in social interaction tasks, we used a validated behavioral coding scheme developed by Beidel and colleagues (Beidel et al., 2000a, 2010; Scharfstein, Beidel, Sims, & Finnell, 2011). We ensured that independent observers had no prior interactions with adolescents in social interaction tasks and were masked to adolescents' referral status and clinical information. Two independent observers provided macro-level ratings of adolescent social anxiety on a 5-point scale ranging from 1 (*Animated*) to 5 (*Severe anxiety*). Independent observers made a total of seven ratings (i.e., one UCT rating, five SSIT ratings, one IST rating) and these ratings were used to compute a social anxiety composite. Inter-rater reliability for the two independent observers' ratings of social anxiety were in the good range (average *ICC* = .76).

Study 2 Data Analytic Plan

Preliminary Analyses

I conducted analyses to determine whether the data used in Study 2 met basic assumptions of parametric statistical tests (i.e., skewness/kurtosis in range of ±2.0). In addition, I conducted bivariate correlation analyses to examine correlations within and between informants' reports of adolescent social anxiety.

Aim 1. Implementing the TFM

I implemented four TFMs using Mplus Version 8.5, including models for caregiver-adolescent reports, caregiver-peer reports, adolescent-peer reports, and caregiver-adolescent-peer reports on the SPS-6 (see Figures 4 and 5). For each TFM, I strictly followed the model building procedures described by Bauer et al. (2013) and used full-information maximum likelihood estimation to calculate likelihood ratio tests and score estimates. I identified each factor through imposing a set of constraints on the factor loadings and factor correlations. First, all informants' ratings were loaded onto the Common Factor. The Common Factor reflects informants' shared variability in item ratings of adolescent social anxiety. Second, I loaded each informant's reports onto the appropriate Perspective Factor (i.e., one per informant). The informants included in this study consistently provide reports that correlate at low-to-moderate magnitudes (Achenbach et al., 1987; De Los Reyes et al., 2015). Further, prior work supports the notion that these informants vary as to their unique contexts and perspectives when rating youth anxiety (Cannon et al., 2020; Deros et al., 2018; Etkin, Leboqitz, & Silverman, 2021a; Etkin, Shimshoni, Lebowitz, & Silverman, 2021b). Thus, I modeled these informants as structurally different. Consistent with Bauer and colleagues' (2013) recommendations for structurally different informants, I equated the intercept and factor loading across informants for the first SPS-6 item (i.e., Item 4) and freely estimated the intercepts and factor loadings for the remaining items. Third, I regressed Specific Factors onto parallel SPS-6 items completed by informants.

I modeled all latent factors (i.e., Common, Perspective, and Specific Factors) as orthogonal to each other. Models that included caregiver reports of adolescent social anxiety represented Conditional TFMs given that they included an informant-level predictor (caregiver self-reports of depressive symptoms) consistent with the depression-distortion hypothesis (Richters, 1992) and Bauer and colleagues' (2013) original application of the TFM. Specifically, I regressed caregiver BDI-II reports onto the Common Factor and caregiver Perspective Factor.

Consistent with Bauer et al.'s (2013) recommendations, I placed additional constraints on the models to set the scale of the latent variables. For the Unconditional TFM, I set the means and variance of the Common, Perspective, and Specific Factors, to 0 and 1, respectively. Doing so allowed for all nonzero factor loading to be estimated and interpreted in terms of their relative magnitude. For the Conditional TFMs, I set the intercepts and residual variances of the Common Factor, Perspective Factors, and Specific Factors, to 0 and 1, respectively.

For the final converged models, I evaluated model fit indices. These indices included the Steiger-Lind Root-Mean-Square Error of Approximation (RMSEA; Steiger, 1990), Bentler Comparative Fit Index (CFI; Bentler, 1990), and Tucker-Lewis index (TLI; Tucker & Lewis, 1973). I considered model fit acceptable if the following model fit criteria were met: RMSEA $\leq .06$, CFI $\geq .95$, and TLI $\geq .95$ (Hu & Bentler, 1999; Kline, 2016; McDonald & Ho, 2002). For converged TFMs with satisfactory model fit, I calculated integrated multi-informant factor scores for use in subsequent analyses. Consistent with Curran et al. (2020), factor scores for the Common, Perspective, and Specific factors were estimated as maximum likelihood expected a posteriori (EAP) scores. This approach up-weights items that are more strongly related to the factor and down-weights those that are less strongly related to the factor.

Aim 2. Implementing the TSSM

I implemented the TSSM in SPSS Version 24 by applying PCA to caregiver, adolescent, and peer confederate SIAS reports. Consistent with Kraemer et al. (2003), I conducted an unrotated PCA with the number of components extracted set to three. I evaluated component weights to determine whether the first component was consistent with a Trait score (i.e., component weights loading strongly and in the same direction). Factor analytic criteria were used to determine whether I would retain the three-component PCA model: Bartlett's Test of Sphericity is significant (Bartlett, 1950), Kaiser-Meyer-Olkin test (KMO) ≥ .60 (Kaiser, 1970), and the first component (i.e., Trait score) eigenvalue > 1 (Pett et al., 2003).

Aim 3. Incremental Validity Relative to Individual Informants' Reports and the Composite Score

I tested the incremental validity of the TFM Common Factor using a series of hierarchical linear and logistic regressions. Consistent with well-established incremental validity conventions (Garb, 2003; Hunsley & Meyer, 2003; Smith et al., 2003), I benchmarked the ability of Common Factor scores to explain variance in criterion variables, relative to well-established strategies for analyzing multi-informant data (i.e., individually, simple average). In separate regressions, I evaluated whether the Common Factor scores explained variance in adolescent referral status and observed adolescent social anxiety, over-and-above the explanatory value of individual informants’ reports or the composite score (i.e., models involved either individual informants’ reports or the composite score, but not both simultaneously). In separate regression models, I added parent (i.e., SPAIC), adolescent (i.e., SPAIC), and peer confederate (i.e., SIAS) reports of adolescent social anxiety in the first step as independent variables. Using a similar design, I ran separate regression models in which the composite score (using averages from informants’ SIAS reports) was added in the first step as an independent variable. I only ran these regression models for scores from TFMs that included the informants’ reports both individually and when averaged. For instance, I included caregiver SPAIC reports in regression models evaluating the incremental validity of the caregiver-adolescent Common Factor score, but not in the regression model evaluating the incremental validity of the adolescent-peer Common Factor score. I included observed adolescent social anxiety and adolescent referral status (0 = community control, 1 = clinic-referred) as dependent variables in these models. I ran separate models for each Common Factor score, alternative integrative strategy, and dependent variable.

Similarly, I tested the incremental validity of the TSSM Trait score relative to individual informants’ reports and the composite score using a series of hierarchical linear and logistic regressions. These regressions evaluated whether the Trait score explained variance in adolescent referral status and observed adolescent social anxiety, over-and-above the explanatory value of individual informants’ reports and the composite score. In separate regression models, I added parent (i.e., SPAIC), adolescent (i.e., SPAIC), and peer confederate (i.e., SPS) reports of adolescent social anxiety in the first step as independent variables. Using a similar design, I ran separate regression models for each composite score (using averages from informants' SPS reports) in the first step as independent variables. I included observed adolescent social anxiety and adolescent referral status (0 = community control, 1 = clinic-referred) as dependent variables in these models. I ran separate models for each alternative integrative strategy and dependent variable. Of note, I previously conducted these analyses using a smaller version of the same sample (i.e., Makol et al., 2020). For this dissertation, I re-ran these analyses to allow for a direct comparison of the effects obtained in criterion-related and incremental validity analyses examining the TFM Common Factor and TSSM Trait scores.

Aim 4. Directly Comparing the Incremental Validity of Scores Derived from the TFM and TSSM

I evaluated the incremental validity of the TFM Common Factor score using caregiver, adolescent, and peer confederate reports relative to the TSSM Trait score using two hierarchical linear and logistic regressions. These regressions evaluated whether the Common Factor score explained variance in adolescent referral status and observed adolescent social anxiety, over-and-above the explanatory value of the Trait score. I entered the Trait score in the first step as an independent variable and the Common Factor score in the second step as an independent variable. I then conducted an additional set of regression models in which I added the Common Factor score in the first step as an independent variable, and the Trait score in the second step as an independent variable. I only compared these two scores, because it allowed for a comparison of scores derived from the TFM and TSSM when both models leveraged the same three informants. Dependent variables included observed adolescent social anxiety and adolescent referral status (0 = community control, 1 = clinic-referred). I ran a separate model for each order of entering the independent variables as well as for each dependent variable.

I interpreted statistical significance for all analyses using a *p*-value threshold of < .05. I inferred magnitudes of effect sizes based on Cohen's (1988) effect size conventions for the effect size *r* and β (low: .10; moderate: .30; large: .50). Unlike the exploratory demographic comparison analyses I reported previously, I did not apply Bonferroni corrections to the analyses reported below. This decision is consistent with recommendations on judicious use of Bonferroni corrections (e.g., Armstrong, 2014; Streiner & Norman, 2011).

Study 2 Results

Preliminary Analyses

I reported descriptive statistics for all Study 2 measures in Table 4. All measures displayed acceptable levels of skewness and kurtosis, with the exception of caregiver self-reports of depression on the BDI-II. To reduce excessive kurtosis, I applied a square root transformation to caregiver BDI-II reports, which reduced kurtosis to tolerable levels. The transformed BDI-II reports were used in the analyses described below.

In Table 5, I reported bivariate correlations among informants' reports on social anxiety measures used in the measurement models. Overall, correlations between informants' reports on the same social anxiety measure were in the moderate-magnitude range. However, consistent with prior work (Deros et al., 2018), the correlation between parent and peer confederate reports on the SPS-6 was non-significant, $p = .17$.

Aim 1. Implementing the TFM

I implemented the TFMs using model-building procedures described by Bauer et al. (2013) and as detailed previously. Across TFMs, I entered SPS-6 reports into the model for each informant. For Conditional TFMs, I regressed the informant-specific predictor (caregiver self-reported depressive symptoms on the BDI-II) onto the Common Factor and caregiver Perspective Factor. When describing model findings, I interpreted the standardized factor loadings given that they are directly comparable when understanding the relative effects of the Common, Perspective, and Specific Factors on informants' SPS-6 items.

Adolescent-Peer TFM. I implemented the Unconditional TFM using adolescent and peer confederate SPS-6 reports as shown in Figure 4 (Panel C). The initial model converged and fit criteria indicated mixed but overall acceptable fit of the model to the data, $\chi^2(38) = 60.79$, $p < .05$; RMSEA = 0.07, 90% CI [0.03, 0.10]; CFI = 0.97; TLI = 0.95. I retained this model and reported raw and standardized intercept and factor loading estimates in Table 6. Both adolescent and peer confederate reports loaded significantly and positively onto the Common Factor, with adolescent self-reports exhibiting stronger loadings (.63-.84 vs. .19-.39). Peer confederate reports loaded positively and strongly onto the peer Perspective Factor (.56-.81). In contrast, adolescent self-reports did not load significantly onto the adolescent Perspective Factor with the exception of Item 4 ("Nervous people are staring"). I observed significant and negative loadings onto Specific Factors for Item

7 (“Worries about shaking or trembling”) and Item 8 (“Gets tense when facing others”) and a significant and positive loading was observed for the Specific Factor for Item 15 (“Worries about attracting attention”).

Caregiver-Adolescent TFM. I implemented the Conditional TFM using caregiver and adolescent SPS-6 reports as shown in Figure 4 (Panel A). The initial model converged and fit criteria suggested good model fit to the data, $\chi^2(46) = 42.72$, $p = .61$; RMSEA = 0.00, 90% CI [0.00, 0.05]; CFI = 1.00; TLI = 1.00. I retained this model and reported raw and standardized intercept and factor loading estimates in Table 7. Adolescent self-reports loaded positively and significantly onto the Common Factor (.57-.93). Caregiver reports loaded positively but less strongly onto the Common Factor (.18-.34), with the exception of Item 7 (“Worries about shaking or trembling”, $p = .051$). Whereas caregiver reports loaded significantly and positively onto the caregiver Perspective Factor (.55-.83), adolescent self-reports did not load significantly onto the adolescent Perspective Factor (-.14-.49). I observed significant and negative Specific Factor loadings for Item 7 (“Worries about shaking or trembling”), Item 15 (“Worries about attracting attention”) and Item 17 (“Feels conspicuous standing in line”). The informant-specific predictor (caregiver self-reported depressive symptoms) was significantly related to the caregiver Perspective Factor (β = .28, standard error [SE] = .09, $p < .01$) but did not relate to the Common Factor (β = -.03, SE = .09, $p = .75$). This suggests that as caregivers’ self-reported depressive symptoms increase, they are more likely to rate their adolescent’s social anxiety higher, which is captured by the predictor’s specific impact on the unique variance in the caregivers’ social anxiety reports that is explained by their Perspective

Factor. One might interpret this effect as reflecting “rater bias” consistent with the depression-distortion hypothesis. The validation tests reported below essentially probe this interpretation. That is, if this model partials out variance reflected by rater bias, then presumably the model should perform quite well not just when compared to the individual informant’s report (i.e., caregiver), but also to the TSSM, which does not treat this unique variance as rater bias.

Caregiver-Peer TFM. I Implemented the Conditional TFM using caregiver and peer confederate SPS-6 reports as shown in Figure 4 (Panel B). The initial model did not converge due to a non-positive definite covariance matrix. I modified model constraints for problematic parameters (including setting the residual variance of the peer confederates’ Item 4 and Item 7 reports and caregivers’ Item 7 reports to a positive near zero value), which allowed the model to converge but resulted in poor model fit, $\chi^2(49) = 122.66$, $p < .001$; RMSEA = 0.11, 90% CI [0.08, 0.13]; CFI = 0.92; TLI = 0.86. I subsequently modified various model constraints to improve model fit, which was unsuccessful. Given these model fit issues, I did not retain the caregiver-peer TFM or use it in subsequent analyses examining the validity of the Common Factor scores.

Caregiver-Adolescent-Peer TFM. I implemented the Conditional TFM using caregiver, adolescent, and peer confederate SPS-6 reports as shown in Figure 5. The initial model did not converge due to a non-positive definite covariance matrix. Given that the variance in peer confederates’ Item 4 reports was almost completely explained by the latent factors in the model, I set the residual variance for this item to a near zero positive value. This resulted in a positive-definite and converged model with overall good model fit with the exception of the chi-square test of model fit, $\chi^2(129) = 169.39$, $p < .01$; RMSEA = 0.05, 90% CI [0.03, 0.07]; CFI = 0.97; TLI = 0.96. I retained this model and reported raw and standardized intercept and factor loading estimates in Table 8. Adolescent and peer confederate reports loaded positively and significantly onto the Common Factor (.29-.77) with the exception of adolescent self-reports on Item 17 ("Feels conspicuous standing in line"). In contrast, caregiver reports did not load significantly onto the Common Factor (-.02-.08), although they loaded positively and significantly onto the caregiver Perspective Factor (.58-.87). Thus, a notable feature of this application of the TFM is that the Common Factor primarily weights adolescent and peer confederate reports, with a non-significant contribution from caregiver reports, which are instead captured primarily in the caregiver Perspective Factor. Both adolescent and peer confederate reports loaded significantly and positively onto their respective Perspective Factors (.33-.93) with the exception of peer confederates' reports on Item 16 ("Tense in elevator"). A total of five SPS-6 items had significant Specific Factor loadings, including positive loadings for Item 4 ("Nervous people are staring"), Item 15 ("Worries about attracting attention"), and Item 17 ("Feels conspicuous standing in line") and negative loadings for Item 8 ("Gets tense facing others") and Item 16 ("Tense in elevator"). Consistent with the Caregiver-Adolescent model findings, caregiver self-reported depressive symptoms were significantly related to the caregiver Perspective Factor ($\beta = .26$, SE = .09, $p < .01$) but did not relate to the Common Factor ($\beta = -.05$, SE = .10, $p = .61$). This suggests that as caregivers' self-reported depressive symptoms increase, they are more likely to rate their adolescent's social anxiety higher, which is captured by the predictor's specific impact on the unique variance in the caregivers' social anxiety reports that is explained by their Perspective Factor. As with the Caregiver-Adolescent TFM, the interpretation that this relation between caregivers' self-reported depressive symptoms and their adolescent social anxiety reports reflects a rater bias effect is evaluated using the validation tests reported below.

Aim 2. Implementing the TSSM

Consistent with procedures described by Kraemer et al. (2003), I implemented the TSSM using caregiver, adolescent, peer confederate SIAS reports (see Table 9). I entered these reports into an unrotated PCA in which I set the number of components extracted to three. I retained this three factor solution given that all factor analytic criteria were met including a statistically significant Bartlett's Test of Sphericity ($\chi^2(3) = 50.41$, $p < .001$), KMO just above the cutoff of .60 (KMO = .64), and eigenvalue >1 for the Trait score component. Similar to prior work (Kraemer et al., 2003; Makol et al., 2020), I qualitatively examined the component loadings and found that they were consistent with Trait (positive and large in magnitude across informants), Context (contrasting loadings for caregiver and peer confederate reports with adolescent reports loading between caregiver and peer confederate reports), and Perspective (contrasting loadings for adolescent reports and caregiver and peer confederate reports) scores. Kraemer and colleagues' assumption that this strategy for integrating informants' reports maximizes the variance explained in the reports is further probed through validation testing below.

Aim 3. Incremental Validity Relative to Individual Informants' Reports and the Composite Score

Incremental Validity of Scores Derived from the TFM Relative to Individual Informants' Reports (Tables 10 and 11). I conducted hierarchical regression analyses to determine the incremental validity of the Common Factor scores in predicting observed adolescent social anxiety, over-and-above individual informants' reports. In the first step of the regressions, individual informants' reports predicted observed adolescent social anxiety and at small-to-large magnitudes. In the second step of the regression, the Common Factor scores varied in the degree to which they displayed incremental validity, depending on the informant's report entered in the first step.

Specifically, the Common Factor scores consistently explained incremental variance in observed adolescent social anxiety, over-and-above caregiver reports. In contrast, the Common Factor scores consistently failed to explain incremental variance, over-and-above peer confederate reports. Findings examining the incremental validity of the Common Factor score over-and-above adolescent reports were mixed. That is, whereas the three-informant Common Factor score explained incremental variance, the caregiver-adolescent Common Factor score did not. Overall, these findings indicate mixed support for the incremental validity of the Common Factor when predicting observed adolescent social anxiety.

I conducted hierarchical logistic regression analyses to determine the incremental validity of the Common Factor score in predicting adolescent referral status, over-and-above individual informants' reports. In the first step of the regressions, individual informants' reports predicted adolescent referral status. As with findings for observed social anxiety, in the second step of the regression, the Common Factor scores varied in the degree to which they displayed incremental validity when predicting adolescent referral status, depending on the informant's report entered in the first step.

Specifically, the caregiver-adolescent Common Factor score explained incremental variance over-and-above caregiver reports and the adolescent-peer TFM explained incremental variance over-and-above peer confederate reports. In contrast, the adolescent-peer and caregiver-adolescent Common Factor scores failed to explain incremental variance in adolescent referral status, over-and-above adolescent reports. The three-informant Common Factor score consistently failed to explain incremental variance in adolescent referral status. Overall, when significant effects were observed for the incremental validity of the Common Factor score when predicting adolescent referral status, adolescents were approximately twice as likely to be in the clinic-referred group, relative to the community control group, for every one-unit increase in the Common Factor score. However, findings in this sample indicate that in most combinations of informants used in the TFM, the Common Factor buys little in terms of incremental validity when predicting whether adolescents were in a community control or clinic-referred group. Stated otherwise, more robust effects may be found when using individual informants' reports, relative to a score integrated with the TFM strategy, when aiming to understand an adolescent's referral status.

Incremental Validity of Scores Derived from the TFM Relative to the Composite Score (Tables 12 and 13). I conducted a series of hierarchical linear regression models to evaluate whether the Common Factor scores explained variance in observed adolescent social anxiety, over-and-above the explanatory value of the composite score (i.e., average of the relevant informants' SIAS reports). In the first step of each regression model, the SIAS composite score explained significant variance in observed adolescent social anxiety, representing large effects. In the second step each regression model, the Common Factor score did not explain incremental variance in observed adolescent social anxiety, over-and-above the composite score.

I conducted a series of hierarchical logistic regression models to determine the incremental validity of the Common Factor score in predicting adolescent referral status, over-and-above the composite score (i.e., average of the relevant informants' SIAS reports). In the first step of each regression model, the composite score explained significant variance in adolescent referral status. In the second step of each regression model, the Common Factor score did not explain incremental variance in adolescent referral status, over-and-above the composite score. Overall, the Common Factor score consistently failed to explain incremental variance in independent criterion variables, over-and-above the composite score.

