

EXAMINATION, PREDICTION, AND OUTCOMES OF INFORMANT DISCREPANCIES
IN BEHAVIORAL RATINGS OF CHILDREN USING
A LATENT MULTITRAIT-MULTIMETHOD MODEL

By

©2015

Joshua J. Turek
B. A., Creighton University, 2001
Ed.S., University of Kansas, 2011

Submitted to the graduate degree program in Educational Psychology and the Graduate Faculty
of the University of Kansas in partial fulfillment of the requirements for the degree of
Doctor of Philosophy.

Dissertation Committee:

Matthew Reynolds (Chairperson)

Steven Lee

Patricia Lowe

Christopher Niileksela

Karrie Shogren

Dissertation defended: July 24, 2015

The Dissertation Committee for Joshua J. Turek

certifies that this is the approved version of the following dissertation:

EXAMINATION, PREDICTION, AND OUTCOMES OF INFORMANT DISCREPANCIES
IN BEHAVIORAL RATINGS OF CHILDREN USING
A LATENT MULTITRAIT-MULTIMETHOD MODEL

Matthew Reynolds (Chairperson)

Date approved: July 24, 2015

ABSTRACT

The use of multiple informants' reports (e.g., mothers, fathers, and teachers) for behavior rating scales (BRS) is common in the psychological assessment of children. Despite widespread use of BRS, discrepancies between informants' ratings are common. The current research was designed to investigate informant discrepancies using a sample of children from the NICHD SECCYD ($n = 784$). Mother, father, and teacher ratings on the CBCL and TRF (Achenbach, 1991) in first, third, and fifth grades were used. Informant discrepancies were modeled as latent mean differences between informants' ratings of aggression, inattention, and anxiety/depression, using a Method Effect with Reference model (Pohl, Steyer, & Kraus, 2008). Several variables were included in models to as predictors of the informant discrepancies, including demographic, intrapersonal, and contextual variables. Discrepancies were also modeled to predict school and diagnostic outcomes. Analysis of latent mean differences showed teachers' ratings of all behaviors were consistently lower than mothers' ratings; fathers' ratings relative to mothers' were dependent on both the type of behavior and the assessment period. Mother-teacher discrepancies were generally larger than mother-father discrepancies. Discrepancies were smaller as levels of behavior increased, particularly for the mother-father dyad. Of the predictor variables, maternal self-reported anger, anxiety, and depression resulted in smaller informant discrepancies in the mother-father dyad; ratings of boys and African Americans resulted in larger discrepancies the mother-teacher dyad, specifically for aggression and inattention. Larger mother-teacher discrepancies were predictive of children's outcomes, including increased referral for special school services and behavioral diagnoses. Finally, some support for the relative longitudinal consistency of discrepancies was found, but dependent on behavior and informant.

ACKNOWLEDGMENTS

Achieving an individual goal is rarely completed without assistance from others. My completion of graduate school and this dissertation is not an exception. First, I wish to thank the five members of my dissertation committee. Karrie Shogren provided a unique perspective, drawing my attention to additional influences on my results, particularly sampling, contextual, and demographic characteristics of the informants. Steve Lee challenged me both during the dissertation process and throughout graduate school to be more critical of research and to consider the implications of research on practice. Chris Niileksela suggested means to better summarize information and strengthen the direct application of the current study. Patricia Lowe has consistently provided guidance, feedback, and instruction in the dissertation process, as well as in the classroom and in navigating the internship and early career process.

I owe a debt of gratitude to Matt Reynolds who served as chairperson for this dissertation and advisor for during my graduate studies. His interest in analytic methods and research sparked my interest in both topics. His constant editing and asking of questions here and in other work has forced me to think more critically and to become a more proficient writer. Thank you to all for your time, feedback, and encouragement.

Second, I have had the opportunity during graduate school to work with an outstanding group of supervisors and colleagues. Several have helped shape me both personally and professionally, including Rene Jamison, Matt Reese, Linda Heitzman-Powell, Lisa Rusinko, Stacy Greenwood, Laura Weber, Carrie Wedel, Robin Sharp, Bo Youngblood, Sue Walker, Jamie Carlisle, and the staff of Tonganoxie High School. Thank you for your dedication to the children and families we had the opportunity to work with and helping me develop the skills to become a better psychologist and researcher.

Words cannot adequately express the thanks that my family deserves. First, my parents, Joyce and John Turek, Jr., have been encouraging my thirst for knowledge since I can remember and have always made it known they were proud of my accomplishments. My siblings and their families, who were encouraging despite never really knowing what it is that I do or why I was still in school. My grandparents, who instilled values of commitment to family, hard work, and taking time to enjoy the good things in life; and especially my grandfather John Turek, Sr., who taught all his grandchildren to seek out and value education. Finally, to my wife, Brighid, who sacrificed to make the completion of my goals possible, all while balancing her own career and education. She has been more supportive than I could have ever imagined, being both mother and father to our young children so I could, in the words of our oldest daughter Maggie, “Go to work”. I am extremely grateful for all you have done to help get me through the last 7 years. I sincerely thank you and love you.

TABLE OF CONTENTS

ABSTRACT.....	iii
ACKNOWLEDGMENTS	iv
TABLE OF CONTENTS.....	vii
LIST OF TABLES	viii
LIST OF FIGURES	ix
 CHAPTER I: Introduction	
Problem Statement	1
Multimethod Assessment.....	2
Informant Discrepancies	2
Multitrait-multimethod Data.....	4
Purpose of the Current Research.....	4
Study Questions	5
 CHAPTER II: Review of Literature	
Behavioral Ratings Scales.....	7
Behavior Rating Informant Discrepancies.....	9
Theories of Informant Discrepancies.....	10
Measurement of Informant Discrepancies	15
Significance and Magnitude of Informant Discrepancies.....	19
Predictors of Informant Discrepancies.....	32
Informant Discrepancies as Predictors.....	35
Multitrait-Multimethod Methods	37
CFA Models for MTMM Data	40
Previous NICHD Studies	51
Importance of the Current Study.....	56
 CHAPTER III: Methods	
Participants.....	57
Measures	59
Missing Data	69
Analytic Plan.....	70
Hypotheses.....	74
Model Evaluation.....	77

CHAPTER IV: Results	
Missing Data Analysis	79
Descriptive Statistics.....	79
Item-Level Measurement Models	81
Measurement Invariance Models	84
Direction and Magnitude of Method Effects	87
Trait-Specific Method Effects.....	91
Relationship of Trait Levels and Method Effects	93
Longitudinal Method Effects	95
Predictors of Method Effects	97
Prediction of Outcomes by Method Effects.....	107
CHAPTER V: Discussion:	
Introduction.....	111
Fathers' Method Effects.....	111
Teachers' Method Effects	116
Relationship of Trait Levels and Informant Discrepancies	119
Prediction of Child Outcomes.....	122
Contributions.....	124
Implications.....	127
Limitations	129
Future Directions	131
References.....	135

List of Tables

Table 1. Summary of reviewed mother-father informant discrepancy studies	27
Table 2. Summary of reviewed mother-father-teacher informant discrepancy studies	32
Table 3. The multitrait-multimethod correlation matrix	39
Table 4. Sample internal consistency (Cronbach's α) for CBCL and TRF ratings	60
Table 5. Descriptive statistics for observed sum scores of CBCL and TRF scales	80
Table 6. Measurement invariance tests for each trait in first grade	82
Table 7. Measurement invariance tests for each trait in third grade	82
Table 8. Measurement invariance tests for each trait in fifth grade	83
Table 9. Longitudinal measurement invariance tests for each trait for mother-father dyad	86
Table 10. Longitudinal measurement invariance tests for each trait for mother-teacher dyad	86
Table 11. Latent covariances, correlations, means, and variances	90
Table 12. Father-ME direction and effect sizes	91
Table 13. Teacher-ME direction and effect sizes	91
Table 14. Tests of equality of means across traits for Father-ME and Teacher-ME	93
Table 15. Tests of longitudinal equality of means and variances for Father-ME and Teacher-ME	96
Table 16. Statistically significant effects of predictors on Aggressive Behavior and related method effects	100
Table 17. Statistically significant effects of predictors on Attention Problems and related method effects	101
Table 18. Statistically significant effects of predictors on Anxious/Depressed and related method effects	102
Table 19. Statistically significant effects of Aggressive Behavior and method effects on outcomes	108
Table 20. Statistically significant effects of Attention Problems and method effects on outcomes	109
Table 21. Statistically significant effects of Anxious/Depressed and method effects on outcomes	110

List of Figures

Figure 1. First order correlated trait-correlated (CTCM) model.....	41
Figure 2. Correlated trait-correlated method (CTCM) model with latent first-order indicators of second order trait and method factors.....	43
Figure 3. Second order Meref model for aggressive behavior with mother as reference	49
Figure 4. Measurement model	71

Chapter I: Introduction

Problem Statement

Children and adolescents who demonstrate aggressive behavior, difficulty sustaining attention, or symptoms of anxiety or depression, are commonly referred for a psychological assessment. Psychological assessment is a problem-solving process designed to obtain behavioral information, identify social-emotional needs or strengths, and inform clinical impressions regarding the problem behaviors or referral question (Merrell, 2007; Meyer et al., 2001). Given the substantial amount of time psychologists spend in assessment and that it often serves as the foundation for providing interventions, critical study of information obtained in the assessment process is imperative to ensure maximum validity of the assessment results. Validity underlies high-quality psychological practice; a lack of validity undermines psychological practice.

Researchers have argued that interpretations that are supported by evidence are valid, those that are not supported are invalid (Kane, 2013). However, psychological assessment is often filled with contradictory evidence. For example, behavior rating scales (BRS), a widely used source of information obtained in psychological assessments (e.g., Merrell, 2007), often result in discrepant ratings made by different informants. As a result, generalization of these data to a valid conclusion is difficult absent direction from the literature. Research is needed to help psychologists better interpret and integrate behavior rating scale information when these discrepancies exist to reach more valid decisions. The current study contributed to the informant discrepancy literature in four areas: 1) clarification of the magnitude and direction of informant discrepancies; 2) identification of predictive variables; 3) the predictive utility of informant

discrepancies; and 4) the longitudinal consistency of magnitude, direction, prediction, and predictive utility of informant discrepancies.

Multimethod Assessment

Best practice in psychological assessment involves the use of a multimethod, multisource, multisetting design (Merrell, 2007). Within this design, information is gathered using different assessment techniques (methods) from a variety of individuals (sources) in numerous contexts (settings). Psychologists then integrate this information to reach diagnostic and treatment decisions.

One aspect of this design, the use of multiple sources (informants), has been advocated in the literature based on the premise that all informants provide accurate information; however, no single informant is completely accurate (Kraemer et al., 2003). Behavioral rating scales (BRS) are one widely used method by which information is provided by multiple informants, with at least one scale completed by each informant. BRS are designed to measure the informant's perception of the traits underlying specific indicators of a child's behavior. They provide a relatively quick and broad description of a child's behavior. Even when ratings are different, it is assumed that each provides a unique contribution to the assessment (i.e., incremental validity; Johnston & Murray, 2003).

Informant Discrepancies

The use of BRS is widespread. Discrepancies between informant's ratings, defined as differences in informants' ratings on parallel measures of behavior (De Los Reyes, Thomas, Goodman, & Kundey, 2013), are similarly widespread (e.g., Achenbach, 2006; Achenbach, McConaughy, & Howell, 1987; Duhig, Renk, Epstein, & Phares, 2000). Such discrepancies exist even when ratings from BRS have been shown to demonstrate considerable validity and

reliability evidence (De Los Reyes, 2011). Systematic research is needed to better explicate the presence of and reasons for informant discrepancies. Psychologists are expected to provide a clear picture of a child's current social-emotional status as a part of a valid psychological assessment. That picture, however, is often clouded by informant discrepancies.

Psychologists do not have clear information about the expected or typical level of discrepancies and relations between informants' ratings (Achenbach et al., 1987); thus, they disagree about how to interpret discrepancies. Some attribute the discrepancies to measurement error (e.g., Krosnick, 1999; McGuire, 1969), which leads to the conclusion that the discrepancies are unavoidable, given that all measurement has error. Others, through recent theoretical developments, have posited alternative explanations for discrepancies, such as differential perceptions of informants and differential behavior dependent on the context (De Los Reyes & Kazdin, 2005; Duhig et al., 2000; Kramer et al., 2003). From these latter perspectives, instead of error, the discrepancies are viewed as additional sources of information that may improve the diagnostic or treatment planning process.

“What informant discrepancies represent” (De Los Reyes, 2011, p. 2) is fundamental to research and practice of psychology. Are discrepancies simply measurement error that can be ignored, or are they are an additional piece of information that can inform the assessment process? Do discrepancies indicate useful information that can inform the assessment? Assessment of the presence of symptoms across multiple settings and informants is part of the current diagnostic criteria or features for several mental health diagnoses, including attention deficit-hyperactivity disorder, oppositional defiant disorder, and conduct disorder (American Psychiatric Association, 2013). Further, rates of diagnoses for childhood disorders depend upon the informant used, influencing both prevalence and traits associated with the disorder (De Los

Reyes & Kazdin, 2005; Langberg et al., 2010). Despite these recent contributions and their implications, few conclusions have been reached regarding informant discrepancies in the assessment process. The answer to these questions has profound importance for the study of psychological disorders and to the psychological assessment of individual children. A better understanding of under what conditions informant discrepancies exist and for what outcome criteria there are implications is needed to better answer the questions.

Multitrait-multimethod Data

The multitrait-multimethod data structure is present when multiple informants rate multiple traits (Campbell & Fiske, 1959). Several studies have demonstrated that substantial proportions of variance in observed scores are due to both the trait being rated and the informant (method) making the ratings (e.g., Konold & Pianta, 2007; Grimm, Pianta, & Konold, 2009). Previous informant discrepancy research has generally ignored the MTMM structure of data. However, ignoring this structure may have drastic implications for findings, such as relations with external variables (Castro-Schilo, Widaman, & Grimm, 2013) and may disregard phenomena, such as the halo effect (Thorndike, 1920).

Purpose of the Current Research

The current research was designed to study informant discrepancies, the prediction of discrepancies, and outcomes related to these discrepancies. The purpose was to utilize recently developed statistical methods to advance the literature regarding the meaningfulness of the expected discrepancies between raters typically observed in multimethod assessment. Specifically, the current study used recently developed statistical models to address four primary remaining gaps in the informant discrepancy literature:

- 1) The magnitude and direction of discrepancies between mother, fathers, and teachers for three common childhood behavioral concerns: aggressive behavior, attention problems, and anxiety/depression.
- 2) The influence of a variety of variables in prediction of informant discrepancies, including demographic and intrapersonal characteristics of informants and contextual information for the setting in which informants observe behavior (i.e., home or school).
- 3) The predictive utility of the discrepancies for both clinical and school referral for services.
- 4) The longitudinal consistency of both the magnitude of discrepancies and explanation of the discrepancies.

To address these gaps, nine questions were answered:

Question 1: What is the informant discrepancy, represented by method effects, for each informant's ratings of the child's behavior?

Question 2: Are these method effects trait-specific? That is, is the size of informant discrepancies dependent upon the trait measured?

Question 3: Are the size of method effects related to levels of trait behavior?

Question 4: Do method effects remain constant over time?

Question 5: Are method and trait effects predicted by SES, ethnicity, and sex?

Question 6: Are method and trait effects predicted by maternal stress, and maternal and paternal depression, anger, and anxiety?

Question 7: Are method and trait effects predicted by independent ratings of parent and teacher sensitivity?

Question 8: Are method effects predicted by ratings of the context in which they occurred?

Question 9: Are method effects predictive of referral to special school services; diagnosed learning disability and attention, behavior, or emotional problems?

Multiple methods (i.e., mother, father, and teacher as informants) were used to measure multiple traits (aggressive behavior, attention problems, and anxious/depressed) using parallel items from two forms from the Achenbach System of Empirically Based Assessment: the Child Behavior Checklist (CBCL; Achenbach, 1991a) and the Teacher Report Form (TRF; Achenbach, 1991b). Latent modeling techniques were used to incorporate both the multitrait-multimethod (MTMM) structure of the data and the influence of measurement error on the measurement of traits. A method effects with reference method model (MEref; Pohl, Steyer, & Kraus, 2010) was used due to a clear conceptualization of method effects (i.e., informant discrepancies). Further, the use of latent variable models allowed for answering the question of whether these discrepancies were actual effects, beyond what was expected from measurement error alone. A sequence of models was fitted to model true score differences between raters within the latent MTMM framework. The findings were expected to advance current methodology measuring informant discrepancies beyond initial descriptions of model specifications and Monte Carlo studies of these methods by applying the models to a national, longitudinal sample of children. Finally, the results from this study may be used by psychologists to more effectively integrate assessment information.

Chapter II: Review of the Literature

Introduction

Informant discrepancies on behavioral ratings scales are common when using multiple raters. The reasons for the discrepancies are not well understood (Youngstrom, Loeber, & Stouthamer-Loeber, 2000) creating problems for the validity of behavioral assessment (Penney & Skilling, 2012). The forthcoming review of the informant discrepancy literature will include a description of behavior rating scales, discussion of theoretical models, statistical measures used in previous research of discrepancies and agreement, literature to date regarding the nature of discrepancies, and recent statistical modeling techniques used to answer questions related to informant discrepancies.

Behavior Rating Scales

Description of BRS. Behavior rating scales (BRS) are the most widely used tools in child behavior assessment (Hunsley & Mash, 2007). These scales typically include several brief items with a question or statement that reference specific behaviors. Item ratings are made by selected informants, often on an ordinal scale (e.g., 0 to 2), with values indicative of the frequency of the behavior (i.e., higher ratings indicate greater frequency of behavior). Ratings of each item within a specific subscale are summed to represent a score for the particular trait (e.g., items asking about frequency of crying, feeling sad, and feeling down are summed to provide a subscale score for depression) or a more general, broadband construct (e.g., externalizing and internalizing) that subsumes several specific subscales. The total raw score is generally converted to an age-based standardized score, providing a comparison of the child's behavior to age-peers.

In general, two types of BRS are available: broadband and narrow-band. Broadband BRS measure multiple traits within the same form, including common childhood behaviors such

as aggression, inattention, and depression. Two broadband scales in particular are used extensively in childhood behavior assessment: the Achenbach System of Empirically Based Assessment (ASEBA; Achenbach, 1991a and 1991b; Achenbach & Rescorla, 2001) and the Behavior Assessment System for Children-Second Edition (BASC-2; Reynolds & Kamphaus, 2006). Broadband BRS often include composite measures of externalizing (outward directed) and internalizing (inward directed) behaviors, in addition to more specific behavior subscales. In comparison, narrow-band BRS focus on a smaller group of closely related traits, such as attention problems and hyperactivity (e.g., Brown Attention Deficit Disorder Scales; Brown, 1996) or depression (e.g., Reynolds Adolescent Depression Scale-2; Reynolds, 2002).

BRS advantages. BRS offer several advantages for childhood behavioral assessment (Merrell, 2007). First, BRS are efficient. Child behavior may be assessed using multiple informants in a timely manner, particularly compared to direct observation, functional analysis, or other techniques available for behavioral assessment. Second, several broadband measures, such as the BASC-2 and ASEBA, are available with substantial reliability and validity evidence across a variety of samples and studies (e.g., Brown & Achenbach, 1993; Greenbaum & Dedrick, 1998; Reynolds & Kamphaus, 2006). Finally, BRS allow clinicians to compare ratings to a comparable age- or grade-norm sample instead of relying on potentially subjective clinical judgment to determine the relative severity of a child's behavior (McClelland & Scalzo, 2006; Reitman, 2006).

BRS disadvantage. Despite the strengths of BRS, a major weakness is that informants' ratings are often discrepant as indicated by score differences between informants (e.g., De Los Reyes & Kazdin, 2005). Informant discrepancies affect diagnostic decisions, depending on a clinician's interpretation of the validity of each informant's ratings and whether the behavior has

reached levels significant enough to warrant diagnosis and intervention. A better understanding of the magnitude of informant discrepancies and why they exist would substantially improve the assessment of childhood behavior.

Behavior Rating Informant Discrepancies

Best practice in behavioral assessment rests on the assumption that the use of multiple informants provides a more comprehensive picture of a child's behavior (Merrell, 2007). However, the commonly found mean differences among informants may obscure the comprehensive picture of behavior. The pervasiveness of informant discrepancies has contributed to researchers concluding that there is no "gold standard" in behavioral assessment that reliably results in correct diagnosis (De Los Reyes & Kazdin, 2005). That is, no one assessment method is consistently relied on and considered consistently accurate for diagnostic purposes. Psychologists are expected to use multiple informants to inform conclusions, despite expected disagreement. This paradox has been described as a "Grand Discrepancy" (De Los et al., 2013).

To resolve this paradox, the psychologist has to make his or her own assumptions. For example, they may assume that multiple informants provide optimal information about child behavior, but then discount any discrepancies between informants as error. Alternatively, the psychologist may believe that the discrepancies in and of themselves are meaningful (De Los Reyes et al., 2013). As a result of these differing approaches, two different psychologists may arrive at two different diagnostic and treatment decisions based on the same information. At its core, the Grand Discrepancy is a validity problem.

The interpretation of informant discrepancies as merely "error" is rooted in converging operations. Converging operations is an approach to understanding data in which several sources

of data are used to arrive at a single conclusion, while systematically eliminating (i.e., “ruling out”) other possible explanations through replication of results (Garner, Hake, & Eriksen, 1956). Converging operations may appeal to psychologists’ desire to use multiple methods in assessment to arrive at the hypothetical “correct” diagnostic decision while ruling out competing diagnoses. Converging operations applied to assessment, however, does not provide an explanation for informant discrepancies and discounts the additional unique information discrepant ratings may provide (De Los Reyes et al., 2013; Lance, Baranik, Lau, & Scharlau, 2009). For example, it is possible that each informant’s perception is valid, but that other variables, such as environmental and inter-personal context, influence behavior. A clear theoretical framework is needed from which the validity of these various interpretations may be evaluated.

Theories of Informant Discrepancies

Two research teams have proposed theories to bring a framework to informant discrepancy research. The first theory, the Multidimensional Validity Theory¹ (MVT), emphasizes the use of multiple informants to accurately triangulate the trait of interest and to provide a method to understand informant discrepancies (Kraemer et al., 2003). The second theory, the Attribution Bias Context theory (ABC; Kazdin & De Los Reyes, 2005), emphasizes explanatory mechanisms to describe the causes of discrepancies (Laird & De Los Reyes, 2013).

Multidimensional Validity Theory. Kraemer and colleagues (2003) proposed the Multidimensional Validity Theory. The theory was designed to provide a framework to understand and aggregate information across multiple informants. In this framework, all informants’ ratings should be indicative of a common trait (i.e., similar to the concept of

¹ Kraemer and colleagues (2003) did not offer a title for their theory. One is provided here for clarity.

convergent operations). For example, ratings made by mothers, fathers, and teachers ideally all converge toward a single point, hypothetically identifying the “true” trait or behavior. The use of independent sources theoretically helps to validly identify the trait, studied extensively through research of construct and convergent validity (i.e., correlations among raters are moderate and positive, thus aggregating the scores will result in more “true” score variance). The problem is that there is rarely a convergence of data, evidenced by small to moderate correlations between different informants’ ratings (Kraemer et al., 2003), and no clear guidelines of what should occur when the data do not converge.

Kraemer and colleagues (2003) described three previous attempts from the literature to solve the lack of convergence in informant reports: identification of an optimal informant, treating all informants as separate and equal outcomes, and aggregating reports. They noted substantial limitations associated with all three, resulting in the development of MVT. The first attempt, identifying an optimal informant, is conducted by using only one informant or using multiple informants but then dismissing conflicting reports from sub-optimal informants (e.g., using mother’s ratings as the most valid ratings). This solution is limited because criteria for the selection of an optimal informant are not available from the literature. The second attempt is to treat all informant data “separately and simultaneously,” specifically in research, by using each informant’s data as a separate outcome. This approach often provides vague results, with limited practical value. For example, if a treatment results in changes in teacher’s ratings, but not parent’s ratings, how do clinicians interpret the results? Is the treatment considered effective because of change in teacher’s ratings, or ineffective due to no change in parent’s ratings? In the third attempt, data are aggregated across ratings from multiple informants. Different approaches have been proposed, including averaging scores across informants or a psychometric approach,

such as weighting based on factor analysis. Kraemer and colleagues argue that these approaches are often arbitrary (e.g., based on clinical judgment) or completed *post hoc* in attempt to interpret discrepant data. They claim that these methods would be more powerful had the aggregation method to be used been determined prior to analysis.

Based on the limitations of the three attempts to solve the problem of informant discrepancies, Kraemer and colleagues (2003) developed the MVT, in which a broader set of four dimensions was proposed to explain the discrepancies (i.e., account for non-convergence). The four dimensions include: 1) the actual trait or symptoms (T); 2) the context in which the behavior is observed (C); 3) the informant's perspective (P); and 4) measurement error (E). Each informant's ratings result from the sum of these four dimensions ($T + C + P + E$). The trait dimension is the behavior of interest that is constant over the studied time span. The context dimension takes into account the setting and circumstances that may have a bearing on the trait behavior. The perspective dimension accounts for informant characteristics that influence ratings. Finally, error includes influences not related to the previous dimensions.

From the MVT perspective, informants' ratings should be correlated more strongly if the informants share a similar context (e.g., mother and father observing behavior in the home) compared to different contexts (e.g., mother observing behavior in the home versus teacher observing behavior in the classroom). The ideal assessment would reduce "extraneous variance" due to different perspectives and contexts, resulting in a maximization of trait variance. To achieve this ideal, different perspectives in the same context should be obtained. For example, the mother (P) tends to see her child largely in the home (C), whereas the teacher (P) tends to see the child largely in school (C). To reduce the extraneous variance due to different perspectives, additional perspectives (i.e., informants) should be used in each context (e.g., father in the home;

additional teacher in the school). This method would reduce the confounding of perspective and context variances (i.e., mother's ratings of behavior in the home treated as synonymous with behavior in the home) when scores are aggregated. Further, to reduce the extraneous variance resulting from context, ratings from the same perspective across different contexts (e.g., mother at home and at school) would be obtained. Ratings gathered in this manner would reduce the variance due to both perspective and context while increasing the proportion of trait variance. However, rarely do parents directly observe their children in the school for extensive periods, and rarely do teachers directly observe children in the home. The ideal is not practical.

Overall, the MVT theory for multi-informant data is designed to maximize trait variance. The theory can be used to guide techniques to integrate informants' reports and as such improve the validity with which traits are measured. This theory and approach, however, is not the only framework proposed to make sense of multi-informant data.

Attribution Bias Context theory. The Attribution Bias Context theory (ABC), proposed by De Los Reyes and Kazdin (2005), was another attempt to bring a coherent framework to the study of informant discrepancies, and specifically, explain the psychological mechanisms by which result in informant discrepancies. Similar to the Multidimensional Validity Theory (Kraemer et al., 2003), in the ABC theory informant discrepancies are theoretically due in part to the variety of contexts in which a child's behavior is observed. Despite this commonality, the two theories differ in emphasis: in ABC it is on the social psychological influences on informants' perspectives; in MVT it is on developing techniques to integrate data to maximize trait variance.

Three social psychological phenomena are used to inform the ABC theory (De Los Reyes & Kazdin, 2005). The first, the actor-observer phenomenon (Jones & Nisbett, 1972), explains

how individuals attribute causes of behavior differently when describing their own behavior compared to describing other's behaviors. Individuals are likely to attribute other's behavior to a dispositional quality, while minimizing the effects of the setting or context. Alternatively, individuals are likely to attribute their own behavior to the effects of the setting or context, while minimizing the effects of a dispositional quality. Because the ratings of others, particularly parents and teachers, are heavily weighted in the assessment of children, the view of the child's problem is therefore more likely to be attributed to within the child, resulting in minimal focus on the context of behavior (De Los Reyes & Kazdin, 2005). Discounting the influence of context on behavior may limit the generalizability of informant ratings across contexts and contribute to discrepancies.

The second phenomenon influencing informant discrepancies in the ABC theory (De Los Reyes & Kazdin, 2005) is biased memory recall. Each informant recalls information differently based on the impression of current behavior. This bias may particularly manifest in the recall of negative events when frustrated or annoyed with the child. As a result, the informant may respond in an overly negative way, more readily recalling negative behaviors when responding. This biased recall results in discrepant ratings when compared to other informants (De Los Reyes & Kazdin, 2005), particularly when more recent or consistent negative events have occurred in interactions between the informant and the child.

The third phenomenon, source monitoring, is the "mechanisms by which people make attributions for how they acquire memories for events" (De Los Reyes & Kazdin, 2005, p. 493). The memories can be derived from heuristics, mental shortcuts used to represent memories, or from systematic use of complex strategies that link different memories. BRS often do not provide context for the questions and instead are based in more global statements of behavior

(De Los Reyes & Kazdin, 2005). This format may then result in informants relying on heuristics, instead of complex factual recall of behavioral events in specific contexts, potentially resulting in more biases.

These three phenomena: the actor-observer phenomenon, biased memory recall, and source monitoring, are hypothesized to influence informant's ratings. However, their effects on informant ratings are largely unknown as the model has rarely been used as the foundation for study.

Theory summary. Two theories were described that provide a framework for understanding informant discrepancies. The Multidimensional Validity Theory emphasizes the use of independent informants and understanding behavior ratings as the sum of trait, perspective (informant), context, and error variances. The ABC theory focuses on three social-psychological phenomena that may result in discrepancies: actor-observer phenomena, memory biases, and source monitoring. Both theories represent plausible frameworks to understand why discrepant reports exist; however, both have limited evidence to support their use due to a paucity of informant discrepancies studies using any theory as a foundation. The current study will focus on hypotheses described in the MVT, including the influence of trait, context, perspective, and error; as well as the memory biases and actor-observer phenomena described in the ABC theory.

Measurement of Informant Discrepancies

A variety of techniques have been used in the literature to measure the agreement and discrepancies observed in multi-informant assessment. The various techniques have often resulted in different outcomes. As a result, understanding these techniques is important to understanding the techniques to be used in the current study.

Measuring informant agreement. Several different statistics have been used to describe the amount of agreement between raters. The most prevalent in the research literature are the Pearson product moment correlation, r , and a related correlation applied to sets of items, q (e.g., Youngstrom et al., 2000). Other measures of agreement have been used sparingly despite potential benefits. For example, the intraclass correlation (ICC) has been described as less dependent on systematic method effects for one rater versus another (e.g., mothers systematically rate children lower than fathers), providing a better measure of agreement (Cicchetti, 1994).

Correlations were used in a widely-cited meta-analysis of informant agreement, in which stronger agreement between parents' ratings ($r = .60$) were reported than between parents' and teachers' ratings ($r = .28$) of behavioral rating scale scores (Achenbach et al., 1987). Additionally, the magnitude of agreement depends on the type of behaviors, such as internalizing versus externalizing. A meta-analysis of 60 studies of maternal and paternal ratings reported moderate agreement ($r = .45$) for internalizing behaviors and strong agreement for externalizing behaviors and total problem behaviors ($r = .63$ and $.70$; Duhig et al., 2000). Although moderate to strong correlations between parents have been reported, these less than perfect correlations also provide evidence for discrepant ratings, even among individuals who are familiar with the child in similar contexts (i.e., complete agreement would result in $r = 1.0$, complete disagreement $r = -1.0$; Achenbach, 2006). Despite multiple available statistics measuring agreement, each is limited in that it only indicates the strength of relationship among ratings.

