Indiana University of Pennsylvania
Knowledge Repository @ IUP

Theses and Dissertations (All)

5-12-2008

Teacher Behavior Ratings of Adolescents with Attention-Deficit Hyperactivity Disorder (ADHD): Interrater Reliability and Sources of Rater Bias

Brandon K. Schultz
Indiana University of Pennsylvania

Follow this and additional works at: http://knowledge.library.iup.edu/etd

Recommended Citation
Schultz, Brandon K., "Teacher Behavior Ratings of Adolescents with Attention-Deficit Hyperactivity Disorder (ADHD): Interrater Reliability and Sources of Rater Bias" (2008). *Theses and Dissertations (All)*. 77.
http://knowledge.library.iup.edu/etd/77

This Dissertation is brought to you for free and open access by Knowledge Repository @ IUP. It has been accepted for inclusion in Theses and Dissertations (All) by an authorized administrator of Knowledge Repository @ IUP. For more information, please contact cclouser@iup.edu, sara.parme@iup.edu.

TEACHER BEHAVIOR RATINGS OF ADOLESCENTS WITH ATTENTION-DEFICIT HYPERACTIVITY DISORDER (ADHD): INTERRATER RELIABILITY AND SOURCES OF RATER BIAS

A Dissertation

Submitted to the School of Graduate Studies and Research

in Partial Fulfillment of the

Requirements for the Degree

Doctor of Education

Brandon K. Schultz

Indiana University of Pennsylvania

May 2008

Indiana University of Pennsylvania
The School of Graduate Studies and Research
Department of Educational and School Psychology

We hereby approve the dissertation of

Brandon Kyle Schultz

Candidate for the degree of Doctor of Education

Joseph F. Kovaleski, D.Ed.
Professor of Psychology, Advisor

William F. Barker, Ph.D.
Professor of Psychology

Lynanne Black, Ph.D.
Assistant Professor of Psychology

Steven W. Evans, Ph.D.
Professor of Graduate Psychology
James Madison University

ACCEPTED

Michele S. Schwietz, Ph.D.
Assistant Dean for Research
The School of Graduate Studies and Research

Title: Teacher Behavior Ratings of Adolescents with Attention-Deficit Hyperactivity Disorder (ADHD): Interrater Reliability and Sources of Rater Bias

Author: Brandon K. Schultz

Dissertation Chair: Dr. Joe Kovaleski

Dissertation Committee Members:

William F. Barker

Steven W. Evans

Lynanne Black

This study examined interrater reliability and potential sources of rater bias among teacher behavior ratings for adolescents with Attention-Deficit Hyperactivity Disorder (ADHD). The intent of the study was two-fold: 1) to assess the consistency between teacher behavior ratings for adolescents with ADHD, and 2) to explore potential sources of bias among teacher raters.

In schools, intervention decisions for children and adolescents with ADHD are often based on rating scale data collected from classroom teachers. However, research has shown that teacher behavior ratings are

oftentimes incongruent, especially at the secondary school level. Furthermore, behavior rating scales are generally viewed to be highly susceptible to rater bias. For example, teacher raters often provide relatively lenient or severe judgments, compared to the judgments of other teachers. While rater inconsistencies and rater bias are occasionally discussed in the professional literature, few studies have directly examined between-teacher reliability in secondary schools and the sources of bias that explain interrater inconsistencies. The present study examined interrater reliability and potential sources of rater bias in teacher ratings of middle school students with ADHD.

ACKNOWLEDGEMENTS

I would like to thank everyone who helped make this project possible, beginning with my dissertation committee for their support and responsiveness. In particular, I would like to thank Dr. Steve Evans for inviting me to participate in the research on the Challenging Horizons Program and for providing thoughtful guidance throughout. I would also like to thank the late Alvin V. Baird, Jr., whose generosity set the groundwork for the Attention and Learning Disabilities Center (ALDC) at James Madison University and made this research possible.

Very special thanks go to my fiancé Debbie for her support and patience during long months of reading, writing, and rewriting. We have sacrificed much in the past two years leading up to our wedding, and I am grateful for her willingness to endure these sacrifices for so long, with such good humor.

And finally, I would like to thank my family and friends for being so incredibly sympathetic. My mother, in particular, has unquestioningly accepted that seven years is a normal timeframe to complete a dissertation. Mom, you were right all along – I will be a student my whole life!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Austin's elementary teachers had noted that he seemed "out of it" and that he often daydreamed, but otherwise he performed well academically. This changed shortly after Austin encountered the new demands of middle school, which included organizing a locker, managing a multiple-course notebook, tracking assignments, and organizing a bookbag. Austin was quickly overwhelmed by his new responsibilities.

By seventh grade, the situation had worsened. Austin frequently forgot assignments, lost his work, and failed tests and quizzes. His grades were falling in several classes and, without intervention, there was a chance he would have to repeat the grade. At the request of Austin's mother, the school psychologist conducted a psychoeducational evaluation including cognitive and academic measures, as well behavior ratings, which were sent to Austin's teachers.

In a few days, the teachers' ratings were returned, but there was very little consistency between teachers regarding the nature and severity of Austin's difficulties. Austin's Social Science and Reading teachers rated his level of inattention as significantly

high, while his science teacher rated his inattention in a range only slightly above that of his same-age peers. A fourth rating scale, sent to Austin's math teacher, suggested that Austin did not exhibit attention problems, but there was some indication that Austin was exceedingly fidgety and restless when compared to his same-age peers. Based on the discrepancies between the teachers' ratings, the school psychologist was uncertain if Austin's behaviors were indicative of a chronic disorder, such as Attention Deficit Hyperactivity Disorder (ADHD), or if each teacher had a unique perspective on Austin's behaviors and interpreted the rating scales in idiosyncratic ways.

Statement of the Problem

One of the most common assessment methods for ADHD is behavior rating scales. Behavior rating scales require respondents to rate the degree to which children exhibit behaviors of interest. There are two general types of ratings scales, including broad- and narrow-band scales. Broad-band scales are designed to assess many potential behavior problems, while narrow-band scales assess behaviors associated with a specific disorder or other clinically relevant phenomena. So, in the assessment of ADHD, a broad-band scale may be used to

assess a wide range of behavior problems including ADHD and potential comorbidities, and narrow-band scales may be used to specifically assess inattention, hyperactivity, and impulsivity. In general, data gathered through broad- and narrow-band rating scales help school psychologists and other professionals in diagnosis, intervention planning, and assessment of treatment outcomes. If, for example, teachers rate a child's level of hyperactivity as significantly high, a school psychologist might suggest an intense behavior modification intervention or perhaps an alternative classroom placement, if warranted. As a result, there is tremendous weight placed upon the data collected from rating scales.

However, behavior rating scales have significant limitations, including the fact that they are highly susceptible to the personal biases of the raters. Rater bias is often evidenced through inconsistencies between raters and through rater-specific response styles. As demonstrated in the scenario of Austin above, school psychologists are often confronted with conflicting rating scales results when multiple sources are used, and this often makes interpretation difficult.

Research Questions

The present study addressed two fundamental questions: First, to what degree are teachers consistent when rating young adolescents with ADHD? Second, can discrepancies between their ratings be, at least to some degree, attributable to teacher characteristics? In other words, are teacher characteristics, such as teaching experience and age, associated with unusually extreme ratings, either high or low, on behavior rating scales when rating young adolescents with ADHD?

Thus, the objective of the present study was two-fold: First, the study evaluated interrater reliability among teachers' behavior ratings of adolescents with Attention-Deficit Hyperactivity Disorder (ADHD). Achenbach, McConaughy, and Howell (1987) conducted a landmark meta-analysis of interrater reliability studies and found average interrater correlations of .64 across all studies, and a trend toward poorer reliability when rating adolescent targets. However, this meta-analysis included only three studies that examined between-teacher reliability when rating adolescents, so the trend toward weaker reliability rates for teachers at the secondary level were unclear. Molina, Pelham, Blumenthal, and Galiszewski (1998) looked specifically at interrater

reliability among secondary teachers' ratings of adolescents with ADHD and found very low reliability (intraclass correlations [ICCs] ranged from .21 to .52) suggesting that interrater reliability among secondary teachers was poorer than that found by Achenbach and colleagues (1987). The first aim of the present study was to replicate the study performed by Molina and colleagues using a similar narrow-band teacher rating scales of ADHD symptomology and items from a broad-band rating scale of impairment.

Second, potential sources of rater bias were assessed by evaluating how well teacher characteristics predicted severe and lenient target ratings relative to that of other teachers. Basic demographic data (e.g., age, sex, years of teaching experience) were collected from teachers who provided ratings of 79 middle school students with ADHD and, using multiple regression, the relative severity and leniency of teacher ratings were regressed onto these data. Similar approaches to assessing sources of rater bias were employed by Hill, O'Grady, and Price (1988) and Hoyt (2002). In Hill and colleagues' study, the authors were largely unsuccessful using this technique, due partly to high interrater reliability prior to analysis. In the proposed study, it

was anticipated that there would be low to moderate interrater reliability (see hypotheses below), thus providing ample variance for the regression analysis.

Hypotheses

Consistent with previous research, it was anticipated that teacher behavior ratings of adolescents with ADHD would be highly inconsistent (e.g., Molina et al., 1998), as evidenced through ICCs similar to those found by Molina and colleagues (i.e., .21 to .52), and less than that found by Achenbach and colleagues (1987). The researcher also anticipated that the differences between teacher ratings would be predicted by individual teacher characteristics, including sex, age, subject taught (specific to target), teaching experience, highest academic degree, whether or not the teacher was a parent, prior experience teaching children with disabilities, and average class size. There was very little prior research to support specific hypotheses as to the direction and influence of potential moderators, so no specific hypotheses regarding the direction of predictor variables were tenable prior to the analysis.

Problem Significance

Rating scales from multiple informants are recommended in the assessment of childhood psychiatric

disorders (American Academy of Child and Adolescent Psychiatry, 1997; American Academy of Pediatrics, 2001). Unfortunately, ratings from multiple sources are often incongruent, due to differences in the expression of the disorder over time, the context of assessment, rater biases, and random measurement error (Kraemer, Measelle, Ablow, Essex, Boyce, & Kupfer, 2003). When the aim of assessment is to measure childhood disorder and not the raters' perceptions of that disorder, such as in the case of clinical diagnosis, rater bias can be a troubling source of interrater discordance. However, the issue of rater bias is rarely addressed in the professional literature. For instance, in an electronic search of the terms "rater bias" and "source bias" in the titles of articles available in PsychINFO and PsychARTICLES databases, only 22 unique results were found, spanning the years from 1963 to 2005, and four were dissertations. Furthermore, school psychologists who rely on rating scales to assess and monitor student progress have little guidance when considering discrepancies in ratings and potential rater biases. The few guidelines that are available typically advise methods of combining or cancelling out discrepant data based on diagnostic concerns (e.g., Hart, Lahey, Loeber, & Hanson, 1994;

Simonoff, Pickles, Hewitt, Silberg, Rutter, Loeber, et al., 1995).

Research on rater bias among secondary teachers has the potential to explain the low interrater reliability rates that have been reported in the literature. Thus, this research has the potential to improve school psychologists' interpretations of inconsistent rating scale results, and to inform methods for weighting or adjusting teacher ratings based on teacher characteristics.

Definition of Terms

Bias – from a psychometric standpoint, this term refers to measurement error variance resulting from disagreement between sources. Disagreements can arise from perceptual errors or from accurate perceptions of differing target behavior. Thus, bias is not necessarily the same as measurement inaccuracy. In the present study, when not otherwise indicated, the term bias will be used to refer to leniency/severity effects (described below). In the professional literature, "rater bias" is sometimes referred to as "source bias" (e.g., Hill, O'Grady, & Price, 1988).

Error – variance in ratings that is uncorrelated with the true score. "Random error" refers to error in ratings

that cannot be accounted for by identifiable factors, and "systematic error" refers to error that can be attributable to identifiable factors.

Halo – the tendency for raters to base their judgments of a target on irrelevant target characteristics (e.g., child is rated less hyperactive than is warranted due to the rater's positive impression of the child's humor). Halo is a source of dyad-specific rater bias because it is thought to vary across rater-target dyads.

Interrater Agreement – the extent to which two or more raters make the exact judgment about a single target, such as when ratings are used to determine diagnosis or non-diagnosis. This concept is also referred to "absolute agreement" at various times in the text.

Interrater Reliability – the degree to which the ratings from two or more raters are proportional in terms of the deviation from their respective means. In other words, interrater reliability refers to the correlation (relative consistency) between rater judgments. Readers should note that the term "between-teacher reliability" is occasionally used to refer to this concept throughout the text when referring to interrater reliability among teacher raters.

Leniency/Severity – the tendency for raters to provide ratings of a target that are overly forgiving (leniency) or harsh (severity). Leniency and severity represent rater-specific biases that are theoretically consistent across targets.

Rater – this term will be used to describe teachers who submit rating scales for adolescents in the study (targets). Elsewhere, the term "rater" is used to describe any adult (e.g., parents, teachers, caregivers) that provide rating scale data; however, the focus of the present study is limited to teacher ratings. Further, readers should note that the terms "rater," "teacher," and "source" are used interchangeably at times in this text.

Target – this term will be used to describe the adolescents for whom behavior ratings were submitted. Readers should note that the term "target," "student," and "adolescent" are used interchangeably at times in the text.

CHAPTER II

REVIEW OF THE RELATED LITERATURE

Attention-Deficit Hyperactivity Disorder (ADHD)

ADHD is one of the most researched childhood-onset psychiatric disorders, with more than 6000 peer-reviewed articles published to date (Barkley, 2006). Thus, a comprehensive review of the existing literature is clearly beyond the scope of this chapter. Instead, this chapter will focus on the research leading up to, and following, the publication of the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV; American Psychiatric Association [APA], 1994) and DSM-IV Text Revision (DSM-IV-TR; APA, 2000). In some instances, the research cited herein is based on previous DSM criteria, but the relatively minor changes to attention deficit nosologies over time allow for "some clinical generalities to be drawn about the literature" (Barkley, 2006, p. 76). In the first section, the chapter will provide an overview of the DSM-IV(-TR) diagnostic criteria for ADHD, as well as the etiology and course of the disorder. Then in the following sections, the discussion will focus on the problems associated with assessing ADHD and the current lack of an indisputable "gold standard" assessment procedure. Finally, this

chapter will examine issues related to behavior rating scales, including problems associated with interrater reliability and rater bias.

DSM-IV(-TR) Criteria for ADHD

According to the DSM-IV(-TR), the cardinal symptoms of ADHD are marked and persistent impairment in attention and/or regulating one's activity level, as compared to other individuals of the same developmental level. Within community samples the disorder is thought to affect 3%-7% of school-aged children, with males being more commonly diagnosed than females at a ratio of 2:1 to 9:1 (American Psychiatric Association [APA], 2000). Among children referred for clinical psychiatric evaluation, the rate of ADHD is much higher, sometimes exceeding 50% (Barkley, 1998). A recent study of child and adolescent outpatient psychiatric patients found that ADHD was the most common diagnosis (43%) and frequently co-occurred with other behavior disorders and anxiety or mood disorders (Staller, 2006).

To be diagnosed with ADHD, individuals must exhibit six or more behavioral symptoms of inattention and/or hyperactivity-impulsivity for longer than six months, in two or more settings (e.g., home and school), with impairment in social, familial, or academic functioning

beginning prior to age seven. The criteria used to define inattention include: failure to give close attention to details in work; difficulty sustaining attention in activities; appearing to not listen when spoken to directly; failure to follow through on instructions; difficulty organizing tasks and activities; avoiding tasks that require sustained mental effort; losing things necessary for tasks; distractibility; and forgetfulness. The criteria used to define hyperactivity include: fidgeting with hands or feet; leaving one's seat inappropriately; running about or climbing excessively; difficulty engaging in leisure activities quietly; appearing "on the go" or acting as if "driven by a motor"; and excessive talking. The criteria used to define impulsivity include blurting out answers before a question can be asked; difficulty waiting for one's turn; and interrupting or intruding on others (APA, 2000).

Subtypes

Based on the relative distribution of symptoms, the DSM-IV(-TR) recognizes three subtypes of ADHD: Predominately Inattentive (ADHD-PI), Predominately Hyperactive-Impulsive (ADHD-HI), or Combined subtype (ADHD-C). ADHD-PI is defined as six or more symptoms of inattention with five or fewer symptoms of hyperactivity-

impulsivity. ADHD-HI is defined as six or more symptoms of hyperactivity and impulsivity (which are combined into one dimension, referred to as hyperactivity-impulsivity) with five or fewer symptoms of inattention. Finally, ADHD-C is defined as six or more symptoms of both inattention and hyperactivity-impulsivity. Additionally, the DSM-IV(-TR) outlines a "Not Otherwise Specified" category (ADHD-NOS), which is used in instances where ADHD symptoms are manifest, but more information is needed (APA, 2000).

Previously in the DSM-III-R, ADHD was described as a unitary disorder without subtypes. However, this conceptualization did not seem to comport with the existing literature or prevailing clinical wisdom that viewed children with ADHD as a highly diverse and heterogeneous population. In the field trials leading up to the DSM-IV, researchers in the Child Disorders Work Group conducted clinical interviews with parents, teachers, and children and examined how their responses to ADHD symptoms related to their functional impairment and the diagnostic opinions of trained clinicians. The results suggested that the DSM-IV criteria for inattention and hyperactivity-impulsivity represented two separate dimensions of behavior, based on differential

impairments (i.e., hyperactivity-impulsivity was associated with global impairments, while inattention was most associated with academic impairment) and the fact that inattentive and hyperactive-impulsive symptoms were differentially related to clinical diagnosis (Lahey, Applegate, McBurnett, Biederman, Greenhill, Hynd, et al., 1994). Further, the items measuring hyperactivity and impulsivity were found to strongly correlate, thus supporting a single hyperactivity-impulsivity dimension in the DSM-IV (Frick, Lahey, Applegate, Kerdyck, Ollendick, Hynd, et al., 1994). The field trials also informed the decision to use six inattentive and six hyperactive-impulsive symptoms as the diagnostic threshold for determining ADHD subtypes, as described above, based on the clinical significance of the associated impairments (Lahey et al., 1994).