Incremental Validity of the TSSM Trait Score Relative to Individual Informants' Reports (Tables 14 and 15). I conducted hierarchical regression analyses to determine the incremental validity of the TSSM Trait score in predicting observed adolescent social anxiety, over-and-above individual informants' reports of social anxiety. In the first step of the regressions, individual informants' reports predicted observed adolescent social anxiety and at moderate-to-large magnitudes. In the second step of the regressions, the Trait score consistently predicted incremental variance in observed adolescent social anxiety, over-and-above the variance explained by individual informants' reports, representing moderate-to-large effects.

I conducted hierarchical logistic regression analyses to determine the incremental validity of the Trait score in predicting adolescent referral status, over-and-above individual informants' reports of social anxiety. In the first step of the regressions, caregiver and adolescent reports predicted adolescent referral status, whereas peer confederate reports did not. As with findings for observed social anxiety, in the second step of the regressions the Trait score predicted incremental variance in adolescent referral status, over-and-above the variance explained by individual informants' reports. Specifically, adolescents were 3 to 7 times more likely to be in the clinic-referred sample, relative to the community control group, for every one-unit increase in the Trait score. Overall, consistent with study hypotheses and my prior work using a smaller version of this sample (Makol et al., 2020), findings support the incremental validity of the Trait score, over-and-above individual informants' reports.

Incremental Validity of the TSSM Trait Score Relative to the Composite Score (Tables 14 and 15). I conducted a hierarchical linear regression model to evaluate whether the TSSM Trait score explained variance in observed adolescent social anxiety, over-and-above the explanatory value of the composite score (i.e., average of caregiver, adolescent, and peer confederate SPS reports). In the first step of the regression, the composite score predicted observed adolescent social anxiety, representing a large effect. In the second step of the regression, the Trait score explained incremental variance in observed adolescent social anxiety, also representing a large effect.

I conducted a hierarchical logistic regression model to determine the incremental validity of the Trait score in predicting adolescent referral status, over-and-above the composite score. In the first step of the regression, the composite score predicted adolescent referral status. In the second step of the regression, the Trait score explained incremental variance in adolescent referral status. Adolescents were over five times more likely to be in the clinic-referred group, relative to the community control group, for every one-unit increase in the Trait score. Overall, consistent with study hypotheses and my prior work using a smaller version of this sample (Makol et al., 2020), findings support the incremental validity of the Trait score over-and-above the composite score.

Aim 4. Directly Comparing the Incremental Validity of Scores Derived from the TFM and TSSM

Preliminary Analyses. To compare the integrative scores obtained from the two measurement models, I conducted bivariate correlation analyses between the Common Factor and Trait scores using all three informants' reports. These analyses revealed a large-magnitude correlation between the Common Factor and Trait scores, $r = .50$, $p < .001$. Thus, these two integrative scores, which were derived from the same informants' reports on two distinct social anxiety measures, contained a fair degree of common variance, but also a fair degree of unique variance.

Incremental Validity of Scores Derived from the TSSM and TFM, Relative to Each Other. I conducted hierarchical linear regression models to evaluate whether the TFM Common Factor score explained variance in observed adolescent social anxiety, over-and-above the explanatory value of the TSSM Trait score, and vice versa. I first evaluated the incremental validity of the Common Factor score. In the first step of the regression model, the Trait score explained significant variance in observed adolescent social anxiety, β = .59, ΔR^2 = .35, Δ*F*(1, 128) = 68.11, *p* < .001. In the second step of the regression model, the Common Factor score did not explain incremental variance in observed adolescent social anxiety, over-and-above the Trait score, β = .08, ΔR^2 = .01, Δ*F*(1, 127) = 1.04, *p* = .31. Next, I evaluated the incremental validity of the Trait score. In the first step of the regression model, the Common Factor score explained significant variance in observed adolescent social anxiety, β = .36, ΔR^2 = .13, Δ*F*(1, 128) = 18.75, *p* < .001. In the second step of the regression model, the Trait score explained incremental variance in observed adolescent social anxiety, over-and-above the Common Factor score, β = .55, ΔR^2 = .23, Δ*F*(1, 127) = 44.12, *p* < .001. Further, the Common Factor score no longer explained significant variance in observed adolescent social anxiety in the second step of the regression, β = .08, *p* = .31. Thus, when predicting observed adolescent social anxiety, the Trait score provided incremental validity relative to the Common Factor score, whereas the reverse was not true.

I conducted hierarchical logistic regression models to evaluate whether the Common Factor score explained variance in adolescent referral status, over-and-above the explanatory value of the Trait score, and vice versa. I first evaluated the incremental validity of the Common Factor score. In the first step of the regression model, the Trait score explained significant variance in adolescent referral status, *b* =

1.28, *OR* = 3.59, *p* < .001. In the second step of the regression model, the Common Factor score did not explain incremental variance in adolescent referral status, over-and-above the Trait score, *b* = -.41, *OR* = .67, *p* = .13. Next, I evaluated the incremental validity of the Trait score. In the first step of the regression model, the Common Factor score did not explain significant variance in adolescent referral status, *b* = 0.35, *OR* = 1.42, *p* = .09. In the second step of the regression model, the Trait score explained incremental variance in adolescent referral status, over-and-above the Common Factor score, *b* = 1.51, *OR* = 4.54, *p* < .001. Overall, these findings indicate that although the Common Factor and Trait scores share a fair degree of overlap, they capture unique information when integrating informants' reports. When entered into the same regression model, the Trait score offers additional predictive power when characterizing adolescent behavior in social interactions and distinguishing adolescents on their referral status.

What About the Other Scores?

When using the TFM and TSSM, questions remain regarding what information is captured (e.g., bias, context, measurement error) within scores beyond the Common Factor and Trait scores. Although fully unpacking this question is outside the scope of this dissertation and what is possible with the criterion variables used in this study, I ran exploratory analyses to examine the relation of other TFM scores (i.e., Perspective Factor scores) and TSSM scores (i.e., Context and Perspective scores) to the independent criterion variables examined in this study. I limited these exploratory analyses to the three-informant measurement models. Using bivariate correlation analyses, I found that all of the TFM Perspective Factor scores were associated with observed adolescent social anxiety (*r*s = .20s, *p*s < .01). In contrast, the TSSM Context and Perspective scores did not relate to observed adolescent social anxiety (*p*s > .14). Using logistic regression analyses, I found that the caregiver and adolescent Perspective Factor scores predicted adolescent referral status (ORs = 2.00, *p*s < .01), whereas the peer Perspective Factor did not (*p* = .33). Further, the TSSM Context score predicted adolescent referral status (OR = 1.81, *p* < .01) whereas the Perspective score did not (*p* = .73).

Study 2 Discussion

Implementing the Measurement Models

In Study 2, I implemented the TFM and TSSM using caregiver, adolescent, and peer confederate reports of adolescent social anxiety. A first step in understanding how these models perform when subjected to ecologically valid assessment conditions involves examining model fit and the loadings of informants' reports onto factors and components. As stated previously, both the TFM and TSSM are guilty of the "naming fallacy" in which factors and components are interpreted as reflecting what they are named without evidence to support the interpretation (Kline, 2016). Further, both purport to optimize reports without evidence to support these claims. Nonetheless, in this section, I describe and interpret scores derived from the TFM and TSSM consistent with Bauer and colleagues' (2013) and Kraemer and colleagues' (2003) original application of the models. I subsequently address whether validity evidence supports these interpretations as well as the assumption that informants' reports are optimized when integrated using each measurement model.

TFM Findings. For TFMs, model fit and factor loadings varied depending on the informants entered into the model, and were driven in part by the extent to which informants' reports on the SPS-6 displayed a high degree of common variance. Caregiver-peer reports were not significantly correlated, and the caregiver-peer TFM exhibited poor model fit. This supports the notion that the TFM works particularly well when applied to informants' reports that share, at minimum, a fair degree of common variance. Further, this may elucidate Study 1 findings that Bauer and colleagues (2013) and subsequent researchers used the TFM with informants who are more likely to converge in their reports (e.g., two caregivers). Taken together, the findings of Studies 1 and 2 point to a key notion: the TFM may only yield adequate model fit under data conditions characterized by significant amounts of common variance among the reports entered into the model.

In contrast, adolescent-caregiver and adolescent-peer SPS-6 reports correlated at moderate levels, and TFMs using these dyads resulted in satisfactory model fit with both informants' reports loading significantly onto the Common Factor. However, in these models, adolescents' reports loaded more strongly onto the Common Factor than the adolescent Perspective Factor, while the reverse was true for caregiver and peer confederate reports. Bauer and colleagues would interpret this as indicating that adolescents' "true" social anxiety contributes less to caregivers' and peer confederates' ratings than does each of these informants' unique biases or perspectives. In contrast, Bauer and colleagues would interpret adolescents' reports as being more "accurate" and not confounded by bias due to non-significant loadings onto the adolescent Perspective Factor. Many symptoms of adolescent social anxiety are covertly expressed (Anderson & Hope, 2009). As such, this finding supports the idea that adolescents may have greater access to signs of social anxiety than do other informants.

However, these interpretations change when entering all three informants into the same TFM. When doing so, I found that adolescent and peer confederate reports loaded significantly onto the Common Factor, whereas the caregivers' reports were almost entirely captured in their Perspective Factor. Further, adolescent and peer confederate reports both loaded significantly onto their respective Perspective Factor. Bauer and colleagues (2013) would interpret this pattern of findings as indicating that caregivers' reports are characterized by such a significant level of bias that they are unable to accurately characterize their adolescent's social anxiety. Stated otherwise, this interpretation leads one to conclude that there is no value in collecting caregiver reports when assessing adolescent social anxiety, a finding that is contradicted by a large evidence-base supporting the psychometric properties of caregivers' anxiety reports (Etkin et al., 2021a). Further, within this sample, caregivers' reports on social anxiety measures relate to clinically relevant independent criterion variables (e.g., observed social anxiety and social skills, referral status; Glenn et al., 2019). Relatedly, Bauer and colleagues would interpret this pattern of factor loadings for the three-informant TFM as indicating that adolescents' and peer confederates' reports each contain accurate information when assessing adolescent social anxiety but that these informants' reports each contain unique biases. Study 2 validity analyses challenge these interpretations of the TFM.

TFMs containing caregiver reports also evaluated an informant-specific predictor, namely caregivers' self-reported depressive symptoms. In both TFMs that included caregiver self-reports of depression, these reports exhibited significant effects on the caregiver Perspective Factor but not the Common Factor. Bauer and colleagues (2013) would interpret this finding as reflecting that caregivers' own depressive symptoms do not relate to "actual" anxiety levels experienced by adolescents, but instead drive caregivers to rate their adolescent's anxiety as greater than other informants do (i.e., their depression biases their ratings). As described previously, this interpretation runs counter to the numerous pathways through which caregiver psychopathology contributes to increased risk for child psychopathology (e.g., genetics, parenting style; Caspi et al., 2004; Goodman & Gotlib, 1999; Lindhiem et al., 2020; Moffitt, 2005; Monroe & Harkness, 2005). Further, methodological features of this study, as in other studies evaluating the depression-distortion hypothesis, prevent me from teasing apart "bias" from shared method variance driven by caregivers providing reports of adolescent social anxiety *and* their own depressive symptoms. Ultimately, this validity testing offers preliminary answers to questions about how to interpret the role of a depressive "bias" in caregivers' reports.

TSSM Findings. As hypothesized and consistent with a prior application of the TSSM in this sample (Makol et al., 2020), I found satisfactory model fit for the TSSM with identified components reflecting patterns consistent with the Trait, Context, and Perspective scores described by Kraemer and colleagues (2003). Specifically, informants' social anxiety reports on the SIAS all loaded positively and substantially onto the Trait score. In contrast, caregivers and peer confederates exhibited contrasting loadings from each other on the Context score, and adolescents exhibited contrasting loadings from caregivers and peer confederates on the Perspective score. Importantly, the process of qualitatively evaluating component scores does not allow for hypothesis testing on the components obtained. Further, much like with TFM findings, evaluating component loadings does not validate that the scores reflect what they purport to. Study 2 analyses further evaluate the validity evidence for scores derived from the TSSM.

Evaluating the Validity of Scores Derived from the TFM and TSSM

As Study 1 demonstrates, a number of studies leverage bias and context models despite significant gaps in our knowledge about the validity of scores derived from them. I know of no prior study that critically evaluates these measurement models, compares and contrasts indices derived from them, and tests the incremental value of these indices, relative to each other. Thus, Study 2 represents an important first step in understanding how these measurement models perform when subjected to ecologically valid assessment conditions.

I found clear answers regarding the validity of scores derived from the TSSM and TFM, relative to each other. In line with prior work and study hypotheses (Makol et al., 2020), I found consistent support for the incremental validity of the Trait score over-and-above individual informants' reports, the composite score, and the Common Factor. Each of these alternative methods represent unique ways of utilizing the information obtained from multi-informant assessments, as well as unique assumptions about what informant discrepancies reflect. Using individual informants' reports aligns with the common practice of selecting an “optimal” informant (De Los Reyes et al., 2015; Loeber et al., 1989, 1990; Marsh et al., 2018), and approaches like the TSSM which assume that each informant provides useful clinical information. In contrast, users of the composite score assume that convergence signals truth or valid psychological phenomena and divergence signals measurement confounds (Borsboom, 2005; Edgeworth, 1888; Martel et al., 2021). This integrative strategy aligns with the TFM’s Common Factor, which also purportedly removes unique variance in the form of caregivers’ depressive bias. Thus, in focusing on the data conditions in which informant discrepancies systematically arise, the TSSM leverages both common and unique variance and appears to offer an effective strategy for integrating informants’ reports. Drawing from empirically supported context models, these findings provide further support for the TSSM as a measurement model for integrating informant discrepancies.

When considering TFM findings in isolation, there is mixed support for the validity of the Common Factor. Specifically, the three-informant Common Factor score predicted observed adolescent social anxiety, representing a moderate effect, but did not distinguish clinic-referred from community control adolescents. The latter finding is particularly surprising given that in all but one instance individual informants’ reports predicted adolescents’ referral status. This finding suggests that the TFM—a measurement model that integrates data from multiple informants—failed to incrementally predict a clinically relevant criterion variable that individual informants could predict on their own. Across TFMs, Common Factor scores inconsistently predicted observed adolescent social anxiety and referral status, over-and-above individual informants' reports. These mixed findings are easily interpreted when considering each informants' factor loadings onto the Common Factors across TFMs. In particular, I observed factor loadings for caregivers' reports at near-zero or relatively smaller magnitudes compared to other informants' loadings onto the Common Factor. Given this lack of overlapping variance, it is unsurprising that Common Factor scores predicted unique variance in observed behavior (and not referral status), over-and-above caregiver reports. In contrast, peer confederates' reports exhibited significant factor loadings in TFMs, and thus the Common Factor score did not provide incremental variance over-and-above the peer confederate reports. Importantly, both the individual peer confederate reports *and* caregiver reports displayed significant relations with the criterion variable. Thus, Common Factor scores performed no better in these models than an individual informant whose report displayed significant loadings onto the Common Factor *and* whose report displayed significant relations with the criterion variable. This is perhaps due to the assumption that variance shared among *multiple* informants' reports ought to outperform unique variance from an *individual* informant's report. Within data conditions that violate these assumptions (e.g., unique variance from informants contains valid information), Common Factor scores appear to hold little incremental value relative to reports from individual informants.

The inability of the Common Factor to predict independent criterion variables over-and-above the composite score also answers important questions about the validity of scores taken from the TFM. That is, although these two integrative strategies overlap in their emphasis on leveraging common variance to remove measurement confounds, the TFM is unique in its inclusion of a caregiver predictor thought to remove caregivers' "depressive bias." Thus, these findings suggest that removing this variance may actually detract from, instead of enhance, the validity of the scores taken from informants' reports. The association between caregivers' own depressive symptoms and the unique information they provide when rating their adolescent's social anxiety is likely best explained by the numerous *valid* ways in which environmental and genetic factors increase risk for the development of child psychopathology (e.g., Caspi et al., 2004; Goodman & Gotlib, 1999; Lindhiem et al., 2020).

Incremental validity findings are also unsurprising when considering the importance of basing theory and measurement in evidence. Relatedly, the sample in which I observed these findings represent a strong "fit" with the "mixing and matching" approach necessitated by Kraemer and colleagues' (2003) TSSM and speaks to the importance of using measurement models that accurately "fit" the underlying data conditions. Nonetheless, Bauer and colleagues (2013) do not offer a theoretical model for selecting informants' reports, and key measurement elements for evaluating their model as described (i.e., item level data from two or more informants, measure of caregiver psychopathology) were available in this study. Overall, my findings do not lend support to Bauer and colleagues' assertion that the TFM "purges" multi-informant reports of subjective bias, including caregivers' depression-related biases. Instead, these findings suggest that the model purges these reports of information that would otherwise enhance measurement validity, namely criterion-related and incremental validity. This is further supported by exploratory analyses finding that TFM Perspective Factors related to the independent criterion variables collected for this study (i.e., observed behavior, referral status). Martel and colleagues (2017a, 2017b) similarly found that Perspective Factors in their ADHD TFMs related to clinician-rated impairment, although this criterion variable was not completely independent from informants' ratings given that it was based on clinical interviews with the informants. Taken together, these findings suggest that the Perspective Factors, by construction, are not simply reflecting measurement confounds like rater biases. Rather, in this study they captured information relevant to understanding adolescents' social anxiety. In contrast, the TSSM, which uses PCA, maximizes the variance explained by the informants' reports. As such, the Trait score is presumed to offer the most precise prediction of independent criterion variables, relative to the Context and Perspective scores. Thus, in addition to asking what is gained by using a measurement model, these findings also speak to the importance of asking what is *lost* when using a model.

Chapter 5: Discussion

Best practices in youth mental health assessment entail collecting and interpreting reports from multiple informants (e.g., parents, teachers, youth, peers). Given that each informant is selected for the unique and incrementally valid information they provide, researchers and clinicians commonly encounter discrepancies between informants' reports (i.e., *r*s = .20s-.30s; Achenbach et al., 1987; De Los Reyes et al., 2015). For decades, these informant discrepancies have created a number of interpretive issues across the full range of research and clinical tasks for which researchers and clinicians collect multi-informant assessment data (e.g., assigning diagnoses, determining prevalence rates of psychopathology, evaluating intervention effects; De Los Reyes & Kazdin, 2005; Hawley & Weisz, 2003). Consequently, there have been repeated calls for empirically supported strategies for reconciling the uncertainties in decision-making when using multi-informant data (e.g., De Los Reyes, 2011; Beidas et al., 2015; Hunsley & Mash, 2007; Offord et al., 1996; Richters, 1992).

Yet, a key premise of this dissertation is that the youth mental health field lacks empirically based guidelines for integrating informants' discrepant reports. The absence of these guidelines stems from the lack of research explicitly linking theoretical models with measurement models. Developers of the prevailing measurement models—bias (e.g., TFM; Bauer et al., 2013) and context (e.g., TSSM; Kraemer et al., 2003) models—derive these models from distinct theoretical traditions. Context models, which are rooted in the concept of situational specificity (Achenbach et al., 1987), have much greater empirical support than bias models, and in particular the depression-distortion hypothesis (Ritchers, 1992). Despite this clear distinction in empirical support, we have much less clarity on how these distinct measurement models perform. In linking these three "pillars" of research on informant discrepancies (i.e., theory, quantitative methods, empirical support), my dissertation advanced research on youth mental health assessments and addressed longstanding conceptual and methodological issues when interpreting multi-informant reports.

Findings from Studies 1 and 2 highlight how the model one selects when using multi-informant reports guides all decisions within the assessment process, ranging from how users of these models select informants to how they obtain and interpret integrative scores. Perhaps most importantly, the range of decisions available to users of multi-informant reports differ widely in their empirical support. Further, in prior work researchers have rarely subjected scores taken from these models to validation testing. Thus, findings from Study 2 revealed key (and sometimes surprising) insights into how scores taken from these models perform when predicting clinically relevant criteria (e.g., observed behavior, referral status).

My dissertation findings suggest that, relative to scores derived from the TFM, scores derived from the TSSM align closest with "best practices" in evidence-based assessment of youth mental health and demonstrate utility as an integrative strategy. Consistent with my hypotheses, research using the TSSM reflects recommended practices when collecting multi-informant reports, and in particular selecting informants who vary in the context and perspective from which they rate youth mental health (Achenbach et al., 1987; De Los Reyes et al., 2015; Hunsley &

Mash, 2007). Further, the Trait score, which its developers contend reflects variance across informants' unique contexts and perspectives, outperformed each individual informant's report when predicting well-established, independent criterion variables (i.e., observed behavior, referral status). This finding supports the notion that integrating multi-informant data using scores taken from the TSSM achieves more in prediction than any one informant can achieve on their own. Additionally, the TSSM-derived Trait score outperformed a strategy one uses when assuming that unique variance reflects measurement confounds (i.e., composite score), as well as a strategy for which the developers contend reflects common variance that "purges" unique biases from individual informants' reports (i.e., TFM-derived Common Factor). Even when using all three informants' reports, the TFM Common Factor score failed to consistently outperform individual informants' reports. In essence, use of the Common Factor score *reduced* measurement validity in two ways. First, it resulted in depressing the incremental value of multi-informant data, relative to individual informants' reports and the *simple average* of these reports. Second, features of the TFM designed to partial out variance reflecting measurement confounds (i.e., Perspective Factors) actually contained information relevant to understanding adolescent social anxiety.