Measuring manifest discrepancies. In addition to statistics designed to measure the amount of agreement between raters, another set of statistics have been used to describe the amount of discrepancy between raters. Measures of agreement and discrepancy potentially answer a similar set of questions but the results from different methods can result in dissimilar

conclusions (De Los Reyes & Kazdin, 2004; Treutler & Epkins, 2003). The two should not be used interchangeably (Duhig et al., 2000), although they often are in the literature. Agreement is measured as a correlation; discrepancies are measured in mean differences in levels of behavior. Discrepancies have been noted to vary in both magnitude and direction (i.e., which informant reports greater levels of behavior; De Los Reyes et al., 2011). A variety of methods can be used to test discrepancies as differences between means, including ANOVA and three types of difference scores.

Analysis of variance (ANOVA) and similar tests subsumed under multiple regression techniques, such as *t*-tests and correlations when comparing only two variables, are often used to estimate differences in ratings. For example, significant main effects (mean differences between informants) were reported in a sample of adolescents using forms from the ASEBA, with mothers and fathers rating children's externalizing behavior higher than teachers, and mothers reporting higher internalizing behavior than teachers (Stanger & Lewis, 1993). Beyond ANOVA, informant discrepancies have been defined in the literature by a variety of additional methods relying on observed variables.

De Los Reyes and Kazdin (2004) outlined three commonly used methods to describe and analyze informant discrepancies: raw difference scores, standardized difference scores, and residual difference scores. Raw difference scores are computed by subtracting one informant's rating from the second informant's rating. Similarly, standardized difference scores are computed by subtracting one informant's standardized score (typically a *z* score; standardized within each informant) from the second informant's standardized score. This method is advantageous in situations in which the scores may be on different scales, potentially reducing differences in variances. A third method, the residual difference score, is computed by using one

informant's rating as an independent variable to predict the second informant's rating in a linear regression. The difference between the predicted and observed score is then standardized resulting in a standardized residual.

Latent difference scores. Thus far the difference score methods described relied on differences in observed scores. These techniques provide valuable information but have shortcomings: primarily the influence of measurement error in the difference scores (McArdle, 2009). All measured indicators have some degree of error; thus, the difference scores are confounded with error. One way to overcome the measurement error problem is to utilize variables that are free of measurement error (i.e., latent variables) through advanced modeling techniques, such as confirmatory factor analysis (CFA) and structural equation modeling (SEM). Latent difference scores (LDS) can then be modeled that have effectively removed measurement error from the constructs of interest

Factor analysis has a long and storied tradition in the measurement of psychological and educational constructs. Charles Spearman's single-factor (1904, 1927) and Louis Thurstone's (1947) multiple-factor models have provided the foundation for the modeling technique. CFA subsequently advanced factor analysis from a psychometric model to a statistical model through the use of computer modeling by several researchers, including Karl G. Joreskog (1969). CFA techniques involve applying a model defining the structure of the data to a sample covariance matrix of observed indicators (items). A common latent factor is modeled to account for common variance (i.e., covariance) among indicators. Unique variance, specific to each indicator (measurement error plus reliable specific variance), is also modeled. The common latent factor and unique variance are typically not allowed to covary. Thus, the common factors contain only the common variance.

These models explicitly decompose measurement error variance from true score variance in observed scores, in line with classical test theory (CTT). Further, in second-order CFA, in which latent factors are used as indicators for a “higher-order” latent factor, the specific variance can be modeled separately from measurement error variance (Marsh & Hocevar, 1988). Several other advantages of CFA have been noted, including the explicit modeling of unique variances, availability of testing invariance across groups, testing the *a priori* theoretical structure of the data (versus exploratory factor analysis), and serving as the measurement model for structural equation models, with directional paths between latent constructs (Brown, 2006).

Given the advantages of CFA, the use of latent difference scores (LDS) in a CFA framework has become more common in psychological measurement. LDS are typically defined as levels of change, with the change defined as the difference between two measurements of true score variables. These difference scores can be computed using common latent factors, which are free of measurement error. Although often applied to longitudinal models to measure a rate of change (i.e., change in time 2 compared to time 1; c.f., McArdle & Hamagami, 2001), these models can be useful in studying latent differences between different informants.

Significance and Magnitude of Informant Discrepancies

Importance of informant discrepancies. The research regarding informant discrepancies has used a variety of informants (i.e., mother, fathers, teachers), and samples (i.e., clinical, community). De Los Reyes (2011) summarized three consistent findings as evidence of informant discrepancies providing important information regarding a child’s behavior beyond the ratings themselves. First, discrepancies exist between informants even when reliable and valid measures are used. This finding suggests discrepancies are not due entirely to unreliability (i.e., measurement error; Hartley, Zakriski, & Wright, 2011) or to a lack of validity (De Los Reyes,

2011). Second, behaviors are expected to be different across contexts (e.g., home and school; De Los Reyes, Henry, Tolan, & Wakschlag, 2009; Kazdin & Kagan, 1994) and between raters in each specific context (e.g., two parents at home, two teachers at school), potentially resulting from different contextual factors in which they observe the child. That is, behaviors are expected to be different across contexts and raters (Mischel, 1968). Third, basic social and psychological theories of memory recall and attributions of behavior, as well as other exogenous variables, may help explain discrepancies (e.g., De Los Reyes & Kazdin, 2005).

A selective body of literature is available that has considered the magnitude, sources, and prediction of informant discrepancies. The current review will focus primarily on studies or sections of studies concerned with mother, father, and teacher ratings, as these raters were used. This review will consider the magnitude of discrepant reports as reported in the literature, focusing primarily on discrepancies (differences) and not agreement (degree of relationship). First, mother and father discrepancies will be considered, followed by parents and teachers. A summary of findings for all reviewed studies is provided in Tables 1 (mother-father) and 2 (mother-father-teacher).

Mother-father discrepancies. Comparisons of ratings by mothers and fathers have yielded a general, yet unequivocal, pattern of results. The mean ratings from mothers are generally higher than those from fathers, indicating greater mother-rated child psychopathology.

A meta-analysis of 60 studies published from 1990 to 1997 summarized correlations (agreement) and mean differences (discrepancies) between mother and father ratings of their children's internalizing and externalizing behaviors using a variety of BRS (Duhig et al., 2000). Of the 60 total studies summarized in this widely-cited meta-analysis, 16 studies reported means and standard deviations for mother and father ratings (the remaining 44 were used to determine

level of agreement). Fifty-five independent effect sizes were calculated from the 16 studies. Positive effects indicated higher ratings endorsed by the mother; negative effects indicated higher ratings endorsed by the father. Average mean differences, computed by averaging Hedge's (1982) g effect sizes between mother and father ratings of behavior for each study, were reported (internalizing behaviors $g = .16$; externalizing behaviors $g = .08$). The overall differences between mother and father ratings were not statistically significant; however, on average, mothers reported more problem behavior. The authors acknowledged the averaging of all observed effects as a limitation of this, and all, meta-analyses. Averaging the effects may limit an understanding of each individual study's effect size. For example, averaging one positive and one negative effect with the same magnitude result in an average effect of zero.

In order to further understand the meta-analysis, I disaggregated the results. Significant differences were noted in 6 of 16 studies, or 38% of studies included. Fifty-five independent effect sizes were calculated across the 16 studies; of these, 10 were statistically significant. The range of statistically significant effect sizes revealed useful information: four significant effect sizes for internalizing behavior ($g = -.54$ to $.26$); two significant effect sizes for externalizing behavior ($g = -.61$ to $.51$); and four significant effect sizes for total problems ($g = .20$ to $.70$). The negative and positive overall effects provided evidence that both mothers and fathers rate children with higher levels of behavior, dependent on the study. In other words, although the overall average effects were negligible, a substantial proportion of studies provided evidence of differences.

Overall, this meta-analysis, which has been treated as a foundational study, revealed important trends in the data, including generally higher ratings by mothers for both internalizing and externalizing behaviors. However, explication of each specific study's sample

characteristics and clearly delineating effects associated with broadband and narrow-band behaviors may provide a more nuanced understanding of discrepancies in mother-father comparisons.

Internalizing behaviors. Three of the studies included in the Duhig and colleagues (2000) meta-analysis reported significant discrepancies in parent ratings of internalizing or related behaviors. In the first study, higher mother ratings ($g = .24$) were reported in a sample of toddlers ($n = 156$; 16 to 24 months of age) but this effect was present only when rating girls (Fagot, 1995). The second study used a small sample of children 6 years of age ($n = 42$). Mother's ratings were significantly lower than father's ratings for girls using the CBCL Internalizing scale ($g = -.54$); mother and father ratings were similar for boys (Crockenberg & Lourie, 1996). The third study researched depressive symptoms in a sample of African American children 9 to 12 years of age ($n = 90$). Mother ratings of the symptoms were higher than father ratings ($g = .26$; Brody et al., 1994). These three studies provide conflicting evidence of which parent report higher levels of children's internalizing behavior and evidence that parent ratings may depend on the child's gender.

Subsequent studies not included in the meta-analysis have provided evidence of higher mother ratings compared to fathers, supporting two studies included in Duhig and colleagues meta-analysis (i.e., Brody et al., 1994; Fagot, 1995). Mother ratings were significantly greater for internalizing behavior in an outpatient clinic sample (Schroeder, Hood, & Hughes, 2010) and in a twin study (Bartels et al., 2003). Despite these examples, other evidence is available that mothers rate lower levels of internalizing behavior than fathers. Lower mother ratings of internalizing behavior were reported in a community sample of 10 to 12 year old children

(Treutler & Epkins, 2003). An additional study noted non-significant differences using a clinical sample (Moreno, Silverman, Saaverdra, & Phares, 2008).

Overall, more evidence is available to provide support for higher mother ratings of internalizing behavior, especially when rating girls. However, contradictory evidence is clearly available. The widely varying results, even in studies included in the most recent meta-analysis, seem limited by small sample sizes, age ranges, or specific demographic characteristics of samples, making generalizing findings difficult. Basic questions such as the direction and magnitude of discrepancies for internalizing behaviors remain unanswered, providing the impetus for continued study of informant discrepancies. The use of a larger, more representative sample will help to clarify the direction and magnitude of the differences (Treutler & Epkins, 2003; Youngstrom et al., 2000).

Externalizing behaviors. Significant informant discrepancies for externalizing behaviors were reported in two studies included in Duhig and colleagues' (2000) meta-analysis. Mother ratings of 6 year old girls were significantly lower than father ratings ($g = -.61$); however, ratings of boys were not discrepant (Crockenberg & Lourie, 1996). These same authors noted similar findings for internalizing behavior. In the second study reporting significant differences in the meta-analysis, mother ratings of externalizing behaviors were significantly higher than father ratings in a sample of adoptive children ages 13 to 19 ($g = .51$; Cohen, Coyne, & Duvall, 1993). Child gender differences were not reported. These two studies demonstrate the indistinct pattern of which informant rates externalizing behavior at higher levels.

Few studies subsequent to the meta-analysis have explicitly studied mean differences in mother and father ratings of externalizing behavior. Similar to Cohen and colleagues (1993), higher mother ratings of externalizing behavior have been reported in an outpatient clinic sample

(Schroeder et al., 2010) and in a twin study (Bartels et al., 2003). Similar to Crockenberg and Lourie (1993), a second study noted fathers reported more externalizing behavior for girls than did mothers (Treutler & Epkins, 2003). The study sample included 100 children ages 10 to 12 years old. Again, the generalizability of the results of these studies are limited by the small sample sizes, limited age ranges included within the studies, or unique demographic characteristics of the sample.

Narrow-band behaviors. A small number of studies have moved beyond the externalizing-internalizing dichotomy of behavior to narrow-band behavior in analyzing mother-father discrepancies. For example, one study found that despite significant correlations between *T*-scores of mother and father ratings of children and adolescents, statistically significant discrepancies existed on six of eight CBCL narrow-band scales in a clinical sample. The significant discrepancies included all three scales used in the current study: attention problems, aggressive behavior, and anxious/depressed (Schroeder et al., 2010). Mothers' ratings were higher on average, similar to the general pattern of internalizing and externalizing behavior. Despite the significance of these findings, the actual magnitude of differences was generally small, between two and three *T*-score points, similar to effect sizes reported by Duhig and colleagues (2000). In other words, although the differences may be statistically significant, practically they may be limited in their utility. However, the utility of these differences remains largely unstudied, with a few exceptions (e.g., Langberg et al., 2010).

Specific studies of depression and anxiety are limited, as they are typically studied as subsumed by the broad internalizing construct. A few studies have provided evidence that mothers rated depressed and anxious behaviors as more severe than fathers when rating their daughters (e.g., Seiffge-Krenke & Kollmar, 1998). However, other studies have reported mean

levels were not discrepant on broadband internalizing ratings and the anxious/depressed scale from the CBCL (e.g., Moreno et al., 2008).

Researchers have also noted limited research focused on ratings of symptoms specific to attention problems, although there is a growing body of literature. In attempts to remedy this gap in the literature, Langberg and colleagues (2010) used a sample of children 7 to 9 years of age with ADHD-combined type from the Multimodal Treatment Study of Children with ADHD (MTA Cooperative Group, 1999). Mothers rated their children higher on all indices of the SNAP-IV scale (Swanson, 1992), which uses ratings similar to DSM-IV-TR ADHD diagnostic criteria (American Psychiatric Association, 2000). In addition, mother's ratings were higher on the CBCL externalizing scale. Standardized difference scores were greater for inattention (.40), hyperactivity/impulsivity (.33), and ADHD (.41) ratings compared to CBCL externalizing problems (.25). Mother and father ratings of externalizing behaviors were highly associated ($r = .58$) but with significantly different levels of behavior ($M = 19.87$ for mothers; $M = 17.83$ for fathers; $ES = .25$).

Several important points are to be gleaned from this study. First, larger effect sizes may be observed with more specific, narrow-band scales compared to general, broadband scales, as well as when clinical samples are used. Second, effects are noted even when two informants' ratings are significantly correlated. Third, the differences are meaningful for diagnostic decisions: 73% of children met criteria for ADHD based on mother ratings, but only 58% based on father ratings. This is an important practical consequence of informant discrepancies, even when effect sizes are not large.

Other studies have also evidenced mean differences in ratings of attention problems. For example, significant discrepancies between mothers and fathers have been noted on ADHD-

inattentive subtype symptom-specific ratings in a small clinical sample, with mothers rating more significant problems (Sollie, Larsson, & Mørch, 2013). However, within the same study differences were not observed on ratings of more general attention problems using the CBCL. Another study supported mother's ratings of ADHD symptoms higher than fathers, and a significant effect on the number of children meeting diagnostic criteria, with 28.5% more meeting criteria based on mother's ratings compared to father's ratings (Caye, Machado, & Rohde, 2013).

Studies using larger samples, however, have noted non-significant differences between parents' ratings of ADHD inattention and hyperactive-impulsive symptoms (Waschbusch, Sparkes, & Northern Partners in Action for Child and Youth Services, 2003). The latter findings could be due to a large portion (> 90%) of the sample exhibiting average or low levels of inattention or hyperactivity. Again, although a general trend of mothers reporting higher levels of behavior is evidenced, it is not consistently present or statistically significant.

Table 1
Summary of Reviewed Mother-Father Informant Discrepancy Studies

Study Authors	Broadband Behavior		Narrow-band Behavior		
	Ext.	Int.	Aggressive Behavior	Attention Problems	Anxious/Depressed
Duhig et al. (2000)	M = F	M = F			
Fagot (1995)		M = F ¹ M > F ²			
Crockenberg & Lourie (1996)	M = F ¹ M < F ²	M = F ¹ M < F ²			
Brody et al. (1994)					M > F
Cohen et al. (1993)	M > F				
Schroeder et al. (2010)	M > F	M > F	M > F	M > F	M > F
Bartels et al. (2003)	M > F	M > F			
Treutler & Epkins (2003)	M = F ¹ M < F ²	M < F			
Seiffge-Krenke & Kollmar (1998)					M = F ¹ M > F ²
Moreno et al. (2008)		M = F			M = F
Langberg et al. (2010)	M > F			M > F	
Sollie et al. (2013)				M > F ^a M = F	
Caye et al. (2013)				M > F ^a	
Waschbush et al. (2003)				M = F ^a	

Note. Greater than (>) or less than (<) used only when statistically significant differences were reported. Ext. = Externalizing; Int. = Internalizing; M = mother ratings; F = father ratings; 1 = boys; 2 = girls; a = ADHD specific symptoms.

Parent-teacher discrepancies. In addition to studies comparing mother and father ratings, studies of informant discrepancies between parents and teachers are available. Tests of agreement have indicated father-teacher and mother-teacher correlations are significantly lower

than mother-father correlations on both internalizing and externalizing scales (Grietens et al., 2004). How these measures of agreement translate to discrepancies is less well understood.

One confounding factor is the common practice to use mother ratings as “parent” ratings, with fathers not included in the study, or included as only a negligible portion (< 10%) of the informants, which has led some to call for increased research using both parents (Penney & Skilling, 2012). For clarity, studies with a significant proportion of mother reports will be included in the mother-teacher section; studies explicitly including fathers will be reviewed following mother-teacher studies. In addition, several studies reported comparisons between informants, but did not explicitly test the differences. These comparisons will be explicitly noted for each study.

Mother-teacher internalizing behavior. Youngstrom and colleagues (2000) reported parents (91% biological, step- or adoptive mothers) endorsed significantly more internalizing behavior than teachers ($d = .57$) using the CBCL and TRF in a sample of boys from the Pittsburgh Youth Study (Loeber, Farrington, Stouthamer-Loeber, & Van Kammen, 1998). Mothers also rated internalizing problems significantly higher than teachers in a small, outpatient clinical sample (Sollie et al., 2013). A cross-sectional study using a sample of Dutch adolescents 11 to 18 years of age ($n = 1122$) indicated parents (95% mothers) rated both girls and boys internalizing symptoms higher than teachers consistently across adolescence (Van der Ende & Verhulst, 2005) although the statistical significance of this difference was not explicitly tested. There was consistent evidence that mother ratings of internalizing behavior were higher than teachers.

Mother-teacher externalizing behavior. Parent ratings of externalizing behavior have been noted to be higher than teachers, similar to evidence with internalizing behaviors. For

example, parents (primarily mothers) endorsed significantly more externalizing problems than teachers ($d = .64$) using the CBCL/TRF in the same sample of boys from the PYS study, similar to the effect observed with internalizing behaviors (Youngstrom et al., 2000).

Beyond these effects, other studies have noted that differences may depend on age of the child. For example, mothers' ratings were higher than teachers' in children 7 years of age (difference = 3.79 *T*-scores); however, when the same sample was assessed again at age 16 teachers rated children higher (difference = 2.61 *T*-scores; Van Dulmen & Egeland, 2011). Contrary to this study, the magnitude of parent-teacher discrepancies for boys' externalizing behavior has been found to be greater in later adolescence, with higher parent ratings on average (Van der Ende & Verhulst, 2005). Specifically, limited discrepancies were noted between parent-teacher ratings of boys prior to age 16. However, beginning at age 16 for boys, parent-teacher discrepancies became larger as teacher ratings decreased substantially and parent ratings decreased less drastically, creating a larger discrepancy. Despite these differences, it is important to note that discussion of the differences was limited, and evidence provided here is based on visual interpretation of graphs reported by Van der Ende and Verhulst (pp. 120 – 121). Both studies (Van Dulmen & Egeland, 2011; Van der Ende & Verhulst, 2005) indicated potential moderation in ratings and discrepancies, dependent on both the age and sex of the child assessed.

In addition to potential age effects, the sex of the target child also appears to influence parent-teacher ratings, and, as a result, discrepancies. For example, parents (primarily mothers) rated boys' externalizing behavior slightly higher than girls consistently across ages 11 to 17, resulting in a consistent discrepancy (Van der Ende & Verhulst, 2005). In contrast, teacher ratings demonstrated a steady decrease in magnitude of the discrepancy between ratings of boys and girls, indicating an age-dependent sex effect.

Mother-teacher narrow-band behavior. Research of mother-teacher discrepancies for narrow-band behaviors is limited. Van der Ende & Verhulst (2005) compared parent-teacher ratings of narrow-band behavior, in addition to externalizing and internalizing behaviors, using the CBCL and TRF across adolescence. Although not explicitly discussed or tested, there seemed to be evidence of an increasing parent-teacher discrepancy for boys' aggressive behavior with parents rating higher levels of behavior at all ages. Conversely, the parent-teacher discrepancies for aggressive behavior became smaller as girls aged in adolescence; parents' ratings were higher at all ages.

No discrepancies were reported for ratings of attention problems when the sample was combined to include both sexes. However, when divided by sex, sex differences were noted: teachers and parents both rated boys higher than girls consistently across adolescence. Further, these discrepancies were dependent on age in addition to sex, with large discrepancies for boys in early adolescence which then became smaller over time. Parent-teacher discrepancies for girls' attention problems were typically negligible. Finally, parent-teacher discrepancies were consistently small for both boys' and girls' anxious/depressed symptoms, with parent ratings higher than teachers across the developmental period. Clearly, inclusion of child's sex and age were needed in the current study, as both influence discrepancies observed between mother and teacher ratings.

A second study (Collishaw, Goodman, Ford, Rabe-Hesketh, & Pickles, 2009) considered discrepancies between parent (primarily mothers) and teacher ratings across three narrow-band scales of the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997, 2001): conduct, hyperactive, and emotional behaviors, which parallel the aggressive behavior, attention problems, and anxious/depressed scales used in the current study. The study used a large sample

($n = 4525$) of children ages 11 to 15 years of age from the United Kingdom. Teacher ratings of conduct problems and hyperactivity were higher than parents for boys, contrary to other studies (e.g., Van der Ende & Verhulst, 2005). However, higher parent ratings of girls were observed for emotional problems, consistent with Van der Ende and Verhulst. Evidence again is inconsistent regarding which informant endorses greater levels of behavior and the differential effects across type of behavior.

Mother, father, and teacher discrepancies. Few studies have included the three adult informants (mother, father, and teacher) to be examined in the current study. The use of both mothers and fathers will differentiate the parent-specific effects when compared to teachers, instead of confounding mothers and fathers together.

Mother, father, and teacher broadband behaviors. One of the few studies to include all three informants occurred over 20 years ago. Both mothers and fathers reported significantly more externalizing behaviors than teachers in a sample of 13 year old adolescents; mothers reported more internalizing behaviors than teachers, whereas fathers and teachers reported similar levels (Stanger & Lewis, 1993). Although differences between informants on the internalizing scales were not large, mother ratings were consistently higher than teacher and father ratings.

Mother, father, and teacher narrow-band behaviors. Discrepancies on narrow-band behaviors are also not widely reported when making parent-teacher comparisons. A noted exception indicated that discrepancies were not present between mother-teacher or father-teacher dyads for ratings of aggression and inattention/hyperactivity in a Norwegian outpatient clinical sample (Sollie et al., 2013). In a Russian community sample, it appeared that teachers endorsed more anxious/depressed and attention problems, but less aggressive behavior compared to both

mothers and fathers on the CBCL/TRF (Grigorenko, Geiser, Slobodskaya, & Francis, 2010).

However, the differences were not explicitly tested.

Table 2

Summary of Reviewed Mother-Father-Teacher Informant Discrepancy Studies

Study Authors	Broadband		Narrow-band		
	Ext.	Int.	Aggressive Behavior	Attention Problems	Anxious/Depressed
Youngstrom et al. (2000)	P* > T	P* > T			
Sollie et al. (2013)	M = T	M > T	M/F = T	M/F = T	
Van der Ende & Verhulst (2005)		P* > T			M/F = T
Van Dulmen et al. (2011)	M > T ^a M < T ^b				
Collishaw et al. (2009)			P* < T ¹	P* < T ¹	P* > T ²
Stanger & Lewis (1993)	M/F > T	M > T F = T			

Note. Greater than (>) or less than (<) used only when statistically significant differences reported. Ext. = Externalizing; Int. = Internalizing; M = mother ratings; F = father ratings; 1 = boys; 2 = girls; a = age 7; b = age 16; * = parent sample consisted primarily of mothers.

Predictors of Informant Discrepancies

A select literature is available regarding several variables that may explain discrepancies in ratings. The variables may be broken down into demographic characteristics (SES, sex, and age of the child), parental and teacher intrapersonal characteristics (psychopathology, stress, and sensitivity), and environmental contexts. These influences align with both the MVT (Kraemer et al., 2003) and the ABC model (De Los Reyes & Kazdin, 2005). Demographic characteristics and intrapersonal characteristics are likely to influence the perspective (P) of MVT theory and the biased memory recall of the ABC model. The actor-observer phenomenon described by

ABC theory, in which informants discount context when rating others' behavior, directly relates to the context (C) described by the MVT.

Socio-economic status. Socio-economic status (SES) has a well-established relationship with children's behavior (e.g., Latourneau, Duffet-Leger, Levac, Watson, & Young-Morris, 2013) and mental health (e.g., McLaughlin et al., 2011; Reiss, 2013). SES also has a general effect on rating levels, with children from lower SES rated higher, independent of the informant (Van der Ende & Verhulst, 2005). In addition to its relation to reported trait levels, SES has been related to discrepancies; however this relation is largely dependent on the method by which SES is measured. SES based on maternal employment and education was not related to mother-teacher discrepancies (Youngstrom et al., 2000) whereas lower family income was related to larger and increased occurrence of discrepancies (Collishaw et al., 2009; Stone, Speltz, Collett, & Werler, 2013).

Sex of the child. Informant-specific sex differences (i.e., larger or smaller discrepancies for informants when rating boys vs. girls) have received a mixed level of support in the literature. Some researchers have claimed that informant discrepancies (including parent-teacher and mother-father) are not typically affected by the sex of the child (De Los Reyes & Kazdin, 2005; Schroder et al., 2010). Several studies, however, have found evidence otherwise (e.g., Collishaw et al., 2009; Van der Ende & Verhulst, 2005).

Age of the child. The age of the child has been inconsistently associated with informant agreement and discrepancies. In their meta-analysis, Achenbach and colleagues (1987) noted statistically significant differences in agreement between parent ratings for children ages 6 to 11 ($r = .51$) compared to ages 12 to 19 ($r = .41$). Similarly, age moderated the discrepancy in parent ratings of attention problems, with younger children's (ages 5 – 8) parents showing smaller

discrepancies than older children's (ages 9 – 13; Schroeder et al., 2010). Subsequent studies have indicated differential effects dependent on the age of the child for discrepancies between parents' and teachers' ratings (Van Dulmen & Egeland, 2011; Van der Ende & Verhulst, 2005). Differences across studies may be due in part to methodological differences, such as small sample size, the use of median split of age groups (De Los Reyes & Kazdin, 2005), or the selection of more extreme groups (e.g., clinical rather than community samples).

Parental psychopathology. Intrapersonal characteristics of the informant have been hypothesized to be related to informant discrepancies. Maternal depression and its relation to behavior ratings has received considerable attention, with the ratings of more depressed mothers expected to have a negative perceptual bias, described as the depression-distortion hypothesis (Richters & Pellegrini, 1989); however, within the study limited evidence was provided to support the hypothesis.

Subsequent studies, however, have provided support for the depression-distortion hypothesis. Chi and Hinshaw (2002) noted that self-reported depressive symptoms were predictive of discrepancies in both ADHD symptoms ($\beta = .28$) and general behavior problems ($\beta = .30$). Youngstrom and colleagues (2000) studied the effects of both mother and father psychopathology on prediction of parent and teacher informant discrepancies. Maternal depressive symptoms was correlated ($r = .23$) with parent-teacher discrepancies (i.e., more depressive symptoms were related to more discrepant ratings). Paternal antisocial behavior and maternal substance use were not related to parent-teacher discrepancies.

In another study, mothers' psychological symptoms, after accounting for parent-child relationship variables, were related to father-mother discrepancies of internalizing behavior ratings ($\beta = -.34$); greater psychological symptoms were related to smaller discrepancies

(Treutler & Epkins, 2003). Both mothers' ($\beta = -.27$) and fathers' ($\beta = .28$) psychological symptoms were related to father-mother discrepancies of externalizing behavior, although in differing directions (i.e., maternal symptoms were related to smaller discrepancies; paternal symptoms were related to larger discrepancies).

Stress. Similar to parental psychopathology, parent stress has been hypothesized to be related to discrepancies and bias in ratings. Maternal stress was related to parent-teacher discrepancies (Youngstrom et al., 2000). Greater discrepancies have also been associated with fathers' ratings of stress ($r = .19$), while controlling for maternal and paternal depression (Mascendaro, Herman, & Webster-Stratton, 2012). Further, differential effects have been observed, dependent on the severity of behavior (Langberg et al., 2010). Fathers with low parental stress rated ADHD and externalizing behaviors lower than mothers; but as stress levels increased so did the severity of father's ratings, surpassing that of mothers. Parent stress has been positively associated with level of agreement on ADHD symptoms, while controlling for parental depressed mood in clinical samples (Oord, Prins, Oosterlaan, & Emmelkamp, 2006). Parent stress may be indicative of other known risk factors, including low SES. As a result, both parent stress and SES were included in the current study.

Informant Discrepancies as Predictors

In addition to conceptualizing informant discrepancies as dependent variables, newer research is treating discrepancies as independent variables that predict a variety of child social and behavioral outcomes. In a recent review, 11 studies from the previous 10 years were identified in which informant discrepancies were identified as predictive of a variety of outcomes, including delinquency, outcomes in treatment, and parent involvement in future

treatment (De Los Reyes, 2011). Two of these studies assessed discrepancies and outcomes between ratings of child psychopathology.

The two studies focused on informant discrepancies between parents and children using the CBCL and Youth Self-Report (YSR) in a community sample of Dutch adolescents and young adults. Discrepant ratings, with parents rating more problems than the children for attention problems, were predictive of a greater likelihood of expulsion from school, school discipline problems, referral for mental health services, and subsequent drug use or risk for drug use by the child (Ferdinand, Van der Ende, & Verhulst, 2004; 2006). These studies were limited, however, in that the parents were not explicitly identified as the mother or father and the use of raw discrepancy scores can be confounded by measurement error, as described previously.

An additional study (Ferdinand, Van der Ende, & Verhulst, 2007) explored parent-teacher discrepancies as predictive of child outcomes. Discrepancies were observed with higher parent ratings of anxious/depressed symptoms. These discrepancies were predictive of increased incidence of mood disorder in the child. Similarly, discrepancies for aggressive behavior (again, with higher parent ratings) was predictive of increased risk for suicide attempts and/or self-mutilation. These effects were above and beyond the effects of CBCL and TRF scores. The researchers proposed that the predictive utility of the discrepancies may be indicators of lack of home or school support and a resulting under- or over-estimation of the child's problems. Other explanations hypothesized included differences in contextual behavior and communication within dyads, potentially placing the child at risk for detrimental outcomes.

Summary

The study of informant discrepancies has seen rapid progress in the last 10 to 15 years. However, several areas of study need continued research to further inform assessment practice.

As stated by Dirks and colleagues (2012), research is still needed to “disentangle the relative contribution of (a) situational variability in behavior, and (b) rater-specific variables” (p. 568).