Since the publication of the DSM-IV, many studies using confirmatory factor analysis have supported the two-factor conceptualization of ADHD when measured by parent and teacher behavior ratings. For example, Burns, Boe, Walsh, Sommers-Flannagan, and Teegarden (2001) found that the two-factor model of ADHD fit mother rating scale data better than four competing models. The two-factor solution was consistent for boys and girls from early

childhood through adolescence. Other disruptive behavior disorders, such as Oppositional Defiant Disorder (ODD), appeared to represent separate but related factors. Similarly, a two-factor solution has been found for teacher ratings of ADHD (DuPaul, Power, Anastopulous, Reid, McGoey, & Ikeda, 1997). As hypothesized, the three items used to measure impulsivity (i.e., blurts out answers, impatience, and interrupts or intrudes on others) loaded strongly on the hyperactivity factor among teacher ratings. Further, the two-factor solution for teacher ratings fit the data in both European and American samples, across rural and urban settings (Wolraich, Lambert, Baumgaertel, Garcia-Tornel, Feurer, Bickman, et al., 2003).

Based on the factor structures of both parent and teacher ratings of ADHD, it appears that the current conceptualization of ADHD is valid. However, more research is needed to assess the potentially unique contributions of impulsivity in explaining long-term outcomes, such as conduct problems and antisocial behavior (White, Moffitt, Caspi, Bartusch, Needles, & Stouthamer-Loeber, 1994), which occur in a minority of ADHD cases (Burns et al., 2001). Impulsivity symptoms appear to load on a separate factor in some samples

(e.g., Amador-Campos, Forns-Santacana, Guàrdia-Olmos, & Peró-Cebollero, 2006), but the conditions that produce a three-factor solution are unclear.

Other research on ADHD has suggested a third subset of symptoms that closely mimic inattention, including increased daydreaming, mental torpidity, tendency toward confusion, and physical hypoactivity. Collectively, these symptoms have been termed Sluggish Cognitive Tempo (SCT; Barkley, 2006). However, field trials for the DSM-IV found that although SCT symptoms were correlated with the Predominately Inattentive subtype, they were not experienced by the majority of children. Hence, SCT criteria were not included as diagnostic criteria (Hartman, Willcutt, Rhee, & Pennington, 2004).

New challenges to the current conceptualization of ADHD are emerging in the neuropsychological literature. For example, although the DSM-IV(-TR) criteria conceptualize inattention in a monolithic fashion (i.e., single class of behaviors), research using neuropsychological tests of attention suggest that it is actually multidimensional. Specific subcomponents of attention, such as set shifting and vigilance, may independently vary within children identified with ADHD (e.g., Levine, 2002). However, such subcomponents of

attention have not been adequately validated by the existing research (Strauss, Thompson, Adams, Redline, & Burant, 2000). Thus, the DSM-IV(-TR) criteria for inattention assume a unitary construct (i.e., no subcomponents), which is differentiated from the hyperactivity-impulsivity factor and does not necessarily include the symptoms associated with SCT. Research on the subcomponents of attention will be discussed in greater detail in the "Objective Measures" section later in this chapter.

Age of Onset

The age-of-onset criterion (AOC), which requires an onset of functional impairment prior to age seven, has come under intense scrutiny. This criterion, which first appeared in the DSM-III, was not supported by research, and specific rationales were not provided for its inclusion. Rather, the AOC has remained in the DSM "more out of tradition" than for any other plausible reason (Barkley & Biederman, 1997, p.1207). Field trials for the DSM-IV examined the AOC, but the results were not available until after publication. When finally examined, the field trial data suggested that 43% of predominately inattentive subtype, 18% of combined subtype, and 2% of predominately hyperactive-impulsive

subtype cases did not experience significant impairment until after age seven (Applegate, Lahey, Hart, Biederman, Hynd, Barkley, et al., 1997). In some instances the AOC may discriminate against girls, who are less likely to exhibit hyperactivity-impulsivity and, as a result, appear to experience ADHD-related impairments later than boys (Cuffe, McKeown, Jackson, Addy, Abramson, & Garrison, 2001). Further, the AOC can be particularly difficult to establish when diagnosing adolescents and adults, due to poor recall among sources. Given such limitations, Barkley and Biederman (1997) conclude that the AOC is "arbitrary, surely discriminatory, and empirically indefensible," and recommend that it be generously interpreted in clinical practice (p. 1208).

Differential Diagnoses

Oftentimes children exhibit inattention, hyperactivity, or impulsivity for reasons not necessarily attributable to ADHD. As a result, the DSM-IV(-TR) lists several diagnoses that must be ruled-out before an ADHD diagnosis is appropriate. For example, if the symptoms are better accounted for by other diagnoses, such as Mood disorder, Anxiety Disorder, Dissociative Disorder, or a Personality Disorder, the diagnosis of ADHD is unwarranted. Further, symptoms cannot occur directly as

a result of Pervasive Developmental Disorder, Schizophrenia, or other Psychotic Disorder (APA, 2000, 1994).

The literature on ADHD discusses additional diagnostic concerns that are not directly addressed by the DSM-IV(-TR) criteria. For example, differential diagnosis appears particularly difficult in early childhood, as some children with mental retardation can exhibit attention-related difficulties. Although the DSM-IV(-TR) allows for children with mental retardation to be diagnosed with ADHD, the problems must be deemed excessive, given the child's mental age. Some research suggests that a lower IQ threshold should be established to exclude behaviors attributable to severe forms of mental retardation (Barkley, 2006), but these concerns are not reflected in the DSM-IV(-TR). In other instances, head injuries or central nervous system damage can mimic the impairments associated with ADHD, so careful screening is required to rule out organic brain or central nervous system damage as a better explanation for inattention or hyperactivity (Evans, Vallano, & Pelham, 1995). At the other end of the intellectual spectrum, gifted children can be misdiagnosed with ADHD due to academic boredom (e.g., daydreaming and off-task

behavior). In other cases, gifted children with ADHD may not be diagnosed because adults cherish their intellectual strengths and overlook symptoms such as disorganization or talkativeness (Webb, Amend, Webb, Goerss, Beljan, & Olenchak, 2005).

When diagnosing ADHD, it is also important to consider the role of potential comorbid psychiatric disorders. For example, children and adolescents with ADHD commonly exhibit comorbid externalized disorders such as Oppositional Defiant Disorder (ODD) or Conduct Disorder (CD), and/or internalized disorders such as anxiety and mood disorders (Jensen, Hinshaw, Kraemer, Lenora, Newcorn, Abikoff, et al., 2001). This complicates the clinical phenomenology of ADHD considerably and makes diagnosis particularly difficult in some cases. The DSM-IV(-TR) specifies that if another psychiatric disorder such as anxiety better explains symptoms of inattention, for example, then the diagnosis of ADHD is unwarranted (APA, 2000). Hence, diagnosis requires adequate screening for other psychiatric conditions and clinical judgment in determining if other psychiatric symptoms are comorbid or a better explanation of the symptoms. Such complications have led to research into the possibility that ADHD and specific comorbidities

represent separate and distinct clinical subtypes (e.g., Jensen, Hinshaw, et al., 2001), but to date, the prevailing nosology only recognizes the three subtypes described above.

Impairment

DSM-IV(-TR) criteria for ADHD require evidence for significant functional impairment in social, academic, or family domains. If the observed symptoms do not result in significant impairment, a diagnosis of ADHD is unwarranted. According to parents and educator reports, the impairments most commonly associated with ADHD include social difficulties, academic underachievement, and disrupted relationships with adults (Evans, Vallano, & Pelham, 1995). Such impairments appear to predict long-term outcomes better than ADHD symptoms alone (Pelham, Fabiano, & Massetti, 2005).

Social difficulties. Over time, the professional literature has increasingly recognized social problems as a serious issue for many children with ADHD (Landau & Moore, 1991), especially among children with hyperactivity-impulsivity (Gadow, Drabick, Loney, Sprafkin, Salisbury, Azizian, et al., 2004; Lahey et al., 1994) and/or aggression (Bagwell, Molina, Pelham, & Hoza, 2001; Hinshaw, Zupan, Simmel, Nigg, & Melnick, 1997).

Children with ADHD-PI can also exhibit social impairments, but are more likely than their hyperactive-impulsive peers to be withdrawn or shy (Hodgens, Cole, & Boldizar, 2000). Interestingly, significant social problems can occur even in the absence of comorbid disorders. For example, ADHD appears to uniquely contribute to peer rejection above that for adolescents with comorbid CD (Bagwell et al., 2001).

In terms of specific social deficits, it is frequently reported that children with ADHD exhibit communication problems, including dysfluent (e.g., shifting and non sequitur) speech patterns. Children with ADHD are also likely to have deficient social problem-solving skills and are more likely than their undiagnosed peers to anticipate desirable consequences for aggressive behavior (Dumas, 1998). Unlike children with severe developmental disabilities where social learning is impeded, children with ADHD learn social skills but are unable to perform them effectively at appropriate times. As a result, social problems are inconsistent for children with ADHD, as the ability to perform up to expectations is adversely impacted by behavioral excesses (Wheeler & Carlson, 1994). Thus, the current literature draws a distinction between social

skill deficits and social performance deficits, with ADHD associated mostly with the latter.

In social interactions, performance deficits commonly lead to two negative outcomes. First, children with ADHD are often actively rejected by their peers. In settings where unfamiliar children are allowed to create their own impressions of one another, children with ADHD are more likely than their normal peers to exhibit poor social skills, resulting in peer rejection. Such rejection can occur quickly, even within the first day that children meet one another (Erhardt & Hinshaw, 1994). In fact, prior to first meetings a social bias pertaining to ADHD may impact initial interactions. For example, when children without ADHD expect that they will soon play with a peer who exhibits ADHD-consistent behavior (e.g., talkativeness, disruptiveness), the quality of their subsequent shared activities are deleteriously impacted. In brief interactions between two unfamiliar children, such expectations result in less reciprocal play and more negative interactions, such as disagreements (Harris, Milich, Johnston, & Hoover, 1990). Unfortunately, once reputation biases develop they appear to persist, even when intense efforts are made to remediate ADHD symptoms through behavioral or

pharmacological interventions (Hoza, Gerdes, Mrug, Hinshaw, Bukowski, Gold, et al., 2005).

Second, children with ADHD (especially boys) appear to have unrealistically positive self-appraisals of their social performances, as compared to the appraisals of their peers (Diener & Milich, 1997; Ohan & Johnson, 2002) and teachers (Hoza, Pelham, Dobbs, Owens, & Pillow, 2002). Overly generous self-appraisals may serve to protect self-esteem, but often complicate intervention efforts because children with ADHD do not recognize a need for change. Interestingly, when children with ADHD are given positive feedback on their social interactions from their peers, the need for this self-protective bias is reduced and subsequent self-appraisals become more self-critical and consistent with that of others. One interpretation of these findings is that children with ADHD are motivated mostly to avoid appearing socially incompetent, and when this concern is assuaged, the self-protective illusory bias lessens (Hoza et al., 2002; Diener & Milich, 1997).

Academic underachievement. Students with ADHD are also likely to exhibit academic underachievement, which can often result in comorbid learning disabilities (LD). Methods of defining LD vary widely and, as a result,

varying rates of comorbid LD are found in the ADHD literature (Barkley, 2006). Higher rates of comorbid LD are often found among school samples as compared to community or clinic samples, as learning disabilities are most commonly diagnosed by school professionals (Staller, 2006), which is not surprising since school professionals are likely to observe the poor study habits, poor class participation, low test grades, and the poor relationships with teachers often associated with students with ADHD (Robin, 1998). Hence, studies utilizing samples from various environments are inconsistent in regards to comorbidity between ADHD and LD. When using a conservative diagnostic procedure requiring a significant discrepancy between IQ and achievement and an academic lag 1.5 standard deviations below the norm-referenced mean, Barkley (1990) found that 19% of children with ADHD had comorbid reading disabilities, 24% had comorbid spelling disabilities, and more than 26% had comorbid math disabilities. Based on such research, it appears that LD occurs much more frequently among children with ADHD than it does in the general population.

Even in instances that do not meet the criteria for LD, students with ADHD are likely to lag behind their

peers in the areas of spelling, reading, and math (Evans, Vallano, & Pelham, 1995; White, Barbour, Schill, Vodra, Garrett, Schultz, et al., 2005). At the elementary school level, academic problems often manifest as failure to complete assignments and less overall productivity relative to peers. By the secondary level, ADHD is associated with lower grades, more special education intervention, and higher rates of grade retention and drop-out as compared to normal peers (Anastopoulos & Shelton, 2001). Despite their challenges, children with ADHD are likely to overestimate their academic performance (Owens & Hoza, 2003), similar to the manner in which social performance is overestimated.

When analyzing the achievement goals of children with ADHD, it appears that unlike their normal peers, children with ADHD prioritize performance-avoidance goals over performance-approach goals. In other words, children with ADHD are generally motivated to avoid appearing incompetent, whereas other children without ADHD appear motivated to outperform their peers. An orientation toward performance-avoidance goals, like that found among children with ADHD, is associated with ineffective learning strategies and an intolerance for

academic challenge and frustration (Barron, Evans, Baranick, Serpell, & Buvinger, 2006).

Strained relationships with adults. Children with ADHD often experience strained relationships with adults. Of particular concern is the relationship between children with ADHD and their parents or guardians, which differ from that of normal peers beginning at early ages. For example, Stormshak and Bierman (2000) found that among a sample of 631 high-risk Kindergartners, hyperactivity was associated with elevated levels of punitive discipline by parents (i.e., threatening child with punishment, yelling, feeling angry when disciplining, spanking or hitting child). Other research suggests that parents of children with ADHD are more likely to resort to aggressive parenting tactics than are parents of normal children (e.g., Edwards, Barkley, Laneri, Feltcher, & Metevia, 2001). It also appears that parents fail to reinforce appropriate behavior and instead focus on punishing inappropriate behavior. As a result, it is hypothesized that some children with ADHD exhibit problem behaviors in an attempt to simply gain adult attention (Kazdin, 1997). For some families, the focus on inappropriate behavior leads to a pattern of harsh punishment that increases in severity over time

(Edwards et al., 2001). Among adolescents, ADHD is associated with more severe parent-adolescent conflict, especially when the child exhibits oppositional behaviors (e.g., arguing with adults, actively defying adults' requests). Thus, it is not surprising that parental measures of family cohesion and family interaction have been shown to negatively correlate with symptoms of ADHD, suggesting that as symptom severity increases, the quality of family functioning declines (Klassen, Miller, & Fine, 2004).

ADHD is also associated with strained relationships with teachers. Teacher-student relationship difficulties may be attributable in part to the general lack of teacher knowledge of issues related to ADHD, especially among preservice teachers and those with little classroom experience (Kos, Richdale, & Jackson, 2004). Research also suggests that teachers perceive students with ADHD as creating more stress in the classroom as compared to their normal peers, especially when ADHD is comorbid with social impairments and oppositional or aggressive behavior. Classroom observations suggest that students with ADHD command significantly more time from their teachers and a high proportion (but not all) of this time is spent in negative interaction (Greene, Beszterczey,

Katzenstein, Park, & Goring, 2002). The logical inference is that students who demand significantly more attention and have the propensity for negative interactions with teachers are likely to be perceived negatively, thus damaging the student-teacher relationship. However, the negative impact may not be limited to the student-teacher dyad, as it appears that such frustration can generalize to other students in the classroom, so that normal classmates experience negative interactions with the teacher as well (Stormont, 2001).

At the secondary school level, teacher relationships with students are further influenced by the school environment. Unlike elementary schools, where students generally interact with the same teachers throughout the entire day, teacher-student relationships in secondary schools are confined by the discrete and unconnected classroom arrangements typically found in these settings. Interestingly, secondary teacher reports of their relationships with students with ADHD suggests that there is more variation on this issue than there is on their ratings of academic performance or ADHD symptoms (Evans, Allen, Moore, Strauss, & Timmins, 2004).

Etiology

Although no definitive cause of ADHD has been found, current research has given rise to neuropsychological theories and potential biological explanations of the disorder that have received attention in the professional literature. The relevant research has also uncovered developmental changes in the disorder that help to predict long-term outcomes.

Theoretical Models

Theories of ADHD help to explain the deficits and impairments associated with the disorder and provide testable hypotheses that explain the nature and mechanisms of the disorder. To date, there is no single, definitive theory of ADHD. Rather, several competing theories are found in the literature, and these theories have stimulated varying paths of research. In general, most contemporary theories of ADHD (e.g., Barkley, 1997) focus on two psychological phenomena: dysregulation in the behavioral inhibition system and deficits in executive functioning.

The behavioral inhibition system (BIS) is posited to limit and control behavioral responses to environmental stimuli. There are several components to this system, which Barkley (2006) describes as the inhibition of

prepotent (i.e., immediately reinforcing) responses, the ability to discontinue behavioral responses based on environmental feedback, and the ability to screen out interfering or distracting stimuli. For children and adolescents with ADHD, all three areas appear to be impaired, but particularly the ability to inhibit prepotent responses (Nigg, 2000).