When considering theoretical models and their empirical support, these findings would only come as a surprise if one ignores the empirical literature on informant discrepancies. That is, the "three pillars" of work on informant discrepancies tell a consistent story: context models rest on strong conceptual and empirical foundations, whereas bias models do not. This distinction in conceptual and empirical support translates to differences in performance when comparing scores derived from these models to one another. Theoretical models and their empirical support inform *who* users of these models select to include in multi-informant assessment batteries. They also inform *how* users conceptualize what informant discrepancies reflect, and *which* strategies they use for integrating them. Importantly, decisions surrounding the strategy used to integrate multi-informant data do not rest on whether the TFM or TSSM are "correct models." Rather, the key determining factor is the degree to which these measurement models "fit" the underlying data conditions. Although well documented in prior research, the findings of this dissertation support the notion that informant discrepancies often reflect valid information. Consequently, and within such data conditions, users should leverage integrative strategies that derive from models that assume informant discrepancies reflect valid information (i.e., TSSM).

Research and Theoretical Implications

What's in a Name?

As mentioned previously, although hundreds of interventions have been developed to address youth mental health needs, few have undergone any kind of empirical testing (Weisz & Kazdin, 2017). That is, intervention developers often disseminate their programs to users and to clients *before* testing their effectiveness. Similarly, a key observation I made in my dissertation is that developers of measurement models often disseminate them to users *before* testing how these models perform in relation to clinically relevant, independent criterion variables. This represents a serious barrier to building consensus guidelines on how and under what circumstances to integrate multi-informant data. A core principle of the evidence-based assessment movement is that developers of assessment instruments need to demonstrate the psychometric properties of scores taken from these instruments (Hunsley & Mash, 2007). A key take-home message from my dissertation is that we must subject scores taken from measurement models to these same standards. The most common data reported in support of using these models (e.g., fit indices, factor and component loadings), do not address whether a given model effectively integrates multi-informant reports. Validation data do. My dissertation only begins to address questions about the validity of integrative scores derived from the TFM and TSSM. That said, my findings point to a method for building a sound evidence-base for measurement models and strategies for integrating multi-informant reports more generally.

A key issue surrounding use of the TFM and TSSM is how model developers approached interpreting scores derived from these models. The researchers who developed these models are both guilty of the "naming fallacy" (Kline, 2016), whereby they interpreted the meaning of measured factors without use of validity evidence to support their interpretations. Consequently, in Study 2 I used independent criterion variables to determine whether the key integrative scores derived from the TFM and TSSM (i.e., Common Factor and Trait, respectively) validly characterized adolescent social anxiety. Future research using other independent criterion variables is needed to fully unpack the interpretive value of scores derived from the TFM and TSSM, such as independent assessments of caregiver mood and observed interactions between adolescents and caregivers. Nonetheless, Study 2 findings suggest that Bauer and colleagues (2013) are inappropriately interpreting the Perspective Factors derived from their model as reflecting informants' biases, including a caregiver rating bias linked to their levels of depressive symptoms. Following Bauer and colleagues' interpretation of the caregiver Perspective Factor, one would conclude that the caregiver's rating bias meaningfully relates to adolescent behavior within peer interactions and referral status. Yet, this interpretation is anathema to how users of these models treat rater biases (i.e., as measurement confounds). By definition, measurement confounds are unrelated to measured domains (see Millsap, 2011), and as such users are compelled to remove variance attributed to bias. My findings support the notion that there are data conditions within which it would be misguided to conclude that Perspective Factor scores reflect rater biases. These findings speak to the need to critically examine integrative scores to unpack the type of information they contain (i.e., measurement confounds vs. valid data) and evaluate their ability to characterize youth mental health. Researchers need to question not only what is gained from leveraging integrative scores taken from measurement models, but also what is lost. These questions can only be addressed by subjecting scores taken from measurement models to validation testing.

Matching Measurement Models to Data Conditions

Very little prior work rigorously examines competing strategies for integrating informant discrepancies, and this dissertation reflects only the beginning of such efforts. My Study 2 findings speak to the necessity of developing measurement models that meet the data conditions underlying their use. In particular, Study 2 illustrates the consequences of developing measurement models informed by conceptualizations of multi-informant assessments (i.e., the depression-distortion hypothesis) that lack a strong evidence-base. In essence, when a model rests on weak theoretical and empirical foundations, then scores derived from the model are at risk of poorly "fitting" the data conditions to which they are applied. For instance, if depression-related rater biases explain minimal-to-no variance in informant discrepancies, my findings indicate that scores taken from a bias model like the TFM will demonstrate poor performance when subjected to validation testing. That said, Study 2 used a single set of data conditions, albeit within a well-characterized sample (e.g., Cannon et al., 2020; Deros et al., 2018; Glenn et al., 2019; Makol et al., 2020). Thus, I recommend that researchers continue to subject scores derived from the TSSM and TFM to validation testing—in addition to scores derived from other measurement models that grew out of the theoretical traditions that informed their development—within varying types of data conditions. These data conditions might include multiple samples of youth from different developmental periods and clinical populations, and studies that, collectively, use diverse sets of multi-informant instruments and criterion variables.

How important is it to vary the data conditions underlying validation tests of integrative scores? My dissertation reveals some initial insights into this larger question. Specifically, model fit and factor loadings for the TFMs varied depending on which informants were included in the model, leading to vastly different conclusions. When using the TFM to integrate multi-informant reports of ADHD symptoms, Martel and colleagues (2017a, 2017b) also found very different model fit and factor loadings depending on which informants were included in the model. In terms of the TSSM, this measurement model includes some key parameters on measurement conditions (e.g., use of three informants). Kraemer and colleagues (2003) also reported unique component loadings depending on which informants were used and, consequently, argued that informants must be selected using their mix-and-match criterion (i.e., varied across relevant contexts and perspectives). Importantly, when varying informants included in the measurement models, future work should select informants who provide psychometrically sound reports of the domain being rated.

Additionally, future research should examine these measurement models across developmental stages. My Study 1 findings indicate that users of the TFM tend to select informants who share substantial common variance (e.g., pairs of informants from the same context; Bauer et al., 2013). My Study 2 findings indicate that this may be due in part to the data conditions needed to obtain satisfactory model fit. I recommend that future research examine whether the TFM may be more aligned with multi-informant assessments in which the data conditions depart from the typical data conditions of youth assessments, namely adult assessments. As mentioned previously, youth mental health assessments tend to systematically select informants who are situated within key contexts (e.g., home and school; Achenbach et al., 1987; De Los Reyes et al., 2015). These data conditions align quite well with the TSSM and in particular, the mix-and-match criterion that guides informant selection. In contrast, in studies of adults, researchers commonly use a mix of "other" informants beyond self-report (e.g., romantic partners, friends, coworkers), typically referred to as *collateral informants* (Achenbach et al., 2005). This "mix" of informants likely precludes the ability of researchers to systematically account for measured factors contributing to discrepancies among informants' reports. As such, these studies do not encapsulate the assessment designs needed to capitalize on the systematic and meaningful variance reflected in informant discrepancies that typically characterize assessments of youth, where all youth in a sample have the same informants rating their behavior. In these respects, multi-informant assessments of adults match characteristics that are consistent with Eid and colleagues' (2008) definition of *interchangeable informants*. The random, interchangeable nature of this multi-informant approach might produce informant discrepancies that reflect measurement confounds. In this sense, common variance estimates such as the TFM Common Factor might be the most appropriate strategy for these data conditions.

Importantly, the notion of selecting interchangeable informants runs counter to the very purpose of collecting multi-informant reports. Either way, when users select two informants with a high degree of overlapping variance, it is unclear whether each informant provides incrementally valuable information, relative to each other, when characterizing mental health. Further, such an approach might result in a "naming fallacy" all its own. That is, in selecting informants whose reports correspond to a considerable degree, one might wrongfully assume that these reports, collectively, provide adequate coverage of the measured domain. Supporting this idea, meta-analytic work demonstrates that adult assessments show similar levels of convergence among informants as do youth assessments (i.e., *r*s = .43-.44; Achenbach et al., 2005). Given that there is relatively little research on how informant discrepancies within adult mental health assessments relate to independent validity criteria, I encourage future work that addresses these important questions.

Ultimately, model users should not assume that any one measurement model applies equally to all data conditions in which one collects and models multi-informant data. Stated otherwise, no single model for integrating these data serves as a panacea for use under all circumstances in which users of these models observe informant discrepancies. In essence, theoretical and measurement models can never be "proven wrong," and as such, data conditions dictate the utility or applicability of these models (Borsboom, 2005). Under a different set of data conditions (e.g., use of informants whose reports display high levels of convergence), I may have observed findings that lent more favorable support for the TFM over the TSSM. Thus, research is needed to understand the conditions in which measurement models, including the TFM and TSSM, enhance prediction, and the conditions in which they do not.

In addition to conducting further validity studies that vary key data conditions, Monte Carlo simulations offer a complimentary approach to accomplishing this task (Paxton, Curran, Bollen, Kirby, & Chen, 2001). Simulations allow a user to freely vary key data conditions (e.g., level of convergence between informants, number of informants), and examine hypotheses about which measurement models perform best within those conditions. Such work may find, for example, that scores taken from the TSSM outperform those taken from the TFM when informants share a relatively low amount of common variance, whereas the scores taken from TFM outperform those taken from the TSSM when the informants share a relatively large amount of common variance. Further, simulation studies of the TFM and TSSM may help users identify *precise thresholds* at which one model versus another might become an optimal strategy for data integration. For example, at what level of common variance might scores taken from the TFM begin to outperform scores taken from the TSSM, in relation to criterion variables? How much unique variance among informants' reports needs to relate to the criterion variable for scores taken from the TSSM to outperform scores taken from the TFM? In sum, future work varying data conditions within a Monte Carlo simulation framework will aid not only in evaluating the validity of scores taken from these and other measurement models, but also in identifying the data conditions in which informants' reports are optimized for the task at hand.

Why Do Bias Models Persist?

Given the limited empirical support for bias models, it is perplexing that they continue to be applied in research and clinical settings and for the most commonly used informants in youth mental health assessment (i.e., caregivers; De Los Reyes et al., 2015). Beyond the lack of rigorous research on theoretical and measurement models of informant discrepancies, a clear reason for their persistence can be found in our paradigms for validating assessment approaches generally. Specifically, Campbell and Fiske's (1959) MTMM matrix is a seminal paradigm that overlaps with Converging Operations, namely the assumption that common variance (i.e., high mono trait-hetero method correlations) reflects "truth" and informant discrepancies (i.e., low mono trait-hetero method correlations) reflect measurement confounds. However, there is clear evidence that informant discrepancies as observed in research and practice (i.e., the data conditions that typify multi-informant assessments of youth), often *do not* reflect measurement confounds. Thus, although the MTMM matrix has advanced measurement in many fields, drawing solely on this paradigm in multi-informant assessment is likely to lead to the persistence of measurement models that rest on faulty assumptions. This creates a tension between the long tradition of focusing on common variance, and more recent research demonstrating the validity of unique variance among informants' reports. We need to refine our validation paradigms, or create new paradigms that fit the data conditions that underlie multi-informant assessment of youth mental health, wherein informant discrepancies often reflect meaningful information.

Avoiding "Either/Or" Approaches

The two measurement models evaluated in this dissertation grew out of distinct conceptual and empirical literatures. Although they were presented as competing strategies for integrating multi-informant reports, they may offer complementary information when assessing youth mental health. Indeed, although a strong evidence-base supports the impact of context on informants' reports, these reports all contain some level of measurement error and at least some of this measurement error may explain the discrepancies between informants' reports (Borsboom 2005; Nunnally & Bernstein, 1994). Might a measurement model accounting for *both* bias and context optimize the information provided by informants' reports? Research examining these questions should first estimate the impact of both bias and context on informants' reports and then test whether the scores derived from a measurement model effectively isolates variance attributable to bias and context effects (i.e., in relation to well-established validity criteria).

However, even when one finds evidence of a rater bias (e.g., depression-distortion effects), this factor may account for relatively little variation within informants' reports and apply to a very limited set of data conditions (e.g., clinical setting in which caregivers' negative moods are systematically "activated" prior to collecting reports and *not* when a caregiver has any lifetime history of depression; Youngstrom at al., 1999). At the same time, it is important to note that although the depression-distortion hypothesis is the most widely studied bias model, other bias models warrant further study. We must subject these bias models to rigorous testing, including models focusing on the impact of racial biases (Fadus et al., 2020; Kang & Harvey, 2020) and cultural perspectives on mental health (Achenbach, 2017; Chen, Ho, Lee, Wu, & Gau, 2017; Lau et al., 2004) on informant discrepancies, among others. This research will be advanced through particular attention to validity issues, and critically examining whether evidence for bias (i.e., measurement confounds) is found.

Relatedly, although the TSSM appears to most immediately offer a measurement tool that effectively integrates informants' reports, the TFM more immediately offers utility in hypothesis testing about the structure of informant discrepancies. Given that variance among informants' reports is broken down into informant-specific and item-specific factors, the model can be used to test hypotheses about whether these factors meaningfully characterize informants' reports. Although Bauer and colleagues (2013) focus on bias when interpreting Perspective Factors, it may very well be that these factor capture unique contextual information about the domain being assessed (e.g., anxiety at home vs. with peers). When a user observes a

Perspective Factor score relating to a well-established validity criterion, such an observation effectively "rules out" the notion that this particular instance or use of the Perspective Factor score reflects a measurement confound. In contrast, PCA is an exploratory technique and interpreting the components is a relatively subjective process, which prevents hypothesis testing regarding the number and structure of components derived from informants' reports. In this way, the TSSM relies more on a priori strategic selection of informants, which is rooted in context models and their empirical support.

Clinical Implications

The theories and measurement models examined in this dissertation grew out of complexities underlying assessments of youth mental health, namely the need to collect and interpret data from multiple informants. Findings have important implications for how users incorporate multi-informant data into their clinical practices, including decisions surrounding which informants to select, how to interpret any discrepancies observed, and how to integrate the data informants provide. Assumptions of the MTMM matrix and Converging Operations are readily observed in routine clinical practice. As previously mentioned, idiosyncratic strategies for reconciling informant discrepancies are commonly used despite the fact that statistical prediction often outperforms clinical prediction (Grove et al., 2000; Meehl, 1954; Rettew et al., 2009; Youngstrom et al., 2018). It is common for assessors who encounter discrepant information from informants to choose the "optimal" informant who is thought to most accurately arrive at the "truth" (Brown-Jacobsen et al., 2011; De Los Reyes et al., 2011, 2015; Hawley & Weisz, 2003;

Loeber et al., 1989, 1990; Marsh et al., 2018). Thus, in focusing on convergence as truth, clinicians gather comprehensive data to gain a fuller picture of their client, only to ignore some information sources when one or more pieces of the picture disagree.

As described by Etkin and colleagues (2021a), taking a bias approach leads to a clinician getting stuck in the "right-versus-wrong conundrum" when interpreting discrepant reports (p. 157). In contrast, taking a context approach facilitates testing hypotheses about why behavior may differ across clinically relevant contexts. Drawing on findings in my dissertation, future work should leverage integrative strategies in routine clinical practice and evaluate whether use of these models enhances care. Further, this work should directly test bias versus context approaches to clinical decision-making, and their impact on assessment and intervention outcomes. These findings can inform *measurement-based care* (i.e., systematic and continuous assessment of mental health throughout treatment for the purpose of monitoring progress and informing clinical decision-making), which has been shown to positively impact therapeutic alliance and treatment outcomes (Jensen-Doss et al., 2020; Scott & Lewis, 2015). For example, future work should examine how integrative scores such as the Trait score can be standardized and normed to provide clinicians with cutoffs and interpretative labels (e.g., non-clinical, borderline, and clinical cutoffs; Achenbach, 2001) that can then be used to inform clinical decision-making (e.g., diagnosis, prognosis, treatment allocation). When developing measurement-based approaches for using multi-informant reports, particular attention should be paid to approaches that enhance use of these reports and the accuracy of clinical decisions beyond current approaches (e.g., selecting a best informant or assuming one informant is more accurate). Further, attention to common barriers for implementing evidence-based assessment practices in routine clinical care will be needed. For example, common barriers to implementing measurement-based approaches can occur at the level of the informants (e.g., time and motivation for completing measures), clinician (e.g., resistance to using reports to inform decisions), and organization (e.g., lack of resources for training clinicians on how to use reports, lack of funding for measures and computer programs for scoring them; Lewis et al., 2019).

Given that validation testing has yet to be conducted in clinical settings, the following vignettes illustrate how use of bias and context models leads to fundamentally different conclusions when a set of informants provides divergent reports. Take for example a mother bringing her preschooler to therapy given concerns about the child's disruptive and oppositional behavior. The clinician administers symptom measures to the child's mother and teacher and finds that the mother reported very elevated levels of externalizing behaviors whereas the teacher did not. Through a clinical interview, the clinician learns that the child's mother experienced depression over the past year, which has contributed to feelings of guilt about her parenting, fewer positive parent-child interactions, and inconsistent use of strategies to address the child's behavior. If the clinician draws from bias models, they may consider disregarding the mother's reports entirely given concerns about a depressive bias that reduces the validity of her reports. The clinician may then turn to the teacher as an "unbiased" expert on the child's behavior who is better equipped to arrive at the "truth." In doing so, the clinician may conclude that the child is not experiencing clinically significant mental health problems and the family would be best served through individual therapy for depression with the child's mother only. Although a clinician is unlikely to disregard the reports of the referral informant bringing the child into treatment, such a decision would directly follow from bias model theory and measurement. This decision, however, does not track with the evidence-base on youth mental health, in particular the robust link between caregiver mental health and the development and maintenance of the mental health of youth in their care (e.g., Carlone & Milan, 2021; Goodman & Gotlib, 1999). Stated another way, what if the mother's depression reflects a feature of the youth's environment that validly relates to the youth's mental health?

Alternatively, if the clinician draws from context models, they may develop hypotheses about where the child's problem behavior is occurring and environmental conditions and antecedents that promote and maintain the behavior. As identified by the child's mother, home may be characterized by high levels of criticism (as opposed to praise), inconsistent discipline, and poor parent-child attachment, risk factors associated with the development of mental health problems in early childhood and in the context of caregiver depression (Goodman & Gotlib, 1999; Lindhiem et al., 2020). The clinician can then consider how to target the child's behavior by reshaping the primary context in which the behavior is occurring, utilizing evidence-based interventions that promote positive parent-child interactions as well as effective and consistent use of discipline (e.g., Child-Parent Psychotherapy or Parent-Child Interaction Therapy; Lieberman et al., 2015; Eyberg & Funderberk, 2011).

As another example, consider a scenario in which a family is seeking an ADHD evaluation for their adolescent daughter. The evaluator collects symptom reports from the adolescent's father and teacher, as well as the adolescent herself. When examining these reports, the evaluator finds that all informants converge in very elevated ADHD symptom reports. However, whereas the adolescent reports moderate levels of anxiety, her father reports no anxiety concerns, and her teacher reports very elevated anxiety concerns as well as some depressive symptoms. If following a bias model approach, the evaluator may consider bias in two forms. First, due to concerns about stigma, the adolescent's father may have a *social desirability bias* in which he wants to minimize the perception that his daughter struggles with anxiety (Rodriguez, Wittig, & Christl, 2019). Second, the teacher's reports may best be understood as suffering from an *evaluative consistency bias* in which the teacher inaccurately endorses a broad range of difficulties due to elevated concerns in only one domain (i.e., ADHD; Dhillon, Bagby, Kushner, & Burchett, 2017). When providing feedback to the family, the evaluator can then consider communicating that either the father or teacher are "missing the mark" on anxiety due to biases contained in their ratings. Such an approach is likely to isolate information-gathering to just a single informant used within the evaluation. From a practical standpoint, this approach begs the question: Why did the clinician select these informants for the expert information they provide, if they were fallible and unreliable as information sources?

Alternatively, if the evaluator adopts a context approach, they may form hypotheses about aspects of the adolescent's social environments that elicit (or do not elicit) anxiety. In a clinical interview with the adolescent, the evaluator may learn that the adolescent's ADHD symptoms are causing significant anxiety about academics and particularly at school (e.g., ADHD-related impairment leads to low self-esteem and anxiety about academic failure). However, the adolescent may experience minimal anxiety at home due to regular organizational support and low levels of pressure from her caregivers on academic tasks (e.g., support on homework, regular praise about academic strengths). When providing feedback to the family, the evaluator can encourage the caregivers to continue their supportive approach and recommend therapy for the adolescent to learn coping strategies to use at school to reduce anxiety (e.g., Cognitive Behavioral Therapy for ADHD; Sprich, Safren, Finkelstein, Remmert, & Hammerness, 2016). Taken together, these clinical vignettes highlight how the act of taking a bias or context approach when interpreting assessment data in clinical care translate to fundamentally different decisions regarding care.