Multimethod-Multitrait Methods

The vast majority of studies describing discrepancies between mothers', fathers', and teachers' ratings of child and adolescent behavior have used observed scores. Hence, most studies have failed to consider the influence of measurement error. Additionally, many studies have ignored the influence of the informants themselves on the ratings (i.e., method effects). In other words, informants may influence test scores due to features of the measurement tool or trait content that results in a consistent response style (Campbell & Fiske, 1959). The few studies that have modeled method effects using latent variable models have focused on correlations between factors or percentages of variance (i.e., covariance structure), but these studies have seldom considered the mean structure (Geiser, Eid, West, Lischetzke, & Nussbeck, 2012). To truly understand the discrepancies between multiple informants (i.e., consistent with the study of discrepancies rather than measures of agreement), and the potential relation of the discrepancies with external variables, the structure of the data should be modeled more closely to the potential population model. The following review will consider the ubiquitous multitrait-multimethod matrix (MTMM; Campbell & Fiske, 1959) and review the development of increasingly sophisticated models from which to choose, including second-order CFA and method effect models.

MTMM matrix. Studies using multiple methods (operationalized as multiple informants as in the current study) often rely on the multitrait-multimethod (MTMM) matrix to describe the data (Campbell & Fiske, 1959). Campbell and Fiske's groundbreaking study has provided the basis for thousands of studies and is one of the most cited in modern psychology research

(Sternberg, 1992). The MTMM is a matrix of correlations between several different measurement methods measuring several different traits. In the MTMM model, the trait-method unit (TMU) is defined as one informant's ratings of one trait. It is assumed the same informant rated other traits, and that there are multiple informants. For example, a mother's rating of aggressive behavior is one TMU; her rating of attention problems is another TMU. More than one trait and more than one method (informant) must be used to study validity and the effects of different methods on trait measurement (e.g., Eid et al., 2008).

Campbell and Fiske (1959) sought to clarify validity criteria for tests within the MTMM matrix. Correlations between two or more measures within the MTMM provide different types of validity evidence: convergent and discriminant. Convergent validity is demonstrated by different measures of a single trait; these measures should strongly correlate. Discriminant validity is demonstrated by measures of different traits; they should correlate weakly, discriminating between traits (Widaman, 1985). An example MTMM matrix, using the different informants as methods and three traits to be used in the current study, is shown in Table 3. Reliabilities of the measures are on the diagonal.

Campbell and Fiske established one criterion for convergent validity and three criteria for discriminant validity.

Convergent validity:

- Correlations of the same trait measured by different methods (mono-trait hetero-method) should be significantly different from zero (i.e., substantial; see diagonal values labeled C. Validity).

Discriminant validity:

- Convergent validity correlations values should be greater than values not having method or trait in common (hetero-trait hetero-method; HTHM). At minimum, values related to different traits or different methods should be less than those measuring a common trait or using a common method (C. Validity values > HTHM values).
- Variables should correlate higher with other measures of the trait than measures of a different trait with the same method (i.e., C. Validity values > HTMM values).
- The pattern of correlations between traits should evidence the same pattern between different trait triangles and both the same method (mono-method) and different method (hetero-method) blocks.

Table 3

The multitrait multimethod correlation matrix

	Trait	Method 1 (Mother)			Method 2 (Father)			Method 3 (Teacher)			
		AGG	ATT	DEP	AGG	ATT	DEP	AGG	ATT	DEP	
Method 1	AGG	Reliability									
	ATT	HTMM	Reliability								
	DEP	HTMM	HTMM	Reliability							
Method 2	AGG	C. Validity	HTHM	HTHM	Reliability						
	ATT	HTHM	C. Validity	HTHM	HTMM	Reliability					
	DEP	HTHM	HTHM	C. Validity	HTMM	HTMM	Reliability				
Method 3	AGG	C. Validity	HTHM	HTHM	C. Validity	HTHM	HTHM	Reliability			
	ATT	HTHM	C. Validity	HTHM	HTHM	C. Validity	HTHM	HTMM	Reliability		
	DEP	HTHM	HTHM	C. Validity	HTHM	HTHM	C. Validity	HTMM	HTMM	Reliability	

Note. Adapted from Campbell & Fiske (1959). AGG = Aggressive Behavior; ATT = Attention Problems; DEP = Anxious/Depressed; HTMM = hetero-trait/mono-method; HTHM = hetero-trait/hetero-method; C. Validity = convergent validity.

In addition to these criteria for convergent and discriminant validity, the MTMM can provide preliminary evidence of method effects. In the MTMM, method effects can be observed in differences between correlations of corresponding values of blocks with one method and

blocks with multiple methods. Evidence of the method effect is observed when correlations between unrelated traits are strong within a single informant. The strength of the correlation is a result of the same informant providing ratings, not because the traits are strongly related.

To elaborate, according to classical test theory (CTT; Novick, 1966; Lord & Novick, 1968), test scores contain systematic variance that represent the traits they were designed to measure (i.e., true score variance) and other sources of variance that they were not designed to measure (e.g., measurement method or other unspecified sources of measurement error).

Although some of these other sources of variances may be error, they may also be related to the effects of specific informants, observed as method variance or method effects. The effects were considered irrelevant by Campbell and Fiske (1959) and, from their perspective, resulted in invalid scores. However, recent views consider these method effects to be quite relevant in both assessment and research (Achenbach, 2006; Pohl & Steyer, 2010). As a result, the MTMM methodology has continued to develop to provide a nuanced understanding of these method effects when using multiple informants' data.

CFA Models for MTMM Data

First-order MTMM-CFA. The traditional MTMM approach of Campbell and Fiske (1959), despite its widespread use, had several limitations. Widaman (1985) outlined three limitations: 1) the inappropriate and arbitrary nature of comparing of correlations, given that they were not independent of each other, 2) estimates of trait and method related variance were not able to be obtained, and 3) observed correlations were influenced by the reliability of measures.

The MTMM confirmatory factory analysis (CFA) approach was developed to overcome these limitations (Figure 1; Widaman, 1985). The use of MTMM-CFA require four data criteria: 1) the use of measures for at least three traits and three methods; 2) each TMU is defined by only

one measured variable and loads only on one factor; 3) one trait and one method factor is possible for each trait and method; and 4) correlations among trait and method factors are estimated, while correlations between them are fixed to zero.

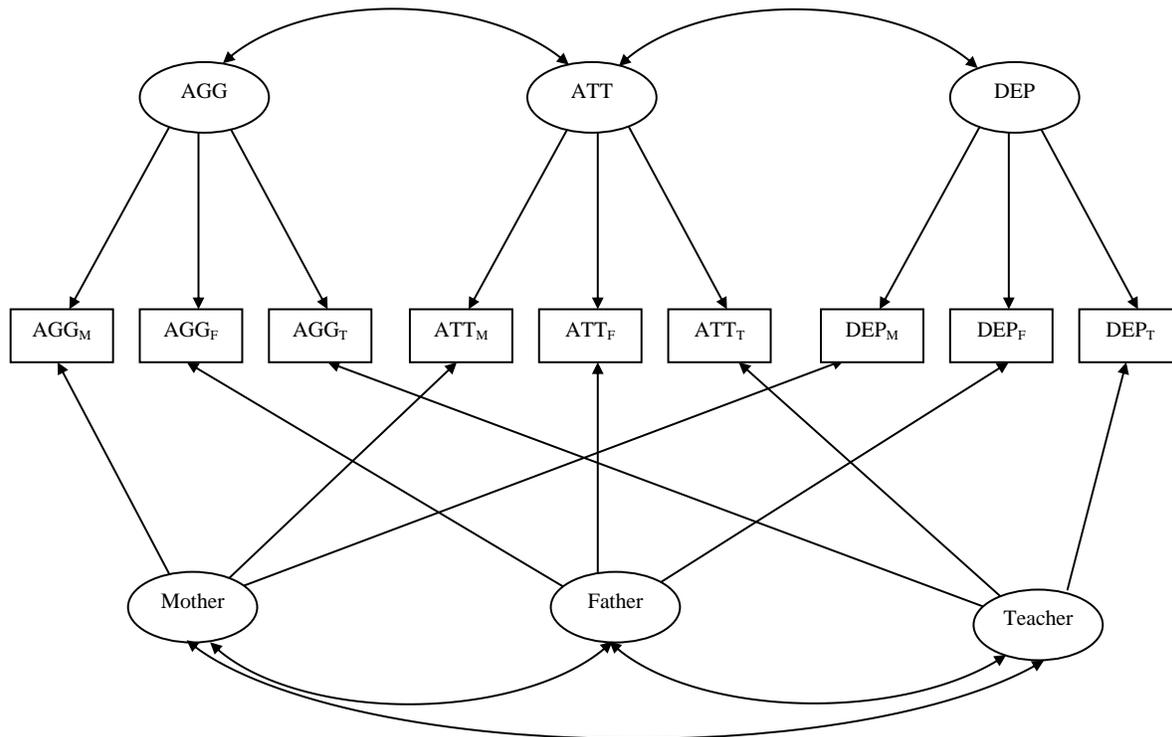


Figure 1. First-order correlated trait-correlated method (CTCM) model. AGG = Aggressive Behavior; ATT = Attention Problems; DEP = Anxious/Depressed; M = Mother; F = Father; T = Teacher. Note: Error variances are not pictured for clarity.

Even this development, however, has been criticized. Foremost in the criticism is that the use of single manifest variables (i.e., use of composite score rather than modeling the individual items) does not allow for disentangling different sources of variance, including random error and reliable specific variance (Marsh & Hocevar, 1988). The use of composite scores as indicators of trait and method effects combines the trait, method, and specific sources of variances into one observed score, effectively discounting their presence or potential influence.

Second-order MTMM-CFA. Marsh and Hocevar (1988) expanded on the MTMM-CFA approach (Widaman, 1985) by describing a higher- or second-order CFA approach to MTMM

data. This model was developed in response to several limitations of the MTMM-CFA, which used a first-order CFA model. First, the first-order MTMM-CFA approach used composite scores (e.g., sum of item responses). Because a composite score is used for each TMU, internal consistency (i.e., Cronbach's α) is typically unknown, although this could be estimated prior to summation using item-level data. Second, measurement error in the MTMM-CFA approach is estimated as the variance left unexplained by covariance among different informants' ratings; that is, the variance not in common between informants. This "error" therefore contains specific reliable variance that may be unique to one specific informant's ratings. Thus, it is not possible to separate specific and random error variance using a single composite score as an indicator for a TMU. Last, in the second-order MTMM-CFA approach, it assumed that a single factor underlies the composite score without actually testing whether that is true.

The second-order MTMM-CFA approach overcomes these three limitations (Marsh & Hocevar, 1988). In this approach, multiple manifest variables (single items or parcels) are used as indicators of first-order factors, resulting in one first-order factor for each TMU defined (e.g., three factors for each of three behaviors as rated by one informant; see Figure 2). These first-order factors are then used as indicators of second-order trait and method factors. Measurement error is the error associated with multiple scores from one informant, thus following an approach more closely aligned with classical test theory (Marsh & Hocevar, 1988). In addition, this approach allows for testing of the factor structure (i.e., significant and substantial loadings on the

latent factor), which is assumed in traditional MTMM and first-order MTMM-CFA approaches.

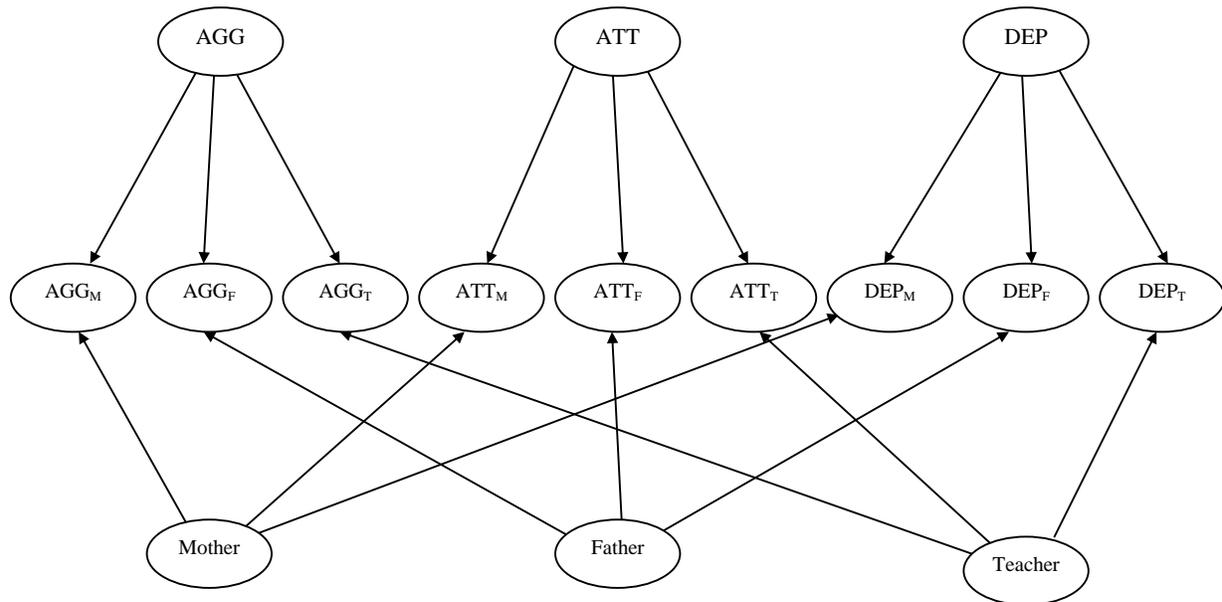


Figure 2. Correlated trait-correlated method (CTCM) model with latent first-order indicators of second-order trait and method factors. AGG = Aggressive Behavior; ATT = Attention Problems; DEP = Anxious/Depressed; M = Mother; F = Father; T = Teacher. Note: all trait factors are correlated with other trait factors (e.g., AGG with ATT); all method factors are correlated with other method factors (e.g., Mother with Father). First order factor correlations and trait-method correlations are fixed to zero. Each first order latent factor is indicated by 3 observed variables (not pictured; see Figure 1). Error variances, observed variables, and correlation paths are not pictured for clarity.

Model Specifications

In addition to the developments in the use of multiple indicators for each TMU, resulting in a second-order model, several specifications of models have been outlined in the literature. Three will be reviewed here: the correlated trait-correlated method (CTCM), correlated trait-correlated-method minus one [(CTC(M-1)], and the Method Effects with reference factor (MEref); the MEref was used in the current study. Each of these models can be used in a first- or second-order framework, depending on the specification of the model.

Correlated trait-correlated method model. The correlated trait-correlated method model (CTCM; Jöreskog, 1974; Widaman, 1985) has been used often in studies with MTMM data (e.g., Konold & Pianta, 2007). Some researchers argue that this model is in line with the

original intent of Campbell and Fiske (Castro-Schilo et al., 2013; Lance, Noble, & Scullen, 2002); others disagree (Geiser, Koch, & Eid, 2014). The CTCM model is specified such that all indicators are influenced by both a trait factor and method factor (Figure 2). The trait factors represent the common variance measured by the indicators from the different informants and the method factors represent the common variance from the informants by which they were obtained. For example, in a study with three informants (mother, father, and teacher) and three behaviors (aggressive behavior, attention problems, and anxious/depressed), the resulting model would have three method factors (informants) and three trait factors (behaviors). The mother's ratings of three types of behaviors would all load on the "mother" method factor. The three informants' ratings of aggressive behavior would all load on the "aggressive" trait factor. Trait factor correlations are freely estimated, as are method factor correlations. However, correlations between trait and method factors are fixed to zero to assist in model identification.

The CTCM model allows for the estimation of method, trait, and error variance when used in higher-order models, allowing for comparisons of each. Despite this strength, the CTCM model is not without limitations. Most notably, many authors have reported problems with estimating CTCM models, as high as almost 70% of models not converging in one recent example (Castro-Schilo et al., 2013), which may be due in part to over-parameterizing of the model (Geiser et al., 2013). The second limitation of the CTCM model is the unclear definition of the trait and method factors, with vague mathematical definitions (Geiser et al., 2012; Pohl & Steyer, 2012). Within this model, trait variance is that which is common to all variables measuring a trait with different methods; method variance is that which is common to all variables measuring different traits by a common method.

Correlated trait-correlated method minus one model. To address the limitations of the CTCM model, the correlated trait-correlated method minus one model (CTC[M -1]; Eid, 2000) and extensions to second-order models with multiple indicators (Eid, Lischetzke, Treierweiler, & Nussbeck, 2003) were developed. In the CTC(M-1) model, all measured trait factors are modeled, but one fewer method factor (i.e., M-1) is modeled compared with the CTCM model. Depending on the study questions, or for substantive and theoretical reasons, mothers or teachers are commonly selected as the reference to which other informant's ratings will be compared. One less method factor in the CTC(M-1) model allows for fewer identification and interpretation problems that are associated with the CTCM model (Eid, 2006). Similar to the CTCM model, the CTC(M-1) specification allows for specific measurements of method and trait variance, allowing researchers to compare the amount of variance related to each.

Trait factors in the CTC(M-1) model are defined as the true score of the manifest variables measuring the trait by the reference method (Pohl, 2012). The trait factor of the second method is regressed on the trait factor of the reference method. The method effect is the residual (error) of this regression; the difference between the second trait factor and the conditional expectation of the second trait given the value of the reference trait (Geiser et al., 2012). Method effects are therefore regression residuals and take on the typical properties of residuals (mean of zero, uncorrelated with the reference trait).

Strengths and limitations of the CTC(M-1) model. The second-order CTC(M-1) model decomposes indicator variance (observed scores) into method, trait, and error variance. Trait factor correlations are thus free of measurement error, providing more accurate estimates of convergent and discriminant validity than the observed correlations in traditional MTMM. The

correlation of trait factors provides a measure of discriminant validity compared to the reference method (i.e., weak correlations indicate discriminant validity).

Despite these strengths, CTC(M-1) models have limitations. For example, the correlation between trait and method effects is fixed to zero. This constraint is needed for the models to converge properly, not because of theoretical reasons (Lance et al., 2002). It is plausible that higher levels of method effects may be related to higher levels of traits (Pohl, Steyer, & Kraus, 2008). For example, a parent may demonstrate higher levels of method effects when rating their child's aggressive behaviors due to frustration with the child or a desire to describe the problem in severe terms so as to receive desired clinical assistance, consistent with the memory bias described in the ABC model (De Los Reyes & Kazdin, 2005).

Another potential limitation of the CTC(M-1) model is the definition of method factors as residuals (Pohl & Steyer, 2010; Pohl et al., 2008). Modeling the effects as residuals carries with it the assumption that the mean of the residuals is zero and the correlation of the method factor with the reference trait factor is also zero (based on the definition of a residual). In the unlikely event that all raters over- or under-estimate trait levels to the same degree the method effect would be zero. If a method effect is zero, it does not indicate that scores between two raters were the same. Instead, it indicates that the observed rating for the second informant does not differ from the expected level, given the reference ratings (Pohl & Steyer, 2010). Conceptually, it seems likely that researchers and clinicians would be interested in the actual differences, not differences from expected scores.

Method effect with reference method (ME_{ref}). Recent MTMM modeling developments have included further re-parameterization of the CFA-MTMM model, in part to address perceived limitations of the CTCM and CTC(M-1) models. One such model, the method

effects with reference method (M_{eref}, Pohl et al., 2008; Latent Difference model, Geiser et al., 2013) overcomes limitations of the CTCM and CTC(M-1) models through different definitions of both method and trait factors, and allows for different research questions to be addressed. The M_{eref} model was the primary model used to address questions in the current study. Similar to the CTC(M-1) model, a reference method is chosen.

Reference method factor definition. Several equations are used to define the method factor. For simplicity, the following equations are for one trait. A manifest indicator Y_{j1} of trait (j) using method 1 is defined as the sum of a true score variable τ_{j1} and an error term ε_{j1} , as prescribed by classical test theory.

$$Y_{j1} = \tau_{j1} + \varepsilon_{j1} \quad (1)$$

If the true scores between two methods measuring the trait are equal (e.g., mother and father's ratings of Aggressive Behavior), the two methods would be considered tau-equivalent and the method effects are zero (no informant discrepancies; Equation 2)

$$Y_{jk} = \tau_{j1} + \varepsilon_{jk} \quad (2)$$

If the true scores are not equal (i.e., there are method effects), then the score of the non-reference indicator can be represented by Equation 3:

$$Y_{jk} = \tau_{j1} + (\tau_{jk} - \tau_{j1}) + \varepsilon_{jk} \quad (3)$$

The term $(\tau_{jk} - \tau_{j1})$ is the method effect of using method k instead of the reference method ($k=1$) to measure trait j . This results in the method effect for trait j using method k being defined as:

$$M_{jk} = \tau_{jk} - \tau_{j1} \quad (4)$$

As seen in equation 4, the method effect represents *the difference in true score variables of the non-reference method and the reference method*. The method effect is a latent difference score

and can be conceptualized as the informant discrepancy in true scores of reference method ratings (mother's) and method k (either father's or teacher's ratings).

Trait factor definition. Trait factors in the MEref model represent two different constructs: the reference trait factor and non-reference trait factor(s). Again, a manifest variable Y_{jk} measures trait j using method k . The reference trait, Y_{j1} is again defined as a true score variable, as seen in equation 1. For non-reference trait factors, Y_{jk} is defined as a true score variable which takes into account the method effect, as defined, substituting equation 4 into equation 3 from the previous section:

$$Y_{jk} = \tau_{j1} + M_{j k} + \varepsilon_{jk} \quad (5)$$

Assumptions. Several assumptions were made in the initial specification of the MEref models described by Pohl and colleagues (2008):

- 1) The method effects loadings are initially assumed equal for each trait across the three traits (e.g., Aggressive Behavior, Attention Problems, and Anxious/Depressed).
However, this assumption can be relaxed to test the equality.
- 2) Error covariances are constrained to zero; true scores and method factors covariances with errors are also constrained to zero.
- 3) Covariances between the trait and method factors are freely estimated, meaning they do not have to be zero as is required in CTCM and CTC(M-1) models.
- 4) Means and variances of the trait or method factors are freely estimated. To identify the latent mean structure, the intercept of each marker variable is constrained to zero.

In addition to these assumptions, by definition (based in classical test theory) the method and trait factors do not correlate with the errors.

Second-order MEref model. Geiser and colleagues (2012) described an extension of the MEref model to a second-order model, with multiple indicators for each TMU (Figure 3).

First, each observed variable in a TMU is defined:

$$Y_{ijk} = \alpha_{ijk} + \lambda_{ijk} T_{jk} + \varepsilon_{ijk} \quad (6)$$

Where α is a constant intercept, λ is a factor loading, T is the common factor for all indicators in the TMU, and ε is the error term. Using MEref specifications for a single trait factor (dropping the j subscript), the non-reference factor is now defined the same as in equation 4, but with common trait factors instead of true scores for a single indicator:

$$T_k = T_1 + (T_k - T_1) \quad (7)$$

Substituting equation 7 into equation 6 results in the equation for each manifest indicator:

$$Y_{ik} = \alpha_{ik} + \lambda_{ik} T_1 + \lambda_{ik} (T_k - T_1) + \varepsilon_{ik} \quad (8)$$

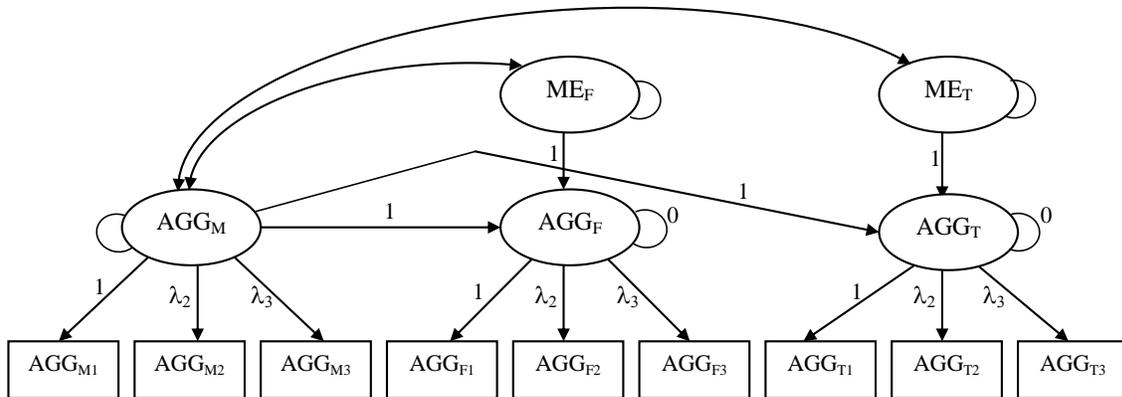


Figure 3. Second-order MEref model for Aggressive Behavior with mother as reference. AGG= Aggressive Behavior; M = Mother; F = Father; T = Teacher; ME = Method Effect. Latent means and variances of AGG_M, ME_F, and ME_T freely estimated. Correlations freely estimated. Strong invariance constraints are assumed. Reference indicator intercepts are fixed to zero. Indicator error variances and intercepts not included for clarity.

The MEref model assumes strong measurement invariance (i.e., equal factor loadings and intercepts) within traits across methods when using multiple indicators per TMU (Pohl & Steyer, 2010).

Strengths and limitations of MEref model. Pohl and colleagues (2008) identified several advantages in the use of the MEref model, as compared to CTCM and the CTC(M-1) models. First, method effects are clearly defined as a latent difference score (differences in true scores). Others have agreed, arguing that this definition is easily understood and intuitive (Geiser et al., 2012). Further, Pohl and colleagues argued their model improves upon the definition of method effects in the CTC(M-1) model (i.e., residuals) and the assumption of uncorrelated method effects and trait factors. The covariances between trait and method factors can be estimated in the MEref model, including between a trait factor and the related trait-specific method factor, unlike CTC(M-1); for example, as the behavior becomes more severe, discrepancies may become smaller.

Despite the strengths, Pohl and colleagues (2008) also identified limitations of the MEref model. First, the model does not allow for trait, method, and error specific variance to be explicitly decomposed as in the CTC(M-1) model due to estimating covariances between method and trait factors. Second, similar to the CTC(M-1) model, the MEref model requires a reference method to be chosen, which raises concerns described earlier in that no “gold standard” method exists. However, the model can easily be re-estimated with a different reference (e.g., teacher instead of mother) with the same model fit (Pohl et al., 2008). The researchers also acknowledged limitations of invariant method effects, as described in the initial first-order specification, as extremely restrictive. The authors suggested that these assumptions could likely be relaxed in future applications.

Summary. The historical development of several advances and specifications in a MTMM framework to analyze data were reviewed. MTMM-CFA models and second-order CFA models have provided the theoretical foundation for recent developments. Although

commonly used, the CTCM model is limited, particularly due to estimation problems. As a result, the CTC(M-1) and Meref model have been proposed in the methodological literature. Each has advantages and disadvantages for MTMM data analysis. The Meref model more closely aligned with the questions of the current study and the focus on true score differences (i.e., informant discrepancies) was advantageous.

Previous NICHD Studies

Three studies have previously used the National Institute of Child Health and Development (NICHD) Study of Early Child Care and Youth Development (SECCYD) CBCL and TRF data using a MTMM framework: Konold and Pianta (2007); Grimm, Pianta, and Konold (2009); and Castro-Schilo and colleagues (2013). The NICHD SECCYD data was used in the current research.

Konold & Pianta (2007) used a correlated trait-correlated method (CTCM) model to measure the effects of trait and method (informant) for CBCL/TRF. The authors selected a sample of first grade students, including only those with complete data from mother, father, and teacher ratings ($n = 562$). The study used composite scores from the Withdrawn, Somatic Complaints, Anxious/Depressed, Delinquent Behavior, and Aggressive Behavior narrow-band scales. Using a model similar to Figure 1, their model included five manifest variables (composite scores from each scale) as reported by the three raters. Each trait factor was indicated by three different informants and five method factors were indicated by each informant's ratings. The current study used two of the five behaviors used by Konold and Pianta (aggressive and anxious/depressed) and therefore only these results will be discussed at length. Attention problems were not studied by Konold and Pianta.

In Konold and Pianta's (2007) model, standardized factor loadings for trait factors were greater for teachers' ratings of Aggressive Behavior ($\lambda = .82$) than for mothers ($\lambda = .37$) or fathers ($\lambda = .29$), indicating a greater degree of convergent validity for teachers' ratings of Aggressive Behavior (i.e., higher loadings on the trait factors). Alternatively, Aggressive Behavior standardized factor loadings for *method* factors indicated greater loadings for mothers ($\lambda = .77$) and fathers ($\lambda = .72$) than for teachers ($\lambda = .29$). Standardized factor loadings on the Anxious/Depressed trait factor were higher for mother ($\lambda = .49$) and fathers ($\lambda = .41$) than teachers ($\lambda = .22$). Loadings on method factors were high for all three informants (mother $\lambda = .69$; father $\lambda = .70$; teacher $\lambda = .69$). In addition, method factors for parents were correlated ($r = .35$), while mother-teacher and father-teacher method factors were not. Although there was overlap between mothers' and fathers' ratings, a significant amount of variance was due to unique perspectives they bring to their ratings. Overall, method factors (i.e., method variance) had a great deal of influence on scores: method loadings were greater than trait loadings for 11 of 15 variables considered.

In the second study in this series, Grimm and colleagues (2009) used data from the first, third, fourth, and fifth grade data collections of the NICHD SECCYD. The researchers examined trait and method stability using a longitudinal CTCM model. Grimm and colleagues also modeled changes in traits using latent growth curve (LGC) modeling with manifest indicators, while controlling for method variance. Mother, father, and teacher report data on the Social Skills Rating System (SSRS; Gresham & Elliot, 1990) and the Internalizing and Externalizing broadband scales of the CBCL/TRF from the ASEBA were used. The CBCL/TRF data contained 93 like-items across both scales, excluding the additional 25 items specific to school.

Standardized factor loadings for trait factor externalizing behavior were significant for mothers ($\lambda = .58$ to $.64$), fathers ($\lambda = .49$ to $.57$), and teachers ($\lambda = .44$ to $.50$). Standardized trait factor loadings for internalizing behavior were statistically significant but consistently smaller compared to externalizing for mothers ($\lambda = .44$ to $.49$), fathers ($\lambda = .30$ to $.42$), and teachers ($\lambda = .30$ to $.37$). These differences across externalizing and internalizing behaviors may be due to the susceptible nature of ratings of internalizing for informant bias, as they are less readily observable (e.g., Loeber & Dishion, 1984). In addition, a large amount of method variance was observed longitudinally across the three informants based on absolute values of the factor loadings (mother $\lambda = .75$ to $.77$; father $\lambda = .76$ to $.86$; teacher $\lambda = .60$ to $.68$), which would include the method variance for social skills, internalizing, and externalizing behaviors.

The evidence from Grimm et al.'s (2009) study supported the previous findings of substantial amounts of method variance (Konold & Pianta, 2007). On average, 38% of the total variance in observed scores was due to the method; more than one-third of variance in observed scores was a result of the informant making the ratings, not the behavior being rated. Percentage of method and trait variance was not dependent upon the grade in which the assessment was completed, providing support for longitudinal consistency. That is, consistency among standardized trait and method factors' loadings indicated convergent validity of the different methods and method-based variance was similar in first to fifth grades.