Executive functioning (EF) is a set of higher order cognitive processes associated with memory, organization, and planning. Although there are competing definitions for EF in the literature, there are some commonalities from which generalizations can be drawn. According to Barkley (2006), EF is generally conceptualized to include nonverbal working memory, verbal working memory (i.e., internalized speech), and planning and foresight around future consequences.

In their review of the relevant literature, Sergeant, Guerts, and Oosterlaan (2002) found evidence for EF deficits in children, adolescents, and adults with ADHD, but the published research provides equivocal results. Specific to children, it was found that while significant deficits in EF were associated with ADHD, similar deficits were observed among children with other disorders, including ODD and CD (Sergeant et al., 2002).

Thus it is clear that more research is needed to advance, modify, or overturn contemporary theories of ADHD. Based on the available research, it appears that EF deficits potentially underlie all disruptive behavior disorders, but especially ADHD. Further, there appears to be a genetic component to EF deficits that is substantially related to ADHD symptoms in particular, and not so much ODD or CD. For example, in a twin study it was found that EF indicators for one twin correlated moderately (r = .66) with the ADHD symptoms of the second twin. This correlation was much weaker (r = .16) among dizygotic twins (Coolidge, Thede, & Young, 2000). Such findings point to a shared genetic component of ADHD that is in some way associated with EF, but this relationship is unclear.

Biological Explanations

The theories described above have sparked research into the biological causes of ADHD, but to date no definitive biological cause has been discovered (Hynd, Voeller, Hern, & Marshall, 1991; Sergeant, 2004). Still, the nascent research provides some compelling evidence that ADHD is heritable and based on identifiable genetic underpinnings. Further, there are observable physical and neuroanatomical differences between individuals with

ADHD and their normal peers that warrant further exploration.

Heritability. Multiple lines of research suggest a genetic basis for ADHD, including family studies (Epstein, Conners, Erhardt, Arnold, Hechtman, Hinshaw, et al., 2000), twin studies (Coolidge et al., 2000; Levy, Hay, McStephen Wood, & Hons, 1997), and molecular genetic research (Waldman & Gizer, 2006). Epstein and colleagues (2000) examined familial aggregation of ADHD by collecting parent reports regarding their own and their partner's behaviors and found that biological parents of children with ADHD identified higher rates of ADHD-consistent impairments than did parents of children without ADHD. Interestingly, nonbiological parents also reported higher levels of ADHD-consistent impairment than did parents of normal children, which may suggest that the expression of ADHD traits has a social learning component that affects family members in unidirectional or reciprocal ways. Or perhaps there may is a "nonrandom selection bias," whereby adults with ADHD select partners and foster/adoptive children with similar qualities (Epstein et al., 2000, p.592). In any event, the significantly high self- and partner-ratings of ADHD impairments among parents of children with ADHD appear

robust, and increases in cases of child ADHD with comorbid conditions (Epstein et al., 2000).

Twin studies, such as that conducted by Levy, Hay, McStephen, Wood, and Hons (1997), suggest that there appears to be an underlying biological liability that predisposes some children to high rates of inattention or hyperactivity. In comparing the ADHD symptoms of monozygotic (MZ) and dizygotic (DZ) twins, Levy and colleagues found a strong correlation (r = .88) for MZ twins and a smaller correlation (r = .49) for DZ twins. Similarly, Coolidge and colleagues (2000) found a strong correlation of ADHD among MZ twins (r = .81) and a smaller correlation for DZ twins (r = .18). While there was considerable overlap in environmental influences on twin symptom concordance rates, the correlations found are believed to be "almost entirely due to genetic influences" (p. 283).

Using genome scans and candidate gene analysis, researchers are beginning to uncover evidence for a genetic basis of ADHD. Studies using genome scans have produced contradictory results. This is not surprising, as genome scans are exploratory in nature and are not particularly powerful in detecting putative genes in complex traits, such as ADHD. However, candidate gene

studies have shown promise in detecting specific genes associated with ADHD symptoms and core cognitive deficits. Candidate gene studies are more targeted and powerful than genome scans, but require specific hypotheses about target genes prior to analysis. Based on research of the neurotransmitters affected by stimulant medications and "knockout" gene studies (i.e., removing specific genes to assess their impact) with mice, investigators have targeted specific genes associated with dopaminergic, adrenergic, and serotonergic systems in the brain. While study results are often mixed, there appear to be associations between ADHD and the dopamine transporter (*DAT1*), dopamine receptor D4 (*DRD4*), and dopamine receptor D5 (*DRD5*) genes. Recent meta-analyses suggest that the *DRD4* and *DRD5* genes play a consistent role, despite the contradictory findings. In the near future, additional meta-analyses may also substantiate the role of the *DAT1* gene as well (Waldman & Gizer, 2006).

However, there are limitations to the existing genetic research. First, the relationship between genes and ADHD is complex, but the methodologies used to examine this relationship are oftentimes crude. On the one hand, it is clear that ADHD does not fit a simple

Mendelian disease model, but rather a polygenetic model with multifaceted variations in genetic penetrance. On the other hand, most gene candidate studies focus on single polymorphisms (i.e., single genetic variables), which is undoubtedly an oversimplified approach. It is highly likely that multiple markers within specific genes are involved in the etiology of the disorder. Second, there are concerns that the genetic expression of ADHD is differentially affected by factors such as age, sex, and environmental influences. Thus, many variables need to be controlled before researchers can confidently identify specific genetic causes. To date, most studies have not addressed these issues adequately. Third, genetic research has been complicated by the heterogeneous nature of ADHD. For example, some candidate gene studies have found associations and linkages for specific subtypes of ADHD, and not for others (e.g., *DAT1* appears to be associated with hyperactivity-impulsivity, but not inattention). As a result, some researchers have attempted to associate genetic variations with ADHD endophenotypes (i.e., the specific deficits that are presumed to underpin the disorder), rather than the DSM-defined symptoms. Endophenotypes, such as EF deficits, are commonly assessed using neurological or

neuropsychological measures (Waldman & Gizer, 2006). Neurological and neuropsychological measurement techniques will be discussed in greater detail later in this chapter.

Neuroanatomy. Attempts to determine physical differences in the brains of children with ADHD compared to their non-affected peers using brain imaging techniques has been wrought with methodological problems and inconsistent findings. While this research has periodically uncovered neuroanatomical differences between ADHD and normal groups, researchers have historically used varying imaging techniques, different means to measure differences, small sample sizes, and different approaches for establishing diagnoses (Hendren, De Backer, & Pandina, 2000).

Such problems are endemic in studies using magnetic resonance imaging (MRI), which is perhaps the most commonly used method for assessing the pathophysiology of mental disorders, including ADHD. Given the high costs associated with MRI and the lack of a standard technique for interpreting the results, the related literature is often based on small sample sizes and the results across studies are often contradictory. However, in their review of the relevant research, Krain and Castellanos

(2006) found that studies relying on MRI scans generally suggest that children with ADHD have less total brain volume than their unaffected peers. Specifically, it appears that an asymmetry in the prefrontal cortex (PFC) seen in normal development is less pronounced for children with ADHD. Further, children with ADHD appear to have lower cerebellar volume and smaller corpus callosums than their normal peers (Durston, 2003; Krain & Costellanos, 2006). However, more research is needed to better understand the role of comorbid conditions and medication on neurophysiology, as these factors may help to explain some (but probably not all) of the neuroanatomical differences among children with ADHD (Seidman, Valera, & Makris, 2005).

Other physical differences associated with ADHD are readily observable. On average, children with ADHD are smaller in height and weight than their normal peers. This problem is thought to be related to growth delays brought about by dysregulated neurotransmitter activity in the brain, which has a temporary growth-stunting effect mediated by the neuroendocrine system. Interestingly, physical disparities seem to disappear over time, as young adults with ADHD and their normal peers are comparable in height and weight (Spencer,

Biederman, & Wilens, 1998). Differences in stature between children with and without ADHD is also attributable to the effects of psychostimulant medications, especially in young children, as research suggests that psychostimulants stunt growth (Swanson, Greenhill, Wigal, Kollins, Stehli, Davies, et al., 2006). However, stimulants cannot explain all of the differences in stature between children with ADHD and their peers (Spencer et al., 1998).

Developmental Course

Like all forms of child psychopathology, ADHD follows a developmental course, with risk factors and symptoms changing with age and growth. There are three aspects of the development and ADHD that are worth noting: the differences between boys and girls, the persistence of ADHD into adolescence, and the potential for a developmental trajectory from ADHD to serious conduct problems.

Gender Differences

As mentioned previously, ADHD is disproportionally diagnosed among boys, at a ratio of approximately 2:1 to 9:1 to girls, depending on how samples are derived (APA, 2000). This difference may reflect true biological differences in the prevalence of the disorder in the

population, or it could reflect referral biases, or some combination of the two. A meta-analysis of gender differences found that, in general, girls are less likely to exhibit hyperactivity, conduct problems, or externalizing behavior problems than boys; however, among clinic-referred samples, girls appear to have more intellectual impairment and greater levels of inattention than their same-sex peers (Gaub & Calson, 1997). Thus, it appears that since girls generally exhibit fewer externalized behavior problems than boys, girls referred to clinics are likely to represent the most severe cases among all girls with ADHD, and studying only clinic-referred girls may provide a skewed picture of girls with ADHD in the larger community. Distorted perceptions of girls with ADHD stemming from referral biases have been referred to as the "paradoxical gender effect" (Waschbusch, 2002, p. 120). In community samples it is clear that boys exhibit more hyperactive-impulsive symptoms (Gaub & Carlson, 1997) and inattention combined with hyperactivity-impulsivity (Hartung, Willcutt, Lahey, Pelham, Loney, Stein, et al., 2002) when compared to girls. Not surprisingly, adult raters generally perceive overactivity as more disruptive than inattention alone (Sciutto, Nolfi, & Bluhm, 2004), which may explain some

of the differential referral rates between boys and girls.

Although differences in symptom expression appear to exist between the sexes, boys are more often referred for assessment and treatment than girls regardless of symptoms. This is commonly referred to as a referral bias. Sciutto, Nolfi, and Bluhm (2004) found that when teachers were asked to rate fictional scenarios of children, where sex varied across scenarios, teachers were more likely to refer boys than girls despite identical symptom descriptions. Sciutto and colleagues estimated that teachers were 1.5 times more likely to refer a boy with hyperactive symptoms as compared to a girl with hyperactive symptoms. Interestingly, the referral bias among teachers appeared consistent for both men and women referral sources. Taken together with the research that suggests boys exhibit more disruptive behaviors, the referral bias helps to explain the increased number of referrals for boys as compared to girls.

Given that girls are referred for diagnosis and treatment less frequently than boys, it is possible that many girls are underdiagnosed. Waschbusch and King (2006) recently examined this possibility and found that

a subgroup of girls with significant impairment relative to their same-age female peers (i.e., > 1.5 *SD* greater impairment) failed to meet the DSM-IV(-TR) criteria of six or more inattentive and/or hyperactive-impulsive symptoms, as assessed by parent and teacher ratings. Among teacher ratings, for example, 4.9% of girls who exhibited significant ADHD combined subtype symptoms relative to their peers failed to meet the DSM-IV(-TR) criteria. Hence, it appears that the DSM-IV(-TR) criteria may indeed underidentify some girls with significant impairments. Regardless, the DSM-IV(-TR) is generally believed to represent an improvement over the DSM-III-R in terms of gender equality (Barkley, 2006). Indeed, research suggests that diagnoses of ADHD among girls are increasing relative to boys (Robison, Skaer, Sclar, & Galin, 2002).

Persistence into Adolescence

While ADHD was once thought to be limited to early childhood, studies suggest that many children continue to exhibit symptoms well into adolescence and adulthood (APA, 2000; Barkley, 2006; National Institute of Mental Health Consensus Forming Panel, 2000; Tucker, 1999). In fact, it is now estimated that somewhere between 2% and 5% of all adolescents in the general population exhibit

symptoms consistent with ADHD (Barkley, 1998). Among adolescents referred for mental health services and special education, the prevalence rate of ADHD is estimated to be about 25% (Tucker, 1999). Barkley, Fischer, Smallish, and Fletcher (2002) studied the persistence of hyperactivity into young adulthood using both DSM criteria and developmentally referenced criterion (i.e., ≥98th percentile on parent behavior ratings of hyperactivity). After eight to ten years, 71% of children diagnosed with ADHD continued to meet the DSM criteria for diagnosis and 83% exhibit persistent hyperactivity as measured by developmentally referenced criterion. In contrast, self-reports of adolescents and young adults seemed to suggest significantly lower rates of ADHD persistence. Taken together, studies suggest that if parent reports are used and diagnostic criteria are adjusted to reflect developmental changes, ADHD appears highly persistent and chronic. In other words, adolescents may outgrow the diagnostic criteria, but not necessarily the disorder (Barkley, 2006). The DSM-IV(-TR) criteria are not adjusted for developmental levels and the same criteria, such as "often runs around and climbs excessively," are used to diagnose children, adolescents, and adults (APA, 2000). Clearly, many DSM-

IV(-TR) behavior criteria are more applicable to elementary-age children than adolescents.

For most adolescents and young adults, the symptoms associated with ADHD appear to attenuate to some degree from those seen in childhood. Specifically, it appears that hyperactivity and, to a lesser extent, impulsivity improve (lessen in severity) relative to inattention. In other words, inattention appears to persist, whereas hyperactivity and impulsivity often do not. When hyperactivity and impulsivity do persist, the problems become less externalized. For example, it is far more likely for an adolescent or young adult to report subjective feelings of restlessness, rather than exhibiting excessive running or climbing behaviors (Barkley, 2006).

Conduct Problems

As mentioned above, children with ADHD are more likely than their normal peers to develop conduct problems, including Oppositional Defiant Disorder (ODD) or Conduct Disorder (CD). It is estimated that between 25% and 75% of adolescents with ADHD exhibit comorbid ODD or CD (Barkley, 1998). Barkley, Fischer, and Edelbrock (1990) conducted a longitudinal study and found that, after 8 years, 43.5% of hyperactive-impulsive children

developed Conduct Disorder, while only 1.6% of non-hyperactive-impulsive comparison children developed similar behavior problems.

Similar findings have been consistently replicated within the research literature, especially among boys (e.g., Vitelli, 1998; White, Moffitt, Caspi, Bartusch, Needles, & Stouthamer-Loeber, 1994), giving rise to a developmental conceptualization of disruptive behavior disorders as beginning in early childhood as ADHD and, in some cases, progressing to include more serious conduct problems. Waschbusch (2002) conducted a meta-analysis of studies examining the comorbidity of ADHD and conduct problems and concluded that ADHD and ODD/CD co-occur more frequently than would be expected by chance, the impairments associated with comorbid ADHD and ODD/CD are far more severe than for the constituent diagnoses alone, and the comorbid ADHD and ODD/CD group experiences the earliest and most persistent behavior problems. The latter finding is particularly troubling, as conduct problems often precede Antisocial Personality Disorder in adulthood, especially among adolescent males with CD (Loeber, Burke, & Lahey, 2002).

Interestingly, there has been little research on girls with severe behavior problems. Based on the scant

research available at this time, it appears that girls with CD may be underdiagnosed because their deviant behaviors are likely to be covert and their aggressive behaviors are likely to be nonphysical. For example, girls are much more likely to use social exclusion, rumors, and gossip as a means to harm their peers. Nonphysical forms of aggression, referred to as *indirect*, *social*, or *relational aggression*, are commonly overlooked by parents and teachers, which might explain why girls are not considered as aggressive as boys (Archer & Coyne, 2005). Thus, little is known about the developmental course of conduct problems in girls at this time (Delligatti, Akin-Little, & Little, 2003).

Family risk factors. Given the serious implications of antisocial behavior, it is not surprising that a growing body of literature has focused on the risk factors that help predict which children with ADHD are more likely to develop conduct problems. Much of this research has focused on family risk factors. For example, noncompliance with parental requests among boys with ADHD appears to play a key role in the development of later antisocial behavior (Lee & Hinshaw, 2004). Also, inconsistent parenting practices (Frick, Christian, & Wootten, 1999), exposure to family violence (Becker &

McCloskey, 2002), and lies, secrecy, and taboo behavior (e.g., incest) among family members (Baker, Tabacoff, Tornusciolo, & Eisenstadt, 2003) appear to predict higher rates of conduct problems among children and adolescents. However, more research is needed in this area as well.

Social risk factors. In addition to familial influences, there are also peer influences that lead to the development of conduct problems. While it is developmentally appropriate for adolescents to prioritize peer interactions over familial interactions, problems can arise when adolescents rely heavily on their peer group for support and guidance (Taffel, 2001). Given the social difficulties associated with ADHD (described above), this developmentally appropriate turn toward peers can introduce new risk factors. Research suggests that children with ADHD commonly gravitate toward deviant peer groups, which seem to welcome and reinforce antisocial behavior. For example, Marshal, Molina, and Pelham (2003) found that children with ADHD reported that their friends were more accepting of deviant behaviors (e.g., substance use), as compared to normal peers who reported their friends were less accepting of deviant behavior. Deviant peer affiliation appears to mediate the relationship between ADHD and high-risk behaviors.

Although the link is unclear, ADHD has been shown to be associated with an increased risk of substance abuse (Lambert & Hartsough, 1998; Molina & Pelham, 2003). ADHD has also been associated with alcohol abuse (Smith, Molina, & Pelham, 2002). Specifically, persistent ADHD (i.e., ADHD that continues into adolescence) has been shown to be associated with significantly increased risk for cigarette and alcohol use. When persistent ADHD is coupled with CD, the risks rise precipitously for cigarettes, tobacco, marijuana, and other illicit drugs (Molina & Pelham, 2003).