Limitations

The findings of this dissertation should be interpreted in the context of its limitations. First, the two measurement models evaluated provide exemplar models given that each draws on bias or context theories about why informant discrepancies occur, and apply a statistical approach to arrive at a quantifiable, integrated multi-informant index. My dissertation does not put to rest conceptual, empirical, or measurement issues when collecting multi-informant reports and encountering informant discrepancies. It is important to refrain from conflating the measurement models of the TFM and TSSM with the theoretical models that informed their development. Within informant discrepancies research, theoretical and measurement models address some shared aims but also many unique aims. Theoretical modeling informs our field's understanding of the construct of informant discrepancies (e.g., how to interpret patterns among informants reports). Measurement models address the need for integrating informant discrepancies to achieve a broad range of research aims when using multi-informant reports (e.g., prediction, characterizing treatment effects). Nonetheless, the approach of examining the three "pillars" of informant discrepancies together is needed to understand how theory informs measurement, and vice versa. Second, I examined how the two measurement models are systematically used in published research in Study 1, but did not examine validity evidence across the numerous research areas in which the measurement models were applied. As previously mentioned, the vast majority of research using the TFM and TSSM does not examine the validity of the measurement models. Thus, I could not obtain meta-analytic effect sizes across the studies examined in Study 1.

Relatedly, a direct comparison of the application of theory to the two measurement models is not possible given that each uses a distinct statistical approach, albeit with unique features that reflect a focus on either common variance and bias or unique variance and context. For example, while PCA as used within the TSSM sets the number of extracted components to three, the TFM portions out variance across numerous factors that are specified a priori (i.e., Common, Perspective, and Specific Factors). In this way, the TFM accounts for additional factors hypothesized to impact variance in informants' reports, including Specific Factors that are not aligned with the depression-distortion hypothesis. For these reasons, I also compared scores taken from the TFM and TSSM to individual informants' reports (emphasizing unique variance) and the composite score (emphasizing common variance) when predicting independent criterion variables.

An additional limitation of Study 2 is the relatively small sample I used to address my aims. I selected the sample for its rich breadth of modalities (e.g., parallel informants' reports on multiple social anxiety measures, observed behavior) and the psychometric evidence supporting use of these modalities. Given the amount of resources required to collect multi-modal assessments with such depth, it is rare to find large datasets that include several informants' reports *and* independent, clinically relevant criterion variables. Nonetheless, future research should address similar aims in a dataset with both "breadth" and "depth." Relatedly, Study 2 lacked independent criterion variables capturing caregivers' mood as well as adolescent behavior within other salient contexts (e.g., home, classroom). Although my use of caregivers' self-reported mood is consistent with empirical studies examining the depression-distortion hypothesis and applications of the TFM (e.g., Curran et al., 2020; Madsen et al., 2020), this shared method variance between caregivers' self-reports of depression and their reports of adolescent social anxiety creates criterion contamination issues (Garb, 2003). Use of rich and truly independent criterion variables are fruitful avenues for future research, and can further elucidate the information captured in the Common Factor and Trait scores, as well as other factors and components within the measurement models.

Finally, the mix-and-match criterion of the TSSM and informants available within my sample prevented me from examining various informant triads in the

TSSM. This is in contrast to modeling for the TFM, which included various informant dyads as well as an informant triad. Thus, more research is needed to identify the range of informants appropriate for the TSSM. In addition, the use of three informants presents practical limitations for researchers. The TSSM may not offer a generalizable tool for routine practice settings in which collecting more than two informants' reports is not feasible due to cost and time limitations. Further, within my study, some factor analytic criteria (i.e., KMO), were just above acceptable thresholds for determining if PCA is an appropriate statistical tool to use.

Chapter 6: Conclusion

Across research and clinical settings, assessors can rarely avoid encountering informant discrepancies, introducing long-standing complexities with using and interpreting assessment data. Most researchers and clinicians advocate for use of multi-informant reports when assessing youth mental health. However, there is a lack of consensus on the best strategy for reconciling informant discrepancies. Across over 50 years of research on this topic, many approaches have been proposed for understanding why discrepancies arise, and relatedly, many strategies for integrating these reports have been developed. The vast majority of these strategies fall within a bias or context approach to informant discrepancies. The implications of assumptions underlying use of these strategies are not trivial. The assumptions users make and the strategies they leverage to address research questions and inform clinical decisions have broad relevance to countless research areas, mental health domains, developmental periods, settings, and even entire disciplines (Achenbach et al., 1987, 2005; De Los Reyes et al., 2015, 2019; Duhig et al., 2000; Gresham et al., 2018; Hou et al., 2019; Jones et al., 2019; Korelitz & Garber, 2016; Narad et al., 2014; Rescorla et al., 2013, 2017; Romano et al., 2018; Stratis & Lecavalier, 2015). Further, across these many areas of research and practice, the strategy used for reconciling discrepant information leads to vastly different conclusions across tasks for which the informants' reports are used (e.g., diagnosis, determining prevalence rates of psychopathology, understanding risk over time, evaluating intervention effects; De Los Reyes & Kazdin, 2005; Hawley & Weisz, 2003).

My dissertation findings suggest that users of strategies for integrating multi-informant data need to make informed decisions, and directly test the assumptions underlying use of these strategies. In particular, the research support for context models of informant discrepancies translate to *use* of informants who provide incremental information and *improved performance* when characterizing adolescent anxiety. Many exciting avenues exist for future research on theoretical and measurement models for reconciling informant discrepancies. These avenues range from the basic science of what informant discrepancies reflect, to the measurement work focused on the data conditions and statistical approaches most appropriate for integrating these reports, to understanding how to best use informants' reports to make clinical decisions within routine practice settings. Across these connected and important tasks, attention to both theory and measurement will advance the field of informant discrepancies, and ultimately, improve clinical care for youth and their families.

Tables

Table 1

Variables Coded in Study 1 Systematic Review of Trifactor Model (TFM) and Trait Score Satellite Model (TSSM) Studies

Variable	Coding Values	Coding Guidelines
Peer Review	1 = Yes, 0 = No	Published in a peer-reviewed journal?
Language	1 = Yes, 0 = No	Is study written in English?
Measurement Model*	1 = Yes, 0 = No	Does the study implement the measurement model as described by the original authors?
Sample Size	Continuous	What is the sample size used for analyses applying the measurement model? If there is more than one time point but the measurement was only implemented at one time point, what is the sample for the time point at which they applied the model? If implemented with different sample sizes, code each separately.
Age	1 = Early childhood (0-4), 2= Childhood (5-12), 3 = Adolescence (13-18), 4 = Adulthood (18 or older), 5 = Multiple age groups (specify), -999 = Not specified	What is the predominant age group of the sample and/or target being rated? Based on best estimate from Mean age, Range of age, grade in school, target group, etc.
Construct(s) Measured	String	What construct or construct(s) were measured in the measurement model?
Number of informants	Continuous	How many informants were included in the measurement model? If implemented with different types and/or total number of informants, code each separately.
Informant 1 Type	1 = Self, 2 = Mother, 3 = Father, 4 = Caregiver, 5 = Teacher, 6 = Peer, 7 = Clinician, 8 = Various, 9 = Other (specify informant)	What type of informant is the first informant included in the measurement model? Code informants in the order they are described.
Informant 2 Type	1 = Self, 2 = Mother, 3 = Father, 4 = Caregiver, 5 = Teacher, 6 = Peer, 7 = Clinician, 8 = Various, 9 = Other (specify informant)	What type of informant is the second informant included in the measurement model? Code informants in the order they are described.
Informant 3 Type	1 = Self, 2 = Mother, 3 = Father, 4 = Caregiver, 5 = Teacher, 6 = Peer, 7 =	What type of informant is the third informant included in the measurement model? Code

	Clinician, 8 = Various, 9 = Other (specify informant)	informants in the order they are described.
Informant 4 Type	1 = Self, 2 = Mother, 3 = Father, 4 = Caregiver, 5 = Teacher, 6 = Peer, 7 = Clinician, 8 = Various, 9 = Other (specify informant)	What type of informant is the fourth informant included in the measurement model? Code informants in the order they are described.

Note. *Coders were provided an extensive guide that included criteria for identifying whether the measurement model was used.

Table 2

Summary of Trifactor Model (TFM) Studies Included in the Study 1 Systematic Review (n = 8)

Study	**Sample Size**	**Developmental Period**	**Construct Measured**	**Number of Informants**	**Informant Context**				
					Home	School	Mix	Peer	Various
Bauer et al. (2013)	626	Early Childhood, Childhood, Childhood, Adolescence	Negative affect	2	✓^{m} ✓^{f}				
Chen et al. (2015)*	1,132	Adulthood	Sexual openness	1			✓^{s}		
Clark et al. (2017)	273	Early Childhood, Childhood	Temperament	2	✓^{m} ✓^{f}				
Curran et al. (2020)	359	Adulthood	Depressive symptoms	2			✓^{s}	✓^{p}	
Haeny et al. (2018)	368	Early Childhood, Childhood, Childhood, Adolescence	Impulsivity	2	✓^{m} ✓^{f}				
Martel et al. (2017a)	725	Childhood, Adolescence	ADHD symptoms	3	✓^{m} ✓^{f}	✓^{t}			
Martel et al. (2017b)	406	Adulthood	ADHD symptoms	2			✓^{s}		✓
von der Embse et al. (2019)	24,094	Childhood, Adolescence	Social, academic, and emotional behavior	2		✓^{t}	✓^{s}		

Note. Early Childhood = 0 to 4 years; Childhood = 5 to 12 years; Adolescence = 13 to 18 years; Adulthood = 18 or older; ✓^{m} = mother; ✓^{f} = father; ✓^{t} = teacher; ✓^{p} = peer; *In this study, four Perspective Factors were included, but all came from self-report.

Table 3

Summary of Trait Score Satellite Model (TSSM) Studies Included in the Study 1 Systematic Review (n = 39)

Study	**Sample Size**	**Developmental Period**	**Construct Measured**	**Number of Informants**	**Informant Context**				
					Home	School	Mix	Peer	Various
Armstrong et al. (2014)	396	Childhood	Internalizing problems, externalizing problems	3	✓^{m}	✓^{t}	✓^{s}		
Buisman et al. (2020)	395	Adolescence, Adulthood	Childhood abuse, childhood neglect	3	✓^{m} ✓^{f}		✓^{s}		
Burk et al. (2008)	238	Childhood	Emotional/behavioral difficulties, social experiences	3	✓^{m}	✓^{t}	✓^{s}		
Burk et al. (2011), Model 1*	362	Early Childhood, Childhood	Temperament	2	✓^{m} ✓^{f}				
Burk et al. (2011), Model 2*	362	Adulthood	Parenting stress	2	✓^{m} ✓^{f}				
Burk et al. (2011), Model 3*	362	Adolescence	Internalizing problems, externalizing problems	3	✓^{m}	✓^{t}	✓^{s}		
Caldwell et al. (2015)	76	Childhood, Adolescence	Externalizing problems	3	✓^{m}	✓^{t}	✓^{s}		
De Pauw et al. (2009) Model 1*	56	Childhood	Self-esteem	2	✓^{c}		✓^{s}		
De Pauw et al. (2009) Model 2*	59	Childhood	Internalizing problems, externalizing problems	2	✓^{c}	✓^{t}			
De Pauw et al.	41	Childhood	Emotional/behavioral difficulties,	2	✓^{c}	✓^{t}			

(2009) Model 3*			Peer problems					
El-Sheikh et al. (2011)	251	Childhood	Marital conflict	3	$\checkmark^{m}$ $\checkmark^{f}$		$\checkmark^{s}$	
Erath et al. (2011)	251	Childhood	Harsh parenting	3	$\checkmark^{m}$ $\checkmark^{f}$		$\checkmark^{s}$	
Essex et al. (2010)	238	Early Childhood, Childhood, Adolescence	Behavioral inhibition	3	$\checkmark^{m}$	$\checkmark^{t}$	$\checkmark^{s}$	
Essex et al. (2011)	96	Childhood, Adolescence	Internalizing problems, externalizing problems	3	$\checkmark^{m}$	$\checkmark^{t}$	$\checkmark^{s}$	
Gardner et al. (2008)	803	Adolescence, Adulthood	Self-regulation, deviant peer affiliation	3	$\checkmark^{m}$	$\checkmark^{t}$	$\checkmark^{s}$	
Goelman et al. (2014)	294	Early Childhood, Childhood	Internalizing problems, externalizing problems	3	$\checkmark^{c}$	$\checkmark^{t}$	$\checkmark^{s}$	
Hatzinger et al. (2007)	102	Early Childhood	Emotional/behavioral difficulties	3	$\checkmark^{c}$	$\checkmark^{t}$	$\checkmark^{s}$	
Hatzinger et al. (2010)	82	Early Childhood, Childhood	Emotional/behavioral difficulties	3	$\checkmark^{c}$	$\checkmark^{t}$	$\checkmark^{s}$	
Houts et al. (2010)	2024	Childhood	Challenging behavior	3	$\checkmark^{m}$ $\checkmark^{b}$	$\checkmark^{t}$		
Keil et al. (2019)	329	Childhood, Adolescence	Emotional/behavioral difficulties, peer problems	3	$\checkmark^{c}$	$\checkmark^{t}$	$\checkmark^{s}$	
Kraemer et al. (2003)	539	Childhood	Emotional/behavioral difficulties, academic problems	3	$\checkmark^{m}$	$\checkmark^{t}$	$\checkmark^{s}$	
Kroenke et al. (2011)	78	Childhood	Internalizing problems, externalizing problems	3	$\checkmark^{c}$	$\checkmark^{t}$	$\checkmark^{s}$	
Makol et al., 2020	127	Adolescence	Social anxiety	3	$\checkmark^{c}$		$\checkmark^{s}$	$\checkmark^{p}$
Noordhof et	2,230	Childhood	Emotional/behavioral	3	$\checkmark^{c}$	$\checkmark^{t}$	$\checkmark^{s}$	

al. (2008)			difficulties						
Obradovic et al. (2010)	338	Childhood	Externalizing problems, prosocial behavior, school engagement, academic competence	3	✓^{c}	✓^{t}	✓^{s}		
Obradovic et al. (2011)	260	Early Childhood, Childhood	Internalizing problems, externalizing problems	3	✓^{c}	✓^{t}	✓^{s}		
Owens & Hinshaw (2016)	140	Childhood, Adolescence	Internalizing problems, externalizing problems	3	✓^{c}	✓^{t}	✓^{s}		
Perren et al. (2006)	168	Childhood	Emotional/behavioral difficulties	3	✓^{c}	✓^{t}	✓^{s}		
Perren et al. (2007)	160	Childhood	Emotional/behavioral difficulties	3	✓^{c}	✓^{t}	✓^{s}		
Rijlaarsdam et al. (2016)	3,136	Childhood	Oppositional behavior, aggression	3	✓^{c}	✓^{t}	✓^{s}		
Roubinov et al. (2018)	338	Childhood	Physical health	2	✓^{c}	✓^{t}			
Roubinov et al. (2020a)	338	Childhood	Externalizing problems	3	✓^{c}	✓^{t}	✓^{s}		
Roubinov et al. (2020b)	338	Childhood	Oppositional defiant disorder symptoms	3	✓^{c}	✓^{t}	✓^{s}		
Ruttle et al. (2011)	96	Childhood, Adolescence	Internalizing problems, externalizing problems	2	✓^{m}	✓^{t}			
Shirtcliff et al. (2007)	294	Childhood, Adolescence	Internalizing problems, externalizing problems	3	✓^{m}	✓^{t}	✓^{s}		
Sierau et al. (2017)	944	Early Childhood, Childhood	Child maltreatment	3	✓^{c}		✓^{s}		✓^{o}

Slattery & Essex (2011)	367	Childhood, Adolescence	Internalizing problems, externalizing problems	3	✓^{m}	✓^{t}	✓^{s}	
Stadelmann et al. (2007)	153	Childhood	Emotional/behavioral difficulties, prosocial behavior	3	✓^{c}	✓^{t}	✓^{s}	
Suh et al. (2016)	392	Adolescence	Internalizing problems, externalizing problems	4	✓^{m} ✓^{f}	✓^{t}	✓^{s}	
Thijssen et al. (2015)	566	Childhood	Aggressive behavior	2	✓^{c}		✓^{s}	
van't Veer et al. (2019)	21	Adulthood	Parental protection	2			✓^{s}	✓
Zaidman-Zait & Hall (2015)	1487	Early Childhood	Internalizing problems, externalizing problems	2	✓^{m} ✓^{f}			
Zhang & Jia (2011)	802	Adolescence, Adulthood	Suicidal intent	2				✓ ✓

Note. Early Childhood = 0 to 4 years; Childhood = 5 to 12 years; Adolescence = 13 to 18 years; Adulthood = 18 or older; ✓^{m} = mother; ✓^{f} = father; ✓^{c} = caregiver; ✓^{t} = teacher; ✓^{p} = peer; ✓^{b} = behavioral observation; ✓^{o} = official record; *These studies included two or more applications of the TSSM that differed in either sample size, number of informants, and/or which informants who were included in the TSSM. Other studies that include more than one application of the TSSM are summarized in a single row if these methodological features do not differ.

Table 4

Descriptive Statistics for Study 2 Caregiver, Adolescent, and Peer Confederate Survey Measures

Variable	***N***	***Mean***	***Standard Deviation***	***Skewness***	***Kurtosis***
SPS-6					
Adolescent Self-Report	134	6.19	5.59	1.25	0.91
Caregiver Report	134	4.77	5.14	1.21	0.45
Peer Confederate Report	131	8.48	5.68	0.41	-0.59
SIAS					
Adolescent Self-Report	134	28.05	16.14	0.84	0.17
Caregiver Report	134	27.04	16.54	0.77	0.02
Peer Confederate Report	132	35.55	17.51	0.09	-0.99
SPAI-C					
Adolescent Self-Report	134	17.39	10.58	0.62	-0.23
Caregiver Report	134	17.45	10.91	0.61	-0.28
SPS					
Adolescent Self-Report	134	21.42	15.41	1.28	1.47
Caregiver Report	134	16.91	14.39	1.18	0.72
Peer Confederate Report	131	25.83	16.63	0.59	-0.25
BDI-II					
Caregiver Self-Report, Raw Score	134	8.69	8.31	1.49	3.04
Caregiver Self-Report, Square Root Transformation	134	2.54	1.50	0.08	-0.38

Not. **SPS-6** = Social Phobia Scale 6-item Short Form; **SIAS** = Social Interaction Anxiety Scale; **SPAIC** = Social Phobia and Anxiety Inventory for Children; **SIAS** = Social Interaction Anxiety Scale; **SPS** = Social Phobia Scale; **BDI-II** = Beck Depression Inventory-II.

Table 5

Correlations among Caregiver, Adolescent, and Peer Confederate Reports Entered into Measurement Models in Study 2

Measure, Informant	1	2	3	4	5	6
1. **SPS-6**, Adolescent Self-Report	–	.31***	.38***	.80***	.34***	.48***
2. **SPS-6**, Caregiver		–	.01	.31***	.73***	.13
3. **SPS-6**, Peer Confederate			–	.29**	.18*	.82***
4. **SIAS**, Adolescent Self-Report				–	.39***	.42***
5. **SIAS**, Caregiver					–	.30***
6. **SIAS**, Peer Confederate						–

Note. Boxes denote correlations between informants on parallel surveys used in Study 2 measurement models. **SPS-6** = Social Phobia Scale 6-item Short Form; **SIAS** = Social Interaction Anxiety Scale; **p* < .05; ***p* < .01; ****p* < .001.

Table 6

Raw and Standardized Intercept and Factor Loading Estimates for the Unconditional Adolescent-Peer Confederate Trifactor Model (TFM) Implemented in Study 2

Raw Estimates

	Intercept		***Common Factor***		***Perspective Factor***		***Specific Factor***
Item	**P**	**A**	**P**	**A**	**P**	**A**	
4	1.65***		.46***	.90***	.86***		-.26
7	1.17***	1.02***	.22*	.82***	.65***	.26	-.35**
8	1.50***	1.06***	.40***	1.05***	.97***	-.01	.27**
15	1.22***	1.00***	.30**	.74***	.72***	.10	.39***
16	1.58***	.92***	.27*	.94***	.89***	-.12	.20
17	1.06***	.66***	.42***	.63***	.83***	-.16	.17

Standardized Estimates

	Intercept		***Common Factor***		***Perspective Factor***		***Specific Factor***[a]
Item	**P**	**A**	**P**	**A**	**P**	**A**	
4	1.30***		.37***	.69***	.68***	.67***	-.21
7	1.00***	.79***	.19*	.63***	.56***	.20	-.30**
8	1.25***	.85***	.33***	.84***	81***	-.01	-.23**
15	1.12***	.91***	.27**	.67***	.66***	.09	.35***
16	1.37***	.80***	.23**	.82***	.77***	-.10	.17
17	.99***	.72***	.39***	.68***	.77***	-.18	.16

Note. [a]All Specific Factor loadings are within .03 between informants. When different, the peer confederate's loadings are reported. **P** = Peer; **A** = Adolescent; *p < .05; **p < .01; ***p < .001.