In the third study from this series, Castro-Schilo and colleagues (2013) used a sample of children in first grade ($n = 775$) with data from mothers, fathers, and teachers on the CBCL or TRF. Data were modeled with the CTCM model, as well with a CTC(M - 1) model, a CTCU model, and a path model. Based on the CTCM model, mothers' trait factor loadings were consistently stronger than those of the other informants. For all three informants, loadings on the

externalizing trait factor were larger compared to internalizing trait factor, consistent with Grimm and colleagues (2009). Smaller loadings on method factors were reported for externalizing than internalizing behaviors, again supporting previous findings that readily observed behaviors may be less susceptible to method effects (Dishion & Loeber, 1984) and demonstrate higher agreement (Duhig et al., 2000).

Method effects exerted significant influence across informants and behavior type. Loadings for method factors were all significant for both externalizing and internalizing behavior (mother externalizing $\lambda = .67$, internalizing $\lambda = .79$; father externalizing $\lambda = .66$, internalizing $\lambda = .89$; teacher externalizing and internalizing $\lambda = .56$). Mother and father method factors were significantly related ($r = .33$), as were mother-teacher ($r = .14$) and father-teacher ($r = .15$) method factors, indicating a small degree of common influence in scores. The common influence was relatively stronger for those observing the child in a similar context (e.g., mother-father).

Castro-Schilo and colleagues (2013) also included explanatory variables to predict variance in the method factors and as outcomes using simulated data and then applied the same model to the NICHD data. Two conditions were tested: 1) positive associations between external variables (as outcomes or predictors) 2) positive associations between external variables and trait factors, but negative associations between external variables and method factors. Results from the CTCM model indicated both father's and mother's depression significantly predicted externalizing behavior and their respective method factors (i.e., mother's depression predicted mother method effects; father's depression predicted father method effects). Teachers generally rated boys' behavior with higher levels than girls. Boys' externalizing behavior was also rated significantly higher than girls.

Results from the CTC(M-1) model, with teachers as the reference group, shared similarities to the CTCM model. Mother and father depression was positively associated with externalizing behavior, and boys demonstrated higher levels of externalizing behavior. Again, both parents' depression predicted their respective method factor. However, differences were noted between the CTCM and CTC(M-1) model. For example, two relations observed in the CTC(M-1) model were not supported in the CTCM model: higher levels of externalizing behavior in African American children and father depression predicted mother method effects. Clearly, the type of model selected can influence results and their interpretation.

Summary. The three studies conducted using a MTMM framework with data from the NICHD SECCYD. First, consistent evidence was found for a substantial amount of method related variance, often more than trait related variance. Grimm and colleagues (2009), expanded on previous work by testing models longitudinally, confirming that the substantial amounts of method variance remain consistent across time. However, the use of CTCM model is not without its weaknesses, particularly with estimation problems: 27% of estimated models in the positive condition and 69% in the differential condition of the Grimm and colleagues study. This may lead to potential instability in the parameter estimates. Finally, Castro-Schilo and colleagues (2013) expanded the evidence to include predictors of the latent trait and method factors.

Importantly, none of these studies considered the mean structure of the data. That is, although they noted a substantial amount of method related variance, they did not consider how this translated to mean differences between raters. Additionally, only the first study (Konold & Pianta, 2007) considered narrow-band behaviors, albeit with composite indicators, which may confound error and method specific variance. Finally, longitudinally effects were part of only one study (Grimm et al., 2009) which used composite scores to model the MTMM data.

Importance of the Current Study

The current study aimed to provide at least four new clear contributions to the literature regarding informant discrepancies. First, the magnitude (i.e., size) and direction (i.e., which informant reports higher levels of behavior) of discrepancies between mothers, fathers, and teachers, is still uncertain. Previous studies of informant discrepancies have typically analyzed discrepancies based on observed scores, rarely considering the discrepancies via the latent mean structure (true scores free from measurement error; Geiser et al., 2012). A modern multi-method SEM technique, the Meref model, was used. Discrepancies from this model are clearly defined as the difference between true scores for a specified informant in comparison to the reference method.

Second, the current study aimed to expand current evidence regarding relations between demographic, contextual, and behavioral variables with informant discrepancies in the assessment of common childhood behaviors (aggressive behavior, attention problems, and anxious/depression) using theoretical frameworks as a foundation (MVT, Kraemer et al., 2003; ABC, De Los Reyes & Kazdin, 2005). Third, in addition to prediction of informant discrepancies, the current study explored evidence of the predictive utility of the discrepancies for relevant outcomes, including both clinical and school referral. Finally, the longitudinal consistency of the magnitude, direction, prediction, and predictive utility of the informant discrepancies was considered.

Chapter III: Methods

Participants

The National Institute of Child Health and Human Development Study of Early Childcare and Youth Development (NICHD SECCYD; NICHD Early Child Care Research Network, 1993) was an extensive longitudinal study conducted across 10 different sites affiliated with area universities in the United States: Temple University, University of Arkansas at Little Rock, Harvard University and Wellesley College, University of California-Irvine, University of Kansas, University of North Carolina-Chapel Hill, University of Pittsburg, University of Virginia, University of Washington-Seattle, and University of Wisconsin-Madison. The original study sample was recruited from hospitals near the 10 sites, beginning in 1991; the study concluded in 2007. Study inclusion criteria required mothers to be 18 years or older, speak English as a primary language, and live in a neighborhood considered safe enough for visits by study staff. Women were initially screened for eligibility following the birth of their child.

The initial screening was conducted with a sample of 8986 women, 5416 of which agreed to be telephoned two weeks after giving birth. Using conditional random sampling based on target goals of inclusion of at least 10% of families from variety of SES, family structure, and race categories, 3015 women were selected for a follow-up phone call. As a result of the telephone interview, 1525 women were considered eligible for the study based upon further inclusion criteria requiring the child to not have been hospitalized for more than seven days, stated intent of the family to stay within the geographical region for the next 3 years, and telephone contact was made in less than 3 attempts. Following this, 1364 mothers completed the one-month home interview, which marked the official start the study.

Original data collection occurred over a lengthy time period, from the age 1 month through 15 years of age for the study child, with a wide variety of behavioral, cognitive, and academic outcomes measured through various means such as parent, teacher, and child report, as well as direct observation and standardized direct assessment (e.g., cognitive and academic achievement). The current study used data from children in first, third, and fifth grade assessment periods. The current study sample ($n = 784$) included only participants with behavioral rating data available from their fathers at any of three assessment periods, similar to previous studies using the same data (Castro-Schilo et al., 2013).

The sample for the study included 400 boys (51%) and 384 girls (49%). The child's ethnicity as reported by mothers was 89.2% White, 4.5% African American, 6.1 % Hispanic, 1.4% Asian/Pacific Islander, 0.3% Native American, and 3.4% identified as "Other"². The education level of the mothers was 4.0% with less than a high school diploma or general equivalency degree, 17.6% with a high school diploma or G.E.D., 31.3% with some college or an associate's degree, 27.9% with a college degree, and 16.6% with some graduate work or master's degree, and 2.8% with a terminal degree higher than a masters (e.g., Ph.D., J.D., M.D).

Comparison of the available data at first grade ($n = 1134$) and the sample selected for the study (those with any valid fathers CBCL data in grades 1, 3 or 5) was conducted using a logistic regression. Hispanic ethnicity, minority status, income to needs ratio in first grade, and maternal education level were independent variables; participants selected as the subsample served as the dependent variable (0 = not selected for study sample, 1 = selected for study sample). Results indicated that income to needs ratio, mother's education level, and racial minority status were significant predictors; neither the child's sex nor Hispanic ethnicity were significant. Participants

² Total ethnicity percentage exceeds 100% due to separate questions regarding ethnicity and Hispanic backgrounds, allowing for membership in two categories (e.g., White and Hispanic).

with higher income to needs ratios, $OR = 1.229$; 95% CI [1.123, 1.345] were more likely to be included in the current study's sample. Participants from racial minority status, $OR = .287$; 95% CI [.195, .422] and lower levels of education, $OR = .344$, 95% CI = [.198, .598] were less likely to be retained in for the sample. As a result of these differences, generalizations of any findings to the original NICHD SECCYD sample are limited.

Measures

Child Behavior Checklist and Teacher Report Form. The Child Behavior Checklist (CBCL; Achenbach, 1991a) and the Teacher Report Form (TRF; Achenbach, 1991b) are widely used problem behavior rating scales for children ages 4 – 18. The CBCL is completed by parents. One hundred items are presented to informants to assess behaviors across eight narrow-band scales: Aggressive Behavior, Attention Problems, Delinquent Behavior, Social Problems, Anxious/Depressed, Somatic Complaints, Withdrawn, and Thought Problems. Two broadband scales: Internalizing (composed of Withdrawn, Anxious/Depressed, and Somatic Complaints) and Externalizing (composed of Aggressive Behavior and Delinquent Behavior scales) are also available when using standard scoring procedures. Informants responded to a short written description of behavior, and rated the study child on a three point scale (0 = *not true*, 1 = *somewhat or sometimes true*, or 2 = *very true or often true*). Higher composite scores indicated higher levels of the trait behavior.

The current study used responses from the study child's mother, father, and teacher on the Aggressive Behavior (AGG), Attention Problems (ATT), and Anxious/Depressed (DEP) narrow-band scales of the CBCL and TRF. Other alternate caregivers (e.g., grandparents as primary caregivers) completed assessments during the study; however, these data were not used to assist with generalization of findings and ease in future replication of the study due to more

clearly defined informants. All scales consisted of similar items across informants to provide for easier comparison as this eliminates differences in the number of items as a source of discrepancies (Treutler & Epkins, 2003). A wide body of evidence is available providing support for the reliability and validity of CBCL and TRF scores (e.g., Achenbach, 1991a and 1991b; Brown & Achenbach, 1993; Greenbaum & Dedrick, 1998). Evidence of both the predictive (e.g., Kasius, Ferdinand, Van den Berg, & Verhulst, 1997) and discriminant validity (Eiraldi, Power, Karustis, & Goldstein, 2000) of the scores on the narrow-band scales is available. Internal consistency (Cronbach's α) for the responses on the three scales by all informants for the current study sample were adequate (Table 4).

Table 4

Sample Internal Consistency (Cronbach's α) for CBCL and TRF Ratings (n = 784)

	Aggressive Behavior			Attention Problems			Anxious/Depressed		
	M	F	T	M	F	T	M	F	T
First Grade	0.84	0.83	0.90	0.75	0.72	0.85	0.73	0.76	0.71
Third Grade	0.84	0.79	0.91	0.78	0.76	0.87	0.79	0.77	0.74
Fifth Grade	0.85	0.87	0.91	0.77	0.77	0.85	0.78	0.81	0.74

Note. M = Mother; F = Father; T = Teacher

Aggressive Behavior. The Aggressive Behavior scale typically includes 20 items on the CBCL and 25 items for the TRF. This scale was reduced to 16 items common to both forms. Items measure various aspects of aggressive behavior, including aggression toward other people (e.g., threatening, attacking others, getting into fights) and toward objects (e.g., destroying other's belongings).

Attention Problems. The Attention Problems scale originally consisted of 11 items on the CBCL and 20 items on the TRF. This scale was reduced to 10 items common to both forms, measuring the child's concentration, impulsivity, and inattention. One item measuring nervousness is included on both the Attention Problems and Anxious/Depressed narrow-band

scales using standard ASEBA scoring procedures. This item was eliminated from analysis to reduce expected cross-loadings (i.e., significant factor loadings on two latent factors).

Anxious/Depressed. The Anxious/Depressed scale originally consisted of 14 items on the CBCL and 18 items on the TRF. Thirteen common items were used, measuring behaviors such as crying, feeling anxious, and feeling sad. The nervousness item included on both Attention Problems and Anxious/Depressed scale was eliminated from analysis.

Explanatory variables

Socio-economic status. Socio-economic status (SES) was measured by the income-to-needs ratio. This ratio was computed from information reported by mothers at each assessment period of the SECCYD via phone interviews and questionnaires. The ratio was computed by dividing total family pre-tax income, including governmental assistance, by the poverty threshold (determined by the year the income is earned, total number of household members, and the number of full-time children living in the home). The poverty threshold was obtained from the US Census Bureau Current Population Survey, as provided in study materials. Higher ratios indicated higher SES. SES is typically variable over time (i.e., income increases or decreases); therefore the income to needs ratio was computed at each assessment period to provide a more accurate representation of concurrent SES.

Ethnicity and race. Ethnicity and race was measured through the use of mother reported data obtained during the initial assessment at 1 month of age for the study child. Several different groups were part of the study; however only White ($n = 699$), African American ($n = 45$), and Hispanic ($n = 48$) groups were large enough to allow for group comparisons. Dummy coded variables were used for comparison: Whites were used as the reference group ($Whites = 0$), and compared to African Americans ($African Americans = 1$), or Hispanics ($Hispanics = 1$).

Parental depression. Mother and fathers completed the Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977), designed to measure depressive symptoms. Ratings were made using a four point scale based on how the informants felt in the last week (0 = *rarely or none of the time/less than once a week*; 3 = *most of or all of the time/5-7 days this week*). Four item scores were reflected. Symptoms assessed including feeling sad, crying, and feeling hopeful about the future. A composite score of these ratings was used, with higher scores indicative of a greater level of depressive symptomology. Ratings of 16 and higher are considered clinically significant based on criteria established on the CES-D. Item scores from the study sample demonstrated a high level of internal consistency (Cronbach's α for mothers = .90 to .91; fathers = .86 to .88).

Parental anger. Mothers and fathers completed ratings of anger in the last week on a 4 point scale (1 = *not at all* to 4 = *very Much*). The scale was titled "My Feelings II" for the purposes of the SECCYD; 10 items were used from the State-Trait Anger scale, including items measuring feeling angry, feeling mad, and feeling like yelling at someone (Spielberger, Jacobs, Russell & Crane, 1983). Higher composite scores indicated higher levels of anger and hostility. Item scores from the sample demonstrated a high level of internal consistency (Cronbach's α for mothers = .90 for all assessment periods; fathers = .90 to .91)

Parental anxiety. Mother and father ratings of anxiety in the last week were made on a 4 point scale (1 = *not at all* to 4 = *very Much*) from the "My Feelings II" scale. Ten items (different items than those measuring anger) from the State-Trait Anxiety Scale (Spielberger, Gorsuch, Lushene, Vagg, & Jacobs, 1983) were used to create a composite score. Items included feeling worried, feeling tense, and feeling at ease (reverse scored). Item scores from the study

sample demonstrated high levels of internal consistency (Cronbach's α for mothers = .86 to .87; fathers = .84 to .86).

Maternal stress. The study child's mother self-reported stressful life events using an adapted version of the Life Events Survey (Sarason, Johnson, & Siegal, 1978) at both the third and fifth grade assessments. The survey was not administered in first grade. Mothers responded to a list of 71 life events, based on their occurrence in the last year. If the event occurred, mothers rated the severity of the impact of the event on a 7 point scale ($-3 = \textit{extremely negative}$; $0 = \textit{no impact}$; $3 = \textit{extremely positive}$). Ratings for the scale demonstrated adequate levels of internal consistency for the entire study sample (Cronbach's $\alpha = .76$ to $.83$). Two ratings were used: sum of negative event ratings and sum of positive event ratings to help differentiate between positive and negative stress in mothers' lives.

Classroom Observation Scale. The Classroom Observation Scale (COS) was developed by the Steering Committee for the SECCYD and was based on a kindergarten observation system. The COS-1, or first grade, was revised extensively for third grade (COS-3), and again slightly for fifth grade (COS-5). The COS provided a direct observation measure of the study child's classroom and was designed to document child behavior, teacher behavior, and overall instructional environment.

The COS was completed by trained study personnel in the spring in each child's classroom. Visits were allowed to be scheduled over 2 days if needed. Observations were completed when the student's regular teacher was present, and the student must have been attending the classroom for at least 2 weeks. Extensive procedures were made available for study personnel to ensure consistency across study sites and personnel, including decision rules when unforeseen circumstances necessitated deviation from expected procedures. Observations

were generally expected to be completed in the morning and to be of sufficient length (> 2 hours) to complete observation cycles (e.g., study child not leaving classroom for other activities).

Teachers were compensated \$50.00 (USD) for the participation in the school observation, and \$50.00 for completion of questionnaires, including the TRF.

The COS used codes for both discrete behaviors and a more subjective, qualitative coding system. Characteristics were rated on a seven-point scale from the qualitative observation (1 = *low or very uncharacteristic of that child or setting*; 7 = *high or very characteristic of that child or setting*). During the qualitative classroom observation, 10 minutes were spent making overall observations of the child, teacher, and classroom environment. This observation was part of the lengthier observation, which typically totaled 160 minutes. Qualitative observation ratings were made following this 10 minute observation. Although this 10 minutes was dedicated to the qualitative ratings, study materials indicated that “everything that occurs” during the total observation cycle could be included to make the rating. Two composite scores were used from COS data: overall Classroom Environment and Teacher Sensitivity.

Classroom environment. Classroom Environment included study personnel’s ratings of positive classroom climate and reflected ratings of negative classroom climate. A positive classroom environment was defined in SECCYD materials as a “safe and respectful environment” in which there is positive emotional support and feedback for the children, and respect for others’ is expected. A negative climate was defined as “hostile, angry, punitive, and controlling,” based on interaction between students and the teacher, as well as overall classroom characteristics.

Teacher sensitivity. Teacher sensitivity was composed of study personnel’s ratings of teacher sensitivity and reflected teacher detachment. Sensitivity was defined in study materials as

“child-centered behavior” from the teacher, including “awareness of each child’s needs, moods, interests, and capabilities.” Detachment was defined as reflecting “a lack of emotional involvement and a lack of awareness of the children’s needs for appropriate interactions with activities, materials, or peers.” Teachers with high detachment ratings were rarely involved with children’s activities or conversations; conversely, low ratings indicated a high level of involvement.

Home environment. Home observations were completed by study personnel based on the H.O.M.E. observation system (Caldwell & Bradley, 1984). Item ratings were made by study personnel based on direct observation and semi-structured interview. Trained study personnel were required to achieve 90% agreement on video tapes prior to actual data collection in the field. The Home Environment score was the sum of four similar items across the two observation periods. Items indicated that the home has no structural or health hazards, there is at least 100 feet of space per person, the home is clean, and the environment is not dark or monotonous in color. Higher scores indicated a healthier or safer physical home environment.

Parental sensitivity. Parental sensitivity was measured by direct observation ratings made by SECCYD study personnel. Ratings were derived from structured interaction activities. These activities occurred with both parents as part of a home visit in first and third grade. Father-child activities in fifth grade were completed in the home and mother-child activities were conducted as part of the laboratory visit in fifth grade. Videotapes of children and their parents were scored on several criteria by study personnel on a 7 point scale (1 = *very low* to 7 = *very high*). A composite score was used based on ratings of both mothers’ and fathers’ behavior. Items included ratings of the supportive presence of the parent, respect for the child’s autonomy, and hostility toward the child (reverse coded). Higher ratings indicated more positive care

giving. Item scores had an adequate level of internal consistency according to study materials (Cronbach's α for mothers = .78 to .85; fathers = .68 to .79).

Extensive description of procedures for the structured interactions was outlined in study materials. Study personnel were required to complete pilot activities to become certified prior to live data collection. Certification was completed by Dr. Ann Ware. The interactions were expected to take approximately 20 minutes. Explicit verbal instructions were provided for each activity to be read by study personnel. During the visits, children earned tickets for each activity they completed, which could then be exchanged for a small gift. Extensive descriptions of ratings and operational definitions of behaviors were made available in study materials. Families were paid \$75 (USD) following completion of home visit activities, including the structured father-child interaction activity in fifth grade. The child was paid an additional \$20 and an additional gift was given at the discretion of each site. During the lab visit, mothers were compensated \$75 and the child \$25.

Three activities were completed as part of the structured interaction during the first grade assessment. First, the mother or father and child operated an Etch-A-Sketch to draw a picture of a sailboat, based on a picture provided by study personnel. One person controlled one knob each, requiring them to coordinate their efforts to draw the picture. Second, the mother or father engaged in a block building activity to replicate designs pictured on Color Cube Task Cards. The parent was encouraged by study personnel to allow the child to attempt to accomplish the task on their own first, and then provide help as needed. Third, the mother or father and child engaged in the playing card game "Slap Jack" in which the participants slapped a "Jack" when turned face up to win the pile of cards below it.

Tasks differed in third grade compared to first grade. Only two tasks were completed by each dyad. In the first task, a rules discussion between parent and child was elicited. This task involved the parent (mother first or father first was counterbalanced by site) and the child discussing family rules. Participants were provided three piles of cards with rules listed on the back of the cards. These piles were identified as rules for kids, rules for parents, and rules of right and wrong behavior. The participants were provided with a spinner which indicated from which pile a card would be selected. The participants were instructed to subsequently discuss what both parent and child thought about the rule. They were told by study personnel that there were not correct answers to the rules. Participants were allowed 7 minutes to complete the activity; study personnel left the room during this task (all tasks were videotaped).

The second task was an errand planning task, completed only by the mother-child dyad. In this task participants were provided a map of a fictional town, with a starting and finishing point labeled “home.” Participants were instructed to complete the list of “errands” to complete using the best route through the town, using a car game piece, and to write down the order with which they completed the tasks. Participants were allowed approximately 8 minutes to complete the task.

The discussion task was also completed with the father, but with different rules on the cards used. The father then completed a different second task than the first. This task involved sorting 18 cards shuffled together that told three different stories. Participants were expected to sort the cards together by story, then to sort them into order telling a story from start to finish. The child was expected to take the lead, with the father providing support as needed. Eight minutes were allowed to complete the task.

In fifth grade, the father-child interaction was completed at home, and the mother-child interaction at the laboratory setting. Two tasks were completed during the each dyad's interaction. The first task (completed by both father-child and mother-child dyads) required participants to discuss issues that are typical sources of conflict in a family, such as homework or chores, using a deck of shuffled "family issues" cards (22 in the set), with brief statements or words written on them. The father and child were instructed to decide together the top 3 topics of disagreement for the dyad, talk about the topics, and try to resolve difficulties for each one. The participants were allowed 7 minutes and study personnel left the room.

The second activity completed by father-child dyads required participants to create a 1 foot tall tower using materials provided by study personnel (1 oz. of "Model Magic", 100 toothpicks, 4 tongue depressors, 4 rubber bands, and 1 12" ruler). Participants were instructed to work together to figure out how to build the tower. They were allowed 7 minutes to complete the task and study personnel left the room.

The second activity completed by mother-child dyads required participants to create a "bungee jump" for a raw egg. Participants were given the following materials: a raw egg, nylon panty hose, a plastic egg, 40 pennies, a 12" ruler, scissors, paper towels, newspaper, roll of masking tape, and 4.5 gallon plastic storage box. Eggs were dropped from a pre-constructed PVC pipe frame approximately 35" by 20" by 17¼", and the interaction was conducted from a low table, approximately 24" high. Children were provided with an instruction card prior to beginning the activity. Participants were allowed 7 minutes to complete the activity.

Child Outcome Variables

School-based special services. The study child's teacher completed a survey including information reporting whether the child received or was to be referred for special services in

school. This broad conceptualization of special services included special education, Title I or other federally funded services, tutoring, social services, occupational or physical therapy, or state funded services. Children who have previously received, currently received, and were to be referred for services were coded as dichotomous variables (0 = *did not receive/will not be referred*; 1 = *received/will be referred*).

Diagnosed learning disability and attention, behavior, or emotional problems.

Information regarding the child's general health was obtained from mothers or primary alternate caregivers during phone interviews. As part of this interview, the respondent was asked if anyone has suggested that they seek professional assistance to address learning problems, emotional problems, or problems with attention, school work, or behavior. If the respondent answered yes, follow up questions were asked. Two questions from this interview were used: 1) If the child has ever had a reading or learning disability as determined by a professional, 2) If the child has ever had an attention, behavior or emotional problem as determined by a professional. Variables were dichotomously coded (0 = *no identified problem*; 1 = *identified problem*).

Missing Data

Missing data is typically characterized by one of three mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), which describe the degree to which researchers understand why the data is missing. A test of MCAR is available (Little, 1988) and was used prior to analysis. It was likely, particularly given the longitudinal nature of the study, that the null hypothesis of MCAR would be rejected (Kline, 2011). Formal tests of MAR and MNAR are not currently available. However, modern data analysis techniques are available to appropriately handle missing data, including the use of maximum likelihood estimators (ML) and related estimators for non-normal data (MLR). In

addition, other techniques can be used, including logistic regression with a binary variable (0 = *active/included in sample*, 1 = *not active/not included in sample*), to compare those selected for the sample with those not selected to determine characteristics associated with leaving the study. This information may be used to inform the assumption of MAR, which assumes that data are missing completely at random after conditioning on variables included in the model.

Analytic Plan

The analysis occurred through several steps: 1) measurement models were used to ensure unidimensionality of latent trait factors 2) continuous item parcels were expected to be created to simplify the models, substantially reducing the number of parameters estimated 3) measurement invariance was tested across raters and time to ensure measurement of the same trait 4) a series of MTMM models, using MRef specifications, were estimated to answer several questions regarding method effects (i.e., informant discrepancies) 5) independent and dependent variables were included in the analyses, to help determine predictors of discrepancies and the predictive nature of the discrepancies. Each step of the analysis is described in further detail below.

Measurement models. Measurement models were fit to the data for each informant, for all traits, within each grade to ensure unidimensionality (i.e., measure a single construct) of latent trait factors. The factor loading for the first indicator was fixed to 1 to scale each latent trait factor. This *reference indicator* scaling technique results in the latent factor measured in the same scale as the reference indicator (Brown, 2006). All remaining manifest indicators' factor loadings were freely estimated on one latent factor based on the expected structure (e.g., aggressive behavior indicators will load on the aggressive behavior latent factor); factor loadings on other latent trait factors were fixed to zero so all variables would load on one factor. All factor variances and indicator thresholds were freely estimated. The measurement model is

shown in Figure 4. Due to the use of categorical manifest indicators, the weighted least square (WLSMV) estimator was used for the measurement models (Muthén & Muthén, 1998-2010). All items were expected to have statistically significant ($p \leq .05$) and salient standardized factor loadings ($\lambda \geq .40$; Brown, 2006) providing evidence of a meaningful level of relation to the latent trait. Items that did not meet these criteria, or demonstrated significant cross-loadings on two factors, as indicated by modification indices, were eliminated from subsequent analysis. Assuming unidimensionality of latent trait factors was tenable, parcels were planned to be created.

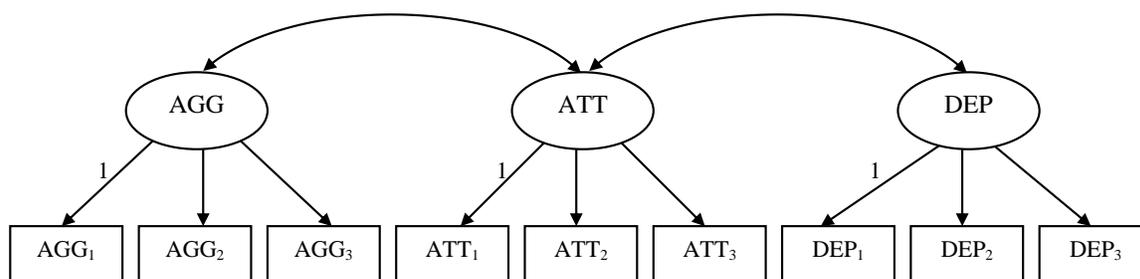


Figure 4. Measurement model. AGG= Aggressive Behavior; ATT = Attention Problems; DEP = Anxious/Depressed. All paths freely estimated unless noted. Indicator error variances not included for clarity. Only three indicators per latent factor included for clarity. The actual model includes: AGG = 16 indicators; ATT = 10 indicators; DEP = 13 indicators.

Parcels. Three parcels for each trait method unit were planned to be constructed using the *balancing approach* (Little, Rhemtulla, Gibson, & Schoemann, 2013). In this approach items are assigned to one of three parcels, based on factor loadings from the measurement model analysis. Loadings from mother's ratings for first grade were planned to be used for parcel assignment for all three assessment periods and all three informants to maintain consistent items in each parcel across time and informant.

Parcel assignments were planned to be made as follows: the item with the highest factor loading would be paired with the item with the lowest factor loading; the item with the second

highest factor loading would be paired with the item with the second lowest factor loading; and so on until all items were allocated to the three parcels for each trait method unit. Items would then be summed to create parceled indicators. Although parcels are not without their shortcomings (Marsh, Lüdtke, Nagengast, Morin, & Von Davier, 2013), the testing of the unidimensionality of the items prior to parceling was expected to address potential concerns of cross-loading items (i.e., items with salient and statistically significant factor loadings on more than one latent trait) and correlated error variances. Additionally, parcels have the advantage of simplifying complex models with large number of items by reducing the number of parameter estimated, potentially increasing reliability, and normalizing potentially non-normal distributions (Little et al., 2013).

Measurement invariance. Following consideration of planned assignment of items to parcels, measurement invariance was tested. Measurement invariance testing explicitly tests the similarity of measurement properties of latent factors across different groups, including different raters or different assessment times, as seen in longitudinal analyses. Testing invariance allows for comparisons to be made across groups, as it provides evidence that the assessment is measuring the same construct in each group. Further, comparison of group means has been noted to be meaningful only in instances when factor loadings and indicator intercepts are invariant (Brown, 2006). Although tests of strict invariance (i.e., equal residual variance) can also be conducted, this test has been described as too stringent in applied research (Brown, 2006; Little, 1997). Tests of invariance occurred both within time (between raters) and across time (within construct).

A specific sequence of constraints in the CFA model can be used to test for measurement invariance (Brown, 2006; Meredith, 1993). These constraints are imposed on the model to test

specific questions. The first step, *configural invariance*, tests whether the basic latent structure is the same across groups, including which manifest indicators load on each latent factor. The second step, *weak factorial invariance*, includes the model structure tested in the configural model, as well as additional equality constraints of like-factor loadings across groups. For example, the second factor loading on factor 1 was equal in the first and second assessment periods. Third, tests of *strong factorial invariance* involve the equality constraint of like indicator intercepts (and allowing for factor mean differences), while maintaining constraints of weak factorial invariance if tenable. Each model is nested within the previous one, allowing for the use of nested model comparisons to assess changes in model fit, guiding decisions as to which model to retain as the most appropriate representation of the data (“best-fitting”), as well as the most parsimonious model.

Method Effect with Reference MTMM Models

Following the estimation of the measurement and invariance models, the M_{ref} MTMM models were tested in a sequence to test the primary study questions. All models continued to use strong invariance constraints (i.e., equal factor loadings and intercepts across informants, within each trait) in the measurement part of the model if tenable. The three latent traits: Aggressive Behavior, Attention Problems, and Anxious/Depressed, were to be indicated by 3 parceled indicators. The mothers’ ratings were the reference method. Models were identified by fixing the first loading for each TMU to 1 (i.e., reference indicator identification). The latent mean structure was identified by fixing the intercept or first threshold with item-level data of each reference indicator to zero.