Unfortunately, the vast majority of research on ADHD has been conducted with elementary age children, leaving a dearth of research focusing exclusively on adolescents. The lack of adolescent research appears to be due in part to the antiquated notion that children would simply "grow out" of the disorder (Evans, Vallano, & Pelham, 1995). Moreover, only a small proportion of adolescents with ADHD receive any mental health care in clinical practice (Kazdin, 1990), due in part to an overall lack of mental health services in most communities (American Psychological Association, 2003), and a growing reluctance among children to voluntarily seek out treatment as they mature into adolescence (Prout & Brown,

1999). Further, adolescents are more likely to resist intervention efforts, as compared to younger children, especially if the techniques are perceived as cliché or disingenuous (Taffel, 2005). Clearly, more research on effective and acceptable treatment options for adolescents with ADHD is needed (Abramowitz & O'Leary, 1991; Evans, Vallano, & Pelham, 1995).

Treatment Outcomes Research

Several treatment options for ADHD have been researched in the literature, but most have proven to be ineffective and, at times, even contraindicated. Some treatment options such as play therapy, dietary restrictions, individual counseling, and relaxation training, are generally not supported in the professional literature (Barkley, 2006). Some potentially promising strategies, such as electroencephalogram (EEG) biofeedback, lack research with adequate scientific rigor to meet the requirements for empirical support (Loo & Barkley, 2005). Currently, only three treatments are empirically supported for the treatment of ADHD: stimulant medications, behavior modification, and the combination of stimulants and behavior modification (Pelham & Hoza, 1996; Pelham, Wheeler, & Chronis, 1998; Smith, Waschbusch, Willoughby, & Evans, 2000). This

section will review the three empirically supported treatment options, with a focus on how outcomes are commonly measured. The techniques by which researchers and clinicians measure treatment outcomes speak directly to the underlying issues addressed by the present study.

Stimulant Medications

Stimulants, such as Methylphenidate (MPH), Dextroamphetamine (DEX), and Pemoline (PEM), are widely researched and generally supported in the professional literature. For example, Spencer and colleagues (1996) reviewed over 127 published studies of stimulants for the treatment of ADHD and concluded that about 70% of children experience some improvement in behavioral functioning. Swanson, McBurnett, Christian, and Wigal (1995) found similar trends, using a large research review conducted for the U.S. Department of Education.

However, the professional literature underscores several limitations of stimulant therapy. First, stimulants do not cure ADHD; they provide temporary improvements in specific symptoms of ADHD (The MTA Cooperative Group, 1999a). Some impairments, such as social skill deficits (described above), are less responsive to stimulant therapy. For example, Hoza, Gerdes, and their colleagues (2005) collected peer

sociometric data for 285 children. Findings suggested that regardless of the treatment used (i.e., stimulants, behavior modification, combinations of stimulants and behavior modification, and community treatments), there were no statistically significant or clinically meaningful benefits. Stating this finding plainly, the researchers reported that, "children with ADHD still experienced significant peer problems at the end of treatment as compared to their classmates, regardless of the treatment they received" (p. 80). Hence, it appears that while stimulant therapy can reduce aggressive or noncompliant behavior, it does not replace problem behaviors with prosocial alternatives, nor do stimulants help children overcome the lasting effects of negative social reputation. In their review of the relevant literature, Landau and Moore (1991) conclude, "these children continue to experience interpersonal difficulties because medication does not generate the socially appropriate behavior necessary for peer acceptance. Thus, the need for adjunctive, nonpharmacological interventions is evident" (p. 247). Although social skills training does not enjoy consistent empirical support in the literature either (Pfiffner & McBurnett, 1997), stimulant medications are clearly

ineffective at ameliorating social deficits in most cases. As such, a psychosocial alternative is often desirable and necessary.

Second, the effectiveness of medications can vary widely, depending on comorbid disorders (Barkley, DuPaul, & Connor, 1999). Children with ADHD are commonly diagnosed with comorbid psychiatric disorders and, as a result, often receive nonstimulant psychotropic medications. In fact, research suggests that a child will receive a stimulant along with a nonstimulant psychotropic medication, such as tricyclic antidepressants (TCAs), selective serotonin reuptake inhibitors (SSRIs), and alpha agonists, in about 16% to 30% of all cases (dosReis, Zito, Safer, Gardner, Puccia, & Owens, 2005; Guevera, Lozano, Wickizer, Mell, & Gephart, 2002).

Third, despite the research supporting stimulant medications, there is controversy surrounding their use. Many have argued, for example, that pressure from the pharmaceutical industry has led to the over-prescription of stimulants (e.g., Breggin, 2001). It is clear that prescriptions for stimulant medications in the United States rose precipitously during the 1990s. In a study of methylphenidate prescription trends from 1990 to 1995,

Safer, Zito, and Fine (1996) found a 2.5-fold increase among youths. Similarly, Olfson, Marcus, Weisman, and Jenson (2002) found a four-fold increase in stimulant use among children from 1987 to 1996. Robison, Skaer, Sclar and Galen (2002) found a 2.8-fold increase for girls and a 2.1-fold increase for boys from 1990 to 1998. The United States Drug Enforcement Agency estimates that there was a seven-fold increase in methylphenidate production from 1990 to 1997, with 90% of the medication sold in the United States. (USDEA, 1999). Although the estimates appear to vary across sources, there is little doubt that prescriptions for stimulant medications for ADHD were on the rise through the 1990s.

Fourth, factors other than medical necessity play a role in the decision to use stimulants to treat ADHD. For example, prescriptions for stimulants are not evenly distributed throughout the United States. Rather, stimulants are more often prescribed for children in the Midwest and South. Further, families with fewer siblings, higher income, and living in predominately white communities are more likely to opt for stimulant therapy (Cox, Motheral, Henderson, & Mager, 2003). Based on such findings, some authors conclude that stimulants are often used for convenience, rather than legitimate

medical necessity (e.g., Breggin, 2001). It may be more accurate to conclude that stimulants are the option of choice when alternative treatments are unavailable, infeasible, or unacceptable.

Fifth, there is a general lack of research on the long-term impact of psychostimulant use (Hechtman & Greenfield, 2003). One result has been a concern around the potential for stimulants to lead to illicit substance abuse in adulthood. Indeed, a study conducted by Lambert and Hartsough (1998) suggested a link between stimulant use and later tobacco and cocaine abuse. However, Lambert and Hartsough's research design has been criticized because potential comorbid conditions such as CD were not adequately assessed (e.g., Hechtman & Greenfield, 2003; Mick, Biederman, & Faraone, 1998). In a meta-analysis of the available research, Wilens, Faraone, Biederman, and Gunawardene (2003) concluded that stimulant medications actually reduce the risk of later substance abuse. In fact, stimulant use was associated with as high as a 1.9-fold risk reduction. However, the potential connection between stimulant medications and later drug use is still debated in the professional literature and in the lay press.

Sixth, medication compliance can become a concern. Swanson (2003) found that parents discontinued stimulants in as much as 45% of cases in as little as ten months. In their review of the literature, Wells and colleagues (2000) conclude that "at least 20-25% of families with a child with ADHD harbor opinions about medication vs. psychosocial treatment firm enough to preclude any willingness to consider one or another type of treatment" (p. 488). Pharmaceutical companies have attempted to redress patient noncompliance issues by producing once-daily medications and providing patients' families with information regarding the relative safety and effectiveness of stimulants. Still, many children express a desire to discontinue due to side effects, such as poor sleep and headaches (Doherty, Frankenberger, Fuhrer, & Snider, 2000), and a significant proportion of their peers report that secondary students sell or give away their medications (Moline & Frankenberger, 2001). Given such trends, it is not surprising that physicians often encounter apprehension among parents in regards to stimulants. In a survey it was found that over half of all parents were hesitant to opt for stimulant medications for their child with ADHD (dosReis et al., 2003). Unfortunately, viable alternatives are not

readily accessible in many communities (Jensen et al., 1999).

Seventh, relatively few medication studies have been conducted with adolescents (Evans et al., 2001; Findling et al., 2001; Smith et al., 2000). While there is some suggestion that the benefits of stimulants in childhood generalize into adolescence (e.g., Smith, Pelham, Gnagy, & Yudell, 1998), academic and behavioral outcomes appear to vary widely at the secondary school level. For example, when using ecologically valid measures of academic performance and classroom behavior (e.g., social studies quizzes), Evans, Pelham, Smith, and colleagues (2001) found that the percentage of adolescents who exhibited improvement varied based on dosage. Most adolescents experienced the greatest benefits at low doses of stimulants and experienced diminishing returns or even performance deterioration at higher doses. At the lowest initial dose, adolescents who exhibited any improvement ranged from 49% to 67%, depending on the dependent measure, which is meaningfully lower than the 70% success rate found among children (Spencer et al., 1996). Thus, the result suggests that either the benefits of stimulants wane as children mature, or

adolescents who continue to warrant stimulant therapy may represent severe (i.e., less responsive) cases.

In summarizing the research on stimulants, it appears that while there is a therapeutic benefit for the majority of children with ADHD, a substantial portion of children do not benefit. Even among stimulant responders, the benefits appear to be limited, and do not adequately ameliorate issues such as social performance deficits. Further, psychostimulant therapy remains a controversial issue and many families find medications an unacceptable treatment option. In other instances, it appears that when alternative treatments are unavailable, infeasible, or unacceptable, stimulants become the first option, resulting in uneven prescription rates around the nation. Also, little is known about the long-term impact of psychostimulant use and potential negative outcomes. Not surprisingly, medication compliance is a common concern for families opting to use stimulants. Finally, as children enter adolescence, there is evidence to suggest that the success rate for stimulants diminishes. Clearly, many factors determine whether stimulants are effective in treating the symptoms of ADHD, and whether families will opt for stimulant therapy. Thus, accurate monitoring of stimulant therapies is vital to help

families make informed decision around this treatment option.

Behavior Therapy

Given the limitations associated with stimulant medications and the trepidation on the part of many parents and children to opt for stimulants, positive behavioral support can represent a more acceptable adjunct or alternative intervention. Oftentimes positive behavioral support is advocated in the school setting to help manage disruptive behaviors. Such efforts commonly include teacher education on issues related to disruptive behavior disorders and their implications for education (Abramowitz & O'Leary, 1991; DuPaul & Stoner, 2002; Webster-Stratton, 1993), as well as ongoing behavioral consultation with a school psychologist or other mental health professional (Evans, Serpell, Schultz, & Pastor, 2007; Schultz & Cobb, 2005; Schultz, Reisweber, & Cobb, 2008; Wells et al., 2000).

The Multimodal Treatment Study (MTA)

To research stimulant and behavioral treatment options, the National Institute of Mental Health (NIMH) and the U.S. Department of Education funded a large, multisite study known as the Multimodal Treatment Study of Children with ADHD (MTA; MTA Cooperative Group 1999a,

1999b, 2004a, 2004b). This study, which included 579 children in first through fourth grades from seven sites across the U.S. and Canada, is the first major clinical trial conducted by the NIMH in child psychopathology. Participants in the MTA met the DSM-IV criteria for ADHD combined subtype, and were randomly assigned to one of four treatment conditions: medication only (MedMgt), behavior modification only (Beh), a combination of medications and behavior modification (Comb), and a community comparison condition (CC) (MTA Cooperative Group, 1999a).

Treatments. The treatment protocols offered to MTA participants who received medication or behavior modification were rigorously designed by experienced clinicians and physicians to best practice standards. Children randomly assigned to the medication arm of the study (including both MedMgt and Comb groups) first went through a medication "wash out" period and then received a double blind, placebo controlled trial of MPH (in this case Ritalin®), which was randomly titrated with multiple dose repeats over 28 days to establish an optimal dose. Children who did not respond well to MPH or who appeared to benefit most from the placebo condition received additional medication trials with alternative stimulants,

including DEX and PEM. Nonstimulant options were made available in cases where stimulants proved ineffective or the side effects were unacceptable. Once an optimal dose was found, participants began a 13-month open-label continuing treatment phase that was closely monitored by physicians in communication with families and teachers. Additional medication trials were initiated for those participants that exhibited problems (e.g., clinically significant symptoms) during the continuing treatment phase. To assess the initial medication trials and monitor progress over time, researchers collected behavior and side effect ratings from parents and teachers, charted these data, and collaboratively problem-solved medication issues (Greenhill, Abikoff, Arnold, Cantwell, Conners, Elliott, et al., 1996).

The behavior modification arm of the MTA was comprised of several components, including parent training, school consultation, and a summer treatment program. Parent training occurred primarily in small group sessions led by doctoral level clinicians for 1.5 to 2 hours at a time for up to 27 sessions. Sessions focused on evidence-based behavioral interventions, such as school-home daily report cards and token economies. To support these efforts, some session topics touched on

issues not typically found in shorter parent training programs, including parental stress management, communication strategies with schools, and helping children improve their social functioning. To individualize and problem-solve the behavior plans that came out of the group meetings, clinicians met with parents individually on a monthly basis (Wells, Pelham, Kotkin, Hoza, Abikoff, Abramawitz, et al., 2000).

The summer treatment component of the MTA was an intensive eight-week behavioral program modeled on Pelham's Summer Treatment Program (STP; Pelham, Gnagy, Breiner, Hoza, Hinshaw, Swanson, et al., 2000). Undergraduate paraprofessionals trained in the STP treatment protocol provided behavior modification, social skills training, sports skills training, and classroom strategies, during eight-hour sessions that met every weekday. In the fall following summer treatment, MTA paraprofessionals provided behavioral consultation for teachers that led to the creation of classroom interventions to address student academic and social needs during the school year (Wells et al., 2000). Given the investment in time and resources in both the medication and psychosocial treatment arms of the MTA, it is clear that the treatment protocol represented a more

intensive approach than typically found in school- or community-based practice (Smith, Barkley, & Shapiro, 2006).

Outcomes. After 14 total months of treatment, the MTA researchers administered outcome measures. While investigators originally used more than 100 instruments for the baseline assessments of participants, this set was reduced to 19 key variables, based on the results of principal component analysis. Of these variables, 10 were found to have significant Treatment X Time interaction effects at the end of treatment. Interestingly, six variables were derived from a single instrument administered to parents and teachers: the Swanson, Nolan, and Pelham, version IV (SNAP-IV) behavior rating scale. Items of the SNAP-IV are derived from the DSM behavioral criteria for ADHD and ODD, with response options that range from *Not At All* to *Very Much* along a four-point scale (Swanson, Kraemer, Hinshaw, Arnold, Conners, Abikoff, et al., 2001).

Results suggest that all four groups (MedMgt, Beh, Comb, and CC) experienced clinically meaningful improvement in ADHD symptoms at 14-month post-treatment. When comparing group means, it appeared that the two groups receiving the intensive medication protocol

(MedMgt and Comb) experienced significantly better outcomes than either the Beh or CC groups. Hence, the MTA Cooperative Group concluded that the medication protocol was the key intervention component that proved beneficial. To a much lesser extent, behavioral therapy was also indicated, but it did not appear to offer significant advantages over the community care condition, or to significantly improve outcomes for the Comb group over the MedMgt group. In other words, MedMgt and Comb groups were not significantly different from one another, and the Beh and CC groups were not significantly different from one another (The MTA Cooperative Group, 1999a). The MTA results can be summarized as follows: Comb ≈ MedMgt > Beh ≈ CC.

Interpretations. Since the initial results of the MTA study were published, several researchers have provided alternative interpretations. It is important to note that the MTA employed an intent-to-treat (ITT) design, whereby participants were randomly assigned to treatments and their outcomes assessed based on original assignments regardless of treatment adherence. In essence, all treatment protocol deviations that occurred after randomization were ignored in a manner that emulated treatment in clinic-referred samples, where

clients may deviate from their prescribed treatments. The ITT strategy effectively maintained the maximum amount of participants in the study without attrition due to methodological concerns. However, there were treatment variations within each group that make the overall results ambiguous and open to multiple interpretations.

Smith and colleagues (2006) pointed out that two-thirds (67%) of the CC group actually received medications in the community. As such, the CC group cannot be thought of as a true control group, per se. Hence, the nonsignificant differences between the Beh and the CC group suggest that behavior modification was roughly as effective as community-based care, where the majority of children received some medication.

Also of note, participants in the Beh condition actually received tapered services after the first nine months, with some only meeting with researchers monthly by the 14-month endpoint. Given that medications were continued up to that point, it may seem like an "unfair comparison" between the Beh condition and Comb and MedMgt conditions. However, the psychosocial treatments used with the Beh were intended to result in coping strategies that would last beyond the study timeframe, so the

comparison is theoretically valid (Taylor, 1999). Indeed, further analysis of the MTA data suggests that dependent measures taken at the nine month point – when behavioral interventions were at their peak "dose" prior to tapering – outcomes were very similar to those reported at 14-months: Comb and MedMgt conditions appeared to outperform Beh and CC conditions on parent and teacher ratings of ADHD and ODD symptoms. However, most of the Beh group participants that eventually started medications (*n* = 38) did so at or following the tapering of behavioral interventions, suggesting that many families attempted to supplement treatment at about the same time that investigators were transferring prime responsibility for intervention implementation to parents and teachers (Arnold, Chuang, Davies, Abikoff, Conners, Elliott, et al., 2004).