Table 7

Raw and Standardized Intercept and Factor Loading Estimates for the Conditional Caregiver-Adolescent Trifactor Model (TFM) Implemented in Study 2

Raw Estimates							
	Intercept		***Common Factor***		***Perspective Factor***		***Specific Factor***
Item	**C**	**A**	**C**	**A**	**C**	**A**	
4	4.24		.30**	.82**	.76***		-.10
7	3.39	4.65	.16	.74**	.46***	.81**	.25*
8	4.27	4.40	.33**	.99***	.81***	.24	-.09
15	3.09	2.55	.26**	.72***	.73****	.41	-.30***
16	5.53	5.53	.33**	1.08***	.90***	-.20	.18
17	3.36	2.21	.35***	.64***	.71***	.13	-.22*

Standardized Estimates								
	Intercept		***Common Factor***		***Perspective Factor***		***Specific Factor***[a]	
Item	**C**	**A**	**C**	**A**	**C**	**A**	**C**	**A**
4	3.56	3.22	.26**	.62***	.66***	.45	-.09	
7	3.93	3.59	.18	.57**	.55***	.49	.29*	.19*
8	3.94	3.52	.31***	.80****	.78***	.15	-.08	
15	2.96	2.30	.25**	.65***	.73***	.29	-.29***	
16	4.86	4.79	.29**	.93***	.83***	-.14	.15	
17	3.32	2.31	.34***	.66***	.73***	.11	-.22*	

Note. [a]When Specific Factor loadings that are within .03 between informants, the caregiver's loadings are reported. When greater than .03, the loadings for each informant are reported. **C** = Caregiver; **A** = Adolescent; **p* < .05; ***p* < .01; ****p* < .001.

Table 9

Principal Component Analysis (PCA) of Caregiver, Adolescent, and Peer Confederate Reports on the Social Interaction Anxiety Scale (SIAS) Implemented in Study 2

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

Table 8

Raw and Standardized Intercept and Factor Loading Estimates for the Conditional 3 Informant Trifactor Model (TFM) Implemented in Study 2

Raw Estimates										
	Intercept			***Common Factor***			***Perspective Factor***			***Specific***
Item	**C**	**P**	**A**	**C**	**P**	**A**	**C**	**P**	**A**	***Factor***
4		2.92**		.10	.40*	.43*		.81***		.23*
7	1.37	2.42**	1.96*	-.02	.51***	.37*	.49***	.29**	.77***	.11
8	1.78*	2.80**	1.50**	.03	.82***	.41*	.87***	.38***	.97***	-.20**
15	2.60*	3.22*	2.47*	.01	.71***	.46***	.78***	.25**	.65***	.29***
16	1.85*	2.88**	1.21**	.01	.84***	.35**	.96***	.23*	.88***	-.22**
17	2.43*	3.15*	1.48*	.06	.80***	.18	.81***	.29**	.57***	.24***
Standardized Estimates										
	Intercept			***Common Factor***			***Perspective Factor***			***Specific***
Item	**C**	**P**	**A**	**C**	**P**	**A**	**C**	**P**	**A**	***Factor***[a]
4	2.43**	2.36**	2.21**	.08	.32*	.32**	.70***	.93***	.62***	.18*
7	1.58	2.10**	1.51*	-.02	.44***	.29*	.58***	.36**	.60***	.09
8	1.63*	2.41**	1.18**	.03	.71***	.32**	.83***	.47**	.78***	-.17**
15	2.49*	2.95*	2.21**	.01	.65***	.41***	.77***	.33*	.60***	.26***
16	1.63*	2.56**	1.03**	.01	.75***	.30**	.87***	.29	.76***	-.19**
17	2.38*	3.03*	1.63*	.06	.77***	.19	.82***	.40*	.64***	.23***

Note. [a]All specific factor loadings that are within .04. When different, the peer confederate's loadings are reported. **C** = Caregiver; **P** = Peer; **A** = Adolescent; _p_ < .05; **_p_ < .01; ***_p_ < .001.

Table 10

Hierarchical Regressions Examining the Criterion-Related Validity of Trifactor Model (TFM) Common Factor Scores in Predicting Observed Adolescent Social Anxiety Relative to Individual Informants' Reports (Study 2)

Step 1: Individual Informants' Report (Caregiver) ΔR^2 = .05, ΔF(1, 132) = 6.58*		Step 1: Individual Informants' Report (Adolescent) ΔR^2 = .22, ΔF(1, 132) = 36.88***		Step 1: Individual Informants' Report (Peer Confederate) ΔR^2 = .30, ΔF(1, 128) = 53.89***	
Variable	***β***	**Variable**	***β***	**Variable**	***β***
SPAIC	.22*	**SPAIC**	.47***	**SIAS**	.54***
Step 2: Common Factor (Adolescent-Peer TFM) n/a		Step 2: Common Factor (Adolescent-Peer TFM) ΔR^2 = .002, ΔF(1, 128) = .26		Step 2: Common Factor (Adolescent-Peer TFM) ΔR^2 = .02, ΔF(1, 127) = 2.81	
Variable	***β***	**Variable**	***β***	**Variable**	***β***
-	**-**	**SPAIC**	.42**	**SIAS**	.47***
-	**-**	**Common Factor**	.06	**Common Factor**	.14
Step 2: Common Factor (Caregiver-Adolescent TFM) ΔR^2 = .11, ΔF(1, 131) = 16.60***		Step 2: Common Factor (Caregiver-Adolescent TFM) ΔR^2 = .002, ΔF(1, 131) = .26		Step 2: Common Factor (Caregiver-Adolescent TFM) n/a	
Variable	***β***	**Variable**	***β***	**Variable**	***β***
SPAIC	.13	**SPAIC**	.42***	**-**	**-**
Common Factor	.34***	**Common Factor**	.06	**-**	**-**
Step 2: Common Factor (3 Informants) ΔR^2 = .12, ΔF(1, 128) = 18.79***		Step 2: Common Factor (3 Informants) ΔR^2 = .06, ΔF(1, 128) = 10.56**		Step 2: Common Factor (3 Informants) ΔR^2 = .001, ΔF(1, 127) = .23	
Variable	***β***	**Variable**	***β***	**Variable**	***β***
SPAIC	.19*	**SPAIC**	.40***	**SIAS**	.58***
Common Factor	.35***	**Common Factor**	.25**	**Common Factor**	-.05

Note. **SPAIC** = Social Phobia and Anxiety Inventory for Children; **SIAS** = Social Interaction Anxiety Scale; **p* < .05; ***p* < .01; ****p* < .001.

Table 11

Hierarchical Regressions Examining the Criterion-Related Validity of Trifactor Model (TFM) Common Factor Scores in Predicting Referral Status Relative to Individual Informants' Reports (Study 2)

Step 1: Individual Informants' Report (Caregiver)		Step 1: Individual Informants' Report (Adolescent)		Step 1: Individual Informants' Report (Peer Confederate)	
Variable	***OR***	**Variable**	***OR***	**Variable**	***OR***
SPAIC	1.10***	**SPAIC**	1.08***	**SIAS**	1.03*
Step 2: Common Factor (Adolescent-Peer TFM) n/a		Step 2: Common Factor (Adolescent-Peer TFM)		Step 2: Common Factor (Adolescent-Peer TFM)	
Variable	***OR***	**Variable**	***OR***	**Variable**	***OR***
-	**-**	**SPAIC**	1.06	**SIAS**	1.01
-	**-**	**Common Factor**	1.38	**Common Factor**	1.98**
Step 2: Common Factor (Caregiver-Adolescent TFM)		Step 2: Common Factor (Caregiver-Adolescent TFM)		Step 2: Common Factor (Caregiver-Adolescent TFM) n/a	
Variable	***OR***	**Variable**	***OR***	**Variable**	***OR***
SPAIC	1.09***	**SPAIC**	1.06*	**-**	**-**
Common Factor	1.99**	**Common Factor**	1.46	**-**	**-**
Step 2: Common Factor (3 Informants)		Step 2: Common Factor (3 Informants)		Step 2: Common Factor (3 Informants)	
Variable	***OR***	**Variable**	***OR***	**Variable**	***OR***
SPAIC	1.10***	**SPAIC**	1.08***	**SIAS**	1.03
Common Factor	1.00	**Common Factor**	1.15	**Common Factor**	.99

Note. **SPAIC** = Social Phobia and Anxiety Inventory for Children; **SIAS** = Social Interaction Anxiety Scale; **OR** = Odds Ratio; *p < .05; **p < .01; ***p < .001.

Table 12

Hierarchical Regressions Examining the Criterion-Related Validity of Trifactor Model (TFM) Common Factor Scores in Predicting Observed Adolescent Social Anxiety Relative to the Composite Score Approach (Study 2)

Step 1: Composite Score (Adolescent-Peer Reports) $\Delta R^2 = .34$, $\Delta F(1, 129) = 66.05$***		Step 1: Composite Score (Caregiver-Adolescent Reports) $\Delta R^2 = .23$, $\Delta F(1, 132) = 38.87$***		Step 1: Composite Score (3 Informants' Reports) $\Delta R^2 = .34$, $\Delta F(1, 129) = 66.51$***	
Variable	***β***	**Variable**	***β***	**Variable**	***β***
Composite Score	.58***	**Composite Score**	.48***	**Composite Score**	.58***
Step 2: Common Factor (Adolescent-Peer TFM) $\Delta R^2 = .01$, $\Delta F(1, 128) = 1.59$		Step 2: Common Factor (Caregiver-Adolescent TFM) $\Delta R^2 = .01$, $\Delta F(1, 131) = 1.03$		Step 2: Common Factor (3 Informants) $\Delta R^2 = .01$, $\Delta F(1, 128) = 1.30$	
Variable	***β***	**Variable**	***β***	**Variable**	***β***
Composite Score	.68***	**Composite Score**	.41***	**Composite Score**	.54***
Common Factor	-.14	**Common Factor**	.10	**Common Factor**	.09

Note. **SIAS** = Social Interaction Anxiety Scale; *$p < .05$; **$p < .01$; ***$p < .001$.

Table 13

Hierarchical Regressions Examining the Criterion-Related Validity of Trifactor Model (TFM) Common Factor Scores in Predicting Referral Status Relative to the Composite Score Approach (Study 2)

Step 1: Composite Score (Adolescent-Peer Reports)		Step 1: Composite Score (Caregiver-Adolescent Reports)		Step 1: Composite Score (3 Informants' Reports)	
Variable	***OR***	**Variable**	***OR***	**Variable**	***OR***
Composite Score	1.12***	**Composite Score**	1.06***	**Composite Score**	1.10***
Step 2: Common Factor (Adolescent-Peer TFM)		Step 2: Common Factor (Caregiver-Adolescent TFM)		Step 2: Common Factor (3 Informants)	
Variable	***OR***	**Variable**	***OR***	**Variable**	***OR***
Composite Score	1.21***	**Composite Score**	1.04	**Composite Score**	1.12***
Common Factor	.95	**Common Factor**	1.41	**Common Factor**	.68

Note. **SIAS** = Social Interaction Anxiety Scale; **OR** = Odds Ratio; **p* < .05; ***p* < .01; ****p* < .001.

Table 14

Hierarchical Regressions Examining the Criterion-Related Validity of the Trait Score Satellite Model (TSSM) Trait Score in Predicting Observed Adolescent Social Anxiety Relative to Individual Informants' Reports and the Composite Score Approach (Study 2)

Incremental Validity Relative to Individual Informants' Reports

Step 1: Individual Informants' Report (Caregiver) ΔR^2 = .06, ΔF(1, 130) = 7.73**		Step 1: Individual Informants' Report (Adolescent) ΔR^2 = .22, ΔF(1, 130) = 35.70***		Step 1: Individual Informants' Report (Peer Confederate) ΔR^2 = .24, ΔF(1, 128) = 39.22***	
Variable	***β***	**Variable**	***β***	**Variable**	***β***
SPAIC	.24**	**SPAIC**	.46***	**SPS**	.48***
Step 2: Trait Score ΔR^2 = .32, ΔF(1, 129) = 65.59***		Step 2: Trait Score ΔR^2 = .14, ΔF(1, 129) = 27.88***		Step 2: Trait Score ΔR^2 = .13, ΔF(1, 127) = 26.63***	
Variable	***β***	**Variable**	***β***	**Variable**	***β***
SPAIC	-.18*	**SPAIC**	.05	**SPS**	.18*
Trait Score	.70***	**Trait Score**	.56***	**Trait Score**	.47***

Incremental Validity Relative to the Composite Score Approach

Step 1: Composite Score ΔR^2 = .28, ΔF(1, 130) = 50.50***	
Variable	***β***
Composite Score	.53***
Step 2: Trait Score ΔR^2 = .07, ΔF (1, 129) = 14.80***	
Variable	***β***
Composite Score	.04
Trait Score	.56***

Note. **SPAIC** = Social Phobia and Anxiety Inventory for Children; **SPS** = Social Phobia Scale; *p < .05; **p < .01; ***p < .001.

Table 15

Hierarchical Regressions Examining the Criterion-Related Validity of the Trait Score Satellite Model (TSSM) Trait Score in Predicting Referral Status Relative to Individual Informants' Reports and the Composite Score Approach (Study 2)

<table>
<tr><th colspan="6">Incremental Validity Relative to Individual Informants' Reports</th></tr>
<tr><td colspan="2">Step 1: Individual Informants' Report (Caregiver)</td><td colspan="2">Step 1: Individual Informants' Report (Adolescent)</td><td colspan="2">Step 1: Individual Informants' Report (Unfamiliar Peer)</td></tr>
<tr><td><b>Variable</b></td><td><b><i>OR</i></b></td><td><b>Variable</b></td><td><b><i>OR</i></b></td><td><b>Variable</b></td><td><b><i>OR</i></b></td></tr>
<tr><td><b>SPAIC</b></td><td>1.11***</td><td><b>SPAIC</b></td><td>1.08***</td><td><b>SPS</b></td><td>1.02</td></tr>
<tr><td colspan="2">Step 2: Trait Score</td><td colspan="2">Step 2: Trait Score</td><td colspan="2">Step 2: Trait Score</td></tr>
<tr><td><b>Variable</b></td><td><b><i>OR</i></b></td><td><b>Variable</b></td><td><b><i>OR</i></b></td><td><b>Variable</b></td><td><b><i>OR</i></b></td></tr>
<tr><td><b>SPAIC</b></td><td>1.06</td><td><b>SPAIC</b></td><td>1.00</td><td><b>SPS</b></td><td>0.95</td></tr>
<tr><td><b>Trait Score</b></td><td>2.64**</td><td><b>Trait Score</b></td><td>3.50***</td><td><b>Trait Score</b></td><td>6.83***</td></tr>
<tr><th colspan="6">Incremental Validity Relative to the Composite Score Approach</th></tr>
<tr><td colspan="6">Step 1: Composite Score</td></tr>
<tr><td colspan="3"><b>Variable</b></td><td colspan="3"><b><i>OR</i></b></td></tr>
<tr><td colspan="3"><b>Composite Score</b></td><td colspan="3">1.09***</td></tr>
<tr><td colspan="6">Step 2: Trait Score</td></tr>
<tr><td colspan="3"><b>Variable</b></td><td colspan="3"><b><i>OR</i></b></td></tr>
<tr><td colspan="3"><b>Composite Score</b></td><td colspan="3">.97</td></tr>
<tr><td colspan="3"><b>Trait Score</b></td><td colspan="3">5.18**</td></tr>
</table>

Note. **SPAIC** = Social Phobia and Anxiety Inventory for Children; **SPS** = Social Phobia Scale; **OR** = Odds Ratio; *$p < .05$; **$p < .01$; ***$p < .001$.

Figures

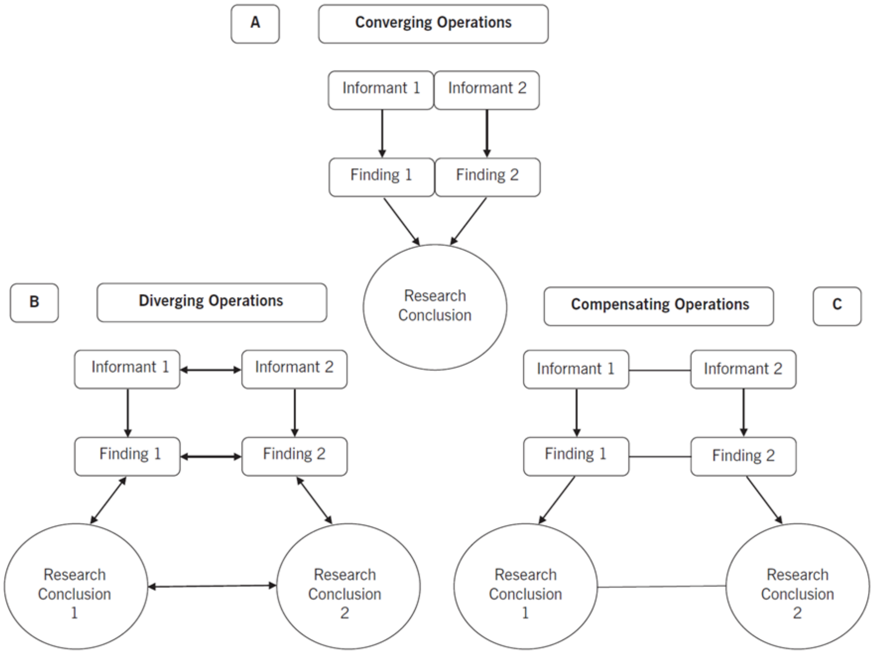

Figure 1. Graphical representation of the research concepts that comprise the Operations Triad Model. The top half (A) represents Converging Operations: a set of measurement conditions for interpreting patterns of findings based on the consistency within which findings yield similar conclusions. The bottom half denotes two circumstances within which researchers identify discrepancies across empirical findings derived from multiple informants' reports and thus discrepancies in the research conclusions drawn from these reports. On the left (B) is a graphical representation of Diverging Operations: a set of measurement conditions for interpreting patterns of inconsistent findings based on hypotheses about variations in the behavior(s) assessed. The solid lines linking informants' reports, empirical findings derived from these reports, and conclusions based on empirical findings denote the systematic relations among these three study components. Further, the presence of dual arrowheads in the figure representing Diverging Operations conveys the idea that one ties meaning to the discrepancies among empirical findings and research conclusions and thus how one interprets informants' reports to vary as a function of variation in the behaviors being assessed. Lastly, on the right (C) is a graphical representation of Compensating Operations: a set of measurement conditions for interpreting patterns of inconsistent findings based on methodological features of the study's measures or informants. The dashed lines denote the lack of systematic relations among informants' reports, empirical findings, and research conclusions. Originally published in De Los Reyes, Thomas, et al. (2013). © Annual Review of Clinical Psychology. Copyright 2012 Annual Reviews. All rights reserved. The Annual Reviews logo, and other Annual Reviews products referenced herein are

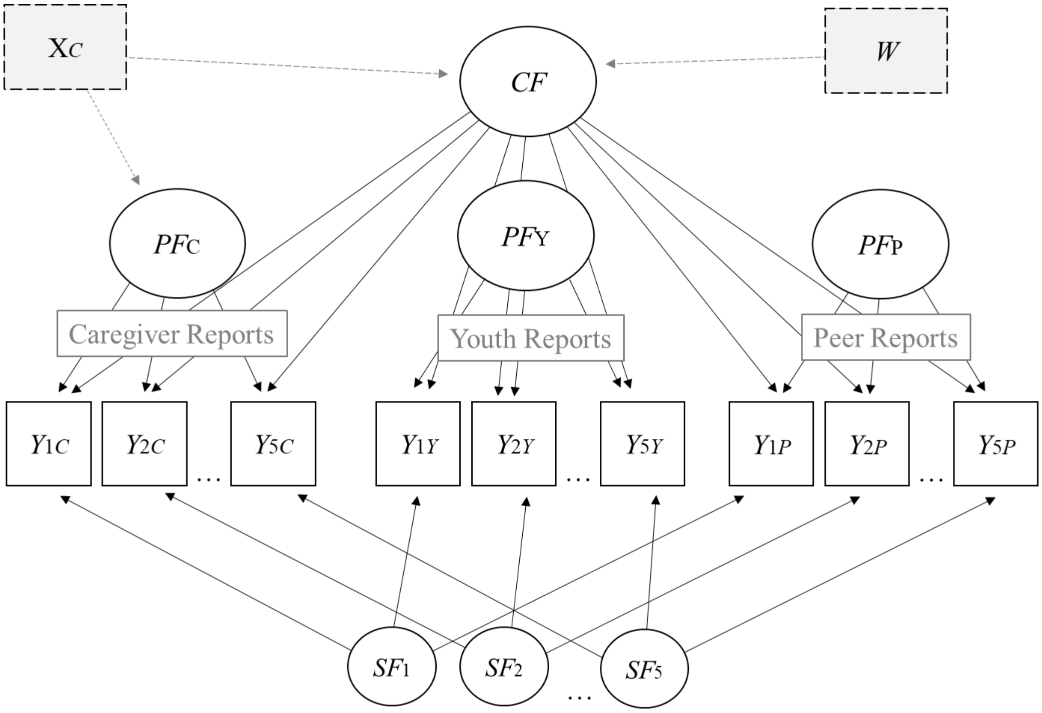

Figure 2. Trifactor Model (TFM) for parallel multi-informant reports on a 5-item measure. Observed item ratings are numbered by item. *C*, *Y*, and *P* subscripts are for caregiver, youth, and peer confederate ratings, respectively. The Common Factor (*CF*), Perspective Factors (*PFs*), and Specific Factors (*SFs*) are modeled from observed item ratings. Dashed lines indicate additional model inputs in the Conditional TFM. Informant-specific predictors are denoted with an *X* and are loaded onto the relevant PF as well as the CF. Target-level predictors are denoted with a *W* and are loaded onto the CF. Figure adapted from Bauer et al. (2013).