The baseline second-order M_{ref} model, described subsequently, is shown in Figure 3 (previously presented on pg. 49). Only one trait is shown due to space constraints. As seen in

Figure 3, each non-reference trait factor (e.g., AGG_F and AGG_T) was regressed on the reference trait factor, with the regression paths constrained to 1. The non-reference trait factor was regressed on its corresponding method effect factor, with the regression path constrained to 1. The variance of the non-reference trait factor was constrained to 0. The latent means and variances of the reference trait factor and method effect latent factor were freely estimated. The method effect factor is the difference between the reference and non-reference trait factor means. All method-trait factors correlations, method-method factor correlations, and trait-trait correlations were freely estimated.

MTMM Model Hypotheses

Question 1: What is the informant discrepancy, represented by method effects, for each informant's ratings of the child's behavior?

Hypothesis 1: Method effects using the Meref model were represented by the mean values of the method factor for each non-reference method. Father method effects were expected to be small, but statistically significant. Teacher method effects were expected to be larger than father method effects and statistically significant. Both father and teacher method effects were expected to be negative in direction, indicating that teachers and fathers endorsed lower levels of behavior than mothers.

Question 2: Are these method effects trait-specific? That is, is the size of informant discrepancies dependent upon the trait measured?

Hypothesis 2: Method effects were expected to be different within informant (mother, father, and teacher) across traits (aggressive behavior, attention problems, and anxious/depressed). Aggressive behavior and attention problems were expected to have smaller method effects compared to anxious/depressed. If the method effects were consistent across traits, it would

indicate the presence of a general “halo effect” in which an informant is consistently higher or lower in their ratings, compared to other informants. This question was tested by comparing a model with means constrained equal across traits to one in which they were freely estimated.

Question 3: Are the size of method effects related to levels of trait behavior?

Hypothesis 3: It was expected that the covariance/correlation between method effects and trait factors would be negative in direction. That is, children with more severe behaviors were likely to be rated in a similar manner (i.e., smaller method effects) than those with more typical levels of behavior.

Question 4: Do method effects remain constant over time?

Hypothesis 4: The method effect means and variances were expected to remain the same across time (first, third, and fifth grade). Specifically, the means and variances of the method effects were constrained to be equal across three measurement periods, using multi-group modeling.

Explanatory and Child Outcome Variables. Following determination of the appropriate model based on fit indices and statistics, models with covariates were estimated using the selected model. Method and trait factors were regressed on these explanatory predictors and conducted in a series of models.

- 1) Inclusion of demographic variables: SES, ethnicity/race, and sex.
- 2) Inclusion of maternal-rated stress and maternal and paternal psychopathology (depression, anger, and anxiety).
- 3) Inclusion of independent ratings of informant sensitivity to the child.
- 4) Inclusion of contextual variables: home and classroom environment scores.
- 5) Regression of receiving special school services, diagnosed learning disability, and diagnosed attention/behavior/emotional problems on latent trait and method factors.

Question 5: Are method and trait effects predicted by SES, ethnicity or race, and sex?

Hypothesis 5a: Lower SES was not expected to be predictive of larger father method effects.

Lower SES was expected to be predictive of larger teacher method effects. Lower SES was also expected to predict higher levels of trait behavior across all informants.

Hypothesis 5b: Similar to SES, ethnic or racial minority status was expected to predict larger teacher method effects. It was not expected to predict father method effects.

Hypothesis 5c: Sex effects were expected to be found, with boys demonstrating higher levels of trait attention problems and aggressive behavior, and girls demonstrating higher levels of anxious/depressed. The influence of sex on the method effects was expected to be non-significant for father's and teacher's method effects for aggressive behavior and attention problems and negative for girls' anxious/depressed (i.e., mothers will rate girls higher).

Question 6: Are method and trait effects predicted by maternal stress, and maternal and paternal depression, anger, and anxiety?

Hypothesis 6: Maternal stress, and maternal and paternal depression, anger, and anxiety were all expected to be related to larger method and trait effects.

Question 7: Are method and trait effects predicted by independent ratings of parent and teacher sensitivity?

Hypothesis 7: Independent ratings of parent and teacher sensitivity were expected to predict larger method effects. That is, those that were observed to be less sensitive rated the child more harshly.

Question 8: Are the method effects predicted by ratings of the context in which they occur?

Hypothesis 8: Lower classroom and home environment ratings (i.e., less supportive or safe) would be predictive of larger trait and method effects.

Question 9: Are the method effects predictive of referral to special school services, diagnosed learning disability, and attention/behavior/emotional problems?

Hypothesis 9: Larger method effects were expected to predict a greater likelihood of referral for special school services, diagnosed learning disability, or attention/behavioral/emotional problems. As these outcomes are dichotomous, these analyses were logistic regressions. As a result, the regression coefficients were interpreted as odds ratios, with ratios greater than 1 indicating an increase in the odds of the outcome; ratios less than 1, indicating a decrease in the odds of the outcome (Tabachnik & Fidell, 2007).

Model Evaluation

Model fit. Several methods are available to evaluate model fit based on different criterion in an SEM framework. Model test statistics provide a test that can be used to determine how close the model implied sample covariance matrix fits the data in the population ($S = \Sigma$; Kline, 2011). The most commonly used model test statistic is the model chi-square (χ^2), or the likelihood ratio chi-square. Chi-square is calculated by multiplying the maximum of the fit function (F_{ML}) by the sample size (N); or ($N-1$) in Mplus. A model $\chi^2 = 0$ indicates perfect fit of the covariance matrix, with values greater than zero indicating increasing degrees of model implied misfit. Model χ^2 , although commonly used, is not without limitations, including being influenced by multivariate non-normality, unreliability of indicators, and sample size (Kline, 2011). A robust maximum likelihood estimator (MLR in Mplus) was planned to therefore be used. Non-normality typically results in larger χ^2 values and in smaller standard errors. The MLR corrects these.

In addition to χ^2 , other methods of model evaluation are available. The standardized root mean square residual (SRMR) is similar to the χ^2 in that it compares the sample and population

correlation matrices. The SRMR is the “average discrepancy between the correlations observed in the input matrix and the correlations predicted by the model” and is computed from the residual correlation matrix (Brown, 2006, p. 82). Values for the SRMR can range from 0 to 1, with values closer to 0 indicating better model fit.

The root mean square error of approximation (RMSEA; Steiger, 1990) is calculated by taking the square root of $(\chi^2 - df)/N$ divided by df . Mplus used N instead of $(N - 1)$, as is also used in practice. As can be seen through examination of the equation, RMSEA involves a correction for the number of parameters. Values for the RMSEA range from 0 to 1, with values closer to 0 indicating better model fit.

The final fit index used was the comparative fit index (CFI; Bentler, 1990), which compares the implied model to a null model. Values can range from 0 to 1, with values closer to 1 indicating better model fit. General guidelines available for acceptable model fit include SRMR values $\leq .08$, RMSEA $\leq .06$, and CFI $\geq .95$ (Hu & Bentler, 1999).

Model comparisons. Nested model comparisons were made using the χ^2 difference test, which compares the χ^2 values between two models based on the change in degrees of freedom (Δdf). Critical χ^2 values of $\alpha \leq .05$ was used as criteria for statistically significant change in model fit. In the event of non-normality of indicators, which is common with behavioral ratings, a correction would be applied to adjust for the non-normality. This Satorra-Bentler adjusted χ^2 (SB χ^2) divides the χ^2 value by a correction factor and can be used in conjunction with the robust maximum likelihood estimator (MLR in Mplus; Satorra & Bentler, 2001). An additional measure of model comparison was used for the tests of measurement invariance, ΔCFI (Cheung & Rensvold, 2002). Criterion was set at $\Delta CFI \geq .01$ as indicating significant change in model fit.

Chapter IV: Results

Missing Data Analysis

Analysis of CBCL and TRF sum scores indicated the amount of missing data ranged from 7.0% of data with mothers' ratings in first grade, to a high of 19.4% fathers' ratings in fifth grade. Additional analysis showed that 15.9% of sample mothers, 37.0% of fathers, and 26.8% of teachers were missing CBCL/TRF data for at least one time point. Little's MCAR test (Little & Rubin, 2002) indicated that the null hypothesis was rejected, meaning data were likely not missing completely at random, $\chi^2(2254) = 2592.535, p < .001$. This result can be expected in applied data sets (Kline, 2011), particularly in longitudinal studies, and given the large sample size.

Based on logistic regression, no demographic predictors (child's sex, Hispanic, African American, or SES) were predictive of missing data at any of the three assessment periods for mothers' behavior ratings. However, statistically significant ($p < .05$) influences were found for fathers' and teachers' behavior ratings. For fathers' behavior ratings, non-African Americans were more likely to have data missing, Odds Ratio (OR) = .454; 95% CI [.229, .990]. For teachers' behavior ratings, females $OR = .1433$; 95% CI [1.004, 2.045] and lower income families $OR = .934$; 95% CI [.874, .999] were more likely to have data missing. However, because some mechanisms of missingness were accounted for by using these variables in the models, analysis was continued under the assumption data missing at random (MAR), using all data.

Descriptive Statistics

Observed sum scores for each informant and behavior were computed (Table 5). Observed mothers' and fathers' mean ratings of aggressive behavior decreased longitudinally;

teachers' ratings were longitudinally stable. All informants' mean ratings of attention problems demonstrated a slight increase from first to third grade, followed by a decrease from third to fifth grade. Mothers' ratings of anxious/depressed increased consistently over time, while both teachers and fathers again demonstrated a peak in third grade, similar to attention problems, followed by a decline in fifth. Variability of scores remained relatively consistent for all behaviors across the measurement periods.

Table 5

Descriptive Statistics for Observed Sum Scores of CBCL and TRF Scales

Grade	Mother			Father			Teacher		
	First	Third	Fifth	First	Third	Fifth	First	Third	Fifth
<i>N</i>	729	727	722	668	636	632	710	692	660
<u>Aggressive Behavior</u>									
Mean	4.86	4.35	3.96	5.13	4.17	3.66	2.26	2.43	2.39
<i>SD</i>	4.02	3.84	3.89	3.97	3.34	3.96	3.76	3.89	3.97
Skew	1.26	1.29	1.45	1.08	0.93	1.78	2.43	2.70	2.62
Kurtosis	2.00	1.93	2.73	1.16	0.79	4.12	6.18	9.33	8.05
Minimum	0	0	0	0	0	0	0	0	0
Maximum	26	23	25	22	19	25	21	27	26
<u>Attention Problems</u>									
Mean	2.28	2.31	2.09	2.40	2.45	2.04	2.57	2.86	2.57
<i>SD</i>	2.36	2.51	2.50	2.30	2.41	2.40	3.41	3.65	3.29
Skew	1.38	1.49	1.56	1.42	1.19	1.66	1.64	1.59	1.45
Kurtosis	2.19	2.80	2.68	3.11	1.45	3.50	2.29	2.25	1.49
Minimum	0	0	0	0	0	0	0	0	0
Maximum	13	15	15	15	13	14	16	19	16
<u>Anxious/Depressed</u>									
Mean	2.33	2.45	2.50	2.23	2.25	2.06	1.55	1.94	1.77
<i>SD</i>	2.39	2.63	2.69	2.50	2.53	2.71	2.06	2.31	2.25
Skew	1.54	1.47	1.84	1.99	1.46	2.30	1.93	1.66	2.07
Kurtosis	2.76	2.33	5.76	5.78	2.01	6.96	4.47	3.94	5.42
Minimum	0	0	0	0	0	0	0	0	0
Maximum	13	15	22	17	13	18	13	16	14

Note. Aggressive Behavior included 16 items; Attention Problems included 10 items; Anxious/Depressed included 13 items. Each item rated on three point scale (0-2), with higher ratings indicating higher levels of behavior.

Item-level Measurement Models

Measurement models were fit to the data to test for statistically significant and salient loadings on the assigned latent trait factor (i.e., unidimensionality). Several items were collapsed from three responses to two (i.e., ratings of 2 recoded to 1) to reduce the number of empty cells, typically as the result of less than 10 total responses in the most extreme category (ratings of 2). Items that were collapsed were done so across all three raters for comparability in subsequent invariance testing, specifically an equal number of thresholds (Sass, 2011).

Models with one trait and three informants were estimated at the three time points (Configural Models; see Tables 6, 7, and 8). Measurement models generally fit the data adequately. All models' RMSEA values indicated acceptable fit ($\leq .05$) and previous literature evidenced strong support for the factor structure of the models (e.g., Achenbach & Rescorla, 2001; Brown & Achenbach, 1993; Greenbaum & Dedrick, 1998). CFI values were lower than predetermined criteria ($CFI \geq .95$) for "good" model fit for some models, but $CFI \geq .90$ has been described as acceptable (Bentler, 1990). Given the evidence in support of the correct model specification, analysis was continued despite suboptimal CFI values.

Table 6

Measurement Invariance Tests for Each Trait in First Grade

	χ^2	df	$\Delta\chi^2$	Δdf	CFI	ΔCFI	RMSEA			WRMR
							Est.	LL	UL	
<u>Aggressive Behavior</u>										
Config.	1620.41*	1077	--	--	.948	--	.026	.023	.028	1.30
Strong	1768.01*	1129	219.08**	52	.939	.009	.027	.025	.030	1.41
<u>Attention Problems</u>										
Config.	929.23*	402	--	--	.929	--	.042	.038	.045	1.46
Strong	946.86*	434	65.73*	32	.931	<.001	.040	.036	.043	1.55
<u>Anxious/Depressed</u>										
Config.	861.28*	699	--	--	.942	--	.018	.013	.021	1.08
Strong	941.42*	731	98.22**	32	.925	.017	.020	.016	.023	1.18
Strong ^a	904.21*	728	55.14*	29	.947	.005	.018	.014	.022	1.13

Note. CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; LL = Lower Limit (90% Confidence Interval); UL = Upper Limit (90% Confidence Interval); WRMR = Weighted Root Mean Square Residual; a = Partial Invariance; * $p < .01$; ** $p < .001$.

Table 7

Measurement Invariance Tests for Each Trait in Third Grade

	χ^2	df	$\Delta\chi^2$	Δdf	CFI	ΔCFI	RMSEA			WRMR
							Est.	LL	UL	
<u>Aggressive Behavior</u>										
Config.	1855.63**	1077	--	--	.912	--	.031	.029	.033	1.46
Strong	1940.98**	1122	155.78**	46	.907	.005	.031	.029	.033	1.53
<u>Attention Problems</u>										
Config.	1242.04**	402	--	--	.909	--	.053	.049	.056	1.66
Strong	1295.09**	432	103.65**	30	.907	.002	.051	.048	.054	1.75
<u>Anxious/Depressed</u>										
Config.	1087.76**	699	--	--	.906	--	.027	.024	.030	1.30
Strong	1160.92**	729	100.38**	30	.896	.010	.028	.025	.031	1.38

Note. CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; LL = Lower Limit (90% Confidence Interval); UL = Upper Limit (90% Confidence Interval); WRMR = Weighted Root Mean Square Residual; * $p < .01$; ** $p < .001$.

Table 8
Measurement Invariance Tests for Each Trait in Fifth Grade

	χ^2	<i>df</i>	$\Delta\chi^2$	Δdf	CFI	ΔCFI	RMSEA			WRMR
							Est.	LL	UL	
<u>Aggressive Behavior</u>										
Config.	1629.70**	1077	--	--	.947	--	.026	.024	.029	1.31
Strong	1743.50**	1129	189.37**	52	.942	.005	.027	.025	.030	1.39
<u>Attention Problems</u>										
Config.	1004.27**	402	--	--	.919	--	.045	.042	.048	1.53
Strong	1098.38**	436	140.09**	34	.911	.008	.045	.042	.049	1.67
<u>Anxious/Depressed</u>										
Config.	973.64**	699	--	--	.932	--	.023	.019	.023	1.18
Strong	1027.26**	739	80.55*	40	.928	.005	.023	.019	.026	1.24

Note. CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; LL = Lower Limit (90% Confidence Interval); UL = Upper Limit (90% Confidence Interval); WRMR = Weighted Root Mean Square Residual; * $p < .01$; ** $p < .001$.

Results from the configural models supported unidimensionality of each latent trait factor. All items had statistically significant ($p < .05$) relations with their respective trait factor. All items also had salient loadings on their respective trait factor ($\lambda \geq .40$, Brown, 2006), with the exception of three items: mothers' ratings of item 46 (related to nervous movements) for Attention Problems in first grade ($\lambda = .38$), fathers' ratings of item 87 (related to moodiness) for Aggressive Behavior in third grade ($\lambda = .39$), and teachers' ratings of item 32 (related to need for perfection) for Anxious/Depressed in fifth grade ($\lambda = .29$). These three items were retained for subsequent analysis due to their statistically significant relation with the expected latent trait and to maintain consistency across raters and time (Sass, 2011). Evidence was observed for the unidimensionality of all traits during all three assessment periods based on the model fit and substantial factor loadings for each trait.

The next step in the analytic plan was to create continuous item parcels and model all of the traits in the same model, both within a single assessment period and longitudinally. However,

when attempts were made to fit multi-trait models using parcels, several models fit the data poorly, due primarily to significant cross-loadings of parcels on two latent factors within time. The presence of cross-loadings have led to researchers to argue that item parceling is typically inappropriate, including when comparing latent means (e.g., Marsh et al., 2013). As a result of these concerns, models using item-level data, instead of parcels, were estimated using the WLMSV estimator to allow for a focus on the substantive questions of the study. That is, all of the research questions in this study could be answered because they did not require that all traits be modeled simultaneously, and additional information (e.g., differential item functioning) regarding invariance can be obtained when invariance is tested at the individual item-level (Crayen, Geiser, Scheithauer, & Eid, 2011).

Measurement Invariance Models

Following determination that all items would be retained for analysis and that item-level data would be modeled, strong factorial invariance was tested for each trait across the three informants within each assessment period. Item factor loadings and thresholds were constrained equal across like-items (Strong Models; see Tables 6, 7, and 8). These parameters were constrained simultaneously (Muthén & Muthén, 1998-2010). Models for latent Aggressive Behavior, Attention Problems, and Anxious/Depressed all demonstrated strong factorial invariance across informants within each grade level for the three assessment periods, with the exception of Anxious/Depressed in first grade.

The test of strong factorial invariance for first grade Anxious/Depressed resulted in a statistically significant increase in χ^2 , [$\Delta\chi^2(32) = 98.22$]. Equality constraints for items 33 and 50 were freed for teachers' ratings in subsequent tests of partial strong invariance based on large modification indices. Item factor loading and threshold constraints were freed simultaneously

(Muthén & Muthén, 1998-2010). Teachers' ratings of item 33 (related to feelings or complaints of not being loved) demonstrated a lower factor loading and higher threshold than the other two informants; teachers' ratings of item 50 (related to guilty feelings) demonstrated a higher factor loading, a higher first threshold, and a lower second threshold.

Following the freeing of these constraints from the strong invariant model, partial strong invariance was supported [$\Delta\chi^2(29) = 55.137$]. Evidence in support of strong and partial strong invariance for all models indicated that the underlying measurement characteristics of the latent trait were similar across informants and comparisons of latent means were valid. The strong or partial strong constraints were maintained for subsequent hypothesis testing of the latent mean structures.

Following tests of within-time invariance across raters, tests of longitudinal invariance were conducted for each dyad (Tables 9 and 10). Configural models in the tests of longitudinal analysis retained within-time strong factorial invariance constraints across raters, with factor loadings and thresholds free to vary across time. Factor loadings and thresholds were then constrained equal across time to test for strong factorial invariance across time. Each model demonstrated longitudinal strong invariance, allowing for latent mean comparisons across time.

Table 9

Longitudinal Measurement Invariance Tests for Each Trait for Mother-Father Dyad

	χ^2	<i>Df</i>	$\Delta\chi^2$	Δdf	CFI	ΔCFI	RMSEA			WRMR
							Est.	LL	UL	
<u>Aggressive Behavior</u>										
Config.	5520.317**	4422	--	----	.948		.018	.016	.019	1.246
Strong	5605.486**	4562	161.734**	45	.946	.002	.018	.016	.020	1.269
<u>Attention Problems</u>										
Config.	2609.768**	1673	--	--	.937	--	.027	.025	.029	1.354
Strong	2633.903**	1697	50.142*	24	.937	<.001	.027	.025	.028	1.373
<u>Anxious/Depressed</u>										
Config.	3436.094**	2877	--	--	.948	--	.016	.014	.018	1.147
Strong	3475.067**	2984	60.912**	29	.947	.001	.016	.014	.018	1.167

Note. CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; LL = Lower Limit (90% Confidence Interval); UL = Upper Limit (90% Confidence Interval); WRMR = Weighted Root Mean Square Residual; a = Partial Invariance; * $p < .01$; ** $p < .001$.

Table 10

Longitudinal Measurement Invariance Tests for Each Trait for Mother-Teacher Dyad

	χ^2	<i>Df</i>	$\Delta\chi^2$	Δdf	CFI	ΔCFI	RMSEA			WRMR
							Est.	LL	UL	
<u>Aggressive Behavior</u>										
Config.	5715.691**	4422	--	--	.935	--	.019	.018	.021	1.360
Strong	5779.990**	4467	131.299**	45	.934	.001	.019	.018	.021	1.375
<u>Attention Problems</u>										
Config.	2939.060**	1673	--	--	.934	--	.031	.029	.033	1.453
Strong	2997.734**	1697	93.346**	24	.932	.002	.031	.029	.033	1.482
<u>Anxious/Depressed</u>										
Config.	3285.902**	2874	--	--	.944	--	.014	.011	.016	1.124
Strong	3337.346**	2901	77.787**	27	.941	.003	.014	.011	.016	1.144

Note. CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; LL = Lower Limit (90% Confidence Interval); UL = Upper Limit (90% Confidence Interval); WRMR = Weighted Root Mean Square Residual; a = Partial Invariance; * $p < .01$; ** $p < .001$.

Direction and Magnitude of Method Effects

Question 1: *What is the informant discrepancy, represented by method effects, for each informant's rating of the child's behavior?*

Latent method effects factors are the difference in trait factor means between the reference informant (based on mothers' ratings) and the non-reference informant (based on fathers' or teachers' ratings). Negative method effects indicated that the fathers' or teachers' latent mean ratings were lower than mothers' (i.e., reference) ratings; positive effects indicated higher ratings by fathers' or teachers'.

Hypothesis 1 stated statistically significant and negative method effects were expected, indicating lower levels of father- or teacher-rated behavior compared to mother-rated behavior. Additionally, fathers' method effects (Father-ME) were expected to be smaller than teachers' method effects (Teacher-ME), based on effects sizes. Effect sizes were computed as standardized latent means of the ME, in which the difference of the latent means (i.e., Father-ME or Teacher-ME) were divided by the square root of the variance, or standard deviation, of the mothers' ratings (Thompson & Green, 2011). This allowed for informants' effects to be on the same scale and for statements regarding the relative effects of fathers or teachers. For example, a Father-ME effect size of .25 provides information that fathers' ratings were approximately .25 standard deviation units higher than mothers (Hancock, 2001; Thompson & Green, 2013). In the absence of criteria more directly related to informant discrepancies research, effect size magnitudes are described based on the following qualitative categories: small effects $\geq .2$ described as small; effects $\geq .5$ described as moderate; and effects $\geq .8$ described as large (Cohen, 1988; Hancock, 2001). Results are summarized in Tables 11, 12, and 13.

Aggressive behavior. Statistically significant Father-ME were observed in first and fifth grades for Aggressive Behavior, indicating statistically significant differences in latent means. However, these two effects were in opposite directions: in first grade, positive Father-ME indicated higher fathers' ratings than mothers'; in fifth grade, negative Father-ME indicated lower fathers' ratings than mothers'. A non-statistically significant positive Father-ME was observed in third grade. In sum, Father-ME were positive in first grade and decreased in magnitude over the three assessment periods until they were negative in fifth grade. The differing effects were contrary to hypothesized negative effects across all assessment periods; however, these findings were consistent with observed sum scores reported in Table 5.

Consistent with hypothesis 1, statistically significant negative Teacher-ME were observed across all three assessment periods, indicating lower teachers' ratings of Aggressive Behavior than mothers'. Further, Teacher-ME for Aggressive Behavior were consistently larger than the Father-ME, indicating greater disagreement in the mother-teacher dyad compared to the father-teacher dyad.

Attention problems. Contrary to hypothesized negative effects, statistically significant positive Father-ME for Attention Problems were observed in first and third grades; fathers' ratings were higher than mothers'. A positive, non-statistically significant effect was observed in fifth grade. In contrast to Father-ME and consistent with hypothesis 1, first and third grade Teacher-ME were statistically significant and negative: teachers' ratings were lower than mothers' ratings. A negative, non-statistically significant Teacher-ME was observed in fifth grade. Contrary to hypothesis 1, the magnitude of effects was similar in third and fifth grade across the two dyads. However, a larger discrepancy was observed between mother-teacher (Teacher-ME) than mother-father (Father-ME) in first grade, as expected.

Anxious/Depressed. Partial support for hypothesis 1 was observed for fathers' ratings of Anxious/Depressed. The Father-ME for Anxious/Depressed was negative but not statistically significant in first grade. Father-ME were also negative in third and fifth grades, but were statistically significant. Fathers' ratings were lower than mothers' ratings. Similar to results for both Aggressive Behavior and Attention Problems, Teacher-ME for Anxious/Depressed were statistically significant and negative (i.e., lower than mothers' ratings) across all three grade levels. Consistent with hypothesis 1, Teacher-ME were consistently larger in magnitude than Father-ME, indicating a greater discrepancy between the mother-teacher dyad compared to the mother-father dyad.

Question 1 summary. Overall results indicated statistically significant method effects dependent upon the type of behavior and assessment period (i.e., first, third, or fifth grades), and for Aggressive Behavior, inconsistent directions (i.e., positive or negative) across assessment periods. Fourteen of eighteen method effects (78%) were statistically significant, providing support for the presence of discrepancies, even when taking measurement error into consideration by using latent variables (Hartley et al., 2011). Of the 14 statistically significant method effects, 11 were negative as hypothesized (76% of significant effects). All three statistically significant positive effects were Father-ME (first grade Aggressive Behavior; first and third grade for Attention Problems).

The majority of effect sizes for Father-ME were small in magnitude. Fathers' effect sizes ranged in absolute value from .08 to .17 for Aggressive Behavior; .12 to .16 for Attention Problems; and .15 to .44 for Anxious/Depressed. Teachers' effect sizes were large for Aggressive Behavior, small to moderate for Attention Problems, and moderate for Depression/Anxiety. Their effect sizes ranged in absolute value from 1.39 to 2.35 for

Aggressive Behavior; .15 to .67 for Attention Problems; and .42 to .71 for Anxious/Depressed.

Results for hypothesis 1 are summarized Tables 11,12, and 13.

Table 11
Latent Covariances, Correlations, Means, and Variances

Agg. Behavior	First Grade			Third Grade			Fifth Grade		
	1	2	3	1	2	3	1	2	3
1. Mother		-0.40	-0.14		-0.32	0.00		-0.35	-0.54
2. Father-ME	-0.56		0.38	-0.59		0.17	-0.30		0.49
3. Teacher-ME	-0.10	0.26		0.00	0.18		-0.22	0.24	
Mean	0.55**	0.15**	-2.02**	0.48**	0.07	-1.51**	0.16*	-0.17*	-1.67**
Variance	0.74	0.67	3.12	0.70	0.41	2.00	1.44	0.99	4.16
Att. Problems	4	5	6	4	5	6	4	5	6
4. Mother		-0.14	0.04		-0.16	0.05		-0.17	-0.08
5. Father-ME	-0.51		0.03	-0.55		0.01	-0.46		0.05
6. Teacher-ME	0.08	0.06		0.09	0.01		-0.12	0.11	
Mean	0.58**	0.09**	-0.38**	-0.68**	0.15**	-0.15*	-0.96**	0.09	-0.12
Variance	0.32	0.23	0.80	0.45	0.17	0.58	0.56	0.25	0.78
Anx./Dep.	7	8	9	7	8	9	7	8	9
7. Mother		-0.27	-0.41		-0.15	-0.34		-0.15	-0.36
8. Father-ME	-0.47		0.30	-0.39		0.14	-0.26		0.21
9. Teacher-ME	-0.56	0.36		-0.63	0.28		-0.50	0.28	
Mean	-1.14**	-0.11	-0.51**	-1.24**	-0.14*	-0.30**	-1.24**	-0.33**	-0.46**
Variance	0.51	0.66	1.05	0.40	0.35	0.76	0.57	0.62	0.94

Note. Covariances above the diagonal, correlations below the diagonal. Mother's were the reference informant; their mean values are the latent mean for the behavior. Father-ME and Teacher-ME mean values are the difference between mother's latent mean and the latent mean of the non-reference informant (father or teacher). Agg. Behavior = Aggressive Behavior; Att. Problems = Attention Problems; Anx./Dep. = Anxious/Depressed; ME = method effects; **statistically significant at $p < .01$; *statistically significant $p < .05$

Table 12
Father-ME Direction and Effect Sizes

	First Grade	Third Grade	Fifth Grade
Aggressive Behavior	Positive (Mother < Father) ES = .17	Positive (Mother < Father) ES = .08	Negative (Mother > Father) ES = -.13
Attention Problems	Positive (Mother < Father) ES = .16	Positive (Mother < Father) ES = .15	Positive (Mother < Father) ES = .12
Anxious/Depressed	Negative (Mother > Father) ES = -.15	Negative (Mother > Father) ES = -.20	Negative (Mother > Father) ES = -.44

Note. ES = Effect Size

Table 13
Teacher-ME Direction and Effect Sizes

	First Grade	Third Grade	Fifth Grade
Aggressive Behavior	Negative (Mother > Teacher) ES = -2.35	Negative (Mother > Teacher) ES = -1.80	Negative (Mother > Teacher) ES = -1.39
Attention Problems	Negative (Mother > Teacher) ES = -.67	Negative (Mother > Teacher) ES = -.15	Negative (Mother > Teacher) ES = -.16
Anxious/Depressed	Negative (Mother > Teacher) ES = -.71	Negative (Mother > Teacher) ES = -.42	Negative (Mother > Teacher) ES = -.61

Note. ES = Effect Size

Trait-Specific Method Effects

Question 2: *Are method effects trait-specific? That is, is the size of informant discrepancies dependent upon the trait measured?*

Hypothesis 2 expected trait-specific ME within each informant. ME for Aggressive Behavior and Attention Problems were expected to be smaller than for Anxious/Depressed, due to more readily observable externalizing behaviors potentially being less influenced by ME (Loeber & Dishion, 1984). However, consistency of ME across all three traits would indicate a

halo effect with consistently higher or lower scores from a specific informant. Models were first tested with all three traits' ME means equal across a specific informant; subsequent models held two traits' ME equal while freely estimating the third (Table 14).

First grade. Consistent with hypothesis 2, Father-ME across the three traits (i.e., behaviors) of interest were not equal in first grade. Father-ME for Aggressive Behavior and Attention Problems were equal and positive (higher ratings than mothers). Father-ME for Anxious/Depressed, as described earlier, was negative but the effect size was similar in absolute magnitude to the other behaviors' effect sizes. Similar to Father-ME, Teacher-ME were also not equal across the three traits: effect sizes for Aggressive Behavior were larger in magnitude than those for both Attention Problems and Anxious/Depressed. However, Teacher-ME for Attention Problems and Anxious/Depressed were equal, not Aggressive Behavior and Attention Problems as hypothesized. All Teacher-ME were negative, indicating lower ratings than mothers.