In their review of the MTA, Conners and colleagues (2001) concluded that the Comb condition contributed to better outcomes in comparison to the MedMgt condition. Post hoc analyses suggested that the participants in the Comb condition exhibited the largest positive effect sizes, as compared to the CC condition. Similar findings were found among children with ADHD attending Pelham's Summer Treatment Program (Pelham et al., 2000). Further,

although the MTA Cooperative Group did not conclude that the data supported a significant benefit of behavior modification, it was noted that parents preferred treatment options that coupled behavior therapy with stimulant medications. Further, 17 participants that were randomly assigned to the MedMgt and Comb conditions discontinued the study to avoid the medications, suggesting dissatisfaction with a pharmacological approach to treatment (The MTA Cooperative Group, 1999a).

Additional analyses of the MTA examined the rating scale data from the SNAP-IV. By setting a clinically derived cut point, combined average teacher and parent ratings on the SNAP-IV were examined to determine the percentage of participants in each treatment condition that achieved clinical "success," which was defined as an overall average SNAP-IV score below 1.0. In other words, success meant behavior ratings of ADHD symptoms below the diagnostic threshold. According to this analysis, the Comb treatment had a small advantage (Cohen's delta [*d*] = .26) over MedMgt alone, which increased the clinical success rate by 12%. However, other questions pertaining to success rates compared to the CC condition were difficult to interpret because the quality of treatments

found in the community appeared to vary widely across study sites (Swanson, Kraemer, et al., 2001).

Children in the MTA study were followed and reevaluated 10 months post-treatment (24-months after starting the study). The results suggest that all participants continued to exhibit significant behavioral benefits over baseline, but the benefits deteriorated by roughly half for the MedMgt and Comb groups (MTA Cooperative Group, 2004a). Further, it appeared that participants who had taken stimulant medications continuously throughout the study exhibited significant height and weight suppression, as compared to the participants that had not taken medications. Unfortunately, the MTA data do not allow for an analysis of whether differences in height and weight were due to pre-existing conditions, and it is still too early to determine if the observed differences will persist into adulthood (MTA Cooperative Group, 2004b).

Implications for treatment. Given the limitations of stimulant medications and behavioral therapies alone, there is a growing consensus that the most promising strategy is a combination of stimulant medications and behavioral therapy (Pelham et al., 2000; The MTA Cooperative Group, 1999a). A comprehensive strategy would

combine several medicinal and behavioral components, including medication monitoring, parent training, systematic reward systems with response cost, and communication between home and school to monitor behavioral interventions.

There is some evidence to suggest that a multimodal approach that includes behavioral interventions may reduce the need for medication. For example, investigators in the MTA noted that at the end of treatment (14-months), children in the Comb condition received less medication on average than the MedMgt group, suggesting that when behavior modification is combined with medication therapy, lower dosages of medication are required to achieve satisfactory results. The apparent medication offset afforded by behavioral interventions is a promising finding, as the side effects associated with stimulants (e.g., growth suppression) or other medications are more likely to occur at higher dosages (MTA Cooperative Group, 2004b).

When multimodal interventions are pursued, many researchers have suggested that school-based interventions are an invaluable component (e.g., Dishion & Kavanagh, 1999; Evans & Weist, 2004; Weist & Evans, 2005). Indeed, various studies have concluded that

school-based services are confronted with fewer obstacles than would be expected with clinic-based services (e.g., Adelman, Barker, & Nelson, 1993; Evans, 1999). Barriers to care are a particular concern in cases of ADHD, as the disorder is chronic and highly resistant to interventions consisting of less than 20 sessions (Robin, 1998). However, in community settings, issues of cost, insurance coverage, and transportation commonly preclude intense long-term treatment (Grove, Evans, Thompson, & Barnett, 2004). As a result, schools - with relatively fewer barriers to care – represent perhaps the most promising vector for successful intervention.

However, research on educational interventions may actually be declining in the professional research. In their review of four leading educational psychology journals from 1983 to 2004, Hsieh, Acee, Chung, Hsieh, Kim, Thomas, and colleagues (2005) found that the overall percentage of intervention studies dropped from 55% to 35%. Further, the more recent studies appeared to lack many of the hallmarks of scientific rigor and quality, suggesting that overall quality has not improved. For example, the percentage of interventions that were analyzed for more than one day dropped from 26% to 16% in the period between 1995 and 2004. During this same

period, the percentage of randomized trials dropped from 34% to 26%.

Assessment of ADHD

As mentioned above, there is controversy surrounding the diagnosis and measurement of ADHD. Research has examined the most effective and efficient means of measuring the disorder; however, many questions still remain. The following discussion will examine the attempts to find objective measures or tests for ADHD, and current "best practice" recommendations for conducting a comprehensive clinical ADHD evaluation.

Objective Measures of ADHD

Clinicians and researchers have attempted to find objective means of diagnosing ADHD, but to date, efforts have not produced measures with adequate sensitivity and specificity to reliably differentiate between ADHD and non-ADHD cases. Several potential candidates for objective measures of ADHD currently do not have enough research support for their widespread use. For example, as was discussed above, studies of genetic markers, neuroanatomy, and response to medication do not provide enough sensitivity or specificity for an indisputable diagnosis of ADHD. Given these limitations, researchers have turned to cognitive, academic, neuropsychological,

and neurological measures, in the hopes of uncovering a pathognomic indicator of ADHD.

Cognitive measures. Research on the cognitive abilities of individuals with ADHD has suggested that the disorder is associated with IQ scores lower than those of normal peers. For example, Frazier, Demaree, and Youngstrom (2004) conducted a meta-analysis of 123 studies, analyzing 137 comparisons of IQ, and found that individuals with ADHD score significantly lower on cognitive measures than their undiagnosed peers (d = 0.61), with a group discrepancy "roughly equivalent to a 9-point difference in [full scale IQ] for most commercial IQ tests" (p. 552). Similar cognitive inefficiencies were found among all three subtypes of ADHD. Interestingly, such differences do not appear attributable to comorbid conditions, such as learning disabilities (Barkley, DuPaul, & McMurray, 1990). However, despite consistent group-level differences, there appears to be significant overlap in the variance among ADHD and non-ADHD groups. Further, of the 137 specific comparisons analyzed by Frazier and colleagues (2004), only 63 reached statistical significance at the .05 alpha level, with one instance where individuals with ADHD actually had significantly higher full scale IQs

than their undiagnosed peers. Similar results were found in a recent meta-analysis of adult studies, with normal adults outperforming adults with ADHD (d = 0.26). Again, significant IQ overlap was found between ADHD and normal groups, with seven of the eighteen examined studies reporting appreciably higher (but not necessarily significantly higher) IQs for adults with ADHD (Bridgett & Walker, 2006). While such findings are interesting, it is clear that IQ is an insufficient indicator of ADHD; lower than average IQ scores may or may not suggest ADHD, while higher than average IQ scores certainly do not exclude ADHD. According to Barkley (2006), "children with ADHD are likely to represent the entire spectrum of intellectual development; Some are gifted, while others have low intelligence, learn slowly, or have mild intellectual retardation" (p. 123).

Achievement measures. As mentioned previously, children with ADHD often exhibit academic underachievement. Frazier and colleagues (2004) examined studies of how children with ADHD perform on standardized, norm-referenced academic achievement tests. Interestingly, studies looking at the effect of ADHD on spelling and mathematics suggested significant underachievement, and these discrepancies were larger

than those found on cognitive measures. However, this phenomenon may be related to the fact that many studies did not take into account the possibility of comorbid LD, which, as discussed above, is not uncommon among children with ADHD. Thus, more research is needed to determine the true relationship between the academic deficits often seen among children with ADHD and the utility of academic measures in assessing ADHD and related factors.

Neuropsychological measures. In addition to IQ and achievement, researchers have turned to neuropsychological measures of attention as a means of assessing ADHD. Neuropsychological instruments attempt to measure the hypothesized components of attention, such as shift and vigilance, mentioned earlier in this chapter. Frazier and colleagues (2004) examined studies looking at ADHD and non-ADHD group performances on commonly used neuropsychological tests and found that only the Continuous Performance Task (CPT) boasted effect sizes higher than that found for full scale IQ.

The CPT is perhaps the most widely researched neuropsychological test of attention. While several versions exist (e.g., Conners, 2000; Freidman, Vaughan, & Erenmeyer-Kimling, 1978; Michael, Klorman, Salzman, Borgstedt, & Dainer, 1981), the CPT generally requires

examinees to discriminately respond to rapidly presented stimuli. For example, Conners' (2000) computerized version of the CPT presents examinees with a series of flashing letters on a computer monitor. Whenever a letter other than "X" appears, examinees are to press a keyboard spacebar as quickly as possible. Target and non-target letters are then randomly flashed for up to 14 minutes, with the sequence of stimuli appearing based on a set ratio. The CPT measures reaction time, vigilance (sustained attention), and error patterns including both omission errors (failing to respond to target stimuli) and commission errors (erroneously responding to a non-target stimuli, or "false alarm").

Research on attention during CPT performance has been mixed, perhaps due to the lack of a standardized CPT format (Börger & van der Meere, 2000). However, a meta-analysis conducted by Losier, McGrath, and Klein (1996) found that children with ADHD generally make more omission and commission errors on the CPT than do their normal (non-ADHD) peers. Further, errors on the CPT can be reduced to some degree when individuals with ADHD use stimulant medications. A separate meta-analysis of CPT studies among adults with ADHD also found significantly more errors, particularly omission errors, as compared to

adults without ADHD (Hervey, Epstein, and Curry, 2004). Although the weighted effect size for omission errors was in the moderate range (*d* ranged from .52 to .76), the observed performance overlap between ADHD and non-ADHD groups precludes reliable diagnosis on the basis of this one measure alone. In other words, while there appears to be aggregate group differences, no specific error pattern or response style on the CPT is pathognomic. The degree of overlap in CPT performance between ADHD and non-ADHD groups does not allow for a truly valid or reliable "cut point" for differential diagnosis (Preston, Fennell, & Bussing, 2005). Rather, it appears that poor performance on the CPT only suggests general central nervous system dysfunction that may or may not be related to attention deficits or impulsivity (Homack & Reynolds, 2005).

In their analysis of the ecological validity of the CPT-II in assessing ADHD (Conners, 2000), Weis and Totten (2004) found that behavioral observations during the task correlated more strongly to parent and teacher ratings than did the actual test results. Thus, Weis and Totten concluded that CPT-II scores are more ecologically valid when augmented with observation data. In related research, Teicher, Ito, Glod, and Barber (1996) used an

infrared motion analysis system to measure the frequency, nature, and distance of examinee body movement during the CPT. Results suggested that children and adolescents with ADHD moved their bodies (particularly their trunks) significantly more than normal controls, with the differences between groups increasing with each successive trial. Interestingly, when body movement data was combined with age and the CPT omission and commission errors, researchers were able to discriminate between youths with and without ADHD with nearly 100% accuracy. However, it should be noted that this study utilized a small sample (n = 29) and has not been reliably replicated in the literature. Further, diagnostic methods such as those used by Teicher and colleagues may not be feasible in clinic or school settings.

Research on the diagnostic utility of EF tasks for ADHD has also been mixed. In their extensive review of the research published between 1990 and 2000, Sergeant, Geurts, and Oosterlaan (2002) examined the stop-signal task, the Stroop, the Wisconsin Card Sorting Test, the self ordered pointing task, tower tasks, and cognitive fluency measures. While there appears to be evidence supporting significant inhibitory deficits among ADHD groups as measured by the Stroop and stop signal tasks,

similar deficits are seen in other clinical populations as well, thus precluding diagnostic differentiation among similar diagnoses. Further, while group means differ, there is considerable overlap between ADHD and non-ADHD groups on EF measures. For example, there appears to be a 60% overlap between ADHD and non-ADHD groups in their reaction times on the stop signal task, which equates to a medium effect size (d = 0.64) in the predicted direction (Sergeant, Geurts, & Oosterlaan, 2002). Similar findings have been replicated in recent meta-analyses of the Stroop test (Homack & Riccio, 2004) and the Wisconsin Card Sorting Test (Romine, Lee, Wolfe, Homack, George, & Riccio, 2004). In both instances, researchers concluded that due to poor sensitivity and specificity, performance on these measures is insufficient to either confirm or disconfirm an ADHD diagnosis. Hence, it appears that even in the best case scenario, tests measuring EF are not specific enough for use in the diagnosis of ADHD due to the fact that not all children with ADHD exhibit EF deficits. As a result, the utility of EF tests appears to be limited to assessing strengths and weaknesses to guide treatment and to measure change over time (Seidman, 2006).

Measures of brain function. As mentioned earlier in this chapter, research on the etiology of ADHD has uncovered interesting but equivocal findings using measures of neuroanatomy (e.g., MRI). Current brain scan technologies are far too expensive to be feasibly used in the routine assessment of ADHD. Recent research has focused on potential alternatives, including the possibility of using EEG profiles to diagnose ADHD. In their review, Loo and Barkley (2005) found somewhat consistent findings that children and adults with ADHD exhibit increased theta power in the frontal lobe when compared to normal controls, suggesting hypoarousal in the frontal cortex. Further, there is some suggestion that the ratio of theta to beta power is unusually weighted toward theta among individuals with ADHD. Given that the leading contemporary theory of ADHD posits that the disorder stems from under-arousal in areas of the brain (e.g., frontal lobe), which is associated with behavioral inhibition, the EEG findings are compelling (Barkley, 2006). It is also interesting to note that some EEG findings suggest a unique subset of individuals with ADHD who may coincide with the sluggish cognitive tempo (SCT) group, described earlier. However, there is currently an unacceptably high rate of misdiagnosis (20

to 30%) based on EEG data, and there is little research on differential diagnosis between ADHD and learning disabilities, depression, or anxiety (Loo & Barkley, 2005).

Given the disappointing results of studies examining objective measures of ADHD, it is not surprising that the DSM-IV(-TR) fails to identify a standard assessment protocol, and the relevant literature has failed to uncover a single "gold standard" assessment procedure (Power & DuPaul, 1996). In fact, the same is true of all childhood psychiatric disorders (De Los Reyes & Kazdin, 2005; Kraemer et al., 2003). However, the objective differences found between individuals with ADHD and their non-ADHD peers lend support to the notion that ADHD is a legitimate physiological disorder that deserves clinical attention (Barkley, 2006).

Clinical Assessment of ADHD

In lieu of a standard assessment protocol or "gold standard" test, ADHD is assessed based on behaviors and impairments (Furman, 2005). There are two main purposes for assessing ADHD, including diagnosis and treatment planning. Diagnostic evaluations of children and adolescents suspected of having problems with attention or hyperactivity-impulsivity utilize the criteria

outlined in the DSM-IV(-TR), described above (see Diagnostic Criteria for details). Typically, a comprehensive diagnostic evaluation includes a diagnostic interview with the primary caregiver, behavior rating scales from caregivers and teachers, psychological and academic tests, and developmental and school histories (DuPaul & Stoner, 1994).

When assessing ADHD, clinicians must weigh the costs and benefits of the various strategies available. For example, clinical interviews offer the possibility of a comprehensive overview of behaviors, impairments, and possible comorbidities, but are time consuming and require intensive training. Naturalistic observations provide ecologically valid assessment of the nature and severity of symptoms, but also are time consuming and require specialized training. Further, low-frequency behaviors, such as aggression, may not be seen during an observation, so more than one session is typically needed. Analogue observation, where clinicians observe behavior in a contrived clinic setting, improves the likelihood of observing relevant, potentially low-frequency behaviors, but lacks ecological validity (Pelham, Fabiano, & Massetti, 2005).

An emerging approach to assessment is functional behavioral assessment (FBA; Olympia & Larsen, 2005). FBA is a strategy of assessing the environmental influences on behavior in an attempt to understand the root causes and behavioral contingencies that affect target behavior. Clearly FBA has potential clinical utility, as the analysis may directly inform intervention. In fact, FBA is required in certain circumstances under the Individuals with Disabilities Education Act (IDEA-97), and was maintained in the most recent iteration of the law (IDEIA-2004). However, FBA requires intensive training and the data that are collected may oftentimes seem ambiguous and difficult to interpret. Further, there is a lack of consensus regarding how FBA should be conducted and how data should be collected and interpreted (Olympia & Larsen, 2005).

Beyond diagnosis, clinical assessment is used to determine eligibility for special services, to inform treatment approaches, and to measure the outcomes of intervention (Pelham, Fabiano, & Massetti, 2005). To these ends, clinicians often use assessment procedures identical to those used for diagnosis, but also techniques to assess short-term outcomes, such as daily report cards (DRC; Evans & Youngstrom, 2006). In

strategies such as the DRC, teachers provide ongoing ratings of child behavior on specific behaviors and then progress is reinforced at both home and school. Similarly, clinicians often ask parents and teachers to provide ongoing behavior ratings on standardized ratings scales to help inform treatment progress and outcomes. This latter strategy is often used in clinical research (e.g., Evans, Serpell, Schultz, & Pastor, 2007), in assessing the effectiveness of medication trials (e.g., The MTA Cooperative Group, 1999a), and is clearly applicable in school settings.

Rating Scales

Rating scales, in comparison to the other clinical measurement techniques mentioned above, offer a relatively efficient, straightforward, and technically precise means of assessing ADHD behaviors. While rating scales are only a small component of the recommended ADHD assessment battery (e.g., Pelham, Fabiano, & Massetti, 2005), ratings provide insight into the perceptions of adults who are close to the target child. In fact, rating scales from multiple informants are recommended as a component of best practice assessment of childhood psychiatric disorders (American Academy of Child and

Adolescent Psychiatry, 1997; American Academy of Pediatrics, 2001).