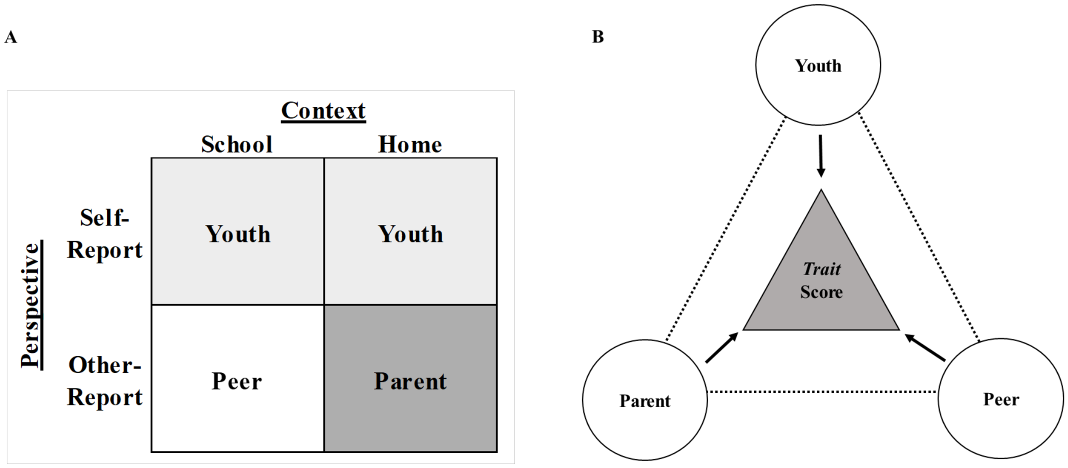

Figure 3. Panel A depicts an example of the "mix-and-match" criterion to identify optimal informants to include in a multi-informant assessment. Informants systematically vary in the perspective and context from which they rate youth mental health symptoms, with the goal of effectively triangulating on a Trait score. Panel B provides a graphical depiction of multi-informant reports triangulating, much like the global positioning system (GPS), to identify the Trait score. Both peer- and parent-reports provide information from an other-perspective, with peers providing information about the school context and parents providing information about the home context. Youth reports provide the self-perspective and information about both the school and home contexts. Figures adapted from Kraemer et al. (2003).

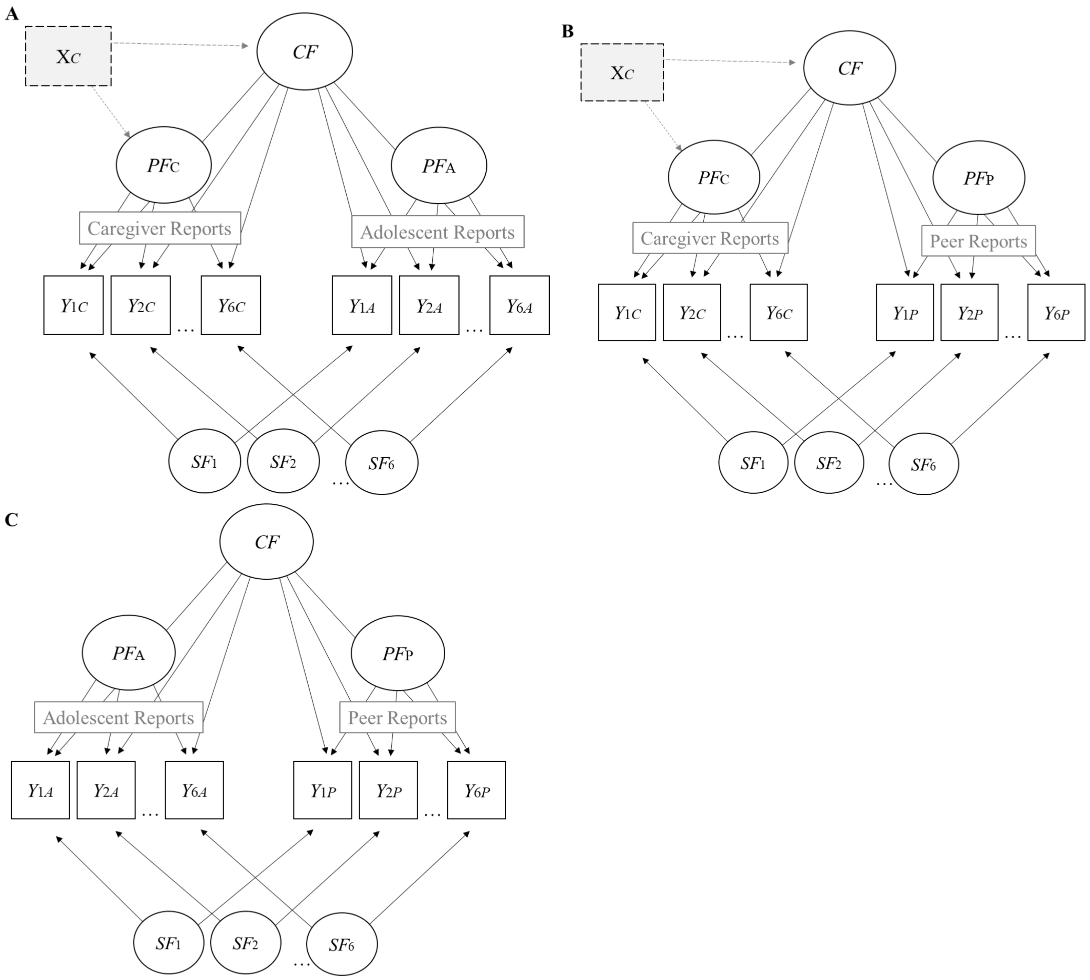

Figure 4. Unconditional and Conditional Trifactor Models (TFMs) using multi-informant reports of adolescent social anxiety on the 6-item short form of the Social Phobia Scale (SPS-6). Panel A represents the caregiver-adolescent report model, Panel B represents the caregiver-peer report model, and Panel C represents the adolescent-peer report model. Observed item ratings are numbered by item. *C*, *A*, and *P* subscripts are for caregiver, adolescent, and peer confederate ratings, respectively. The Common Factor (*CF*), Perspective Factors (*PFs*), and Specific Factors (*SFs*) are modeled from observed item ratings. Dashed lines indicate additional model inputs in the Conditional TFMs (Panel A and Panel B). The informant-specific predictor, caregiver self-reported depressive symptoms, is denoted with an X_C and is loaded onto the caregiver PF as well as the CF. Figures adapted from Bauer et al. (2013).

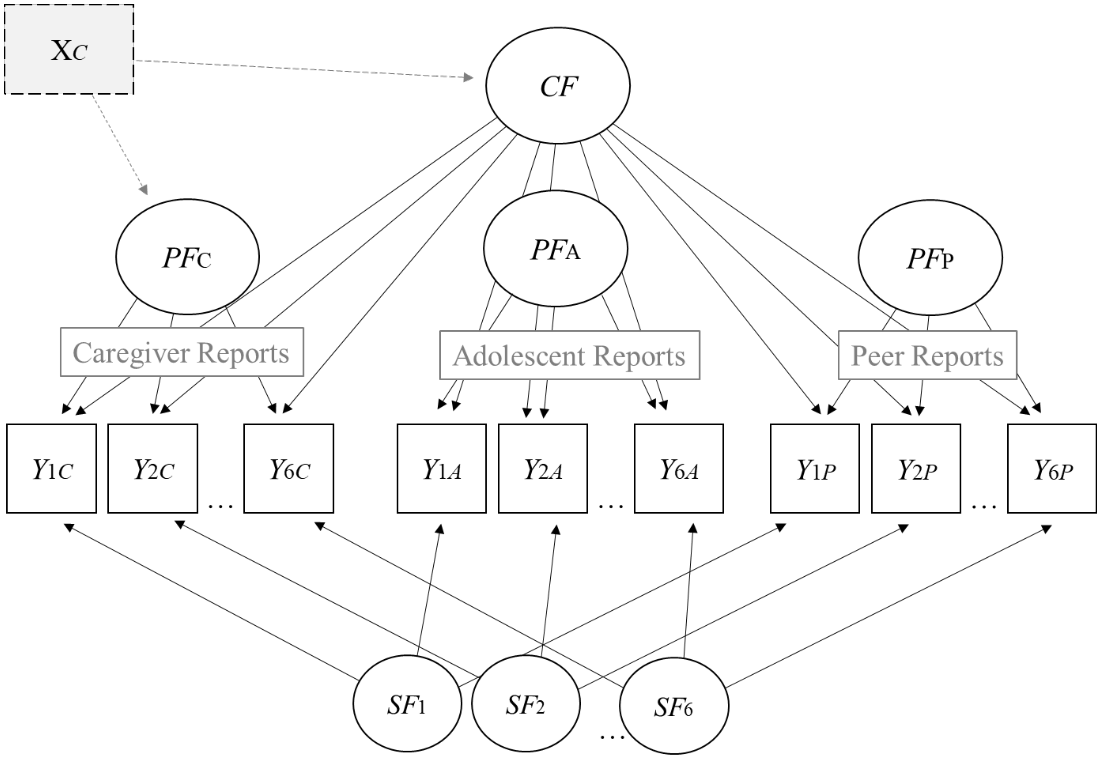

Figure 5. Conditional Trifactor Model (TFM) using caregiver, adolescent, and unfamiliar peer reports of adolescent social anxiety on the 6-item short form of the Social Phobia Scale (SPS-6). Observed item ratings are numbered by item. *C*, *A*, and *P* subscripts are for caregiver, adolescent, and unfamiliar peer ratings, respectively. The Common Factor (*CF*), Perspective Factors (*PFs*), and Specific Factors (*SFs*) are modeled from observed item ratings. The informant-specific predictor, caregiver self-reported depressive symptoms, is denoted with an X_C and is loaded onto the caregiver PF as well as the CF. Figures adapted from Bauer et al. (2013).

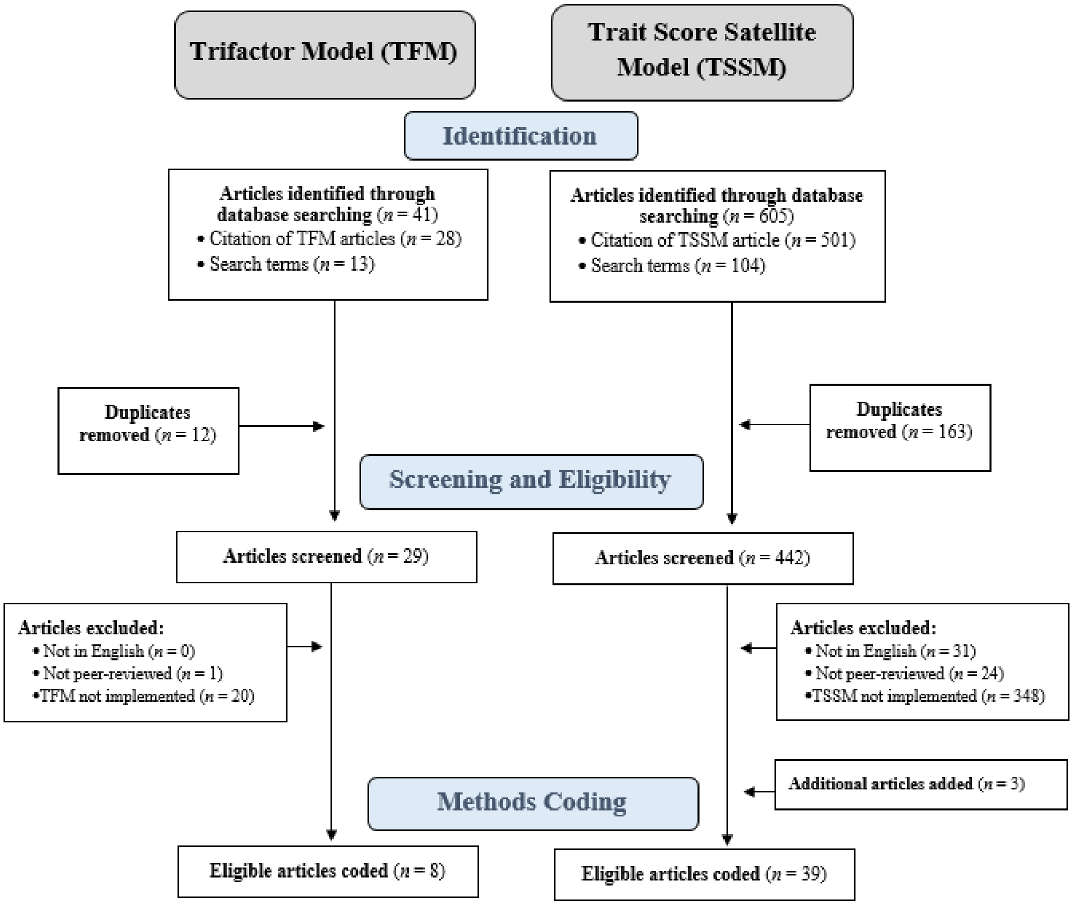
Trifactor Model (TFM)
Trait Score Satellite Model (TSSM)
Identification
Articles identified through database searching (n = 41)
• Citation of TFM articles (n = 28)
• Search terms (n = 13)
Articles identified through database searching (n = 605)
• Citation of TSSM article (n = 501)
• Search terms (n = 104)
Duplicates removed (n = 12)
Duplicates removed (n = 163)
Screening and Eligibility
Articles screened (n = 29)
Articles screened (n = 442)
Articles excluded:
• Not in English (n = 0)
• Not peer-reviewed (n = 1)
•TFM not implemented (n = 20)
Articles excluded:
• Not in English (n = 31)
• Not peer-reviewed (n = 24)
•TSSM not implemented (n = 348)
Additional articles added (n = 3)
Methods Coding
Eligible articles coded (n = 8)
Eligible articles coded (n = 39)

Figure 6. Flowchart of studies included in the Study 1 systematic review.

References

** = Study 1 Systematic Review References Using Either the TFM or TSSM*

Ablow, J. C., Measelle, J. R., Kraemer, H. C., Harrington, R., Luby, J., Smider, N., ... & Essex, M. J. (1999). The MacArthur Three-City Outcome Study: Evaluating multi-informant measures of young children's symptomatology. *Journal of the American Academy of Child and Adolescent Psychiatry*, *38*(12), 1580-1590.

Achenbach, T.M. (2005). Advancing assessment of children and adolescents: Commentary on evidence-based assessment of child and adolescent disorders, *Journal of Clinical Child and Adolescent Psychology*, *34*(3), 541-547, doi: 10.1207/ s15374424jccp3403_9

Achenbach, T. M. (2001). *Child behavior checklist for ages 6 to 18.* Burlington: University of Vermont, Research Center for Children, Youth, and Families.

Achenbach, T. M. (2017). Future directions for clinical research, services, and training: Evidence-based assessment across informants, cultures, and dimensional hierarchies. *Journal of Clinical Child & Adolescent Psychology*, *46*(1), 159-169. doi: 10.1080/15374416.2016.1220315

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*, 213-232. doi: 10.1037/0033-2909.101.2.213

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.

Affrunti, N. W., & Woodruff-Borden, J. (2015). The effect of maternal psychopathology on parent–child agreement of child anxiety symptoms: A hierarchical linear modeling approach. *Journal of Anxiety Disorders*, *32*, 56-65. doi: 10.1016/j.janxdis.2015.03.010

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (*5th ed.*) (*DSM-V*). American Psychiatric Association.

Anderson, E. R., & Hope, D. A. (2009). The relationship among social phobia, objective and perceived physiological reactivity, and anxiety sensitivity in an adolescent population. *Journal of Anxiety Disorders*, *23*, 18-26. doi: 10. 1016/j.janxdis.2008.03.011

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics*, *34*, 502-508. doi: 10.1111/opo.12131

*Armstrong, J. M., Ruttle, P. L., Klein, M. H., Essex, M. J., & Benca, R. M. (2014). Associations of child insomnia, sleep movement, and their persistence with mental health symptoms in childhood and adolescence. *Sleep*, *37*(5), 901-909. doi: 10.5665/sleep.3656

Azad, G. F., Reisinger, E., Xie, M., & Mandell, M. (2016). Parent and teacher concordance on the Social Responsiveness Scale for children with autism. *School Mental Health*, *8*, 368-376. doi: 10.1007/s12310-015-9168-6

Barbot, B., Bick, J., Bentley, M. J., Balestracci, K. M., Woolston, J. L., Adnopoz, J. A., & Grigorenko, E. L. (2016). Changes in mental health outcomes with the intensive in-home child and adolescent psychiatric service: A multi-informant, latent consensus approach. *International Journal of Methods in Psychiatric Research*, *25*(1), 33-43. doi: 10.1002/mpr.1477

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology*, *3*, 77-85.

*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. doi: 10.1037/a0032475

Beauchaine, T. P., & Hinshaw, S. P. (2020). RDoC and psychopathology among youth: Misplaced assumptions and an agenda for future research. *Journal of Clinical Child & Adolescent Psychology*, *49*(3), 322-340. doi: 10.1080/15374416.2020.1750022

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). Beck Depression Inventory—Second Edition manual. The Psychological Corporation.

Beidel, D. S., Turner, S. M., Hamlin, K., & Morris, T. L. (2000a). The Social Phobia and Anxiety Inventory for Children (SPAI-C): External and discriminative validity. *Behavior Therapy*, *31*, 75-87. doi:10.1016/S0005-7894(00)80005-2.

Beidel, D. C., Turner, S. M., & Morris, T. L. (2000b). Behavioral treatment of childhood social phobia. *Journal of Consulting and Clinical Psychology*, *68*(6), 1072-1080. doi: 10.1037/0022-006X.68.6.1072

Beidel, D. C., Rao, P. A., Scharfstein, L., Wong, N., & Alfano, C. A. (2010). Social skills and social phobia: An investigation of DSM-IV subtypes. *Behaviour Research and Therapy*, *48*(10), 992-1001. doi: 10.1016/j.brat.2010.06.005

Beidas, R. S., Stewart, R. E., Walsh, L., Lucas, S., Downey, M. M., Jackson, K., ... & Mandell, D. S. (2015). Free, brief, and validated: Standardized instruments for low-resource mental health settings. *Cognitive and Behavioral Practice*, *22*(1), 5-19. doi: 10.1016/j.cbpra.2014.02.002

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238-246. doi: 10.1037/0033-2909.107.2.238

Borsboom, D. (2005) *Measuring the Mind.* Cambridge University Press.