Third grade. In third grade, Father-ME across the three traits were again unequal, similar to results observed in first grade. Consistent with first grade models and hypothesis 2, Father-ME for Aggressive Behavior and Attention Problems were positive and equal to one another. Anxious/Depressed was a negative effect, but the effect size was similar in absolute magnitude to Attention Problems. Teacher-ME demonstrated the same results as that observed in first grade, with effects for Teacher-ME for Aggressive Behavior larger than the equal effects for Attention Problems and Anxious/Depressed.

Fifth grade. In fifth grade, Father-ME were not equal across the three traits, as in first and third grade. However, instead of equal effects across Aggressive Behavior and Attention Problems, Father-ME for Aggressive Behavior and Anxious/Depressed were equal. Both effects were negative and effect sizes were larger in absolute magnitude than for Attention Problems.

Teacher-ME for Attention Problems and Anxious/Depressed were equal, consistent with first and third grades. Again, the largest Teacher-ME effect size was for Aggressive Behaviors.

Question 2 summary. Support for the halo effect was not found: no models fit better with all three traits' ME constrained equal. ME-Father for Aggressive Behavior and Attention problems were equal in first and third grade as hypothesized; ME-Father for Aggressive Behavior and Anxious/Depressed were equal in fifth grade. ME-Teacher for Attention Problems and Anxious/Depressed were consistently equal across the three assessment periods, contrary to hypothesis 2. ME-Teacher were typically larger than ME-Father, as expected.

Table 14

Tests of Equality of Means Across Traits for Father-ME and Teacher-ME

Grades Traits Constrained Equal	First			Third			Fifth		
	$\Delta\chi^2$	<i>df</i>	<i>p</i>	$\Delta\chi^2$	<i>df</i>	<i>p</i>	$\Delta\chi^2$	<i>df</i>	<i>P</i>
<u>Father-ME</u>									
All Traits	16.17	2	<.001	13.65	2	.001	17.21	2	<.001
Agg.-Att.	2.84	1	.090	2.28	2	.131	8.08	1	.005
Agg.-Dep.	16.24	1	<.001	6.48	1	.011	3.26	1	.071
Att.-Dep.	9.19	1	.002	12.15	1	<.001	16.52	1	<.001
<u>Teacher-ME</u>									
All Traits	66.97	1	<.001	89.17	2	<.001	66.13	1	<.001
Agg.-Att.	57.57	1	<.001	91.03	1	<.001	55.39	1	<.001
Agg.-Dep.	23.25	1	<.001	42.09	1	<.001	32.97	1	<.001
Att.-Dep.	0.80	1	.370	2.26	1	.130	0.09	1	.770

Note. ME= Method effects; Agg. = Aggressive Behavior; Att. = Attention Problems; Dep. = Anxious/Depressed.

Relationship of Trait Levels and Method Effects

Question 3: *Are the size of method effects related to levels of trait behavior?*

Hypothesis 3 expected the covariances/correlations (Table 11) between ME and the reference trait factor (mother-rated behavior) would be negative in direction, indicating smaller

ME associated with greater levels of behavior. That is, the more extreme behavior, the smaller the discrepancy between informants' ratings.

Aggressive Behavior. In first grade, the correlation between mothers' ratings and Father-ME was statistically significant and negative ($r = -.56$), demonstrating that as mother-rated trait levels of Aggressive Behavior increased, the size of the discrepancy decreased. This correlation was also statistically significant in third ($r = -.59$) and fifth grades ($r = -.30$). The correlation between mothers' ratings and the Teacher-ME was not statistically significant in first or third grades. In fifth grade the statistically significant negative correlation ($r = -.22$) indicated that as trait levels increased, the size of the discrepancy between mothers' and teachers' ratings decreased, similar to the mother-father dyad.

Father- and Teacher-ME had statistically significant positive correlations within first ($r = .26$), third ($r = .18$), and fifth grades ($r = .24$). This relation indicated the discrepancies between mother and teacher ratings increased as the discrepancies between the mother and father ratings increased. That is, the size of the discrepancy between the two dyads moved in the same direction, indicating some consistency in the direction of disagreement.

Attention Problems. The correlation between mothers' ratings of Attention Problems and Father-ME was statistically significant and negative in first ($r = -.51$), third ($r = -.55$), and fifth grades ($r = -.46$). As the level of mother-rated Attention Problems increased, the size of the discrepancy decreased, similar to the relation observed with Aggressive Behavior. The correlations between mothers' ratings and Teacher-ME, and between Father-ME and Teacher-ME, were not statistically significant, although they too were positive.

Anxious/Depressed. The correlation between mothers' ratings and Father-ME was statistically significant and negative in first ($r = -.47$), third ($r = -.39$), and fifth grades ($r = -.26$).

These results are similar to the relations observed in Aggressive Behavior and Attention Problems. The correlation between mothers' ratings and Teacher-ME was also negative across all three grades ($r = -.56, -.63, \text{ and } -.50$). Similar to the correlations observed in Aggressive Behavior, a positive relation was observed across all three times for Father-ME and Teacher-ME ($r = .35, .28, \text{ and } .28$) meaning that discrepancies between each dyad's ratings increased as the other's discrepancies increased.

Question 3 summary. A consistent, negative relation between mothers' ratings and Father-ME provided support for hypothesis 3: as trait levels of behavior increase, the discrepancy is smaller. This relation was observed across all three behaviors and across all three assessment periods within the mother-father dyad. Similarly, the negative relation between mothers' ratings and Teacher-ME was consistent across time for Anxious/Depressed and one assessment period for Attention Problems. However, no support for a relation between the mothers' ratings and Teacher-ME for Attention Problems was found. Finally, although no hypothesis was made regarding the relation between Father-ME and Teacher-ME, a significant positive relation between the two factors was observed for Aggressive Behavior and Anxious/Depressed, but not Attention Problems.

Longitudinal Method Effects

Question 4: *Do method effects remain constant over time?*

The method effect means and variances were expected to be equal across time according to hypothesis 4. The equality of means and variance of method effects were tested across the three types of behavior, with only two assessment periods in each model, due to convergence issues when attempting to conduct the tests across all three assessment periods. Equality of means and variances were tested within the same model by constraining respective means and

variances to be equal. If the $\Delta\chi^2$ was statistically significant ($p < .05$) when compared to the Meref model, in which all means and variances were freely estimated, subsequent tests were conducted. First means and then variances were constrained equal in separate models (Table 15).

Table 15

Tests of Longitudinal Equality of Means and Variances for Father-ME and Teacher-ME

Grades	Father-ME				Teacher-ME			
	First-Third		Third-Fifth		First-Third		Third-Fifth	
	$\Delta\chi^2$	Δdf						
<u>Aggressive Behavior</u>								
Means/Variances	6.03*	2	7.32*	2	1.42	2	11.61**	2
Variances	5.02*	1	4.08*	1	2.63	1	1.71	1
Means	3.17	1	7.60**	1	2.12	1	4.27*	1
<u>Attention Problems</u>								
Means/Variances	3.22	2	5.59	2	4.97	2	1.52	2
Variances	1.16	1	2.68	1	1.20	1	0.03	1
Means	1.21	1	3.67	1	6.47*	1	0.06	1
<u>Anxious/Depressed</u>								
Means/Variances	3.14	2	2.45	2	2.22	2	7.17*	2
Variances	2.74	1	1.48	1	0.79	1	4.36*	1
Means	0.06	1	7.06**	1	0.63	1	0.13	1

Note. Each comparison model was nested within a strong invariant model with means and variances freely estimated. * $p \leq .05$ ** $p \leq .01$

Aggressive Behavior. The $\Delta\chi^2$ for the omnibus test for Father-ME across first and third grade was statistically significant. Subsequent testing indicated that Father-ME were equal in first and third grades; variances (i.e. variability in the dyad's discrepancies) across this same span were significantly different. Greater variability in discrepancies between mothers' and fathers' ratings was observed in first grade than third. Neither means nor variances were equal across third and fifth grades: larger discrepancies and greater variability were observed in fifth grade.

Longitudinal comparisons between Teacher-ME across first and third grades indicated equal means and variances. However, Teacher-ME variances were equal across third to fifth grades but means were not: larger discrepancies were observed in fifth grade.

Attention Problems. Means and variances of Father-ME were equal from first to third grades, and from third to fifth grades. Teacher-ME demonstrated equal variances across first and third grades, as well as third and fifth grades. Mean discrepancies were significantly smaller in third than first grade.

Anxious/Depressed. Father-ME means and variances were equal across first and third grades. Variances were equal across third and fifth grades. Means differed significantly from the third to fifth grade, with larger mean Father-ME in fifth grade. Teacher-ME means and variances were equal from first to third grade. Means were also equal from third to fifth grade. However, variance for Teacher-ME was greater in fifth grade than in third.

Demographic Predictors of Method Effects

Question 5: *Are method and trait effects predicted by SES, race/ethnicity, and sex?*

Hypothesis 5a stated that SES was not expected to have statistically significant effects on Father-ME, but negative effects were expected on Teacher-ME. In addition, it was stated that SES would have a negative relation with trait levels of behavior. Hypothesis 5b stated that ethnic or racial minority status was not expected to be predictive of Father-ME, but was expected to have a positive effect on Teacher-ME. Finally, hypothesis 5c stated that boys were expected to demonstrate higher levels of mother-rated Attention Problems and Aggressive Behavior, and girls to demonstrate higher levels of Anxious/Depressed. Sex was expected to have non-significant effects on Father- or Teacher-ME for Aggressive Behavior and Attention Problems, and negative effects of Anxious/Depression. All statistically significant demographic predictors

were retained for subsequent analysis. Results from final trimmed models are reviewed below and reported in Tables 16, 17, and 18.

Aggressive Behavior. Consistent with hypothesis 5a, SES had negative effects on mothers' ratings of Aggressive Behavior in first and third grades; children from higher SES families had lower ratings of aggression. However, this effect was not statistically significant in fifth grade. SES did not have statistically significant effects on Father-ME during any assessment, consistent with hypothesis 5a. SES had a statistically significant effect on Teacher-ME in first grade, as expected, but not in any other grade.

In regards to ethnicity and race, African American children were rated similarly by their mothers as White children. In addition, no relation was observed between African Americans and Father-ME. However, African Americans had larger mother-teacher discrepancies for third and fifth grade. Hispanic ethnicity did not have statistically significant effects on mother-rated Aggressive Behavior, Father-ME, or Teacher-ME. The child's sex had statistically significant effects on Teacher-ME across all assessment periods, indicating boys had larger mother-teacher discrepancies than girls.

Attention Problems. SES had statistically significant negative effects on mothers' ratings of Attention Problems in first and third grades. Children from higher SES families had lower ratings of attention related concerns. In regards to the child's sex, statistically significant effects on mothers' ratings were observed across all assessment periods: mothers rated more boys' Attention Problems than girls'.

Statistically significant effects were also observed for child's sex on Teacher-ME, indicating larger mother-teacher discrepancies for boys. In addition, African American children also had larger Teacher-ME, across all three time periods, as hypothesized; these results are

similar to those observed with Aggressive Behavior. No effects of SES on Father-ME or Teacher-ME for Attention Problems were observed. Further, no differences in mothers' ratings or Father-ME were observed for African Americans.

Anxious/Depressed. SES had a positive effect on mothers' ratings in fifth grade: mothers from higher SES families rated their child's Anxious/Depressed higher. No demographic variables were predictive Father-ME or Teacher-ME in first, third, or fifth grades, after controlling for the other predictors in the model.

Table 16
Statistically significant effects of predictors on Aggressive Behavior and related method effects

	First				Third				Fifth										
	β	<i>b</i>	LL	UL	SE	<i>p</i>	β	<i>b</i>	LL	UL	SE	<i>p</i>	β	<i>b</i>	LL	UL	SE	<i>p</i>	
<u>Aggression on</u>																			
SES	-0.132	-0.009	-0.013	-0.005	.002	<.001	-0.102	-0.021	-0.039	-0.004	.009	.018	--	--	--	--	--	--	--
Mother's Anger	.284	.014	.008	.020	.003	<.001	.305	.067	.043	.091	.012	<.001	.307	.112	.070	.153	.021	.021	<.001
Mother's Depr.	.153	.004	.002	.006	.001	.003	--	--	--	--	--	--	--	--	--	--	--	--	--
Negative Stress	--	--	--	--	--	--	.089	.011	.002	.020	.005	.018	--	--	--	--	--	--	--
Positive Stress	--	--	--	--	--	--	.145	.015	.005	.025	.005	.002	--	--	--	--	--	--	--
Mother's Sens.	--	--	--	--	--	--	-0.231	-0.090	-0.119	-0.061	.015	<.001	-.189	-0.112	-0.175	-0.049	.032	.032	<.001
Father's Sens.	--	--	--	--	--	--	--	--	--	--	--	--	-.149	-0.087	-0.152	-0.023	.033	.033	.008
<u>Father-ME on</u>																			
Mother's Anger	--	--	--	--	--	--	-0.220	-0.040	-0.065	-0.016	.013	.001	-0.271	-0.091	-0.134	-0.048	.022	.022	<.001
Mother's Anx.	-0.251	.014	.010	.018	.002	<.001	--	--	--	--	--	--	.330	.100	.047	.153	.027	.027	<.01
Positive Stress	--	--	--	--	--	--	-0.126	-0.019	-0.021	-0.001	.005	.032	--	--	--	--	--	--	--
Mother's Depr.	-0.263	-0.007	-0.011	-0.003	.002	<.001	--	--	--	--	--	--	--	--	--	--	--	--	--
<u>Teacher-ME on</u>																			
African-Am.	--	--	--	--	--	--	--	1.095	.259	1.931	.427	.010	--	1.959	.084	3.833	.956	.956	.041
Sex	-0.102	-0.053	-0.089	-0.027	.018	.003	--	-0.571	-0.977	-1.666	.207	.005	--	-0.747	-1.511	.017	.390	.390	.055
SES	.108	.009	.003	.015	.003	.004	--	--	--	--	--	--	--	--	--	--	--	--	--
Mother's Anger	-0.223	.014	.008	.020	.003	<.001	--	--	--	--	--	--	--	--	--	--	--	--	--
Mother's Depr.	-0.142	-0.005	-0.009	-0.001	.002	.002	--	--	--	--	--	--	--	--	--	--	--	--	--
Father's Sens.	--	--	--	--	--	--	-0.167	-0.126	-0.221	-0.030	.049	.010	--	--	--	--	--	--	--

Note: β = standardized regression coefficient; *b* = unstandardized regression coefficient; LL = unstandardized regression coefficient lower limit (95% confidence interval); UL = unstandardized regression coefficient upper limit (95% confidence interval); SE = standard error; Depr. = depression; Sens. = sensitivity; Anx. = anxiety.

Table 17

Statistically significant effects of predictors on Attention Problems and related method effects

	First 95% Conf. Int.			Third 95% Conf. Int.			Fifth 95% Conf. Int.					
	β	<i>b</i>	<i>p</i>	β	<i>B</i>	<i>p</i>	β	<i>b</i>	<i>p</i>	β	<i>b</i>	<i>p</i>
<u>Attention on</u>												
SES	-.148	-.025	.009	-.127	-.178	.006	-.127	-.178	.006	-.127	-.178	.006
Sex	--	-.185	<.001	--	-.032	.012	--	-.231	.002	--	-.231	.002
Mother's Ang.	.250	.030	.007	--	--	<.001	--	.058	.001	--	.058	.001
Father's Depr.	--	--	--	--	--	--	--	.026	--	--	.026	--
Home Environ.	--	--	--	--	--	--	--	--	--	--	--	--
Positive Class	--	--	--	-.141	-.034	.025	--	--	--	--	--	--
Father's Sens.	-.142	-.029	.010	--	--	.005	--	--	--	--	--	--
<u>Father-ME on</u>												
Father's Depr.	--	--	--	.383	.023	<.001	--	--	--	--	--	--
Mother's Ang.	-.118	-.013	.007	--	--	.052	--	-.051	--	--	-.051	.001
Positive Class	--	--	--	.163	.028	.027	--	--	--	--	--	--
<u>Teacher-ME on</u>												
African-Am.	--	.728	.120	--	.628	.034	--	1.675	.008	--	1.675	.008
Sex	--	-.286	.299	--	-.286	.023	--	-.772	.022	--	-.772	.022

Note. β = standardized regression coefficient; *b* = unstandardized regression coefficient; LL = unstandardized regression coefficient lower limit (95% confidence interval); UL = unstandardized regression coefficient upper limit (95% confidence interval); SE = standard error; Ang. = anger; Depr. = depression.

Table 18

Statistically significant effects of predictors on Anxious/Depressed and related method effects

	First Grade 95% Conf.					Third Grade 95% Conf.					Fifth Grade 95% Conf.								
	β	b	LL	UL	SE	p	β	b	LL	UL	SE	p	β	b	LL	UL	SE	p	
<u>Anx/Depr. on</u>																			
SES	--	--	--	--	--	--	--	--	--	--	--	--	.076	.105	.007	.203	.050	.036	
Sex	--	--	--	--	--	--	--	--	--	--	--	--	-.102	-.017	-.031	-.002	.007	.025	
Mother's Ang.	.168	.034	.009	.058	.013	.007	.167	.029	.009	.049	.010	.004	.281	.048	.026	.069	.011	<.001	
Mother's Depr.	.131	.014	.001	.026	.006	.029	.148	.019	.006	.033	.007	.004	.123	.016	.004	.027	.006	.007	
Positive Stress	--	--	--	--	--	--	.093	.007	.001	.014	.003	.019	--	--	--	--	--	--	
<u>Father-ME on</u>																			
Mother's Ang.	--	--	--	--	--	--	-.164	-.030	-.055	-.004	.013	.023	-.256	-.043	-.066	-.021	.011	.034	
Father's Depr.	.328	.062	.028	.097	.018	<.001	--	--	--	--	--	--	--	--	--	--	--	--	
<u>Teacher-ME on</u>																			
Mother's Ang.	--	--	--	--	--	--	-.137	-.036	-.068	-.004	.016	.028	-.150	-.156	-.065	-.002	.016	.004	
Father's Sens.	-.111	-.084	-.167	.000	.043	.050	--	--	--	--	--	--	--	--	--	--	--	--	

Note. β = standardized regression coefficient; b = unstandardized regression coefficient; LL = unstandardized regression coefficient lower limit (95% confidence interval); UL = unstandardized regression coefficient upper limit (95% confidence interval); SE = standard error; Ang. = anger; Depr. = depression; Sens. = sensitivity.

Parental Psychopathology and Stress as Predictors of Method Effects

Question 6: *Are method and trait effects predicted by maternal stress, and maternal and paternal depression, anger, and anxiety?*

Hypothesis 6 expected stress and the three types of parental psychopathology to have positive effects on both trait and method factors, resulting in greater levels of trait behavior and larger discrepancies. Results are summarized in Tables 16, 17, and 18 previously presented.

Aggressive Behavior. In first grade, mothers' self-reported anger and depression had positive effects on mother-rated Aggressive Behavior. Mothers' self-reported anxiety had negative effects on Father-ME, and mothers' self-reported anger and depression had negative effects on Teacher-ME. That is, as the mothers' anxiety, anger, or depression increased, the specific ME decreased. All of these effects were controlling for the other predictors in the model.

In third grade, mothers' self-reported anger, and both positive and negative stress, all demonstrated positive effects on mother-rated Aggressive Behavior, predicting higher ratings. Mothers' self-reported anger and positive stress predicted smaller Father-ME. No relations were observed between Teacher-ME and any type of parental psychopathology or stress in third grade, contrary to effects of mothers' anger and depression observed in first grade.

In fifth grade, mothers' self-reported anger demonstrated a statistically significant positive effect on mother-rated Aggressive Behavior, consistent with findings in first and third grades. Statistically significant negative effects for mothers' self-reported anger and anxiety on Father-ME were observed, predicting smaller mother-father discrepancies. Similar to third grade, no relationships were observed between Teacher-ME and parental psychopathology and stress variables.

Summary of effects on Aggressive Behavior. Contrary to hypothesized positive effects for all parental psychopathology and stress variables on trait and ME factors, effects were inconsistent. The only significant predictor at all three assessment periods for trait levels of Aggressive Behavior was mothers' self-reported anger. Other significant effects on trait behavior were positive as expected (i.e., depression, positive and negative stress), although only at one assessment period.

Effects on both Father- and Teacher-ME were inconsistent with hypothesis 6; all effects were negative or not significant. Effects of mothers' self-reported anxiety and anger were statistically significant at two time points, but resulted in smaller, not larger, Father-ME. Likewise, mothers' self-reported anger and depression were statistically significant only in first grade, and again resulted in smaller Teacher-ME. All remaining predictors were not statistically significant at any time point, after controlling for the other variables in the model.

Attention Problems. In first grade, a significant positive effect of mothers' self-reported anger was observed on mother-rated Attention Problems. Negative effects of mother self-reported anger on Father-ME was also retained in the analysis ($p = .052$). No type of parental psychopathology was predictive of Teacher-ME, after controlling for all of the other variables in the model. In third grade, only one type of parental psychopathology had statistically significant effects on Attention Problems and related ME: fathers' self-reported depression, which predicted larger Father-ME.

In fifth grade, mothers' self-reported anger and fathers' self-reported depression had statistically significant and positive effects on mother-rated Attention Problems. Mother's self-reported anger had statistically significant negative effects on Father-ME. With the inclusion of these variables, SES was no longer a significant predictor of Attention Problems and was

trimmed from subsequent models. All other parental psychopathology and stress-related predictors were unrelated to mothers' ratings or the method effects.

Anxious/Depressed. In first grade, mothers' anger and depression had positive effects on mother-rated Anxious/Depressed. Fathers' self-reported depression had positive effects on Father-ME. No effects were noted for parental psychopathology on Teacher-ME in first grade. In third grade, mothers' self-reported anger and depression again had positive effects on mother-rated Anxious/Depressed, similar to first grade. In addition, positive stress also demonstrated a small positive effect. Mothers' self-reported anger also had negative effects on both Father- and Teacher-ME. Finally, in fifth grade, effects were similar to those in third grade, with the exception of positive stress on trait levels of Anxious/Depressed, which was no longer statistically significant.

Sensitivity and Method Effects

Question 7: *Are method effects predicted by independent ratings of parent and teacher sensitivity?* Ratings of sensitivity were expected to predict greater levels of method effects according to hypothesis 7. Results are summarized in tables 16, 17, and 18 previously presented.

Aggressive Behavior. Ratings of mothers' sensitivity did not demonstrate statistically significant effects on mothers' ratings of Aggressive Behavior, Father-ME, or Teacher-ME, with the exception of a negative effect in fifth grade on mothers' ratings. Similarly, fathers' sensitivity did not demonstrate statistically significant effects, except for negative effects on Teacher-ME in third grade and mothers' ratings in fifth grade. In other words, more sensitive fathers' were predictive of smaller mother-teacher discrepancies, and lower mothers' ratings of aggression. Teachers' sensitivity did not have effects at any point on any outcome.

Attention Problems and Anxious/Depressed. Mothers', fathers', or teachers' sensitivity did not typically demonstrate significant effects on the trait or method effect latent variables for both Attention Problems and Anxious/Depressed. The two exceptions were: 1) significant negative effects of fathers' sensitivity on mothers' ratings of Attention Problems in first grade (i.e., more sensitive fathers results in lower ratings of attention concerns by mothers); and 2) Teacher-ME for Anxious/Depressed in first grade (i.e., more sensitive fathers result in smaller mother-teacher discrepancies). All other effects on the latent variables were not statistically significant.

Question 8: *Are method effects predicted by rating of the context in which they occur?*

Hypothesis 8 predicted a negative effect between ratings of context and both trait and method effects. Results are summarized in tables 16, 17, and 18 previously presented.

Aggressive Behavior and Anxious/Depressed. Ratings of classroom context and home environment were not statistically significant predictors of mothers' ratings of Aggressive Behavior or Anxious/Depressed and any related ME during any assessment period.

Attention Problems. A positive classroom environment demonstrated a statistically significant negative effect on mothers' ratings of Attention Problems in third grade: the more positive classroom environment, the lower the ratings. The third grade positive classroom environment variable had a statistically significant and positive effect on Father-ME, with a more positive classroom associated with larger mother-father discrepancies. Ratings of the home environment had a statistically significant negative effect on mothers' ratings of attention in fifth grade.

Prediction of Outcomes by Method Effects

Question 9: *Are method effects predictive of referral to special school services, diagnosed learning disability, or attention, behavior, or emotional problems?* Larger method effects were expected to be predictive of increased rates of referral to special school services, diagnosed learning disability, and attention/behavior/emotional problems.

Aggressive Behavior. Mother-rated Aggressive Behavior was predictive of diagnosed learning disabilities and attention/behavior/emotional problems in third grade, diagnosed attention/behavior/emotional problems in third and fifth grades, and diagnosed learning disabilities in fifth grade (Table 19). Even when controlling for the effects of mothers' ratings, larger Teacher-ME were predictive of referral to special school services in first and fifth grades; and attention/behavior/emotional problems in fifth grade. All remaining regression paths were not significant.

Attention Problems. Mother-rated Attention Problems were predictive of referral to special school services in first and third grades, diagnosed attention/behavior/emotional problems in third and fifth grades, and diagnosed learning disabilities in third grade (Table 20). Again, while controlling for mothers' ratings, Teacher-ME were predictive of referral to special school services in first and fifth grades, similar to results observed with Aggressive Behavior. Father-ME were not predictive of any of the three outcomes.

Anxious/Depressed. Mother-rated Anxious/Depressed was predictive of referral to special school services and diagnosed attention/behavior/emotional problems in fifth grade (Table 21). Teacher-ME was a significant predictor of referral to special school services in school in first grade and fifth grade, and diagnosed attention/behavior/emotional problems in fifth grade. Father-ME was not predictive of any of the three outcomes.

Table 19
 Statistically Significant Effects of Aggressive Behavior and Method Effects on Outcomes

	First Grade				Third Grade				Fifth Grade				
	β	b	95% Conf. Int.		β	b	95% Conf. Int.		β	b	95% Conf. Int.		
			LL	UL			LL	UL			LL	UL	
<u>Referred/Receive</u>													
<u>Special Education</u>													
Mother ratings	--	--	--	--	0.201	0.253	0.078	0.429	0.089	0.005	--	--	--
Father-ME	--	--	--	--	--	--	--	--	--	--	--	--	--
Teacher-ME	0.198	0.069	0.013	0.122	0.028	0.014	--	--	--	-0.208	-0.105	-0.197	0.047
													0.025
<u>Diagnosed</u>													
<u>Attention/Behavior</u>													
<u>Problems</u>													
Mother ratings	N/A	N/A	N/A	N/A	N/A	0.449	0.621	0.345	0.898	0.141	<.001	0.446	0.393
Father-ME	N/A	N/A	N/A	N/A	N/A	--	--	--	--	--	--	0.186	0.6
Teacher-ME	N/A	N/A	N/A	N/A	N/A	--	--	--	--	--	--	0.279	0.153
												0.028	0.279
												0.064	0.016
<u>Diagnosed Learning</u>													
<u>Disability/ Problems</u>													
Mother ratings	N/A	N/A	N/A	N/A	N/A	--	--	--	--	--	--	0.252	0.206
Father-ME	N/A	N/A	N/A	N/A	N/A	--	--	--	--	--	--	0.324	0.072
Teacher-ME	N/A	N/A	N/A	N/A	N/A	--	--	--	--	--	--	--	--

Note. β = standardized regression coefficient; b = unstandardized regression coefficient; LL = unstandardized regression coefficient lower limit (95% confidence interval); UL = unstandardized regression coefficient upper limit (95% confidence interval); SE = standard error.

Table 20

Statistically Significant Effects of Attention Problems and Method Effects on Outcomes

	First Grade				Third Grade				Fifth Grade										
	β	b	LL	UL	SE	p	β	b	LL	UL	SE	p	β	b	LL	UL	SE	p	
<u>Referred/Receive</u>																			
<u>Special Education</u>																			
Mother ratings	0.212	0.463	0.127	0.798	0.171	0.007	0.269	0.394	0.162	0.625	0.118	0.001	--	--	--	--	--	--	--
Father-ME	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Teacher-ME	0.338	0.378	0.159	0.596	0.112	0.001	--	--	--	--	--	--	0.218	0.171	0.016	0.327	0.079	0.031	0.031
<u>Diagnosed</u>																			
<u>Attention/Behavior</u>																			
<u>Problems</u>																			
Mother ratings	N/A	N/A	N/A	N/A	N/A	N/A	0.625	1.113	0.638	1.713	0.298	<.001	0.525	0.724	0.418	1.031	0.156	<.001	<.001
Father-ME	N/A	N/A	N/A	N/A	N/A	N/A	--	--	--	--	--	--	--	--	--	--	--	--	--
Teacher-ME	N/A	N/A	N/A	N/A	N/A	N/A	--	--	--	--	--	--	--	--	--	--	--	--	--
<u>Diagnosed Learning</u>																			
<u>Disability/Problems</u>																			
Mother ratings	N/A	N/A	N/A	N/A	N/A	N/A	.534	0.889	0.435	1.344	0.232	<.001	--	--	--	--	--	--	--
Father-ME	N/A	N/A	N/A	N/A	N/A	N/A	--	--	--	--	--	--	--	--	--	--	--	--	--
Teacher-ME	N/A	N/A	N/A	N/A	N/A	N/A	--	--	--	--	--	--	--	--	--	--	--	--	--

Note. β = standardized regression coefficient; b = unstandardized regression coefficient; LL = unstandardized regression coefficient lower limit (95% confidence interval); UL = unstandardized regression coefficient upper limit (95% confidence interval); SE = standard error.

Table 21
Statistically Significant Effects of Anxious/Depressed and Method Effects on Outcomes

	First Grade				Third Grade				Fifth Grade										
	β	b	LL	UL	SE	p	β	b	LL	UL	SE	p	β	b	LL	UL	SE	p	
<u>Referred/Receive Special Education</u>																			
Mother ratings	--	--	--	--	--	--	--	--	--	--	--	--	0.209	0.317	0.049	0.585	0.137	0.02	
Father-ME	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Teacher-ME	0.183	0.139	0.011	0.267	0.065	0.033	--	--	--	--	--	--	0.215	0.263	0.042	0.484	0.113	0.02	
<u>Diagnosed Attention/Behavior Problems</u>																			
Mother ratings	N/A	N/A	N/A	N/A	N/A	N/A	--	--	--	--	--	--	0.55	0.921	0.511	1.33	0.209	<.001	
Father-ME	N/A	N/A	N/A	N/A	N/A	N/A	--	--	--	--	--	--	--	--	--	--	--	--	--
Teacher-ME	N/A	N/A	N/A	N/A	N/A	N/A	--	--	--	--	--	--	0.244	0.263	0.053	0.607	0.141	0.02	
<u>Diagnosed Learning Disability/ Problems</u>																			
Mother ratings	N/A	N/A	N/A	N/A	N/A	N/A	--	--	--	--	--	--	--	--	--	--	--	--	--
Father-ME	N/A	N/A	N/A	N/A	N/A	N/A	--	--	--	--	--	--	--	--	--	--	--	--	--
Teacher-ME	N/A	N/A	N/A	N/A	N/A	N/A	--	--	--	--	--	--	--	--	--	--	--	--	--

Note. β = standardized regression coefficient; b = unstandardized regression coefficient; LL = unstandardized regression coefficient lower limit (95% confidence interval); UL = unstandardized regression coefficient upper limit (95% confidence interval); SE = standard error.