Advantages of Rating Scales

As a clinical assessment technique, rating scales have several advantages. First, rating scales are relatively easy to administer and, in most cases, require only a brief time for respondents to complete. Second, rating scales are typically standardized, so that raters respond to identical items, presented in a uniform manner. Thus, direct comparisons between respondents are appropriate. Third, there is little cost associated with rating scales, as they are often sold in large packs and each individual record form is reasonably priced (Evans, Williams, Schultz, & Weist, 2004; Pelham, Fabiano, & Massetti, 2005). Fourth, most published, norm-referenced rating scales boast technical precision based on large normative samples. Responses to the scale can be compared to the normative sample and standard scores can be derived that reflect the clinical meaningfulness of the ratings relative to a population of other respondents. Such comparisons can be highly precise, provided that clinicians carefully examine the norm group for each rating scale and assure that it is representative of the target, as large norm groups do not

necessarily guarantee suitable comparisons (Reid & Maag, 1994).

Given the advantages of rating scales, it is not surprising that they are widely used in ADHD assessment, including diagnosis, intervention planning, and treatment outcomes. In fact, it appears that school psychologists are increasingly relying on behavior rating scales in their assessment of school children (Shapiro & Heick, 2004). Behavior rating scales are appropriate for such purposes, provided that several inherent assumptions are met: 1) raters share an understanding of the behaviors rated, 2) raters can distinguish between occurrence and nonoccurrence of the behaviors, and 3) raters share a common metric by which to judge behaviors (Cairns & Green, 1979). However, as Reid and Maag (1994) point out, the assumptions underlying behavior rating scales are not always met.

Disadvantages of Rating Scales

The assumptions underlying rating scales are sometimes threatened by how raters individually interpret and perceive the target's behaviors. For example, in the assessment of ADHD, different raters may have conflicting impressions of what constitutes "fidgeting" behavior, based on subjective judgments of what fidgeting is, when

fidgeting is occurring, and how intense the fidgeting has to be before it becomes remarkable. Such inconsistencies arise because rating scales do not directly assess behavior, as is the case in direct observation or analogue assessment. Rather, as pointed out above, rating scales assess the perceptions of adults familiar with the target child. Consequently, the data provided by rating scales are indirect and dependent upon the idiosyncrasies of the raters (Reid & Maag, 1994).

Clinicians have long recognized that rater perception is subject to unexplainable error, and research has been conducted to study those conditions that help to improve the validity of rater data. In their meta-analysis of observer bias studies, Hoyt and Kearns (1999) found several moderators that increased variance attributable to rater error, including limited rater training time with the rating system (i.e., less than five hours) and little or no overlap in shared observations between the raters (i.e., minimal observation of the target in the same setting at the same time). Further, ratings that required greater rater inference (e.g., behavior not explicitly tied to scale items) was also associated with increased rater error.

Another limitation is that most narrow-band rating scales focus solely on the symptoms of interest to the exclusion of other vital data. For example, other factors such as functional impairment (Pelham, Fabiano, & Massetti, 2005) and the quality of life (Klassen, Miller, & Fine, 2004) are often ignored. Thus, as pointed out above, rating scales are best suited as one component of a clinical assessment and cannot completely replace other techniques, such as interviews and direct observations.

Analyzing Variance in Rating Scales

Due in part to the differing perceptions of raters across time and setting, it is often found that ratings are inconsistent. For example, the same rater may provide seemingly incongruous ratings on two separate instruments designed to measure the same construct, or multiple raters may provide equivocal results regarding the same target. Studies examining the behavior of raters have noted cases of extreme inconsistency within and between raters (e.g., Achenbach et al., 1987), and this has led to several ways of conceptualizing rater reliability.

In classical measurement theory, reliability is generally conceptualized as having two components: the true score and measurement error. The true score is the

component of the rating that is based objectively upon the behavior of the target and, conceptually, is the average rating from a large set of randomly selected raters (O'Brien, O'Brien, Packman, & Onslow, 2003). In other words, the true score is the component of the rating that can be considered accurate. The error is the component of the rating data that is variance due to influences that are independent of the target's behavior (Hoyt & Melby, 1999; O'Brien et al., 2003). For example, measurement error occurs when a rater misinterprets a specific item and responds inaccurately.

Researchers have discussed several types of rater reliability based on classical measurement theory. For example, test-retest reliability involves the degree of intra-rater consistency across multiple applications of the same instrument. Split-halves reliability involves the degree of intra-rater consistency within the same instrument. Alternate-forms reliability involves the degree of intra-rater consistency across two versions of the same instrument designed to measure the same construct. In each instance, a reliability coefficient in computed as a ratio of true score variance to actual score variance (true score variance plus the error variance). Thus, coefficients close to 1 suggest there

is little difference between observed and true scores, whereas coefficients close to 0 suggest high rates of difference, or low reliability (O'Brien et al., 2003).

Interrater reliability involves consistency between two or more raters using similar instruments. A lack of interrater reliability suggests that individual rater biases are not adequately controlled (Danforth & DuPaul, 1996). Oftentimes researchers and clinicians are concerned with the *relative reliability* between the raters, or the degree to which responses from one rater correlate with those of another. In other instances, measurement decisions may depend on the level of *absolute agreement* between raters. This can occur on rating instruments that have specific score thresholds ("cut-points") that define a clinically meaningful phenomenon (Hoyt & Melby, 1999). For example, as discussed previously, ratings of inattention and hyperactivity-impulsivity are often compared to the DSM-IV(-TR) criteria that six or more symptoms of either are present. In such instances, clinicians and researchers may be most interested in the degree of absolute agreement between multiple raters.

However, the classical measurement conceptualization of reliability seems overly simplistic, given the

complexities of measurement error in psychological research (Brennan, 2001; Shavelson & Webb, 1991). Kraemer and colleagues (2003) posit that there are at least four sources of variance in ratings, including target trait (T), context (C), perspective (P), and random error (E) (p. 1569). The T dimension refers not to stable traits, but rather the variance in symptomology observed over time. For example, among children with ADHD, there is behavioral inconsistency, even though the underlying disorder is theoretically always present. Thus, some of the variance observed in ratings is directly attributable to this inconsistency and is not attributable to unsystematic error. The C dimension refers to the environmental circumstances that influence the behavior, also referred to as *situational specificity*. For children with ADHD, there appears to be a significant impact of environmental cues on behavior. For example, hyperactivity appears to be most pronounced in non-stimulating environments, such as drab waiting rooms, and is not as apparent in situations that include stimuli such as televisions and videotapes (Antrop, Roeyers, Van Oost, & Buysse, 2000). Thus, C may or may not be considered systematic error, depending on the aims of the assessment. The P dimension refers to variables

specific to the raters that influence their ratings. For example, parents vary in how well they manage frustration. As a result, one parent's interpretation of the significance of hyperactivity is different than the interpretation of others, based on differing frustration tolerances. As Kraemer and colleagues (2003) explain, the P dimension is an important source of error, which necessitates various strategies to reduce the impact on data interpretation. The P dimension is also associated with some of the variance found in ratings from multiple informants. The E dimension refers to the error naturally associated with any instrumentation that is beyond the examiner's control. This error term (E) includes factors such as the misreading of specific items, situations that occurred just prior to the rating, and other uncontrollable influencing events (Kraemer et al., 2003).

Teacher Ratings

Due in part to the current trends toward greater mental health services in schools (Center for Mental Health in Schools, 2003; Flaherty & Osher, 2003; President's New Freedom Commission, 2003), teachers are increasingly called upon to provide vital psychological information relevant to the academic and behavioral

performance of their students. In fact, guidelines for the diagnosis of ADHD specifically identify that impairment must occur in two settings (APA, 1994, 2000), which is often interpreted as home and school. As such, teachers are often sources of information in diagnosis and treatment outcomes evaluation (e.g., American Academy of Child and Adolescent Psychiatry, 1997; American Academy of Pediatrics, 2000).

Teacher-parent reliability. There are limitations in the utility of teacher ratings when assessing ADHD. For example, studies using C-theory approaches have typically found only modest reliability and agreement between parent and teacher ratings of the same child (e.g., Mitsis, McKay, Schulz, Mewcorn, & Halperin, 2000; Wolraich, Lambert, Bickman, Simmons, Doffing, & Worley, 2004). As a result, it seems clear that each informant provides unique information (Achenbach et al., 1987). The implication is that teacher ratings cannot substitute for parent ratings, and vice versa. Hence, behaviors that may impact academic functioning are probably best rated by a teacher familiar with the child (Loeber, Green, & Lahey, 1990).

There are several strategies for interpreting inconsistent ratings from multiple informants, but each

appears to have potential flaws. The extant literature highlights two common approaches: 1) combining the ratings in some logical fashion, or 2) selecting the one rating that seems most accurate (Hart, Lahey, Loeber, & Hanson, 1994; Simonoff, Pickles, Hewitt, Silberg, Rutter, Loeber, et al., 1995). When researchers and school psychologists combine ratings, often a single indicator is produced. For example, if each rater's perspective is seen as valid, then all endorsed symptoms might be totaled (Cohen, Velez, Kohn, Schwab-Stone, & Johnson, 1987). This strategy was used in the MTA study described above (Lahey et al., 1994). However, when using this strategy, situation-specific symptoms will result in inflated symptom counts, leading to higher diagnostic prevalence rates and possible Type I errors. For example, Mitsis and colleagues (2000) found that this strategy resulted in inflated rates of ADHD Combined Subtype, and significantly fewer cases of nondiagnosis.

An alternative strategy for combining incongruent data is to count only those symptoms reported by both parents and teachers (i.e., raters in separate settings). There are potential benefits to this approach, as aggregated data from both parents and teachers is useful in positively identifying students with ADHD. When both

parent and teacher ratings suggest ADHD-consistent symptoms, the overall positive prediction rate appears to improve incrementally (Power, Andrews, Eiraldi, Doherty, Ikeda, DuPaul, et al., 1998). However, symptom-wise cancellation based on incongruent parent-teacher ratings results in significantly lower prevalence rates, which could lead to Type II errors. Interestingly, when the DSM-IV(-TR) criteria for symptoms and impairments in two settings is strictly interpreted, symptom-wise cancellation based on incongruent ratings is the logical approach (Wolraich et al., 2004).

The other general strategy involves choosing the rater (or raters) that appears most credible. In general, data from single reporters appear to inflate the identification of childhood disorders and the addition of other raters appears to reduce overidentification in most cases (Cluett, Forness, Ramey, Ramey, Hsu, Kavale, et al., 1998). Still, school psychologists often "weight" ratings when making clinical decisions, but the literature does not discuss this practice in detail. It appears that school psychologist decisions are made on a case-by-case basis, and the little research that is available offers only general guidelines. For example, research suggests that parents and children generally

provide valid ratings of CD symptoms, but are not particularly valid raters of ADHD and ODD symptoms (Hart et al., 1994). Conversely, researchers and mental health professionals generally view teachers as the most useful informants on the impact of child behavior on academic and social outcomes (Loeber, Green, & Lahey, 1990). Teacher ratings are a logical choice when academic and social outcomes are of importance as teachers generally have more opportunities to observe children's academic performance and social interactions with same-age peers.

Still, there are questions about the clinical utility of teacher ratings at the secondary level. Robin (1998) suggests that given the discrete classroom arrangements of secondary schools, where students will interact with multiple teachers in separate classrooms, "secondary education teachers do not have as comprehensive a picture of the average adolescent as do elementary school teachers, who see the younger student in a variety of activities" (p. 101). As a result, secondary teacher ratings of students are likely to be based on limited interactions, and the impact is apparent when examining between-teacher reliability rates at the secondary level. When considering the moderators of rater error (see Hoyt & Kearns, 1999), it seems that

secondary teacher ratings represent perhaps the worst-case scenario, as teachers generally receive inadequate training on child mental health issues (Weist, 2005), there is little or no observation overlap due to separate classrooms, and ratings of classroom behavior commonly require a great deal of rater inference.

Between-teacher reliability. It is common for between-teacher ratings to vary considerably, especially among secondary school teachers. For example, in previous studies interrater reliability among secondary teachers' ratings of child behavior were found to be appreciably poorer than those of elementary teachers (e.g., Achenbach et al., 1987). When interrater reliability is low, the utility of teacher ratings is questionable.

Molina, Pelham, Blumenthal, and Galiszewski (1998) examined interrater reliability among secondary teachers' ratings of adolescents with ADHD and found very low interrater reliability, with reliability coefficients appreciably below that by Achenbach and colleagues (1987). Molina and colleagues used three separate narrow-band rating instruments for disruptive behavior disorders, including inattention, hyperactivity,

aggression, and delinquency. Intraclass correlations (ICCs) for all the measures ranged from .13 to .52.

In their examination of middle school teacher ratings over the course of one school year, Evans, Allen, Moore, and Strauss (2005) concluded that ICCs among teacher ratings were the lowest from October to December (*M* = .27), especially on items measuring symptoms of inattention and social impairments. Based on these findings, the investigators concluded that collecting data on inattention prior to January "may not be worth the effort" (p. 704). Following January, however, ICCs improved (*Mean* ICC = .41) to a level comparable to that found in other studies for middle school samples (e.g., Achenbach et al, 1987; Molina et al., 1998). Thus, there may be a "time effect," whereby middle school teachers have a limited understanding of student challenges in the first semester of the school year, based on the discrete classroom arrangement of secondary schools and limited student contact, and are forced to rely on incomplete data when rating student performance until later in the school year (Evans, Allen, et al., 2005, p. 702).

Between-teacher reliability can be improved in specific settings. For example, Danforth and DuPaul (1996) found significant interrater reliability

coefficients (*p*s < .01) when ratings were collected from special education co-teachers who worked together and shared almost complete overlap in observations. While the correlations were significant, a large proportion of the variance was still unaccounted for. Thus, even in highly overlapped observations with trained special education teachers, school psychologists can expect to find a high degree of interrater inconsistency. In an attempt to interpret these findings, the authors concluded that, "characteristics of the teacher are a considerable source of error variance in ADHD rating scales" (p. 233).

Types of Rater Bias

Rater bias, or bias variance, refers to the degree of interrater disagreement that can be attributable to observer misperception or to accurate perceptions of different target behavior (as often occurs in non-overlapped observations). By this definition, bias is not necessarily the same as inaccuracy, and it can "be considered as a systematic source of variability in ratings and may be an object of study in its own right" (Hoyt & Kearns, 1999, p. 420).

All rating scales are susceptible to rater bias. Oftentimes interrater variance is based on individual

rater tendencies to be overly lenient or critical (Evans, Williams, et al., 2004). Thus, clinicians must interpret rating scale data by making clinical judgments regarding source or setting effects (DuPaul, 2003). This judgment is necessary for both broad- and narrow-band instruments, as source and setting effects are relative to each rater and differ from instrumentation and target effects.

Hoyt (2000) described several types of rater bias. In general terms, rater bias can be categorized into either rater-specific or dyad-specific bias. Rater-specific bias refers to cases where raters have varying perspectives on the construct measured. For example, among teachers rating students with ADHD, it is possible that some will have varying interpretations of what is ADHD-related behavior; one teacher may interpret poor class work as an exemplar of inattention, while another may attribute poor class work to a lack of ability unrelated to inattention. When there is inconsistency between raters in this fashion, it contributes to measurement error. The terms leniency or severity have been used to describe rater-specific bias that is consistent across targets, either in a forgiving or critical direction, respectively (Evans, Williams, et al., 2004).

In dyad-specific bias, individual raters are influenced by target-specific characteristics, such as target attractiveness (e.g., similarities between the target and rater in race or ethnic background) or agreeableness (e.g., good manners). From a measurement standpoint, dyad-specific bias is a much more serious form of error, as it will vary from target to target and from rater to rater, making it extremely difficult to estimate and correct in the analysis. For example, a teacher may rate one student's level of inattention as within a normal range, due to their agreeable demeanor. However, a different student exhibiting the same behaviors may be rated as much more impaired by the same teacher, due to characteristics unrelated to inattention. The term *halo* has been used to describe dyad-specific biases, and this effect can result in either positive (lenient) or negative (severe) ratings (Evans, Williams, et al., 2004).

Sources of Rater Bias

While types of potential rater bias are known, sources of bias are not well understood (Hill, O'Grady, & Price, 1988). As mentioned previously, few studies have been conducted to look at potential sources of bias. It is clear, however, that observers are largely unaware of

their own susceptibility to bias. Even when observers are made aware of various biases (e.g., attribution bias), there appears to be a tendency to deny one's own susceptibility to bias while concluding that others are highly susceptible (Pronin, Lin, & Ross, 2002). Clearly, more research is needed in order to understand the nature of rater bias and how bias-related error can be accounted for in psychological measurement. Of the studies that have been conducted, the majority appear to examine dyad-specific bias, although in practice there is a great deal of overlap between dyad- and rater-specific bias.

Sources of Dyad-Specific Bias

Evidence supporting dyad-specific bias in teacher behavior ratings has been discussed around issues of child race and ethnicity, gender, and comorbid behavior problems. In terms of ethnicity, research suggests that African-American children are rated as exhibiting higher rates of ADHD symptoms than their Caucasian peers by their teachers (e.g., Reid, DuPaul, Power, Anastopoulos, Rogers-Adkinson, Noll, et al., 1998). Similarly, teachers in the MTA study (described above) rated African-American children as exhibiting more symptoms of ADHD than Caucasian participants (Epstein, Willoughby, Velencia, Tonev, Abikoff, Arnold, et al., 2005).