Boyle, M. H., & Pickles, A. (1997). Maternal depressive symptoms and ratings of emotional disorder symptoms in children and adolescents. *Journal of Child Psychology and Psychiatry*, *38*(8), 981-992. doi: 10.1111/j.1469-7610.1997.tb01615.x

Briggs-Gowan, M. J., Carter, A. S., & Schwab-Stone, M. (1996). Discrepancies among mother, child, and teacher reports: Examining the contributions of maternal depression and anxiety. *Journal of Abnormal Child Psychology*, *24*(6), 749-765. doi: 10.1007/BF01664738

Brown-Jacobsen, A. M., Wallace, D. P., & Whiteside, S. P. H. (2011). Multimethod, multi-informant agreement, and positive predictive value in the identification of child anxiety disorders using the SCAS and ADIS-C. *Assessment*, 18, 382-392. doi: 10.1177/1073191110375792

*Buisman, R. S., Pittner, K., Tollenaar, M. S., Lindenberg, J., van den Berg, L. J., Compier-de Block, L. H., ... & van IJzendoorn, M. H. (2020). Intergenerational transmission of child maltreatment using a multi-informant multi-generation family design. *PloS one*, *15*(3), e0225839. doi: 10.1371/journal.pone.0225839

*Burk, L. R., Armstrong, J. M., Park, J.-H., Zahn-Waxler, C., Klein, M. H., & Essex, M. J. (2011). Stability of early identified aggressive victim status in elementary school and associations with later mental health problems and functional impairments. *Journal of Abnormal Child Psychology*, *39*(2), 225-238. doi: 10.1007/s10802-010-9454-6

*Burk, L. R., Park, J. H., Armstrong, J. M., Klein, M. H., Goldsmith, H. H., Zahn-Waxler, C., & Essex, M. J. (2008). Identification of early child and family risk factors for aggressive victim status in first grade. *Journal of Abnormal Child Psychology*, *36*(4), 513-526. doi: 10.1007/s10802-007-9196-2

Byrne, B. M. (2013). Structural equation modeling with Mplus: Basic concepts, applications, and programming. Routledge.

*Caldwell, J. Z., Armstrong, J. M., Hanson, J. L., Sutterer, M. J., Stodola, D. E., Koenigs, M., ... & Davidson, R. J. (2015). Preschool externalizing behavior predicts gender-specific variation in adolescent neural structure. *PloS one*, *10*(2), e0117453.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105. doi: 10.1037/h0046016

Carleton, R. N., Thibodeau, M. A., Weeks, J. W., Teale Sapach, M. J. N., McEvoy, P. M., Horswill, S. C., & Heimberg, R. G. (2014). Comparing short forms of the Social Interaction Anxiety Scale and the Social Phobia Scale. *Psychological Assessment*, *26*(4), 1116-1126. doi: 10.1037/a0037063

Carlone, C., & Milan, S. (2021). Maternal depression and child externalizing behaviors: The role of attachment across development in low-income families. *Research on Child and Adolescent Psychopathology*, *49*, 603-614. doi: 10.1007/s10802-020-00747-z

Casey, B. J., Oliveri, M. E., & Insel, T. (2014). A neurodevelopmental perspective on the research domain criteria (RDoC) framework. *Biological Psychiatry*, *76*(5), 350-353. doi: 10.1016/j.biopsych.2014.01.006

Caspi, A., Moffitt, T. E., Morgan, J., Rutter, M., Taylor, A., Arseneault, L., ... & Polo-Tomas, M. (2004). Maternal expressed emotion predicts children's antisocial behavior problems: Using monozygotic-twin differences to identify environmental effects on behavioral development. *Developmental Psychology*, *40*(2), 149-161.

Chen, Y. Y., Ho, S. Y., Lee, P. C., Wu, C. K., & Gau, S. S. F. (2017). Parent-child discrepancies in the report of adolescent emotional and behavioral problems in Taiwan. *PLoS One*, *12*(6), e0178863. doi: 10.1371/journal.pone.0178863

*Chen, X., Wang, Y., Li, F., Gong, J., & Yan, Y. (2015). Development and evaluation of the Brief Sexual Openness Scale—A construal level theory based approach. *PloS one*, *10*(8), e0136683. doi: 10.1371/journal.pone.0136683

Cheung, K., Aberdeen, K., Ward, M. A., & Theule, J. (2018). Maternal depression in families of children with ADHD: A meta-analysis. *Journal of Child and Family Studies*, *27*(4), 1015-1028. doi: 10.1007/s10826-018-1017-4

Cheung, K., & Theule, J. (2019). Paternal depression and child externalizing behaviors: A meta-analysis. *Journal of Family Psychology*, *33*(1), 98-108. doi: 10.1037/fam0000473

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284-290. doi: 10.1037/1040-3590.6.4.284.

*Clark, D. A., Durbin, C. E., Donnellan, M. B., & Neppl, T. K. (2017). Internalizing symptoms and personality traits color parental reports of child temperament. *Journal of Personality*, *85*(6), 852-866. doi: 10.1111/jopy.12293

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

Cooper, P. J., Fearn, V., Willetts, L., Seabrook, H., & Parkinson, M. (2006). Affective disorder in the parents of a clinic sample of children with anxiety disorders. *Journal of Affective Disorders*, *93*(1-3), 205-212. doi: 10.1016/j.jad.2006.03.017

Curran, J. P. (1982). A procedure for the assessment of social skills: The Simulated Social Interaction Test. In J.P. Curran & P.M. Monti (Eds.), *Social skills training: A practical handbook for assessment and treatment* (pp. 348–373). New York, NY: Guilford.

*Curran, P. J., Georgeson, A. R., Bauer, D. J., & Hussong, A. M. (2020). Psychometric models for scoring multiple reporter assessments: Applications to integrative data analysis in prevention science and beyond. *International Journal of Behavioral Development*, *35*, 40-50. doi: 10.1177/0165025419896620

Dhillon, S., Bagby, R. M., Kushner, S. C., & Burchett, D. (2017). The impact of underreporting and overreporting on the validity of the Personality Inventory for DSM-5 (PID-5): A simulation analog design investigation. *Psychological Assessment*, *29*(4), 473-478. doi: 10.1037/pas0000359.

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. doi: 10.1080/15374416.2011.533405

De Los Reyes, A., Augenstein, T. M., & Aldao, A. (2017). *Assessment issues in child and adolescent psychotherapy.* In J. R. Weisz & A. E. Kazdin (Eds.), *Evidence-based psychotherapies for children and adolescents* (p. 537-554). The Guilford Press.

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. doi: 10.1037/a0038498

De Los Reyes, A., Cook, C. R., Gresham, F. M., Makol, B. A., & Wang, M. (2019). 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. doi: 10.1016/j.jsp.2019.05.005

De Los Reyes, A., Kundey, S. M., & Wang, M. (2011). The end of the primary outcome measure: A research agenda for constructing its replacement. *Clinical Psychology Review*, *31*(5), 829-838. doi: 10.1016/j.cpr.2011.03.011

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. doi: 10.1007/s10802-009-9307-3

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin*, *131*(4), 483-509. doi: 10.1037/003.3-2909.131.4.483

De Los Reyes, A., & Kazdin, A. E. (2009). Identifying evidence-based interventions for children and adolescents using the range of possible changes model: A meta-analytic illustration. *Behavior Modification*, *33*(5), 583-617. doi: 10.1177/0145445509343203

De Los Reyes, A., Goodman, K. L., Kliewer, W., & Reid-Quiñones, K. (2010). The longitudinal consistency of mother-child reporting discrepancies of parental monitoring and their ability to predict child delinquent behaviors two years later. *Journal of Youth and Adolescence, 39,* 1417-1430. doi: 10.1007/s10964-009-9496-7

De Los Reyes, A., Lerner, M. D., Keeley, L. M., Weber, R. J., Drabick, D. A., Rabinowitz, J., & Goodman, K. L. (2019). Improving interpretability of subjective assessments about psychological phenomena: a review and cross-cultural meta-analysis. *Review of General Psychology*, *23*(3), 293-319. doi: 10.1177/1089268019837645

De Los Reyes, A., Thomas, S. A., Goodman, K. L., & Kundey, S. M. (2013). Principles underlying the use of multiple informants' reports. *Annual Review of Clinical Psychology*, *9*, 123-149. doi: 10.1146/annurev-clinpsy-050212-185617

*De Pauw, S. S., Mervielde, I., De Clercq, B. J., De Fruyt, F., Tremmery, S., & Deboutte, D. (2009). Personality symptoms and self-esteem as correlates of psychopathology in child psychiatric patients: Evaluating multiple informant data. *Child Psychiatry and Human Development*, *40*(4), 499-515. doi: 10.1007/s10578-009-0140-2

Deros, D. E., Racz, S. J., Lipton, M. F., Augenstein, T. M., Karp, J. N., Keeley, L. M., ... & De Los Reyes, A. (2018). Multi-informant assessments of adolescent social anxiety: Adding clarity by leveraging reports from unfamiliar peer confederates. *Behavior Therapy*, *49*(1), 84-98. doi: 10.1016/j.beth.2017.05.001

DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment*, *23*(3), 225-241. doi: 10.1177/073428290502300303

Drabick, D. A., Gadow, K. D., & Loney, J. (2007). Source-specific oppositional defiant disorder: Comorbidity and risk factors in referred elementary schoolboys. *Journal of the American Academy of Child & Adolescent Psychiatry*, *46*(1), 92-101. doi: 01.chi.0000242245.00174.90

Drabick, D. A. G., & Kendall, P. C. (2010). Developmental psychopathology and the diagnosis of mental health problems among youth. *Clinical Psychology: Science and Practice*, *17*(4), 272-280. doi: 10.1111/j.1468-2850.2010.01219.x

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. doi: 10.1093/clipsy.7.4.435

Dunteman, G. H. (1989). *Principal components analysis (No. 69).* Sage.

Edgeworth, F.Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society*, *51*, 598-635.

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. doi: 10.1037/a0013219

*El-Sheikh, M., Hinnant, J. B., & Erath, S. (2011). Developmental trajectories of delinquency symptoms in childhood: The role of marital conflict and autonomic nervous system activity. *Journal of Abnormal Psychology*, *120*(1), 16-32. doi: 10.1037/a0020626

*Erath, S. A., El-Sheikh, M., Hinnant, J. B., & Cummings, E. M. (2011). Skin conductance level reactivity moderates the association between harsh parenting and growth in child externalizing behavior. *Developmental Psychology*, *47*(3), 693. doi: 10.1037/a0021909

*Essex, M. J., Armstrong, J. M., Burk, L. R., Goldsmith, H. H., & Boyce, W. T. (2011). Biological sensitivity to context moderates the effects of the early teacher–child relationship on the development of mental health by adolescence. *Development and Psychopathology*, *23*(1), 149-161. doi: 10.1017/S0954579410000702

Essex, M. J., Klein, M. H., Miech, R., & Smider, N. A. (2001). Timing of initial exposure to maternal major depression and children's mental health symptoms in kindergarten. *The British Journal of Psychiatry*, *179*(2), 151-156.

*Essex, M. J., Klein, M. H., Slattery, M. J., Goldsmith, H. H., & Kalin, N. H. (2010). Early risk factors and developmental pathways to chronic high inhibition and social anxiety disorder in adolescence. *American Journal of Psychiatry*, *167*(1), 40-46. doi: 10.1176/appi.ajp.2009.07010051

Etkin, R. G., Lebowitz, E. R., & Silverman, W. K. (2021a). Using evaluative criteria to review youth anxiety measures, Part II: Parent-report. *Journal of Clinical Child & Adolescent Psychology*, *50*(2), 155-176. doi: 10.1080/15374416.2021.1878898

Etkin, R. G., Shimshoni, Y., Lebowitz, E. R., & Silverman, W. K. (2021b). Using evaluative criteria to review youth anxiety measures, Part I: Self-report. *Journal of Clinical Child & Adolescent Psychology*, 1-20. doi: 10.1080/15374416.2020.1802736

Eyberg, S. M., & Funderburk, B. (2011). Parent-child interaction therapy protocol. Gainesville, FL: PCIT International.

Fabrigar, L. R., & Wegener, D. T. (2011). Exploratory factor analysis. Oxford University Press.

Fadus, M. C., Ginsburg, K. R., Sobowale, K., Halliday-Boykins, C. A., Bryant, B. E., Gray, K. M., & Squeglia, L. M. (2020). Unconscious bias and the diagnosis of disruptive behavior disorders and ADHD in African American and Hispanic youth. *Academic Psychiatry*, *44*(1), 95-102. doi: 10.1007/s40596-019-01127-6

Fergus, T. A., Valentiner, D. P., Kim, H. S., & McGrath, P. B. (2014). The Social Interaction Anxiety Scale (SIAS) and the Social Phobia Scale (SPS): A comparison of two short-form versions. *Psychological Assessment*, *26*(4), 1281-1291. doi: 10.1037/a0037313

Fergusson, D. M., Boden, J. M., & Horwood, L. J. (2009). Situational and generalised conduct problems and later life outcomes: evidence from a New Zealand birth cohort. *Journal of Child Psychology and Psychiatry*, *50*(9), 1084-1092. doi: 10.1111/j.1469-7610.2009.02070.x

Fergusson, D. M., Lynskey, M. T., & Horwood, L. J. (1993). The effect of maternal depression on maternal ratings of child behavior. *Journal of Abnormal Child Psychology*, *21*(3), 245-269. doi: 10.1007/BF00917534

Garb, H. N. (2003). Incremental validity and the assessment of psychopathology in adults. *Psychological Assessment*, *15*, 508-520. doi: 10.1037/1040-3590.15.4.508

*Gardner, T. W., Dishion, T. J., & Connell, A. M. (2008). Adolescent self-regulation as resilience: Resistance to antisocial behavior within the deviant peer context. *Journal of Abnormal Child Psychology*, *36*(2), 273-284. doi: 10.1007/s10802-007-9176-6

Gartstein, M. A., Bridgett, D. J., Dishion, T. J., & Kaufman, N. K. (2009). Depressed mood and maternal report of child behavior problems: Another look at the depression–distortion hypothesis. *Journal of Applied Developmental Psychology*, *30*(2), 149-160. doi: 10.1016/j.appdev.2008.12.001

*Goelman, H., Zdaniuk, B., Boyce, W. T., Armstrong, J. M., & Essex, M. J. (2014). Maternal mental health, child care quality, and children's behavior. *Journal of Applied Developmental Psychology*, *35*(4), 347-356. doi: 10.1016/j.appdev.2014.05.003

Gonzales, N. A., Coxe, S., Roosa, M. W., White, R. M., Knight, G. P., Zeiders, K. H., & Saenz, D. (2011). Economic hardship, neighborhood context, and parenting: Prospective effects on Mexican–American adolescent's mental health. *American Journal of Community Psychology*, *47*(1-2), 98-113. doi: 10.1007/s10464-010-9366-1

Goodman, S. H., & Gotlib, I. H. (1999). Risk for psychopathology in the children of depressed mothers: A developmental model for understanding mechanisms of transmission. *Psychological Review*, *106*(3), 458-490.

Gould, J. W., Rappaport, S. R., & Flens, J. R. (2018). Use of psychological tests in child custody evaluations: Effects of validity scale scores on evaluator confidence in interpreting clinical scales. In R. Rogers & S. D. Bender (Eds.), *Clinical assessment of malingering and deception* (pp. 497-513). The Guilford Press.

Gravetter, F. J., & Wallnau, L.B. (2013). Statistics for the behavioral sciences (9th ed.). Belmont: Wadsworth Cengage Learning.

Gresham, F. M., Elliott, S. N., Metallo, S., Byrd, S., Wilson, E., & Cassidy, K. (2018). Cross-informant agreement of children's social-emotional skills: An investigation of ratings by teachers, parents, and students from a nationally representative sample. *Psychology in the Schools*, *55*(2), 208-223. doi: 10.1002/pits.22101

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19-30. doi: 10.1037/1040-3590.12.1.19

*Haeny, A. M., Littlefield, A. K., Wood, P. K., & Sher, K. J. (2018). Method effects of the relation between family history of alcoholism and parent reports of offspring impulsive behavior. *Addictive Behaviors*, *87*, 251-259. doi: 10.1016/j.addbeh.2018.07.022

*Hatzinger, M., Brand, S., Perren, S., von Wyl, A., von Klitzing, K., & Holsboer-Trachsler, E. (2007). Hypothalamic-pituitary-adrenocortical (HPA) activity in kindergarten children: Importance of gender and associations with behavioral/emotional. *Journal of Psychiatric Research*, *41*(10), 861-870. doi: 10.1016/j.jpsychires.2006.07.012

*Hatzinger, M., Brand, S., Perren, S., Stadelmann, S., von Wyl, A., von Klitzing, K., & Holsboer-Trachsler, E. (2010). Sleep actigraphy pattern and behavioral/emotional difficulties in kindergarten children: Association with hypothalamic-pituitary-adrenocortical (HPA) activity. *Journal of Psychiatric Research*, *44*(4), 253-261. doi: 10.1016/j.jpsychires.2009.08.012

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. doi: 10.1037/0022-006X.71.1.62

Hou, Y., Benner, A. D., Kim, S. Y., Chen, S., Spitz, S., Shi, Y., & Beretvas, T. (2019). Discordance in parents' and adolescents' reports of parenting: A meta-analysis and qualitative review. *American Psychologist, 75*(3), 329-348. doi: 10.1037/amp0000463

*Houts, R. M., Caspi, A., Pianta, R. C., Arseneault, L., & Moffitt, T. E. (2010). The challenging pupil in the classroom: The effect of the child on the teacher. *Psychological Science*, *21*(12), 1802-1810. doi: 10.1177/0956797610388047

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. doi: 10.1037/pas0000737

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

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment*, *15*(4), 446-455. doi: 10.1037/1040-3590.15.4.446

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. doi: 10.1080/10705519909540118.

Jensen-Doss, A., Douglas, S., Phillips, D. A., Gencdur, O., Zalman, A., & Gomez, N. E. (2020). Measurement-based care as a practice improvement tool: Clinical and organizational applications in youth mental health. *Evidence-Based Practice in Child and Adolescent Mental Health*, *5*(3), 233-250. doi: 10.1080/23794925.2020.1784062

Jones, J. D., Boyd, R. C., Calkins, M. E., Ahmed, A., Moore, T. M., Barzilay, R., ... Gur, R. E. (2019). Parent-adolescent agreement about adolescents' suicidal thoughts. *Pediatrics*, *143*, 1-10. doi: 10.1542/peds.2018-1771

Jouriles, E. N., & Thompson, S. M. (1993). Effects of mood on mothers' evaluations of children's behavior. *Journal of Family Psychology, 6,* 300-307. doi: 10.1037/0893-3200.6.3.300

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika*, *35*(4), 401-415.

Kang, S., & Harvey, E. A. (2020). Racial differences between Black parents' and White teachers' perceptions of attention-deficit/hyperactivity disorder behavior. *Journal of Abnormal Child Psychology*, *48*(5), 661-672. doi: 10.1007/s10802-019-00600-y

*Keil, J., Perren, S., Schlesier-Michel, A., Sticca, F., Sierau, S., Klein, A. M., ... & White, L. O. (2019). Getting less than their fair share: Maltreated youth are hyper-cooperative yet vulnerable to exploitation in a public goods game. *Developmental Science*, *22*(3), e12765. doi: 10.1111/desc.12765

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

Kim-Cohen, J., Moffitt, T. E., Taylor, A., Pawlby, S. J., & Caspi, A. (2005). Maternal depression and children's antisocial behavior: Nature and nurture effects. *Archives of General Psychiatry*, *62*(2), 173-181. doi: 10.1001/archpsyc.62.2.173

Korelitz, K. E., & Garber, J. (2016). Congruence of parents' and children's perceptions of parenting: A meta-analysis. *Journal of Youth and Adolescence*, *45*(10), 1973-1995. doi: 10.1007/s10964-016-0524-0

*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. doi: 10.1176/appi.ajp.160.9.1566

*Kroenke, C. H., Epel, E., Adler, N., Bush, N. R., Obradović, J., Lin, J., Blackburn, E., Stamperdahl, J. L., & Boyce, W. T. (2011). Autonomic and adrenocortical reactivity and buccal cell telomere length in kindergarten children.

Psychosomatic Medicine, *73*(7), 533-540. doi: 10.1097/PSY.0b013e318229acfc

Lahey, B. B., Piacentini, J. C., McBurnett, K. E. I. T. H., Stone, P., Hartdaghn, S., & Hynd, G. (1988). Psychopathology in the parents of children with conduct disorder and hyperactivity. *Journal of the American Academy of Child & Adolescent Psychiatry*, *27*(2), 163-170. doi: 10.1097/00004583-198803000-00005

Lau, A. S., Garland, A. F., Yeh, M., Mccabe, K. M., Wood, P. A., & Hough, R. L. (2004). Race/ethnicity and inter-informant agreement in assessing adolescent psychopathology. *Journal of Emotional and Behavioral Disorders*, *12*(3), 145-156. doi: 10.1177/10634266040120030201

Lewis, C. C., Boyd, M., Puspitasari, A., Navarro, E., Howard, J., Kassab, H., ... & Kroenke, K. (2019). Implementing measurement-based care in behavioral health: A review. *JAMA Psychiatry*, *76*(3), 324-335. doi: 10.1001/jamapsychiatry.2018.3329

Lewis, K. J., Mars, B., Lewis, G., Rice, F., Sellers, R., Thapar, A. K., …Thapar, A. (2012). Do parents know best? Parent-reported vs. child-reported depression symptoms as predictors of future child mood disorder in a high-risk sample. *Journal of Affective Disorders*, *41*(2), 233-236. doi: 10.1016/j.jad.2012.03.008.