Chapter V: Discussion

Discrepancies between informant ratings are common in research and clinical assessment (Achenbach, 2006; Duhig et al., 2000). The purpose of the current study was to answer several questions regarding the presence and prediction of informant discrepancies on a behavior rating scale, and their utility in identifying students for potential clinical or school services. A Method Effects Reference (MEref) model (Geiser et al., 2013; Pohl et al., 2008) was used to model multiple-informant longitudinal data from the NICHD SECCYD. Informant discrepancies were modeled as latent mean difference between mothers' ratings (reference informant) and either fathers' (Father-ME) or teachers' ratings (Teacher-ME). The magnitude, direction, and longitudinal consistency of those discrepancies on behavior rating scales were investigated. Influences that explained informant discrepancies were also explored, including intrapersonal and demographic characteristics of the informant and child as well as independent ratings of context and the informant. Finally, the predictive utility and relevance of informant discrepancies on specific child outcomes was studied to help inform assessment practices. The Discussion is organized to summarize the findings related to informant discrepancies (method effects) 1) between two dyads of raters: mother-father and mother-teacher across three behaviors (Aggressive Behavior, Attention Problems, and Anxious/Depressed; 2) the predictors and outcomes associated with the father and teacher informant discrepancies; and 3) the implications of current findings for assessment practices.

Informants' Method Effects

Fathers' method effects. Fathers' method effects (Father-ME), defined as latent mean differences between mothers' and fathers' ratings, depended on the behavior (i.e., Aggressive Behavior, Attention Problems, and Anxious/Depressed) and assessment period (i.e., first, third,

and fifth grades). The inconsistent direction of method effects (i.e., mothers' or fathers' with higher ratings) was contrary to the expectation that fathers' ratings would be consistently lower than mothers (e.g., Brody et al., 1994; Schroeder et al., 2010). In fact, only fathers' ratings for Anxious/Depressed were consistently lower than the mothers' ratings, with fathers' endorsing relatively fewer symptoms related to depression and anxiety across first to fifth grades. Small Father-ME were found for Aggressive Behavior. The effects for aggression differed in direction over time, mothers had lower ratings in first grade whereas fathers had lower ratings in fifth grade; there was not a statistically significant difference in those ratings in third grade. Father-ME for Attention Problems were small, and fathers' ratings were consistently higher than mothers' ratings, contrary to expected negative effects (e.g., Caye et al., 2013; Langberg et al., 2010). The inconsistent direction of discrepancies provides evidence that a halo effect was not present for parent ratings because neither parent consistently rated all behaviors higher or lower within the same time period.

Despite the inconsistent direction of effects, patterns in the discrepancies emerged. Fathers' ratings of Aggressive Behavior and Attention Problems were higher than mothers' for five of six effects (four effects were statistically significant), whereas mothers' ratings of Anxious/Depressed were consistently higher. Although these differences were statistically significant, the effects were generally small. Absolute latent Cohen's *ds* were all less than or equal to .20, except for Anxious/Depressed in fifth-grade ($d = .44$). Informants sharing a context are expected to have smaller discrepancies than those who do not (Kraemer et al., 2003), as shared context reduces "extraneous variance" related to context. Current findings supported this expectation: Father-ME were typically smaller in magnitude than Teacher-ME for the same behavior at the same time. However, even with a shared context, statistically significant

differences still existed between mothers and fathers. Practically, the size of these discrepancies may have limited influence on assessment decisions, with the size related to differences in T-scores of approximately 1.5 to 4 points, on average. However, previous studies have noted substantial differences in meeting diagnostic criteria dependent on informant, based on effects of approximately 3 to 4 points (Langberg et al., 2010).

Longitudinally, Father-ME were stable for all three behaviors from first to third grades; Father-ME for Attention were also stable from third to fifth grades. Previous research has provided support for stable discrepancies in parent-child dyads over the course of up to 2 years (De Los Reyes, Alfano, & Beidel, 2010; De Los Reyes, Goodman, Kliewer, & Reid-Quiñones, 2010). The current findings further confirm that there is some stability to be expected in informant discrepancies, regardless of the dyad. However, in this study there was more stability at younger ages and for attention problems.

There were two instances of instability in Father-ME. The instability in mother-father discrepancies from third to fifth grades for Aggressive Behavior and Anxious/Depressed seemed to be explained by different patterns in ratings. All parent ratings of aggression decreased, on average, over the assessment periods, but fathers' decreased more. In contrast, the discrepancy between fathers' and mothers' ratings of Anxious/Depressed increased due to both increases in mothers' ratings and decreases in fathers' ratings of those symptoms over that period of time. Aged-based differences in the magnitude of discrepancies (Schroder et al., 2010; Van de Ende & Verhulst, 2005) and in agreement (Achenbach et al., 1987) have been previously reported, but were limited to these two behaviors across these two time points in this study.

The instability in direction and magnitude of mother-father discrepancies over time may result from changes in behavioral expectations. Certain behaviors may be more socially

acceptable or tolerated at a given age (Ehrensaft, Cohen, Chen, & Berenson, 2007); perhaps this shift in age-based behavioral expectations occurs at different times for mothers and fathers. For example, fathers' more rapid decline in ratings of aggression may reflect changes in expectations or tolerance of aggression, particularly given the relatively slower decline observed in mothers. Interestingly, the shift in expectations did not occur with attention problems, it may be that age-based behavioral expectations for attention problems shift in tandem for mothers and fathers.

Fathers' generally higher ratings for more disruptive behaviors and lower ratings for more internalizing behaviors may be a result of fathers' being more sensitive to aggressive and inattentive behaviors, and less sensitive to internalizing behaviors (e.g., Anxious/Depression), or a combination of the two (e.g., Karver, 2006). Mothers may be more aware of internalizing problems simply due to a greater amount of time spent talking with the child (Treutler & Epkins, 2003) or as a result of interpersonal exchanges between parent and child (Collins, 1990). These differences manifested themselves in a larger discrepancy for Anxious/Depressed ratings in fifth-grade. The influence of mother-child personal exchanges or time spent with their child may result in a cumulative effect on mother's ratings or may reflect a fifth-grade child's improved ability to verbalize internalizing concerns. The shifts in behavioral expectations are supported by current evidence; however, parent self-reported ratings of intrapersonal characteristics at each time point provided further systematic explanation of differences in mother-father discrepancies.

Prediction of Father-ME. Several variables, including parent and child characteristics, and contextual ratings, were included in the MEref models to determine if they had unique influences on the mother and father discrepancies. Several variables emerged as unique predictors. Mothers' self-reported anger, anxiety, and depression were all predictive of smaller mother-father discrepancies when rating Aggressive Behavior. In other words, mothers with

more self-reported symptoms had lower levels of disagreement with fathers on the child ratings. Of these symptoms, however, only anger predicted mother-father discrepancies when rating Attention Problems and Anxious/Depressed, controlling for the other predictors.

Maternal psychological symptoms have previously been predictive of smaller discrepancies for internalizing and externalizing behaviors (Treutler & Epkins, 2003). Treutler and Epkins (2003) conjectured that different types of parent symptoms may affect discrepancies differentially. Here, to some extent that was true, but mothers' self-reported anger symptoms consistently emerged as a predictor (controlling for all of the other predictors) across all three behaviors that were rated.

Researchers have extensively discussed why mothers' psychological symptoms might result in rating discrepancies between mothers and fathers. The depression-distortion hypothesis is perhaps most widely studied (e.g., Chi & Hinshaw, 2002; Youngstrom et al., 2000). In this hypothesis, the ratings of depressed mothers have a negative perceptual bias, resulting in higher ratings from mothers and larger informant discrepancies. Both Richters (1992) and Hay et al. (1999) argued if mothers' ratings are actually biased, larger mother-father discrepancies would be observed for dyads with distressed mothers compared to dyads with non-distressed mothers. Their ratings would reflect the bias, not the accurate level of behavior. In the current study, although mothers with higher levels of distress rated their child's symptoms higher, maternal distress was associated with smaller discrepancies, not larger ones. The depression-distortion hypothesis was not supported.

The lack of support for the depression-distortion hypothesis found in this study provides evidence that mothers who have more symptoms may also have children with more symptoms; the higher ratings do not reflect bias (Richters, 1992). For example, mother's psychological

distress may be a causal influence on increased levels of maladaptive child behavior and their increased ratings validly represent higher levels of behavior. Indeed, parent's psychological distress has been related to child outcomes, including conduct problems (e.g., McMahon, Wells, & Kotler, 2006), depression (e.g., Stark et al., 2006), and overall child adjustment (Hay et al., 1999). It could also work in reverse; the child's increased behaviors might result in more psychological symptoms from mothers. Last, there are known genetic effects on ratings of both externalizing and internalizing behavior (Bartels et al., 2003). Thus the relation between mothers' symptoms and mothers' ratings may reflect shared genes. Considered together, mothers with more psychological distress may simply have children who show more symptoms, and they may not show perceptual bias in their ratings.

Teachers' method effects. Teachers' method effects (Teacher-ME), defined as latent mean differences between mothers' and teachers' ratings, indicated teachers' ratings were consistently lower than mothers (i.e., negative) and all but one discrepancy was statistically significant. The consistently lower teacher ratings are in stark contrast to the variety of higher and lower ratings from fathers, relative to mothers. Further, effect sizes for teachers were typically larger, and often much larger, than for fathers. Effect sizes for Teacher-ME were moderate to large, with the largest differences in ratings of Aggressive Behavior. Larger discrepancies for aggression may suggest context-specific differences in these types of behavior (i.e., greater levels of aggression at home), particularly given the relatively smaller differences between mothers and fathers' ratings of aggression. The differences in effect sizes across behaviors within the same time also suggest that there was not a halo effect.

Previous studies have reported various results in regard to mother-teacher discrepancies, including negative effects for broadband (Stanger & Lewis, 1993; Youngstrom et al., 2000) and

narrow-band behaviors (Van der Ende & Verhulst, 2005); and equal or positive effects for narrow-band behavior (Collishaw et al., 2009; Sollie et al., 2013). Given the substantial effect sizes found in this study, especially with aggression, it is surprising that there is not more consistency in the literature. Some of the inconsistency and large effects found in this study could be due to restricting the mother and teacher items to those that were common across settings. Items specific to home or school were eliminated from analysis which may limit the breadth of behavior reported related to the underlying latent trait; although it might also be argued that using only the common items would result in more consistency, not less. The influence of only using common items may have particularly influenced teachers' ratings for Aggressive Behavior and Attention Problems, as both scales included items specific to school, such as disrupting other students or failing to complete tasks. Elimination of these items may have reduced the latent mean, resulting in larger than expected mother-teacher discrepancies.

Another interesting finding, especially given different teachers as informants across time, was the stability for Teacher-ME for Aggressive Behavior from first to third grade, Attention Problems from third to fifth, and Anxious/Depressed across the three assessment periods. The longitudinal stability across two to four years among different teachers suggests that discrepancies may function as a result of the type of rater (e.g., teacher), not necessarily the specific informant. Teachers are more likely to agree with other teacher's ratings than other informants (Epkins, 1995). Despite the longitudinal stability, it is of note that effect sizes seem to generally decrease in magnitude over time largely as a result of changes in mothers' ratings. Future research should extend the time period studied to include adolescence to determine if this general trend continues, although some evidence is available that discrepancies may actually

increase as children age, particularly for externalizing behaviors (Van der Ende & Verhulst, 2005).

Prediction of Teacher-ME. Prediction of Teacher-ME was attributed to symptoms of maternal psychopathology, father sensitivity, and demographic characteristics. The influence of intrapersonal characteristics was similar to the influence observed on Father-ME. The findings further support arguments that mothers' distress is associated with increases in child behavior, resulting in smaller discrepancies in ratings. However, the influence of intrapersonal characteristics on mother-teacher discrepancies generally occurred during only one assessment period, whereas there were intrapersonal influences across the assessment periods observed with mother-father discrepancies.

Child characteristics also influenced Teacher-ME. Mothers' and teachers' ratings of boys and African Americans were predictive of larger discrepancies across the majority of assessment periods for Aggressive Behavior and Attention Problems. Other demographic variables, including SES and Hispanic ethnicity, were not predictive of mother-teacher discrepancies; further, no differences were noted for boys or African Americans for Father-ME for any behavior. Previous studies have noted consistent differences between mothers and teacher ratings of African American students by teachers (Youngstrom et al., 2000) but not in ratings made by mothers and fathers (Duhig et al., 2000), congruent with present findings. The differences could be due in part to cultural differences in behavioral expectations and norms (Terry & Irving, 2010) or the potential cultural mismatch between student and teacher (e.g., Brown-Jeffy & Cooper, 2011; Ladson-Billings, 1995).

The influence of the sex of the child on rating discrepancies has received mixed support in the past (c.f., De Los Reyes, & Kazdin, 2005; Collishaw et al., 2009). Schroeder et al. (2010)

noted that agreement (although not necessarily discrepancies) may be attributed to sex-specific behavior expectations. For example, boys are more frequently associated with attention problems. Informants make ratings based on expected behaviors (i.e., boys are expected to be inattentive) rather than actual behaviors, in line with the actor-observer phenomenon described in the ABC theory (De Los Reyes & Kazdin, 2005). In this phenomenon, informants make ratings based on dispositional characteristics, rather than taking context into consideration. The actor-observer phenomenon is typically observed in ratings of others versus self-reports. However, perhaps the actor-observer phenomenon has a more distinct influence on mothers than teachers, although this hypothesis has not been explicitly tested and was not tested here.

The observed demographic influences on rating differences have practical implications, given the strength of relationships and that these effects are above and beyond the influence of included predictors. These effects resulted in effects sizes of approximately 1 to 2 for African Americans and 0.5 to 1 for boys, indicating that the differences may result in substantial differences for interpretation and diagnostic purposes. Clinicians and school psychologists are more likely to observe larger discrepancies between mother and teacher ratings of Aggressive Behavior and Attention Problems for boys and African Americans. Discrepancies have been noted to be meaningful regarding interpretation of meeting or failing to meet diagnostic criteria (Caye et al., 2013; Langberg et al., 2010) and may result in ratings that are in different qualitative categories (e.g., “at-risk” versus “average) when using BRS. These differences may influence decisions regarding diagnosis, treatment options, or eligibility for services.

Relation of Trait Levels and Informant Discrepancies

Higher levels of mother-rated behaviors were negatively related to Father-ME for all behaviors. These correlations were substantial for Aggressive Behavior ($r = -.30$ to $-.56$),

Attention Problems ($r = -.46$ to $-.55$), and Anxious/Depressed ($r = -.26$ to $-.47$). Similar substantial negative correlations were observed between mothers' ratings and Teacher-ME for Aggressive Behavior ($r = -.22$) in fifth grade and for Anxious/Depression in all three assessment periods ($r = -.50$ to $-.63$). The level of mother-rated Attention Problems was not related to the size of Teacher-ME. In other words, the magnitude of mother-teacher discrepancies was consistent across different levels of mother-rated attention.

Considered together, the findings show that greater mother-reported levels of behavior tended to result in smaller discrepancies, with some exceptions. Although this relation was not consistent across different types of behavior or informants, a positive relation was never observed: higher levels of mother-rated behavior never resulted in larger discrepancies. It is evident that as mothers' ratings of behavior ratings decrease, greater discrepancies can be expected. Previous research regarding this relation is limited, as the preponderance of studies use a CTC(M-1) model, in which the relation cannot be estimated due to the definition of the ME as residual (Pohl & Steyer, 2010; Pohl et al., 2008).

Theoretically, the relationship between higher ratings and smaller discrepancies may be due to less frequent use of heuristics when completing ratings, consistent with one influence outlined in the Attribution Bias Context theory (De Los Reyes & Kazdin, 2005). Informants would less frequently use source monitoring and associated heuristics assuming that higher levels of behaviors are less ambiguous and more readily observable (Karver, 2006). However, BRS often provide limited context for ratings (De Los Reyes & Kazdin, 2005), and absent high levels of behavior, informants rely on heuristics to make more general ratings, potentially resulting in larger discrepancies. For example, it seems likely that behaviors will be more evident when a child frequently cries or threatens others. In other words, ratings may be more accurate as

more frequently occurring behavior is more readily recalled, in the absence of specific or contextual examples to recall when completing ratings. Similar arguments regarding less ambiguity have been made to explain differences in informant biases between externalizing and internalizing behaviors (Loeber & Dishion, 1984).

Pragmatically, this finding may indicate psychologists will find smaller discrepancies to reconcile in clinical settings or samples (i.e., the higher the level of behavior, the smaller the discrepancy). Studies using clinical samples have noted significant discrepancies between mothers and fathers, with effect sizes of approximately .2 to .3 for narrow-band behavior (Schroeder et al., 2010 ; Sollie et al., 2013), which are generally larger than those described here. However, the relation between trait level and the size of discrepancy was not explicitly studied in the research, and there may have also been smaller discrepancies with increases in mothers' ratings within those samples. Regardless of the correlation found in this study, there is a likelihood for one parent rating to be elevated (i.e., outside the average range) whereas the others is not (e.g., Caye et al., 2013; Langberg et al., 2010). But, optimistically, if more behaviors result in smaller discrepancies between raters, BRS demonstrate sensitivity to the behaviors they purport to measure (e.g., the presence of behaviors is identified by more than one informant), providing support for their validity and for their continued widespread use. Future research using a clinical or at-risk sample, and drawing comparisons to a more typical sample, is needed to confirm the current findings.

Father-ME and Teacher-ME correlations. The positive relationship between Father-ME and Teacher-ME for both Aggressive Behavior and Anxious/Depressed indicated that discrepancies between the two dyads move in a similar direction; however, this relationship was not observed for Attention Problems. Previous research, using a sixth grade sample from the

SECCYD, also noted significant correlations between father and teacher method factors for externalizing and internalizing behavior, as well as for ADHD symptoms (Low, Keith, & Jensen, 2015). Other studies have reported a similar positive relation between method factors (e.g., Eid et al., 2008; Grimm et al., 2009)

The positive relationship between the two ME for Aggressive Behavior and Depression/Anxiety remained statistically significant even with the inclusion of predictors. A unique set of mechanisms, perhaps still unaccounted for, may be influencing mothers' ratings (e.g., Dirks et al., 2012; Dumenci et al., 2011). In other words, mothers' ratings are not necessarily invalid or inaccurate (Trentler & Epkins, 2003), but perhaps mothers present a unique view from fathers and teachers. Given this, each informant should be treated as providing valid, although perhaps not equivalent, information (e.g., Achenbach et al., 1987). As such, the incremental validity of information provided by multiple informants is important to consider for clinical applications, with psychologists attempting to balance discrepant information with accurate diagnosis and treatment (Hunsely & Meyer, 2003).

Prediction of Child Outcomes

Teacher-ME for Aggressive Behavior and Anxious/Depressed predicted increased rates of receiving/referral for special school services and diagnosed attention/behavior/emotional problems at various times; Teacher-ME for Attention Problems also predicted increased rates of special school services. The larger the discrepancy between the ratings from the child's mother and teacher, the more likely the child was to be referred for school services or to have been diagnosed with a behavior-related concern. These effects occurred beyond the effects of mothers' ratings, which were also included in the model. Father-ME did not predict any of the three studied outcomes, nor did Teacher-ME predict learning disabilities at any time.

The current findings expand the body of literature on the utility of ME, particularly in the school-based setting. A review of previous studies noted informant parent-child discrepancies as predictive of outcomes including delinquency, response to mental health treatment, and parental involvement in treatment (De Los Reyes, 2010). Others have noted parent-child discrepancies are predictive of increased rates of expulsion from school, school discipline problems, and referral for mental health services (Ferdinand, van der Ende, & Verhulst, 2004; 2006). One additional study focused on parent-teacher discrepancies, in which support was found for discrepant ratings of aggressive behaviors associated with increased risk for suicide attempts and/or self-mutilation, 14 years after initial ratings (Ferdinand et al., 2007). These same researchers hypothesized the utility of Teacher-ME was a result of three potential influences: 1) lack of home or school support; 2) contextual differences in behavior; or 3) poor home-school communication. All three may contribute to explanations of the present findings.

First, a child with minimal support either from home or school may be more at-risk to develop school-related or emotional problems. The lack of support may result in the parents or teachers (or both) seeking additional clinical or school services to intervene with the child. Second, as discussed previously, contextual differences may play a role in discrepancies. Given that teachers' ratings were, on average, lower than mothers, the discrepancies provide support for contextual differences in behavior. In fact, research has indicated behaviors that are most notable to parents do not occur across settings (Karver, 2006). However, it is counter-intuitive to think that behavior in one setting is so problematic to warrant diagnosis or services, while not being problematic in the other. It seems more likely that pervasive behaviors would be perceived as problematic. In fact, the presence of behaviors across settings serves as criteria for specific mental health diagnoses (American Psychiatric Association, 2013). One informant's opinion,

nevertheless, is often enough to result in referral for additional support (Ferdinand et al., 2007), so only one informant, not two, would have to be concerned for there to be a referral. Third, poor communication between home and school may result in these outcomes. As conjectured by Ferdinand and colleagues, perhaps the discrepancy is indicative of social isolation, resulting in increased negative outcomes.

Study Contributions

Rarely have informant discrepancies been studied using latent means, instead using observed scores. By using a Meref model, in which discrepancies are modeled as differences in latent means (i.e., true scores free from measurement error), evidence was found that informant discrepancies are not simply measurement error (e.g., Achenbach, 2011; De Los Reyes, 2011). In other words, informant discrepancies between mother-father and mother-teacher dyads exist because of differences in their perceptions of similarly conceived underlying trait behaviors of the child. Previous research has proved inconclusive as to the magnitude and direction of the discrepancy (i.e., which informant had higher or lower ratings). Even with the use latent means, similarly inconclusive directions were found for mother-father discrepancies, possibly due to a number of influences including changes in ratings over time or smaller discrepancies in general. Mother-teacher informant discrepancies were larger or equal to mother-father discrepancies in absolute magnitude for all nine effects measured (three behaviors by three assessment periods); and in some cases, particularly for aggression, effects were substantially larger than those observed in the mother-father dyad.

Given the current evidence that informant discrepancies are robust phenomenon between commonly used informants of childhood behavior, even beyond the effects of measurement error, it was important to begin to understand the reasons that discrepancies exist and the

potential implications on outcomes. The study of several types of variables, including both mother and father intrapersonal characteristics, child and family demographic characteristics, and contextual ratings helped to clarify which variables influence discrepant ratings, above and beyond the influence of the other variables. Most notable among the influences were mothers' intrapersonal characteristics and child demographic variables (i.e., boys and African Americans). Other variables did not typically predict discrepancies, including fathers' intrapersonal characteristics, independent ratings of informant sensitivity, and independent ratings of classroom and home contexts, while controlling for the other predictors. Previous exploration of a variety of different types of influences (intrapersonal, contextual, and demographic) within one study is limited. The implications here direct the research to a focus on the intrapersonal characteristics of informants, as well as differences between groups. Although contextual variables were not significant predictors, there was considerable evidence for contextual based differences in behavior, particularly given the difference in magnitude of discrepancies between mothers-fathers and mothers-teachers. It is possible the contextual differences were validly captured in parents' or teachers' ratings. These contextual differences resulting in larger discrepancies warrant further study.

Support for the predictive utility of informant discrepancies receiving school based services and diagnosed attention/behavior/emotional problems was also evidenced, building upon a foundation of previous studies that noted outcomes associated with informant discrepancies. Notably, only three studies have considered the influence of parent-teacher discrepancies on outcomes (Ferdinand et al., 2007), or considered school based outcomes (Ferdinand et al., 2004; 2006). Larger discrepancies between mothers and teachers predicted a greater likelihood for referral to school-based services and diagnosed

attention/behavioral/emotional problems, similar to previous findings using parent-child discrepancies (Ferdinand et al., 2004; 2006). That is, discrepancies themselves are important predictors of outcomes, even beyond the effects of mothers' ratings or measurement error.

The study also contributed to the informant discrepancy literature by exploring the longitudinal stability of informant discrepancies two dyads with limited previous research. Previous studies were limited in the time span in which stability was explored and were typically limited to the parent-child dyad. Mother-father discrepancies were stable from first to third grades for all behaviors, and across all grades for attention concerns. Mother-teacher discrepancies also showed a degree of stability: from first to third grade for aggression, third to fifth grade for attention concerns, and across all grades for anxious/depressed. Beyond the stability of effects themselves, variables that predicted discrepancies were often statistically significant at more than one time point, as were the utility of the discrepancies to predict outcomes. There is a degree of stability in the influences predictive of discrepancies, at least during a majority of the elementary school years.

Finally, several methodological decisions helped to contribute to the growing body of research on informant discrepancies. First, the inclusion of mother, fathers, and teachers as informants for ratings of child's behavior within the same study is surprisingly rare. In addition, clearly distinguishing between mothers and fathers is also rare, but allowed for study of any unique influences associated with each informant. Second, the study of different types of narrow-band behaviors versus broadband behavior is also limited, particularly within the same study. The inclusion of three common childhood behavior problems helped to provide an understanding of whether discrepancies and the influences are similar across behavior type. Finally, as discussed previously, the use of a latent variable model to study the mean structure

has rarely occurred beyond Monte Carlo studies or brief examples of model use. The application of these recently developed models to a national sample is perhaps a step toward more widespread use of these modeling advances to study informant discrepancies.

Implications

Psychologists can expect consistent discrepancies in informant behavior ratings of elementary school students, with larger differences in behavioral ratings between mothers and teachers than between mothers and fathers. The use of additional assessment techniques to obtain information, as suggested in a multi-source assessment, may be especially vital when attempting to understand discrepant ratings. Given the potential implications of informant discrepancies associated with demographic characteristics and important outcomes, additional clinical interviewing or direct observation has been suggested to make sense of the discrepancy (e.g., Achenbach, 2011; Smith, 2007). As described in the introduction and literature review, a “Grand Discrepancy” (De Los Reyes et al., 2013) exists, in which the clinician are expected to use multiple informants to inform conclusions, despite the discrepancies. However, it is important to consider that perhaps all discrepancies do not need to be reconciled. In fact, BRS may be useful because they result in discrepancies (Youngstrom et al., 2000); the BRS may be sensitive to differences in behavior across contexts (e.g., in this study there were clearly more differences across the school and home contexts). Discrepancies on BRS may not be due to error, but provide meaningful information that has the potential to enhance the validity of decisions reached in assessment.

Several studies have offered suggestions for how to reconcile the differences or which rater should be more heavily weighted under circumstances (e.g., Schroeder et al., 2010; Smith, 2007). However, a desire to reconcile the differences seems to be based largely in the need to

arrive at a decision or diagnosis to gain eligibility for school services, or for insurance purposes in the clinic (Achenbach, 2011; Penney & Skilling, 2012). Making a yes/no decision is based in converging operations: the psychologist decides whether evidence agrees or disagrees with the potential diagnosis. This “categorical versus quantitative judgment” (Achenbach, 2011) occurs on two levels: informants are tasked to determine if a behavior occurs or not; clinicians are tasked to determine if the ratings warrant a diagnosis or not. These categorical (yes/no) decisions are made despite the expansive body of evidence that behavior is much more contextual and nuanced than a simple dichotomous decision.

Moving beyond the diagnostic dichotomy, clinicians that are aware of the complex influences resulting in ratings and discrepancies potentially have a better understanding of why the behavior may be occurring and under which circumstance (De Los Reyes & Kazdin, 2005). For example, discrepant ratings indicating high levels of aggression at home but not in school provide information as to contextual and interpersonal influences on the child’s behavior. Particularly if the parent dyad has small discrepancies and the parent-teacher dyad has larger discrepancies, it may be indicative of contextual differences in behavior (i.e., home vs. school). If the discrepancy is a result of influences such as parental anger or depression, shifts in behavioral expectations, or parent-child exchanges, as discussed previously, appropriate treatment options would include addressing those concerns. However, practically speaking, determining the presence of these influences noted here is a difficult task.

Clinicians are often limited by time constraints to fully assess all potential influences on ratings. Further, the ability to obtain information on potential reasons for disagreement depends on the willingness of informants to offer personal information. Despite this, the current findings support the view discrepancies are valuable information (Hunsley & Mash, 2007) given the

predictive utility of informant discrepancy for outcomes. Attempts at making sense of discrepancies should be used not to eliminate differences, but instead to determine why they occur for a particular individual. The discrepancies themselves may be a symptom or indicator of underlying problems, such as poor communication among informants, different behavior across contexts, or the influence of parent distress on the child's behavior. As a result, the discrepancy information can help to inform treatment decisions. For example, they may provide guidance on a method to measure outcomes of treatment. That is, if there are large differences between ratings across contexts (i.e., home and school), and all ratings are assumed to be valid, perhaps outcomes could be measured in reduction of the discrepancy. A change in behavior ratings for informants primarily in the problematic context while maintaining baseline levels of behaviors in the non-problematic context would indicate effective treatment, with limited unplanned effects in other settings.

Limitations

The generalizability of findings from this study is restricted by its limitations. First, neither the complete NICHD SECCYD sample, nor the sample used for the current study, is reflective of the current U.S. population's ethnic and racial composition. The selected study sample was 89% White, while current U.S. Census Bureau data (2011) reports 72% of the population as White, 13% as African American, and 16% as Hispanic. Generalizability from the current study is difficult given the small sample size of both groups ($n < 50$ for both Hispanics and African Americans). As a result, conclusion regarding effects associated with African American children should be tempered. A more representative sample that includes larger groups of Hispanics and African Americans could clarify the role ethnic and racial group membership play in informant discrepancies. Similarly, the sample's fathers may not be representative of

fathers in the general population. Data were selected to include only those children who had available father CBCL data at any assessment period. Clearly, there may be differences in a sample in which fathers are still present throughout a child's elementary school years.

Information regarding the cohabitation of parents, marital status, and time spent with the child could provide more clarity to the specific presence and role played by the father. Additionally, as discussed in the methods chapter, the selected sample differed from the original SECCYD sample, including having higher SES, less minorities and lower levels of maternal education.

A second limitation was the use of raw scores for the ratings. Although their use allowed for direct comparisons across informants, and in particular, comparisons across time, their use may inhibit the direct application of findings to assessment practices. Raw scores are typically ignored during actual assessments, in favor of scale scores, such as the T-scores used on the CBCL/TRF. Further, since differences exist in the items on the CBCL and TRF as published, only items common across forms were selected for study, resulting in a potential loss of context specific behavior information. As a result of these methodological decisions, findings may not directly transfer to assessment practices. In practice, informant-specific items would be included in summed raw scores, and then calculated into form (CBCL/TRF) specific T-scores (Achenbach, 2011). These informant-specific scale scores can result in different T-scores even with the same raw score, with the potential to create a discrepancy with different magnitudes or direction, or completely erase the discrepancies observed here. Further, items on the CBCL or TRF unique to the home or school were eliminated from analysis. The elimination of items may have disproportionately affected teachers' ratings, as many items eliminated, such as failing to finish tasks, messy work, or disrupting class, resulting in a downward bias of their ratings.