However, there are several potential alternative explanations for this finding other than dyad-specific bias. First, it is possible that African-American children are more overactive than their Caucasian peers due to cultural differences. Second, African-American children may experience a referral bias (to school and/or community services) similar to that of girls, where it is possible that only the most impaired are referred. Or third, classes with predominately African-American students are more active that those of predominately Caucasian students. In examining these questions using the MTA data, Epstein and colleagues (2005) found some support for the latter explanation and recommended interpreting elevated teacher ratings of African-American students in the context of their classroom environment. However, Epstein and colleagues did not use objective measures to contrast with teacher ratings and, as such, were unable to draw definitive conclusions. In other research, it appears that when teacher and student are both African American, the tendency for severe teacher ratings disappears (Downey & Pribesh, 2004).

In a study that compared teacher ratings against objective measures of physical movement (actigraph), it was found that teachers rated minority students, in this

case Asian students attending British schools, as more physically overactive than was warranted from the objective measure (Sonuga-Barke, Minocha, Taylor, & Sandberg, 1993). Thus, while there is some evidence for a potential dyad-specific rater bias based on child ethnicity, the nature of this bias is unclear and more research using appropriate experimental comparisons is needed.

Other research has examined the impact of student gender on teacher behavior ratings. Using teacher ratings of the seriousness of student problem behaviors, Kokkinos, Panayiotou, and Davazoglou (2005) found that inexperienced teachers rated non-stereotypic behaviors (e.g., boys with depressed symptoms) as more problematic than their experienced counterparts. In contrast, experienced teachers rated stereotypic behaviors (e.g., boys exhibiting aggression) as more problematic than their inexperienced counterparts, suggesting that gender stereotypes influence teacher appraisals, but this influence changes with experience. The researchers concluded that inexperienced teachers appear to enter the profession with preconceived notions of what is acceptable behavior for boys and girls. Thus, it can be predicted that inexperienced teachers would tend to rate

stereotype-inconsistent behaviors as more severe than stereotype-consistent behaviors.

Other research suggests that child behaviors unrelated to ADHD can influence teacher ratings of inattention and hyperactivity/impulsivity. For example, teachers who examined videotapes of children exhibiting oppositional behaviors were likely to then rate targets as substantially inattentive and hyperactive/impulsive, even though these latter problems were not depicted (Abikoff, Courtney, Pelham, & Koplewicz, 1993; Stevens, Quittner, & Abikoff, 1998). Such findings suggest that conduct problems unrelated to ADHD can create a negative halo, whereby teachers perceive ADHD-consistent behaviors where they do not exist. More recent research suggests a potential gender-by-behavior halo interaction, whereby the tendency for teachers to rate depictions of oppositional behavior as consistent with ADHD was significantly greater for boys than for girls (Jackson & King, 2004). However, the studies summarized here utilized videotaped dramatizations of behavior and may not generalize to actual classroom settings.

Sources of Rater-Specific Bias

While there are few studies that exclusively examine rater-specific teacher bias, there are some potentially

informative studies of rater-specific bias under similar conditions. For example, Chi and Hinshaw (2002) examined parent ratings and discovered that mothers with significant depressive symptoms rated their children as more impaired than did teachers or the children themselves. The authors concluded that this relationship supported a "*Depression → Distortion* hypothesis" (Richters, cited in Chi & Hinshaw, 2002, p. 388). Chi and Hinshaw's finding underscores the idiosyncratic role of rater perception: Depressed mothers perceived their child's behaviors as problematic, perhaps as a function of their own distress, whereas other raters did not. Interestingly, depressed mothers also rated their own parenting styles as more negative than did independent raters observing mother-child interactions in a clinic setting.

Research examining observer bias in process ratings of counseling and psychotherapy is also informative. Hoyt (2002) examined observer ratings of therapist effectiveness as depicted in videotaped interactions. The results suggest that variance attributable to raters accounted for the vast majority of variance in the overall model, ranging from 21% to 32%. Hoyt then assessed the impact of rater effects by regressing rating

deviation scores computed by subtracting the overall mean rating for each therapist from each rater's rating, adjusting for order-of-presentation effects, onto individual rater differences (e.g., personality traits). The results suggest that raters with positive self-perceptions were likely to provide favorable impressions of the therapists. The author speculated that this may be related to a correlation between positive self-views and positive views of others.

In terms of teacher ratings, similar sources of rater-specific bias clearly exist. For example, evidence for rater-specific bias among teachers is found in the normative data for the ADHD-RS(-IV) (DuPaul et al., 1998). DuPaul and colleagues found that teachers overidentified ADHD in every age group when compared to the prevalence rates found in community samples. Similarly, Glass and Wegar (2000) found that teachers identified as much as 15% of their students with ADHD-consistent symptoms on behavior rating scales, despite the DSM-IV(-TR) estimated prevalence rate of 3% to 7%, suggesting a general tendency toward severity among teacher ratings of ADHD behavior.

Research examining the severity phenomenon suggests a possible relationship between class size and the

tendency for teachers to overidentify ADHD. In one study, overidentification occurred in classes with above average class size, but not in smaller classes (Havey, Olson, McCormick, & Cates, 2005), which may suggest that teachers are sensitized to problem behavior in large classrooms because of the demands and stress created by increasing class size. However, the opposite trend occurred in a study examining private school classrooms, where overidentification was associated with smaller class sizes (Glass & Wegar, 2000). Given such inconsistencies, Havey and colleagues (2005) concluded, "continued study of the relationship between class size and ADHD is important to determine the degree to which ADHD, a disorder with clear biological connections, is being identified because of environmental factors" (p. 124).

Other research has examined the effect of experience on teacher ratings of student behavior. In a study of the impact of teacher experience on their perceptions of various student behaviors, Kokkinos, Panayiotou, and Davazoglou (2004) found that while all teachers rate externalized behaviors (e.g., hyperactivity) as more serious than internalized behaviors (e.g., inattention), the severity of ratings of externalized behaviors seems

to attenuate with teaching experience. In other words, it appeared that experienced teachers become increasingly tolerant of externalized behavior problems over time. Conversely, inexperienced teachers appear generally insensitive to internalized behavior problems. It could be predicted that teachers with little experience would provide severe ratings of externalized problems and lenient ratings of internalized problems, as compared to their experienced counterparts.

Changes in teacher tolerances for classroom behavior problems over time may be related to stress and occupational burnout. Teacher burnout is characterized by emotional exhaustion and feelings of frustration and inadequacy, which in turn can affect appraisals of student behavior (Schamer & Jackson, 1996). Kokkinos and colleagues (2005) found that experienced teachers who self-reported burnout appeared less tolerant of antisocial and oppositional classroom behavior than their non-burned out counterparts, suggesting that teacher perceptions of student behavior change with experience, but that these changes may interact with feelings of stress. Specifically, it appears that when experienced teachers feel overly stressed, they are more likely to rate student problem behaviors in a severe manner than

teachers who are not overly stressed. This finding appears analogous to the impact of depression on parent ratings of ADHD cited above, whereby parental depression appears to result in more severe perceptions of child problem behavior.

Conclusion

This chapter has examined the DSM-IV(-TR) criteria for ADHD, the theories surrounding its etiology and course, and the current best practice methods of assessing children and adolescents with the disorder. Of particular importance are the diagnostic criteria for ADHD. The behaviors that comprise the disorder are delineated in the DSM-IV(-TR), along with additional criteria used to differentiate ADHD from other similar disorders. Unfortunately, attempts to measure ADHD symptoms objectively (e.g., IQ tests, continuous performance tasks, and executive functioning measures) have proven inadequate and unreliable. In fact, it appears that no single assessment technique or measure has adequate sensitivity and specificity to reliably differentiate between ADHD and non-ADHD groups. Thus, best practice assessment requires multiple techniques, with information from multiple sources. In general, this approach typically includes parent interviews, parent and

teacher rating scales, classroom or analogue observations, psychological and academic testing, and functional behavioral assessment. Of these, the most efficient and technically precise method appears to be behavior rating scales. As noted throughout, much of the research cited in this chapter has relied heavily on the behavior ratings of adults familiar the child.

Teachers are considered to be a valid and reliable source, given their close role with children and their unique ability to assess a child's performance in the school environment. However, questions surround the utility of teacher reports for adolescents at the secondary school level. For example, between-teacher reliability at the secondary level is substantially weaker than at the elementary level, and appears to fluctuate throughout the school year. Further, rating scales are susceptible to variations in target behavior, situational error, random error, and rater perspective or bias.

Given the multiple sources of variation, it is not surprising that school psychologists are often confronted with conflicting teacher data. In such cases, it is not always clear how these discrepancies are best interpreted, and several strategies present themselves.

For example, divergent ratings could be aggregated in some fashion, or the school psychologist may decide to choose what appears to be the most accurate rating. In any event, the examiner must interpret the discrepant data based on clinical judgment.

One important contributing factor to error variance is rater bias, but unfortunately there is dearth of research on this topic. There are several potential sources of dyad-specific bias, including similarity between student-teacher race, student gender, and student comorbid behavior problems. Among rater-specific biases, there is some suggestion that class size, teacher experience, and teacher burnout influence teacher appraisals of student behavior, but these relationships are unclear.

CHAPTER III

METHODS

Introduction

This chapter describes the design and procedures used in the present study. As described in the introduction, the aims of this study were twofold: First, this study examined interrater reliability among middle school teacher ratings of adolescents with Attention-Deficit Hyperactivity Disorder (ADHD). Second, potential sources of rater bias were explored by evaluating how well teacher characteristics predicted lenient and severe ratings. Most of the data analyzed in this study were archival; as a result, much of this chapter will describe characteristics of the archived sample, as well as the procedures used to gather those data. In addition, this chapter will discuss the instruments and specific procedures used to address the second aims of the present study, which required additional data collection. Finally, this chapter provides descriptions of the statistical procedures used to analyze the data, as well as the associated power and sample size analyses. Complications that arose during the course of the study and the adjustments taken to address those complications will be discussed in the following chapter.

Design

The first aim of the study was to assess interrater reliability among teachers rating the ADHD-related symptoms and impairments of middle school youth. To answer this question, the researcher relied on archived data from the Challenging Horizons Program – Consultation Model (CHP-C; Evans, Serpell, Schultz, & Pastor, 2007), a school-based study of the effectiveness of school consultation for young adolescents with ADHD. (The CHP-C study is described in greater detail later in the chapter and site coordinator letters of permission are reproduced in Appendix A.) Since the CHP-C was a field study, experimental manipulation of the relationships between the teacher raters and the targets (hereafter, the terms *targets* and *students with ADHD* are used interchangeably) was impossible. Thus, the resulting analyses were static group comparisons, consistent with a pre-experimental design.

In the CHP-C study, groups of teachers rated middle school students with ADHD based on the natural relationships occurring in the school setting. As a result, not all teachers rated all targets. Rather, teachers only rated those targets selected by the schools to be in their classrooms. As a result, data in the

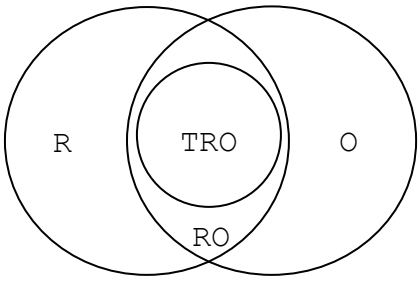
CHP-C study were collected in a pattern referred to as a partially-crossed, incomplete block design (Hoyt, 2000), where raters rated a varying number of targets and not all targets were rated by all raters. Thus, the variance attributable to raters and the variance attributable to targets were partially nested and, from an analysis standpoint, completely confounded with measurement error (Brennan, 2001). Figure 1 provides a visual representation of the CHP-C study design, and sources of variance in the CHP-C study are depicted in Figure 2. Due to the complex design of the CHP-C study, it was necessary to assess reliability among teacher target ratings on the basis of between- and within-target variances. Figure 3 provides an overview of the research design used to assess reliability among teacher raters in the first aim of this study.

In the second part of the study, potential sources of rater bias were examined. Given that this analysis focused on teacher characteristics, such as demographics and experiences, no manipulations of independent variables were possible. Thus, the second aim of the study also utilized a static group comparison, pre-experimental design. The results from the analysis, therefore, are descriptive in nature and speak only to

	O1								O2							
	R1	R2	R3	R4	R5	R6	R7	R8	R1	R2	R3	R4	R5	R6	R7	R8
T1	+	+	+	+					+	+	+	+				
T2	+		+	+					+	+	+	+				
T3		+	+	+					+	+		+				
T4	+		+							+	+					
T5					+	+		+					+	+		+
T6					+	+	+	+						+		+
T7					+	+		+					+			
T8						+	+	+					+	+		+

T = Target, R = Rater, O = Occasion, and + = rating

Figure 1. Diagram of the CHP-C measurement design.



T = Target, R = Rater, and O = Occasion

Figure 2. Venn diagram of variance sources in the CHP-C measurement design.

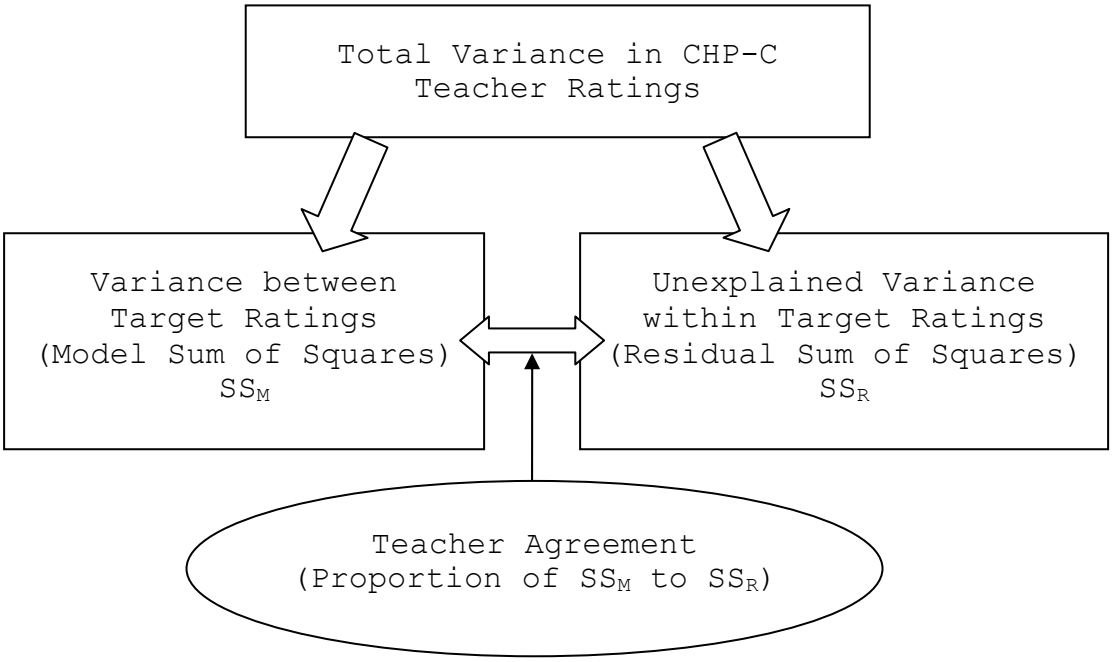
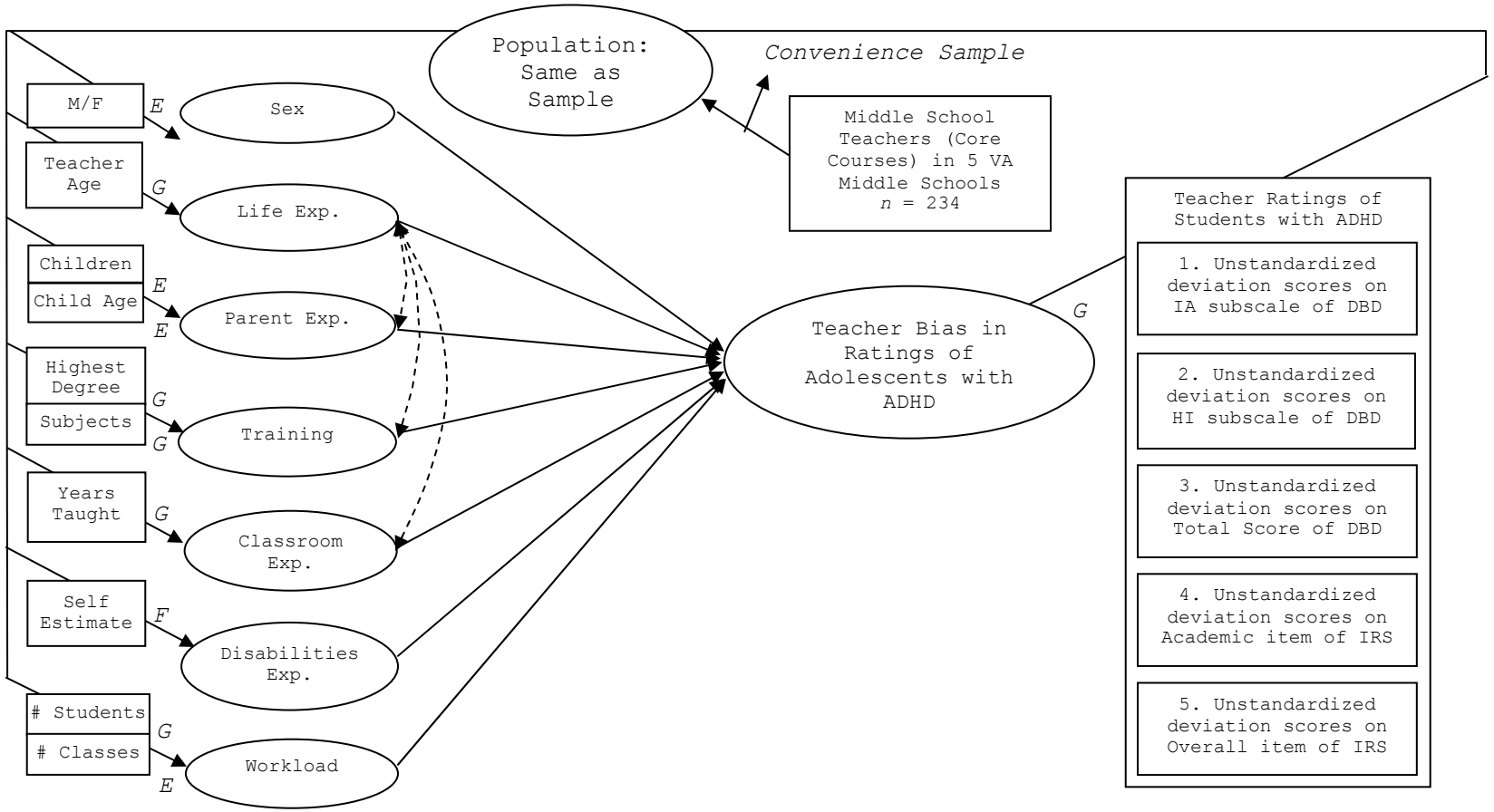


Figure 3. Diagram of the first aim of the present study: Interrater reliability among teacher groups on ratings of student behavior.

the relationships occurring within this setting. Based on the resulting analysis, there can be no assumptions of cause and effect, only association. An overview of the design of the second aim of the study is provided in Figure 4.