Lieberman, A. F., Ghosh Ippen, C., & Van Horn, P. J. (2015). Don’t hit my mommy: A manual for child-parent psychotherapy with young witnesses of family violence (2nd ed.). Washington, DC: Zero to Three Press.

Lindhiem, O., Vaughn-Coaxum, R. A., Higa, J., Harris, J. L., Kolko, D. J., & Pilkonis, P. A. (2020). Development and validation of the Parenting Skill Use Diary (PSUD) in a nationally representative sample. *Journal of Clinical Child and Adolescent Psychology*, *50*(3), 400-410. doi: 10.1080/15374416.2020.1716366

Lippold, M. A., Greenberg, M. T., & Collins, L. M. (2013). Parental knowledge and youth risky behavior: A person oriented approach. *Journal of Youth and Adolescence*, *42*(11), 1732-1744. doi: 10.1007/s10964-012-9893-1

Lippold, M. A., Greenberg, M. T., & Collins, L. M. (2014). Youths' substance use and changes in parental knowledge-related behaviors during middle school: A person-oriented approach. *Journal of Youth and Adolescence*, *43*(5), 729-744. doi: 10.1007/s10964-013-0010-x.

Loeber, R., Green, S. M., & Lahey, B. B. (1990). Mental health professionals' perception of the utility of children, mothers, and teachers as informants of childhood psychopathology. *Journal of Clinical Child Psychology, 19,* 136-143. doi: 10.1207/s15374424jccp1902_5

Loeber, R., Green, S. M., Lahey, B. B., & Stouthamer-Loeber, M. (1989). Optimal informants on childhood disruptive behaviors. *Development and Psychopathology, 1,* 317-337. doi: 10.1017/S095457940000050X

Lohaus, A., Rueth, J. & Vierhaus, M. (2020) Cross-informant discrepancies and their association with maternal depression, maternal parenting stress, and mother-child relationship. *Journal of Child and Family* Studies, *29*, 867-879 doi: 10.1007/s10826-019-01625-z

Madsen, K. B., Rask, C. U., Olsen, J., Niclasen, J., & Obel, C. (2020). Depression-related distortions in maternal reports of child behaviour problems. *European Child & Adolescent Psychiatry*, *29***,** 275-285. doi: 10.1007/s00787-019-01351-3

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 & Human Development*, *52*(1), 49-62. doi: 10.1007/s10578-020-00985-8

Makol, B. A., De Los Reyes, A., Ostrander, R. S., & 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. doi: 10.1007/s10802-019-00540-7

Makol, B. A., Polo, A. J. Parent-child endorsement discrepancies among youth at chronic-risk for depression. (2018). *Journal of Abnormal Child Psycholology*, *46*, 1077-1088. doi: 10.1007/s10802-017-0360-z

*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. doi: 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*, 856-879. doi:10.1037/a0023678.

Marsh, J. K., De Los Reyes, A., & Lilienfeld, S. O. (2018). Leveraging the multiple lenses of psychological science to inform clinical decision making: Introduction to the special section. *Clinical Psychological Science*, *6*(2), 167-176. doi: 10.1177/2167702617736853

Martel, M. M., Eng, A. G., Bansal, P. S., Smith, T. E., Elkins, A. R., & Goh, P. K. (2021). Multiple informant average integration of ADHD symptom ratings predictive of concurrent and longitudinal impairment. *Psychological Assessment*, *33*(5), 443-451. Doi: 10.1037/pas0000994

*Martel, M. M., Markon, K., & Smith, G. T. (2017a). Research Review: Multi-informant integration in child and adolescent psychopathology diagnosis. *Journal of Child Psychology and Psychiatry*, *58*(2), 116-128. doi: 10.1111/jcpp.12611

*Martel, M. M., Nigg, J. T., & Schimmack, U. (2017b). Psychometrically informed approach to integration of multiple informant ratings in adult ADHD in a community-recruited sample. *Assessment*, *24*(3), 279-289. doi: 10.1177/1073191116646443

Mattick, R. P., & Clarke, J. C. (1998). Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. *Behaviour Research and Therapy*, *36*(4), 455-470. doi: 10.1016/S0005-7967(97)10031-6

McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64-82. doi: 10.1037//1082-989X.7.1.64

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* University of Minnesota Press.

Millsap, E. (2011). *Statistical methods for studying measurement invariance*. Taylor & Francis.

Moffitt, T. E. (2005). The new look of behavioral genetics in developmental psychopathology: Gene-environment interplay in antisocial behaviors. *Psychological Bulletin*, *131*(4), 533-554. doi:10.1037/0033-2909.131.4.533.

Monroe, S. M., & Harkness, K. L. (2005). Life stress, the “kindling” hypothesis, and the recurrence of depression: Considerations from a life stress perspective. *Psychological Review*, *112*(2), 417-445.

Najman, J. M., Williams, G. M., Nikels, J., Spence, S., Bor, W., O’Callaghan, M., . . . Andersen, M. J. (2000). Mothers’ mental illness and child behavior problems: Cause–effect association or observation bias? *Journal of the American Academy of Child & Adolescent Psychiatry, 39,* 592-602. doi:10.1097/00004583-200005000-00013

Narad, M. E., Garner, A. A., Peugh, J. L., Tamm, L., Antonini, T. N., Kingery, K. M., ... & Epstein, J. N. (2015). Parent–teacher agreement on ADHD symptoms across development. *Psychological Assessment*, *27*(1), 239-248. doi: 10.1037/a0037864

*Noordhof, A., Oldehinkel, A. J., Verhulst, F. C., & Ormel, J. (2008). Optimal use of multi-informant data on co-occurrence of internalizing and externalizing problems: The TRAILS study. *International Journal of Methods in Psychiatric Research*, *17*(3), 174-183. doi: 10.1002/mpr.258

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

*Obradović, J., Bush, N. R., & Boyce, W. T. (2011). The interactive effect of marital conflict and stress reactivity on externalizing and internalizing symptoms: The role of laboratory stressors. *Development and Psychopathology*, *23*(1), 101-114. doi: 10.1017/S0954579410000672

*Obradović, J., Bush, N. R., Stamperdahl, J., Adler, N. E., & Boyce, W. T. (2010). Biological sensitivity to context: The interactive effects of stress reactivity and family adversity on socioemotional behavior and school readiness. *Child Development*, *81*(1), 270-289. doi: 10.1111/j.1467-8624.2009.01394.x

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 & Adolescent Psychiatry*, *35*(8), 1078-1085. doi: 10.1097/00004583-199608000-00019

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*. Advance online publication.

*Owens, E. B., & Hinshaw, S. P. (2016). Childhood conduct problems and young adult outcomes among women with childhood attention-deficit/hyperactivity disorder (ADHD). *Journal of Abnormal Psychology*, *125*(2), 220-232. doi: 10.1037/abn0000084

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, *8*(2), 287-312. doi: 10.1207/S15328007SEM0802_7

Peters, L., Sunderland, M., Andrews, G., Rapee, R. M., & Mattick, R. P. (2012). Development of a short form Social Interaction Anxiety (SIAS) and Social Phobia Scale (SPS) using non- parametric item response theory: The SIAS-6 and the SPS-6. *Psychological Assessment*, *24*, 66-76. doi: 10.1037/a0024544

Pérez, J. C., Coo, S., & Irarrázaval, M. (2018). Is maternal depression related to mother and adolescent reports of family functioning? *Journal of Adolescence*, *63*, 129-141. doi: 10.1016/j.adolescence.2017.12.013

*Perren, S., Stadelmann, S., Von Wyl, A., & Von Klitzing, K. (2007). Pathways of behavioural and emotional symptoms in kindergarten children: What is the role of pro-social behaviour? *European Child & Adolescent Psychiatry*, *16*(4), 209-214. doi: 10.1007/s00787-006-0588-6

*Perren, S., Von Wyl, A., Stadelmann, S., Bürgin, D., & von Klitzing, K. (2006). Associations Between Behavioral/Emotional Difficulties in Kindergarten Children and the Quality of Their Peer Relationships. *Journal of the American Academy of Child & Adolescent Psychiatry*, *45*(7), 867-876. doi: 10.1097/01.chi.0000220853.71521.cb

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Sage.

Qasmieh, N., Makol, B.A., Augenstein, T.M., Lipton, M.F., Deros, D.E., Karp, J. , . . . De Los Reyes, A. (2018). A multi-informant approach to assessing safety behaviors among adolescents: Psychometric properties of the Subtle Avoidance Frequency Examination. *Journal of Child and Family Studies*, *27*, 1830-1843. doi: 10.1007/s10826-018-1040-5

Rausch, E., Racz, S. J., Augenstein, T. M., Keeley, L., Lipton, M. F., Szollos, S., ... & 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 & Youth Care Forum*, *46*, 661-683. doi: 10.1007/s10566-017-9403-4

Renouf, A. G., & Kovacs, M. (1994). Concordance between mothers' reports and children's self-reports of depressive symptoms: A longitudinal study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *33*(2), 208-216. doi: 10.1097/00004583-199402000-00008

Rescorla, L. A., Ewing, G., Ivanova, M. Y., Aebi, M., Bilenberg, N., Dieleman, G. C., ... & Steinhausen, H. C. (2017). Parent–adolescent cross-informant agreement in clinically referred samples: findings from seven societies. *Journal of Clinical Child & Adolescent Psychology*, *46*(1), 74-87. doi: 10.1080/15374416.2016.1266642

Rescorla, L. A., Ginzburg, S., Achenbach, T. M., Ivanova, M. Y., Almqvist, F., Begovac, I., ... & Döpfner, M. (2013). Cross-informant agreement between parent-reported and adolescent self-reported problems in 25 societies. *Journal* *of Clinical Child & Adolescent Psychology*, *42*(2), 262-273. doi: 10.1080/15374416.2012.717870

Rettew, D. C., Lynch, A. D., Achenbach, T. M., Dumenci, L., & Ivanova, M. Y. (2009). Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *International Journal of Methods in Psychiatric Research*, *18*, 169-184. doi: 10.1002/mpr.289

Rodriguez, C. M., Wittig, S. M. O., & Christl, M.-E. (2019). Psychometric evaluation of a brief assessment of parents' disciplinary alternatives. *Journal of Child and Family Studies*, *28*(6), 1490-1501. doi: 10.1007/s10826-019-01387-8

Romano, E., Weegar, K., Babchishin, L., & Saini, M. (2018). Cross-informant agreement on mental health outcomes in children with maltreatment histories: A systematic review. *Psychology of violence*, *8*(1), 19-30. doi: 10.1037/vio0000086

Richters, J. E. (1992). Depressed mothers as informants about their children: A critical review of the evidence for distortion. *Psychological Bulletin, 112,* 485-499. doi: 10.1037/0033-2909.112.3.485

*Rijlaarsdam, J., Tiemeier, H., Ringoot, A. P., Ivanova, M. Y., Jaddoe, V. W. V., Verhulst, F. C., & Roza, S. J. (2016). Early family regularity protects against later disruptive behavior. *European Child & Adolescent Psychiatry*, *25*(7), 781-789. doi: 10.1007/s00787-015-0797-y

Rimm-Kaufman, S. E., Curby, T. W., Grimm, K. J., Nathanson, L., & Brock, L. L. (2009). The contribution of children's self-regulation and classroom quality to children’s adaptive behaviors in the kindergarten classroom. *Developmental Psychology*, *45*(4), 958-972. doi: 10.1037/a0015861

*Roubinov, D. S., Boyce, W. T., & Bush, N. R. (2020b). Informant-specific reports of peer and teacher relationships buffer the effects of harsh parenting on children's oppositional defiant disorder during kindergarten. *Development and Psychopathology*, *32*(1), 163-174. doi: 10.1017/S0954579418001499

*Roubinov, D. S., Bush, N. R., Hagan, M. J., Thompson, J., & Boyce, W. T. (2020a). Associations between classroom climate and children’s externalizing symptoms: The moderating effect of kindergarten children’s parasympathetic reactivity. *Development and Psychopathology, 32*(2), 661-672. doi: 10.1017/S095457941900052X

*Roubinov, D. S., Hagan, M. J., Boyce, W. T., Adler, N. E., & Bush, N. R. (2018). Family socioeconomic status, cortisol, and physical health in early childhood: The role of advantageous neighborhood characteristics. *Psychosomatic Medicine*, *80*(5), 492-501. doi: 10.1097/PSY.0000000000000585

*Ruttle, P. L., Shirtcliff, E. A., Serbin, L. A., Fisher, D. B. D., Stack, D. M., & Schwartzman, A. E. (2011). Disentangling psychobiological mechanisms underlying internalizing and externalizing behaviors in youth: Longitudinal and concurrent associations with cortisol. *Hormones and Behavior*, *59*(1), 123-132. doi: 10.1016/j.yhbeh.2010.10.015

Scharfstein, L. A., Beidel, D. C., Sims, V.K., & Finnell, L.R. (2011). Social skills deficits and vocal characteristics of children with social phobia or Asperger’s disorder: A comparative study. *Journal of Abnormal Child Psychology*, *39*, 865-875. doi: 10.1007/ s10802-011-9498-2.

*Shirtcliff, E., Zahn-Waxler, C., Klimes-Dougan, B., & Slattery, M. (2007). Salivary dehydroepiandrosterone responsiveness to social challenge in adolescents with internalizing problems. *Journal of Child Psychology and Psychiatry*, *48*(6), 580-591. doi: 10.1111/j.1469-7610.2006.01723.x

Scott, K., & Lewis, C. C. (2015). Using measurement-based care to enhance any treatment. *Cognitive and Behavioral Practice*, *22*(1), 49-59. doi: 10.1016/j.cbpra.2014.01.010

*Sierau, S., Brand, T., Manly, J. T., Schlesier-Michel, A., Klein, A. M., Andreas, A., Garzón, L. Q., Keil, J., Binser, M. J., von Klitzing, K., & White, L. O. (2017). A multisource approach to assessing child maltreatment from records, caregivers, and children. *Child Maltreatment*, *22*(1), 45-57. doi: 10.1177/1077559516675724

*Slattery, M. J., & Essex, M. J. (2011). Specificity in the association of anxiety, depression, and atopic disorders in a community sample of adolescents. *Journal of Psychiatric Research*, *45*(6), 788-795. doi: 10.1016/j.jpsychires.2010.11.003

Smith, G.T., Fischer, S., & Fister, S. M. (2003). Incremental validity principles in test construction. *Psychological Assessment*, *15*(4), 467-477. doi: 10.1037/1040-3590.15.4.467

Southam-Gerow, M. A., & Prinstein, M. J. (2014). Evidence base updates: The evolution of the evaluation of psychological treatments for children and adolescents. *Journal of Clinical Child & Adolescent Psychology*, *43*(1), 1-6. doi: 10.1080/15374416.2013.855128

Sprich, S. E., Safren, S. A., Finkelstein, D., Remmert, J. E., & Hammerness, P. (2016). A randomized controlled trial of cognitive behavioral therapy for ADHD in medication-treated adolescents. *Journal of Child Psychology and Psychiatry*, *57*(11), 1218-1226. doi: 10.1016/j.cbpra.2015.01.001

*Stadelmann, S., Perren, S., Von Wyl, A., & Von Klitzing, K. (2007). Associations between family relationships and symptoms/strengths at kindergarten age: what is the role of children's parental representations? *Journal of Child Psychology and Psychiatry*, *48*(10), 996-1004. doi: 10.1111/j.1469-7610.2007.01813.x

Stratis, E. A., & Lecavalier, L. (2015). Informant agreement for youth with autism spectrum disorder or intellectual disability: A meta-analysis. *Journal of Autism and Developmental Disorders*, *45*(4), 1026-1041. doi: 10.1007/s10803-014-2258-8

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173-180. doi: 10.1207/s15327906mbr2502_4

Streiner, D. L., & Norman, G. R. (2011). Correction for multiple testing: Is there a resolution? *Chest*, *140*, 16-18. doi: 10.1378/chest.11-0523

*Suh, G. W., Fabricius, W. V., Stevenson, M. M., Parke, R. D., Cookston, J. T., Braver, S. L., & Saenz, D. S. (2016). Effects of the interparental relationship on adolescents’ emotional security and adjustment: The important role of fathers. *Developmental Psychology*, *52*(10), 1666-1678. doi: 10.1037/dev0000204

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

Teglasi, H., Ritzau, J., Sanders, C., Kim, M. J., & Scott, A. (2017). Explaining discrepancies in assessment protocols: Trait relevance and functional equivalence. *Psychological Assessment*, *29*(12), 1517-1530. doi: 10.1037/pas0000447

*Thijssen, S., Ringoot, A. P., Wildeboer, A., Bakermans-Kranenburg, M. J., El Marroun, H., Hofman, A., ... & White, T. (2015). Brain morphology of childhood aggressive behavior: a multi-informant study in school-age children. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(3), 564-577. doi: 10.3758/s13415-015-0344-9

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1-10. doi: 10.1007/BF02291170

*van't Veer, A. E., Thijssen, S., Witteman, J., van IJzendoorn, M. H., & Bakermans-Kranenburg, M. J. (2019). Exploring the neural basis for paternal protection: An investigation of the neural response to infants in danger. *Social Cognitive and Affective Neuroscience*, *14*(4), 447-457. doi: 10.1093/scan/nsz018

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine, 37*(5), 360-363.

*von der Embse, N., Kim, E. S., Kilgus, S., Dedrick, R., & Sanchez, A. (2019). Multi-informant universal screening: Evaluation of rater, item, and construct variance using a trifactor model. *Journal of School Psychology*, *77*, 52-66. doi: 10.1016/j.jsp.2019.09.005

Wakschlag, L. S., Briggs-Gowan, M. J., Carter, A. S., Hill, C., Danis, B., Keenan, K., ... & Leventhal, B. L. (2007). A developmental framework for distinguishing disruptive behavior from normative misbehavior in preschool children. *Journal of Child Psychology and Psychiatry*, *48*(10), 976-987. doi: 10.1111/j.1469-7610.2007.01786.x

Wall, K., Ahmed, Y., & Sharp, C. (2019). Parent-adolescent concordance in borderline pathology and why it matters. *Journal of Abnormal Child Psychology*, *47*(3), 529-542. doi: 10.1007/s10802-018-0459-x

Wang, Y. P., & Gorenstein, C. (2013). Psychometric properties of the Beck Depression Inventory-II: A comprehensive review. *Revista Brasileira de Psiquiatria*, *35*(4), 416-431. doi: 10.1590/1516-4446-2012-1048

Weissman, M. M., Wickramaratne, P., Warner, V., John, K., Prusoff, B. A., Merikangas, K. R., & Gammon, G. D. (1987). Assessing psychiatric disorders in children: Discrepancies between mothers' and children's reports. *Archives of General Psychiatry*, *44*(8), 747–753. doi: 10.1001/archpsyc.1987.01800200075011

Weisz, J. R., & Kazdin, A. E. (2017). The present and future of evidence-based psychotherapies for children and adolescents. In J. R. Weisz & A. E. Kazdin (Eds.), *Evidence-based psychotherapies for children and adolescents*, 3rd ed. (pp. 577-595). The Guilford Press.

Weisz, J. R., Kuppens, S., Ng, M. Y., Eckshtain, D., Ugueto, A. M., Vaughn-Coaxum, R., ... & Weersing, V. 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. doi: 10.1037/a0040360

Youngstrom, E. A., Halverson, T. F., Youngstrom, J. K., Lindhiem, O., & Findling, R. L. (2018). Evidence-based assessment from simple clinical judgments to statistical learning: Evaluating a range of options using pediatric bipolar disorder as a diagnostic challenge. *Clinical Psychological Science*, *6*(2), 243-265. doi: 10.1177/2167702617741845

Youngstrom, E., Loeber, R., & Stouthamer-Loeber, M. (2000). Patterns and correlates of agreement between parent, teacher, and male adolescent ratings of externalizing and internalizing problems. *Journal of Consulting and Clinical Psychology*, *68*, 1038-1050. doi: 10.1037/0022-006X.68.6.1038

Youngstrom, E. A., Izard, C., & Ackerman, B. (1999). Dysphoria-related bias in maternal ratings of children. *Journal of Consulting and Clinical Psychology, 67,* 905-916. doi: 10.1037/0022-006X.67.6.905

Youngstrom, E. A., & Van Meter, A. (2016). Empirically supported assessment of children and adolescents. *Clinical Psychology: Science and Practice*, *23*(4), 327-347. doi: 10.1111/cpsp.1217

*Zaidman-Zait, A., & Hall, W. A. (2015). Children's night waking among toddlers: Relationships with mothers' and fathers' parenting approaches and children's behavioural difficulties. *Journal of Advanced Nursing*, *71*(7), 1639-1649. doi: 10.1111/jan.12636

*Zhang, J., & Jia, C. (2011). Suicidal intent among young suicides in rural China. *Archives of Suicide Research*, *15*(2), 127-139. doi: 10.1080/13811118.2011.56526