A final limitation is the lack of reliability and validity evidence for some measures used as predictors, including context and parental sensitivity, beyond that provided by study materials. Specifically, measures with lower reliability are limited in the strength of the relation that they can have with dependent variables due to the proportion of true variance to total variance. Given these limitations, it is possible that effects with more reliable measures may result in different relations between the predictors and the method effects. Further, additional research for evidence of validity for measures of anxiety and anger was not provided.

Future Directions

Future studies should focus on expanding the current findings longitudinally, including downward extensions to pre-school samples and upward to include adolescents or adults. Previous research has noted that comparing multiple informants needs to take into consideration the child's developmental level (Achenbach, 2011) and have noted higher levels of agreement as a child ages (Achenbach et al., 1987). The current research supports a degree of longitudinal stability for both Father and Teacher-ME. Others have noted stable method effects across elementary school years using a CTC(M-1) model (Grimm et al., 2009), which models method effects as the residuals of latent variables. In collaboration with expanding the time period assessed, the use of latent variables will help to parse out the effects of true score variables, free from the influence of measurement error. For example, multimethod latent state trait models, allows for the parsing of influences from the traits, time period, and simultaneous or longitudinal effects, may prove beneficial (Courvoisier, Nussbeck, Eid, Geiser, & Cole, 2008). However, it should be noted, the expense of obtaining measures from multiple informants at multiple times may become prohibitive and analysis increasingly complex (Courvoisier et al., 2008).

A second focus for future study should be the testing of other broadband behavioral rating scales, including the BASC-2 (Reynolds & Kamphaus, 2004) and its pending revision (BASC-3, Reynolds & Kamphaus, 2015), a popular scale commonly used in school-based assessments (e.g., Myers, Bour, Sidebottom, & Murphy, 2010). Informant discrepancy research has largely utilized forms from the ASEBA (e.g., Grigorenko et al., 2010; Penney & Skilling 2012; Van der Ende & Verhulst, 2005). It is not expected that the observed results would differ largely based on the behavior rating scale utilized due in part to the consistent presence of discrepancies despite using reliable scales (De Los Reyes, 2011). However, research has indicated mean differences between different scales in specific populations (e.g., preschool-age children; Myers et al., 2010). Testing the generalizability of the informant discrepancy phenomenon as not restricted to the ASEBA is important to further research.

A third focus for future study should include the use of self-report measures, particularly if the sample extends the age range to adolescence. Previous research has noted the presence of differences between self-report and other report (e.g., Van der Ende & Verhulst, 2005; Grigorenko et al., 2010), consistent with the actor-observer phenomenon (Jones & Nisbett, 1972), one aspect of De Los Reyes and Kazdin's (2005) Attribution Bias Context theory. Further, self-reported ratings for internalizing behaviors have been described as "essential" (p. 340, Merrell, 2008). Particularly in adolescents, some have noted the adolescent themselves provide valuable information on otherwise unobservable behavior (Van de Ende & Verhulst, 2005). However, self-report ratings were unavailable for the selected assessment periods in the NICHD SECCYD. The use of child or adolescent self-report may help to clarify context based differences, observed in part in the larger mother-teacher discrepancies.

Future study should also focus on teachers' characteristics as contributing factors that may influence their ratings. Despite the wide body of literature exploring the influence of parents' (particularly mothers) characteristics on ratings of their child, limited research has been conducted on how teacher characteristics may influence discrepancies. Recent study, using a sample from the SECCYD, reported teachers' self-efficacy and years of experience were related to more consistency between mothers and teachers' ratings (Low et al., 2015). That study provides preliminary evidence that teacher characteristics likely influence ratings, similar to findings from the current study of the influence of mothers' intrapersonal characteristics on discrepancies.

The positive relationship between levels of behavior and discrepancies provides potential evidence that more assessments conducted on a clinical population could expect smaller discrepancies. The comparison of clinical, sub-clinical, and normative populations would further assist clinicians, informing them of expected differences between raters. However, research has rarely, if ever, considered different types of populations with the same study or compared the strength of the relationship in a clinical sample. In addition, continued study of the practical influence of informant discrepancies on diagnostic decisions (i.e., do the discrepancies result in different rates of diagnosis; e.g., Langberg et al., 2010) are needed to determine the implications of informant discrepancies for diagnostic decisions. Despite their statistical significance, this continued study of the practical significance is vital to practical application.

Finally, study of the bidirectional influence of a child's behaviors on parents' ratings and psychological symptoms, and vice versa, would clarify the effects of parent symptoms on child behavior. The relation between parents symptoms have on behavior is well-established, but understanding the direction of influence, whether it be unidirectional or bidirectional, would

inform both the informant discrepancy and developmental psychopathology literature. The use of a longitudinal cross-lagged panel model with parent symptoms and parents ratings of child behaviors would be conducive to this type of model.

References

- Achenbach, T. M. (1991a). *Manual for the Child Behavior Checklist/ 4-18 and 1991 profile*. Burlington: University of Vermont, Department of Psychiatry.
- Achenbach, T. M. (1991b). *Manual for the Teacher's Report Form and 1991 profile*. Burlington: University of Vermont, Department of Psychiatry.
- Achenbach, T. M. (2006). As others see us: Clinical and research implications of cross-informant correlations for psychopathology. *Current Directions in Psychological Science, 15*, 94-98.
- Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implication of cross-informant correlations for situational specificity. *Psychological Bulletin, 101*, 213-232.
- Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth & Families.
- Adamson, R. M., & Wachsmuth, S. T. (2014). A review of direct observation research within the past decade in the field of emotional and behavioral disorders. *Behavioral Disorders, 39*, 181-189.
- American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.
- American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: Author.
- Bartels, M., Hudziak, J.J., Boomsma, D. I., Rietveld, M. J. H., Van Beijsterveldt, T. C. E. M., & Van den Oord, E. J. C. G. (2003). A study of parent ratings of internalizing and externalizing problem behavior in 12-year-old twins. *Journal of the American Academy*

- of Child and Adolescent Psychiatry*, 42, 1351-1359. doi:
10.1097/01.CHI.0000085755.71002.5d
- Bentler, P. M. (1990). Comparative fit indexes in structural equation models. *Psychological Bulletin*, 107, 238-246.
- Brody, G. H., Stoneman, Z., Flor, D., McCrary, C., Hastings, L., & Conyers, O. (1994). Financial resources, parent psychological functioning, parent co-caregiving, and early adolescent competence in rural two parent-African American families. *Child Development*, 65, 590-605.
- Brown, J., & Achenbach, T. M. (1993). Bibliography of published studies using the Child Behavior Checklist and related materials: 1993 edition. Burlington: University of Vermont, Department of Psychiatry.
- Brown-Jeffy, S., & Cooper, J.E. (2011). Toward a conceptual framework of culturally relevant pedagogy: An overview of the conceptual and theoretical literature. *Teacher Education Quarterly*, 38, 65-84.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.
- Brown, T. E. (1996). *Brown Attention-Deficit Disorder Scales*. San Antonio, TX: The Psychological Corporation.
- Caldwell, B., & Bradley, B. (1984). *Home observation for measurement of the environment*. Little Rock, AR: University of Arkansas at Little Rock.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.

- Castro-Schilo, L., Widaman, K. F., & Grimm, K. J. (2013). Neglect the structure of multitrait-multimethod data at your peril: Implications for associations with external variables. *Structural Equation Modeling, 20*, 181-207. doi: 10.1080/10705511.2013.769385.
- Caye, A., Machado, J. D., & Rohde, L. A. (2013). Evaluating parental disagreement in ADHD diagnosis: Can we rely on a single report from home? *Journal of Attention Disorders*. Advanced online publication. doi: 10.1177/1087054713504134
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255.
- Chi, T. C., & Hinshaw, S. P. (2002). Mother-child relationships of children with ADHD: The role of maternal depressive symptoms and depression-related distortions. *Journal of Abnormal Child Psychology, 30*, 387-400.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284-290.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, N. J., Coyne, J., & Duvall, J. (1993). Adopted and biological children in the clinic: Family, parental and child characteristics. *Journal of Child Psychology, Psychiatry, and Allied Disciplines, 34*, 545-562.
- Collins, W. A. (1990). Parent-child relationships in the transition to adolescence: Continuity and change in interaction, affect and cognition. In R. Montemayor, G. Adams, & T. Gullotta (Eds.), *Advances in adolescent development: From childhood to adolescence: A transitional period?* (Vol. 2, pp. 85–106). Beverly Hills, CA: Sage.

- Collishaw, S., Goodman, R., Ford, T., Rabe-Hesketh, S., & Pickles, A. (2009). How far are associations between child, family, and community factors and child psychopathology informant-specific and informant-general? *Journal of Child Psychology and Psychiatry*, *50*, 571-580. doi:10.1111/j.1469-7610.2008.02026.x
- Courvoisier, D. S., Nussbeck, F. W., Eid, M., Geiser, C., & Cole, D. A. (2008). Analyzing the convergent and discriminant validity of states and traits: Development and applications of multimethod latent state-trait models. *Psychological Assessment*, *20*, 270-280. doi: 10.1037/a0012812
- Crayen, C., Geiser, C., Scheithauer, H., & Eid, M. (2011). Evaluating interventions with multimethod data: A structural equation modeling approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*, 497-524. doi: 10.1080/10705511.2011.607068
- Crockenberg, S., & Lourie, A. (1996). Parents' conflict strategies with children and children's conflict strategies with peers. *Merrill-Palmer Quarterly*, *42*, 495-518.
- De Los Reyes, A. (2011). Introduction to the special section: More than measurement error: Discovering meaning behind informant discrepancies in clinical assessments of children and adolescents. *Journal of Clinical Child & Adolescent Psychology*, *40*, 1-9. doi: 10.1080/15374416.2011.533405
- De Los Reyes, A., Alfano, C. A., & Beidel, D. C. (2010). The relations among measurements of informant discrepancies within a multisite of treatments for childhood social phobia. *Journal of Abnormal Child Psychology*, *38*, 395-404. doi: 10.1007/s10802-009-9373-6
- De Los Reyes, A., Goodman, K. L., Kliewer, W., & Reid-Quiñones, K. R. (2010). The longitudinal consistency of mother-child reporting of parental monitoring and their

- ability to predict child delinquent behaviors two year later. *Journal of Youth and Adolescence*, *39*, 1417–1430. doi: 10.1007/s10964-009-9496-7
- 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*, 637-642. doi: 10.1007/s10802-009-9307-3
- De Los Reyes, A., & Kazdin, A. E. (2004). Measuring informant discrepancies in clinical child research. *Psychological Bulletin*, *16*, 330-334. doi: 10.1037/1040-3590.16.3.330
- De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin*, *131*, 483-509. doi: 10.1037/0033-2909.131.4.483
- De Los Reyes, A., Thomas, S. A., Goodman, K. L., & Kundey, S. M. A. (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 Los Reyes, A., Youngstrom, E. A., Pabón, S. C., Youngstrom, J. K., Feeney, N.C., & Findling, R. L., (2011). Internal consistency and associated characteristics of informant discrepancies in clinic referred youths age 11 to 17 years. *Journal of Clinical Child & Adolescent Psychology*, *40*, 1-9. doi: 10.1080/15374416.2011.533402
- Dirks, M. A., De Los Reyes, A., Brigg-Gowan, M., Cella, D., & Wakschlag, L. S. (2012). Annual research review: Embracing not erasing contextual variability in children's behavior-theory and utility in the selection and use of methods and informants in developmental psychopathology. *Journal of Child Psychology and Psychiatry*, *53*, 558-574. doi:10.1111/j.1469-7610.2012.02537.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 & Practice, 7*, 435-453.
- Dumenci, L., Achenbach, T. M., & Windle, M. (2011). Measuring context-specific and cross-contextual components of hierarchical constructs. *Journal of Psychopathology and Behavior Assessment, 33*, 3-10. doi: 10.1007/s10862-010-9187-4
- Ehrensaft, M. K., Cohen, P., Chen, H., & Berenson, K. (2007). Developmental transitions in youth behavioral opposition and maternal beliefs in social ecological context. *Journal of Child and Family Studies, 16*, 577-588. doi: 10.1007/s10826-006-9108-z
- Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika, 65*, 241-261.
- Eid, M. (2006). Methodological approaches for analyzing multimethod data. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 223-230). Washington, DC: American Psychological Association.
- 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*, 230-253. doi: 10.1037/a0013219
- Eiraldi, R. B., Power, T. J., Karustis, J. L., & Goldstein, S. G. (2000). Assessing ADHD and comorbid disorders in children: The Child Behavior Checklist and the Devereux Scales of Mental Disorders. *Journal of Clinical Child Psychology, 29*, 3-16.
- Epkins, C. (1995). Teachers' ratings of inpatient children's depression, anxiety, and aggression: A preliminary comparison between inpatient-facility and community-based teachers'

- ratings and their correspondence with children's self-reports. *Journal of Clinical Child Psychology*, 24, 63-70.
- Fagot, B. I. (1995). Classification of problem behaviors in young children. A comparison of four systems. *Journal of Applied Developmental Psychology*, 16, 95-106.
- Ferdinand, R. F., Van der Ende, J., & Verhulst, F. C. (2004). Parent-adolescent disagreement regarding psychopathology in adolescents from the general population as a risk factor for adverse outcomes. *Journal of Abnormal Psychology*, 113, 198-206. doi: 10.1037/0021-843X.113.2.198
- Ferdinand, R. F., Van der Ende, J., & Verhulst, F. C. (2006). Prognostic value of parent-adolescent disagreement in a referred sample. *European Child & Adolescent Psychiatry*, 15, 156-162.
- Ferdinand, R. F., Van der Ende, J., & Verhulst, F. C. (2007). Parent-teacher disagreement regarding psychopathology in children: A risk factor for adverse outcome? *Acta Psychiatrica Scandinavica*, 115, 48-55. doi: 10.1111/j.1600-0447.2006.00843.x
- Garner, W. R., Hake, H. W., & Eriksen, C. W. (1956). Operationism and the concept of perception. *Psychological Review*, 63, 149-159.
- Geiser, C., Eid, M., West, S. G., Lischetzke, T., & Nussbeck, F. W. (2012). A comparison of method effects in two confirmatory factor models for structurally different methods. *Structural Equation Modeling*, 19, 409-436. doi: 10.1080/10705511.2012.687658
- Geiser, C., Koch, T., & Eid, M. (2014). Data-generating mechanisms versus constructively-defined latent variables in multitrait-multimethod analysis: A comment on Castro-Schilo, Widaman, and Grimm (2013). *Structural Equation Modeling*, 21, 1-15. doi: 10.1080/10705511.2014.919816

- Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry, 38*, 581-586.
- Goodman, R. (2001). Psychometric properties of the Strengths and Difficulties Questionnaire (SDQ). *Journal of the American Academy of Child and Adolescent Psychiatry, 40*, 1337-1345.
- Greenbaum, P. E., & Dedrick, R. F. (1998). Hierarchical confirmatory factor analysis of the Child Behavior Checklist/4-18. *Psychological Assessment, 30*, 149-155.
- Gresham, F. M., & Elliott, S. N. (1990). *The Social Skills Rating System*. Circle Pines, MN: American Guidance Service.
- Grigorenko, E. L., Geiser, C., Slobodskaya, H. R., & Francis, D. J. (2010). Cross-informant symptoms from CBCL, TRF, and YSR: Trait and method variance in a normative sample of Russian youths. *Psychological Assessment, 22*, 893-911. doi: 10.1037/a0020703
- Grimm, K. J., Pianta, R. C., & Konold, T. (2009). Longitudinal multitrait-multimethod models for developmental research. *Multivariate Behavioral Research, 44*, 233-258. Doi: 10.1080/002731709027944230
- Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika, 66*, 373-388.
- Hartley, A. G., Zakriski, A. L., & Wright, J. C. (2011). Probing the depths of informant discrepancies: Contextual influences on divergence and convergence. *Journal of Clinical Child & Adolescent Psychology, 40*, 54-66. doi: 10.1080/15374416.2011.533404
- Hay, D. F., Pawlby, S., Sharp, D., Schmücker, G., Mills, A., Allen, H., & Kumar, R. (1999). Parents' judgments about young children's problems: Why mothers and fathers might

- disagree yet still predict later outcomes. *Journal of Child Psychology and Psychiatry*, *40*, 1249-1258.
- Hedges, L.V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, *92*, 490-499.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.
- 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
- Johnston, C., & Murray, C. (2003). Incremental validity in the psychological assessment of children and adolescents. *Psychological Assessment*, *15*, 496-507. doi: 10.1037/1040-3590.15.4.496
- Jones, E. E., & Nisbett, R. E. (1972). The actor and the observer: Divergent perceptions of the causes of behavior. In E. E. Jones, D. E. Kanouse, H. H. Kelly, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 79 –94). Morristown, NJ: General Learning Press.
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*, 183-202.
- Jöreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In R.C. Atkinson, D. H. Krantz, R. D. Luce, & P. Suppes (Eds.), *Contemporary Developments in Mathematical Psychology* (pp. 1-56). San Francisco, W. H. Freeman.
- Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, *42*, 448-457.

- Karver, M. C. (2006). Determinants of multiple informant agreement on child and adolescent behavior. *Journal of Abnormal Child Psychology*, *34*, 251-262. doi: 10.1007/s10802-005-9015-6
- Kasius, M. C., Ferdinand, R. F., Van den Berg, H., & Verhulst, F. C. (1997). Associations between different diagnostic approaches for child and adolescent psychopathology. *Journal of Child Psychology and Psychiatry*, *38*, 625-632.
- Kazdin, A. E., & Kagan, J. (1994). Models of dysfunction in developmental psychopathology. *Clinical Psychology: Science and Practice*, *1*, 35-52.
- Kazdin, A. E., & Petti, T. A. (1982). Self-report and interview measures of childhood and adolescent depression. *Journal of Child Psychology and Psychiatry*, *23*, 437-457.
- Kline, R. B. (2011). *Principles and practice of structural equation modeling*. (3rd ed.). New York: Guilford Press.
- Konold, T. R., & Pianta, R. C. (2007). The influence of informants on ratings of children's behavioral functioning: A latent variable approach. *Journal of Psychoeducational Assessment*, *25*, 222-236. doi: 10.1177/07434282906297784
- Kraemer, H. C., Measell, 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*, 1566-1577.
- Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537-567.
- Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, *32*, 465-491.

- Laird, R. D., & De Los Reyes, A. (2013). Testing informant discrepancies as predictors of early adolescent psychopathology: Why difference scores cannot tell you what you want to know and how polynomial regression may. *Journal of Abnormal Child Psychology*, *41*, 1-14. doi: 10.1007/s10802-012-9659-y
- Lance, C. E., Baranik, L. E., Lau, A. R., & Scharlau, E. A. (2009). If it ain't trait it must be method: (Mis) application of the multitrait-multimethod design in organizational research. In C. E. Lance, & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends* (pp. 337–360). New York, NY: Routledge.
- Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait-correlated method and correlated uniqueness models for multitrait-multimethod data. *Psychological Methods*, *7*, 228. doi: 10.1037//1082-989X.7.2.228
- Langberg, J. M., Epstein, J. N., Simon, J. O., Loren, R. E. A., Arnold, E., ... Wigal, T. (2010). Parent agreement on ratings of children's attention deficit/hyperactivity disorder and broadband externalizing behaviors. *Journal of Emotional and Behavior Disorders*, *18*, 41-50. doi: 10.1177/1063426608330792
- Latourneau, N. L., Duffet-Leger, L., Levac, L., Watson, B., & Young-Morris, C. (2013). Socioeconomic status and child development: A meta-analysis. *Journal of Emotional and Behavioral Disorders*, *21*, 211-224. doi: 10.1177/1063426611421007
- Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*, 1198–1202.
- Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, *32*, 53-76.

- Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods, 18*, 285-300. doi: 10.1037/a0033266
- Loeber, R., & Dishion, T. J. (1984). Boys who fight at home and school: Family conditions influencing cross-setting consistency. *Journal of Consulting and Clinical Psychology, 52*, 759-768. doi: 10.1037/0022-006X.52.5.759
- Loeber, R., Farrington, D. P., Stouthamer-Loeber, M., & Van Kammen, W. B. (1998). *Antisocial behavior and mental health problems*. Hillsdale, NJ: Erlbaum.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley Publishing Company.
- Low, J. A., Keith, T. Z., & Jensen, M. (2015). What predicts method effects in child behavior ratings. *Journal of Psychoeducational Assessment, 33*, 177-187. doi: 10.1177/0734282914544922
- Marsh, H. W., & Hocevar, D. (1988). A new, more powerful approach to multitrait-multimethod analyses: Application of second-order confirmatory factor analysis. *Journal of Applied Psychology, 73*, 107-117.
- Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S., & Von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right – camouflaging misspecification with item parcels in CFA models. *Psychological Methods, 18*, 257-284. doi: 10.1037/a0032773
- Mascendaro, P. M., Herman, K. C., & Webster-Stratton, C. (2012). Parent discrepancies in ratings of young children's co-occurring internalizing symptoms. *School Psychology Quarterly, 27*, 134-143. doi: 10.1037/a0029320

- McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60*, 577-605. doi: 10.1146/annurev.psych.60.110707.163612
- McArdle, J. J., & Hamagami, F. (2001). Linear dynamic analyses of incomplete longitudinal data. In L. Collins & A. Sayer (Eds.), *Methods for the analysis of change* (pp. 137–176). Washington, DC: APA Press.
- McClelland, M. M., & Scalzo, C. (2006). Social skills deficits. In M. Hersen (Ed.). *Clinician's handbook of child behavioral assessment*. (pp. 313-336). Burlington, MA: Elsevier Academic Press.
- McGuire, W. J. (1969). Suspiciousness of experimenter's intent. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 15-47). New York: Academic Press.
- McLaughlin, K. A., Breslau, J., Greif Green, J., Lakoma, M. D., Sampson, N. A., Zaslavsky, A. M., & Kessler, R. C. (2011). Childhood socio-economic status and the onset, persistence, and severity of DSM-IV mental disorders in a US national sample. *Social Science and Medicine, 73*, 1088-1096. doi:10.1016/j.socscimed.2011.06.011
- McMahon, R. J., Wells, K. C., & Kotler, J. S. (2006). Conduct problems. In E. J. Mash & R. A. Barkley (Eds.), *Treatment of childhood disorders* (3rd ed., pp. 137-268). New York, NY: Guilford.
- Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*, 525-543.
- Merrell, K. W. (2007). *Behavioral, social and emotional assessment of children and adolescents* (3rd ed.). New York, New York: Lawrence Erlbaum Associates.

- Merydith, S. P., Prout, H. T., & Blaha, J. (2003). Social desirability and behavior rating scales: An exploratory study with the Child Behavior Checklist/4-18. *Psychology in the Schools, 40*, 225-235.
- Meyer, G.J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R.,...Reed, G.M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*, 128-165. doi: 10.1037//0003-066X.56.2.128
- Mischel, W. (1968). *Personality and assessment*. New York: Wiley.
- Moreno, J., Silverman, W. K., Saaverdra, L. M., & Phares, V. (2008). Fathers' ratings in the assessment of their child's anxiety symptoms: A comparison to mothers' ratings and their associations with paternal symptomatology. *Journal of Family Psychology, 22*, 915-919. doi: 10.1037/a0014097
- MTA Cooperative Group. (1999). A 14-month randomized clinical trial of treatment of attention deficit hyperactivity disorder (ADHD). *Archives of General Psychiatry, 56*, 1073-1086.
- Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.
- Myers, C. L., Bour, J. L., Sidebottom, K. J., & Murphy, S. B. (2010). Same constructs, different results: Examining the consistency of two behavior-rating scales with referred preschoolers. *Psychology in the Schools, 47*, 205-216. doi: 10.1002/pits.20465
- NICHD Early Child Care Research Network. (1993). *The NICHD Study of Early Child Care: A comprehensive longitudinal study of young children's lives*. (ERIC Document Reproduction Service No. ED 353 0870).
- Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*, 1-18.

- Oord, S. V. D. , Prins, P. J. M., Oosterlaan, J., & Emmelkamp, P. M. G. (2006). The association between parenting stress, depressed mood, and informant agreement in ADHD and ODD. *Behaviour Research and Therapy*, *44*, 1585–1595. doi:10.1016/j.brat.2005.11.011
- Penney, S. R., & Skilling, T. A. (2012). Moderators of informant agreement in the assessment of adolescent psychopathology: Extension to a forensic example. *Psychological Assessment*, *24*, 386-401. doi: 10.1037/a0025693
- Pohl, S., & Steyer, R. (2010). Modeling common traits and method effects in multitrait-multimethod analysis. *Multivariate Behavioral Research*, *45*, 45-72. doi: 10.1080/00273170903504729
- Pohl, S., & Steyer, R. (2012). Modeling traits and method effects as latent variables. In S. Salzborn, E. Davidov, & J. Reinecke (Eds.), *Methods, theories, and empirical applications in the social sciences: Festschrift for Peter Schmidt* (pp. 57-65). Berlin: Springer VS.
- Pohl, S., Steyer, R., & Kraus, K. (2008). Modeling method effects as individual causal effects. *Journal of the Royal Statistical Society*, *171*, 41-63.
- Radloff, L. S. (1977). The CES-D scale: A self report depression scale for research in a general population. *Applied Psychological Measurement*, *1*, 385-401.
- Reiss, F. (2013). Socioeconomic inequalities and mental health problems in children and adolescents: A systematic review. *Social Science and Medicine*, *90*, 24-31. doi: 10.1016/j.socscimed.2013.04.026
- Reitman, D. (2006). Overview of behavioral assessment with children. In M. Hersen (Ed.). *Clinician's handbook of child behavioral assessment*. (pp. 4-24). Burlington, MA: Elsevier Academic Press.

- Reynolds, C. R., & Kamphaus, R. W. (2006). *BASC-2: Behavior Assessment System for Children, Second Edition*. Upper Saddle River, NJ: Pearson Education, Inc.
- Reynolds, C. R., & Kamphaus, R. W. (2015). *BASC-3: Behavior Assessment System for Children, Third Edition*. Upper Saddle River, NJ: Pearson Education, Inc.
- Reynolds, W. M. (2002). *Reynolds Adolescent Depression Scale – Second Edition: Professional Manual*. Odessa, FL: Psychological Assessment Resources.
- Richters, J., & Pellegrini, D. (1989). Depressed mothers' judgments about their children: An examination of the depression-distortion hypothesis. *Child Development, 60*, 1068-1075.
- Sarason, I., Johnson, J., & Siegel, L. (1978). Assessing the impact of life changes: Development of the Life Experiences Survey. *Journal of Consulting and Clinical Psychology, 46*, 932-946.
- Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment, 29*, 347-363. doi: 10.1177/0734282911406661
- Satorra, A., & Bentler, P. M. (1999). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507-512.
- Schroeder, J. F., Hood, M. M., & Hughes, H. M. (2010). Inter-parent agreement on the syndrome scales of the Child Behavior Checklist (CBCL): Correspondence and discrepancies. *Journal of Child & Family Studies, 19*, 646-654. doi: 10.1007/s10826-010-9352-0
- Seiffge-Krenke, I., & Kollmar, F. (1998). Discrepancies between mothers' and fathers' perceptions of sons' and daughters' problem behaviour: A longitudinal analysis of parent-adolescent agreement on internalising and externalising problem behaviour. *Journal of Child Psychology and Psychiatry, 39*, 687-697.

- Sollie, H., Larsson, B., & Mørch, W. (2013). Comparison of mother, father, and teacher reports of ADHD core symptoms in a sample of child psychiatric outpatients. *Journal of Attention Disorders, 17*, 699-710. doi: 10.1177/1087054711436010
- Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology, 15*, 201-293.
- Spearman, C. (1927). *The abilities of man*. New York, New York: Macmillan.
- Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory (Form Y)*. Palo Alto, CA: Consulting Psychologists Press.
- Spielberger, C. D., Jacobs, G. A., Russell, S. F., & Crane, R. S. (1983). Assessment of anger: The State-Trait Anger Scale. In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment: Volume 2* (pp. 161-189). Hillsdale, NJ: Erlbaum.
- Stanger, C., & Lewis, M. (1993). Agreement among parents, teachers, and children on internalizing and externalizing behavior problems. *Journal of Clinical Child Psychology, 22*, 107-115.
- Stark, K. D., Sander, J., Hauser, M., Simpson, J., Schnoebelen, S., Glen, R., & Molnar, J. (2006). Depressive disorders during childhood and adolescence. In E. J. Mash & R. A. Barkley (Eds.), *Treatment of Childhood Disorders* (3rd ed., pp. 336-407). New York, NY: Guilford.
- Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*, 173-180.
- Sternberg, R. J. (1992). Psychological Bulletin's Top 10 "hit parade." *Psychological Bulletin, 112*, 387-388.

- Stone, S. L., Speltz, M. L., Collett, B., & Werler, M. M. (2013). Socioeconomic factors in relation to discrepancy in parent versus teacher ratings of child behavior. *Journal of Psychopathology and Behavioral Assessment, 35*, 314-320. doi: 10.1007/s10862-013-9348-3
- Swanson, J. M. (1992). *School based assessments and interventions for ADD students*. Irvine, CA: K.C.
- Tabachnik, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson Education, Inc.
- Terry, N. P., & Irving, M. A. (2010). Cultural and linguistic diversity: Issues in education. In R. P. Colarusso & C.M. O'Rourke (Eds.), *Special education for all teachers*, (5th ed., pp. 109-132). Dubuque, IA: Kendall Hunt Publishing Company.
- Thompson, M. S., & Green, S.B. (2013). Evaluating between-group differences in latent variable means. In G. R. Hancock and R. O. Mueller (Eds.), *Structural Equation Modeling: A Second Course*. (163-218). Charlotte, NC: Information Age Publishing.
- Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*, 25-29.
- Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago, Illinois: University of Chicago.
- Treutler, C. M., & Epkins, C. C. (2003). Are discrepancies among child, mother, and father reports on children's behavior related to parents' psychological symptoms and aspects of parent-child relationships? *Journal of Abnormal Child Psychology, 31*, 13-27.
- United States Census Bureau. (March 24, 2011). 2010 Census shows America's diversity (Press Release CB11-CN. 125). Retrieved from <https://www.census.gov/2010census/news/releases/operations/cb11-cn125.html>

- Van der Ende, J., & Verhulst, F. C. (2005). Informant, gender and age differences in ratings of adolescent problem behavior. *European Child and Adolescent Psychiatry, 14*, 117-126. doi:10.1007/s00787-005-0438-y
- Van Dulmen, M. H. M., & Egeland, B. (2011). Analyzing multiple informant data on child and adolescent behavior problems: Predictive validity and comparison of aggregation procedures. *International Journal of Behavioral Development, 35*, 84-92. doi: 10.1177/0165025410392112
- Waschbusch, D. A., Sparks, S. J., & Northern Partners in Action for Child and Youth Services. (2003). Rating scale assessment of attention-deficit/hyperactivity disorder (ADHD) and oppositional defiant disorder (ODD): Is there a normal distribution and does it matter? *Journal of Psychoeducational Assessment, 21*, 261-281. doi: 10.1177/073428290302100303
- Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*, 1-26. doi: 10.1177/014662168500900101
- 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