Population

The CHP-C was a school-based study of middle school students with ADHD that was conducted from the 2003-2004 to 2005-2006 school years. The five middle schools that participated in the CHP-C study are located in Rockingham and Augusta Counties of Virginia. At the time of the 2000 U.S. Census Bureau, Rockingham County had a total population of 67,725 and an average of 79 residents per square mile. The vast majority of Rockingham County residents identified themselves as white (96.6%), followed by African American (1.4%). Median household income in 1999 for Rockingham County was estimated at $40,748. According to the 2000 U.S. Census Bureau, Augusta County had a population of 65,615, with an average of 68 residents per square mile. The majority of Augusta County residents identified themselves as white (95.0%), followed by African American (3.6%). Median household income in 1999 for Augusta County was estimated at $43,045.



Note. *E* = Excellent, *G* = Good, and *F* = Fair estimated validity and reliability

Figure 4. Proposed regression model for the second aim of the present study: Teacher bias analysis.

Sample

Given the aims of the present study, the analysis required data from two groups of CHP-C study participants: middle school students with ADHD (i.e., targets) and their teachers. Since characteristics of both groups are important in interpreting the results of the present study, each will be described in detail.

Middle School Student Participants

The CHP-C study utilized a longitudinal, two-wave cohort design, with one cohort of student participants enrolled during their sixth through eighth grade years, followed by a second cohort enrolled in the study in their sixth and seventh grade years. An overview of the student participants' demographics is provided in Table 1. In total, the CHP-C study enrolled 79 middle school students between the ages of 10 and 14 (M = 11.93 at the time of intake). The majority of the CHP-C study sample was boys (77.2%); hence, the sex ratio in the CHP-C study was roughly equivalent with the estimated proportion of boys to girls with ADHD in the population (APA, 2000). Parents of the participants identified the majority as Caucasian (93.7%), and most families reported a total yearly income less than %60,001 (65.8%).

Table 1

Summary of Intake Data for Student Participants: Basic Demographic Information

	Variable	Number	Percentage
Sex	Male	61	77.2
	Female	18	22.8
Race	Caucasian	74	93.7
	Latino / Hispanic	3	3.8
	Did not specify	2	2.5
Income	0 - $20,000	13	16.5
	$20,001 - $40,000	23	29.1
	$40,001 - $60,000	17	21.5
	$60,001 - $80,000	14	17.7
	$80,001 - $100,000	5	6.3
	Above $100,000	2	2.5

Note. Income refers to yearly family income.

It was vital for the internal and external validity of the CHP-C study to establish that targets met the diagnostic criteria for ADHD. To that end, the CHP-C study researchers utilized a recruitment and intake screening process. To begin recruitment, parents of all students attending the five participating schools were sent study announcement flyers at the start of the 2003-2004 and 2004-2005 school years. The CHP-C study flyers provided an overview of the study and encouraged interested families to telephone researchers at the Alvin V. Baird Attention and Learning Disabilities Center (ALDC) at James Madison University. Respondents answered a phone screen, administered by a trained research assistant, which inquired about the presence or absence of ADHD symptoms based on DSM-IV(-TR) criteria. Respondents who indicated the presence of six or more ADHD symptoms were then scheduled for a clinical evaluation at the ADLC (Evans et al., 2007).

The clinical evaluations occurred on separate days and were administered by trained graduate students under the supervision of a certified school psychologist (Brandon Schultz) and licensed clinical child psychologist (Steve Evans). Each evaluation began with informed consent procedures, which were approved by the

James Madison University IRB. The timing of each participant's evaluation depended on the cohort that he or she entered (either the 2003-2004 school year or the 2004-2005 school year). In general, intake assessments were completed by January of each school year, although in a few instances the assessments were conducted later in the school year due to scheduling issues or if the student was replacing a former student who had dropped out of the study.

In the evaluations several variables were examined, including cognitive ability, academic performance, and behavioral functioning. To assess cognitive ability, participants were administered the Kaufman Brief Intelligence Test (K-BIT; Kaufman & Kaufman, 1990). The K-BIT is a brief intelligence scale designed to efficiently estimate intellectual functioning, and is comprised of only two subtests. The K-BIT subtests are intended to measure both verbal and nonverbal reasoning. Potential CHP-C study recruits whose estimated full-scale IQ scores (i.e., overall score comprised of both verbal and nonverbal performance tasks of the K-BIT) fell below a standard score of 80 were excluded from the CHP-C study, as it was anticipated that students with scores below this threshold would not benefit from the cognitive

and behavioral interventions targeted in the study (Evans et al., 2007).

To assess academic performance, researchers administered the Wechsler Individual Achievement Test, Second Edition (WIAT-II; The Psychological Corporation, 2002). For the purposes of the CHP-C study, only the abbreviated version of the WIAT-II was administered (Word Reading, Math Computation, and Spelling subtests), with the addition of the Written Expression subtest. Student performance on this test did not inform the diagnosis of ADHD, but it was deemed important to gather information regarding the impact that ADHD symptoms may have had on the students' ability to benefit from classroom instruction. No additional diagnoses (e.g., learning disabilities) were made based on the scores, but the information was shared with teachers once students began the CHP-C program (Evans et al., 2007). A summary of the intake data from the K-BIT and WIAT-II are presented in Table 2.

Diagnostic decisions were based on parent responses during a structured interview, the Diagnostic Interview Schedule for Children - IV (DISC-IV; Shaffer, Fisher, Lucas, Dulcan, & Schwab-Stone, 2000), and teacher and parent responses to both broad- and narrow-band rating

Table 2

Summary of Intake Data for Student Participants: Standard Scores on Cognitive (K-BIT) and Academic (WIAT-II) Measures

Instrument	Min	Max	Mean	SD
K-BIT				
Full Scale IQ	81	132	104.0	11.8
Verbal IQ	72	126	100.2	11.8
Performance IQ	83	136	106.9	12.8
WIAT-II				
Word Reading	46	133	99.2	13.4
Numerical Operations	46	135	94.2	14.9
Spelling	49	134	96.9	14.4
Written Expression	48	126	96.9	14.6

scales. Behavioral ratings at the time of intake were collected from parents and teachers using the Parent and Teacher Rating Scales of the Behavior Assessment System for Children (BASC; Kamphaus & Frick, 1998), Home and School Versions of the Disruptive Behavior Disorders scale (DBD; Pelham, Gnagy, Greenslade, & Milich, 1992) and the Impairment Rating Scale (IRS; Fabiano, Pelham, Gnagy, Waschbusch, Lahey, Chronis, et al., 2006).

All middle school student participants who were accepted into the study met the diagnostic criteria for one of the subtypes of ADHD, based on the criteria set forth by the Diagnostic and Statistical Manual – Fourth Edition – TR (DSM-IV-TR; APA, 2000). Diagnostic decisions were made by clinical consensus between a licensed clinical psychologist and a school psychologist using the data collected during the intake evaluation. Specifically, the researchers compared the parent responses to the structured interview to the results of parent, teacher, and child rating scales. Per best practice recommendations (e.g., Pelham, Fabiano, & Massetti, 2005), the researchers emphasized the reports of parents and teachers in establishing the presence of ADHD symptoms in two or more settings. Parent reports of developmental history (also gathered in the structured

interview) were used to establish DSM-IV(-TR) criteria for chronicity (i.e., age of onset prior to age seven, more than six months of continuous symptoms). Finally, parent and teacher ratings of impairment along with school records were used to establish significant impairment (i.e., failing grades, clinically significant ratings of social, academic, or familial functioning). In other instances, measures of behavioral functioning helped to verify the presence of behaviors pertinent to either ADHD or other potential disorders, such as anxiety or depression (Evans et al., 2007). An overview of the final diagnostic decisions for participants in the CHP-C study is provided in Table 3.

Teacher Participants

At the outset of the present investigation, much less was known about the teachers who participated in the CHP-C study than was known about the student participants. All teacher participants in the present study taught at the schools participating in the CHP-C study, but the CHP-C researchers were unable to collect individual teacher information prior to participation. However, it was assumed that most teachers resided in either Rockingham or Augusta County in Virginia, or in the local rural and suburban communities. Throughout the

Table 3

Summary of Intake Data for Student Participants: ADHD Subtypes and Comorbidities

	Diagnosis	Percent
ADHD	Inattentive Subtype	35.4
	Hyperactive-Impulsive Subtype	0.0
	Combined Subtype	64.6
Comorbid Diagnoses	Oppositional Defiant Disorder	46.2
	Conduct Disorder	16.9
	Mania/Hypomania	3.1
	Major Depression	3.1
	Dysthymic Disorder	1.5

Note: Comorbid diagnoses were based on the results of a structured interview with the primary caregiver. Anxiety disorders were assessed using separate rating scale data and are not included in this table.

three-year CHP-C study, a total of 234 teachers provided behavior ratings of the student participants, with 107 teachers in the Augusta County sites (89% of teaching faculty) and 127 teachers in the Rockingham County sites (75% of teaching faculty).

Assignment

As mentioned above in the Design section, all analyses in the present study were static group comparisons based on teacher demographics and experiences, as well as target ratings collected in the CHP-C study. Thus, experimental conditions were not assigned in the present study and all of the participants in the CHP-C study were entered into the analysis. However, it is important to note that in the CHP-C study, two of the participating schools (one from Augusta and one from Rockingham Counties) were randomly pre-assigned to a treatment condition that received teacher training and school consultation (Student *n* = 43, Teacher *n* = 113), and the remaining three schools were pre-assigned to a community care condition where families were encouraged to pursue treatments otherwise available at their schools and in their local communities (Student *n* = 36, Teacher *n* = 121). Teachers in the treatment schools received four to five hours of training on ADHD and the

procedures of the program prior to the start of each school year. The CHP-C study teacher trainings included information pertaining to the clinical diagnosis, prognosis, and academic relevance of ADHD. Since the CHP-C study employed a consultation model, whereby the teacher-consultant relationship was vital, specific attempts were made to establish cooperative partnerships with the participating teachers beginning with the first training session (see Schultz & Cobb, 2005; Schultz, Reisweber, & Cobb, 2008). As such, specific attempts were made to build consensus around the idea that targets were "at risk" for long-term academic and social problems. Teacher questions regarding ADHD and its treatment were answered by the researchers based on the current research.

During the initial trainings, the CHP-C researchers also introduced the rating scales that were used as dependent measures in the present study (described below). Much of the focus was on explaining the technical aspects of completing and submitting the scales, which proved problematic for many teachers at the start of each school year. Otherwise, no attempts were made to provide operational definitions of the individual rating scale items (e.g., defining "fidgetiness") or to

build consensus about how specific behaviors should be rated. Rather, teachers were allowed to interpret rating scale items independently.

Following this training, a certified school psychologist provided ongoing behavioral consultation for teachers and school counselors, based on a set of psychosocial interventions -- outlined in a treatment manual -- that have shown promise for middle school students with ADHD (see Evans, Langberg, Raggi, Allen, & Buvinger, 2005). To support implementation, teachers were encouraged to develop one-to-one "mentorships" with participants, in collaboration with the school consultant (see Schultz & Cobb, 2005). In addition, school counselors were recruited to implement social skills groups for the study participants. While the psychosocial components of the CHP-C study were provided through the mediation of teacher consultees and school counselors, community-based pediatricians were available to supervise medication trials at the request of the participants' legal guardians. All interventions focused solely on ameliorating the academic and social difficulties commonly exhibited by young adolescents with ADHD.

Participants attending schools that were randomly assigned to the community care condition were encouraged to seek services otherwise available in their schools and communities. To support these efforts, teachers at community care sites received the same trainings provided to the teachers at the treatment sites, but were not provided with ongoing consultation. Instead, researchers provided community care families with a list of local community-based resources and offered assistance finding appropriate resources by phone.

As a result of the experimental conditions in the CHP-C study, it is possible that teacher perception of participant behavior was differentially influenced across the comparison groups, thus introducing a potential complication to the present study. For example, teachers in the treatment condition may have been motivated to rate participant behavior more leniently, based on their desire to please the school consultant. I will address this possibility in the Results section.

Measurement

To answer the research questions of the present study, it was vital to gather data relative to teacher ratings of target behavior and data concerning teacher characteristics. In terms of teacher ratings, this study

focused on teacher perception of ADHD symptoms, based on the current DSM-IV(-TR) description, and teacher perception of academic and overall target impairment. The present study also examined teacher demographics and indicators of personal and professional experiences to assess how well these variables predicted rater bias. In this section, the definition and measurement of these variables are described in detail.

Variables

In the present study there were several latent variables related to the research questions, including the degree of interrater reliability on teacher ratings of middle school students with ADHD and potential sources of rater bias. Hence, the first variable of interest was between-teacher reliability. Specifically, the researcher was primarily interested in the degree of consistency among groups of teachers and among teacher dyads. In terms of teacher groups, the present study examined consistency in ratings from as many as four teachers, each from separate classes (four core courses, including reading, social studies, math, and science), across multiple measurement occasions. In terms of teacher dyads, the researcher was interested in rating consistency between any two teachers from the larger

teacher group. Consistency in this aspect is theoretically different from absolute agreement, where teacher ratings on continuous scales would match exactly, or where teacher ratings would agree in terms of predetermined cutpoints for diagnosis or no diagnosis, for example. Instead, the researcher was interested in examining the degree to which teacher ratings were consistent for each target; for example, the degree to which severe ratings from one teacher would be replicated from additional teachers. Or conversely, the degree to which relatively lenient ratings would be replicated by other teachers.

The second latent variable of interest was teacher rater bias. In the assessment literature, the term rater bias is defined simply as the disagreement between raters that is based on unique interpretations of the instruments or of the target under observation. In most instances rater bias is considered a component of measurement error and is not treated as a source of systematic variance (Hoyt, 2000). In the present study, teacher bias was viewed as a potential source of systematic variance in ratings of ADHD symptoms and impairment by defining bias as disagreement between one teacher rating of a target and the average teacher rating

of that same target. When conceptualized in this manner, the range of potential discrepancies is determined by the scales used, with the maximum discrepancy limited to the range of acceptable responses minus one. Thus, teacher bias was conceptualized as one dimensional, ranging from the lowest possible discrepancy score to the highest possible discrepancy score allowed by the respective instrument. In this manner, low discrepancy scores suggested relative rater leniency and high discrepancy scores suggested relative rater severity. It is important to note, however, that the *cause* of leniency or severity could be attributable to accurate perceptions of classroom behavior, dyad-specific bias based on student-teacher match (e.g., halo effects), or inaccurate perception of student behavior by the teacher. The design of the present study did not permit the researcher to parse potential causes of bias or to study such causes separately. Thus, in the present study, bias was simply conceived as systematic variance in teacher ratings, and not necessarily rater error.

Additional variables of interest included several teacher characteristics that were assessed for their ability to predict teacher bias. Specifically, the researcher assessed the leniency and severity effect

related to demographic variables, such as teacher sex and age, as well as several experiential variables, such as teacher training, classroom experience, experience with student disabilities, parental experience, and workload. Conceivably, women and men perceive child behavior problems in unique ways, and similarly, younger and older raters may perceive child behavior problems in disparate ways. Further, experiences such as teaching experience, parenting experience, experience with student disabilities, and stress created by workload may potentially affect rater perception of child behavior problems. These variables are depicted in Figure 3 and the measures used to assess them are explained in detail below.

Instruments

As mentioned above, teacher perception of target behavior in the CHP-C study was monitored using five measures taken across two separate teacher rating scales: one narrow-band instrument that measures ADHD symptoms and one broad-band instrument that measures functional impairment.

Disruptive Behavior Disorders Scale

The Disruptive Behavior Disorders Scale (DBD; Pelham, Gnagy, Greenslade, & Milich, 1992) is a narrow-

band scale that was created to assess disruptive behavior disorders, as defined by the DSM. The DBD was originally designed around the DSM-III-R criteria for Attention Deficit Disorder (ADD), Oppositional Defiant Disorder (ODD), and Conduct Disorder (CD). To accurately capture these data, the DBD includes items that reflect the DSM-III-R behavioral criteria virtually verbatim. Using teacher ratings of a random sample of boys, Pelham and colleagues (1992) found that the DBD has excellent internal consistency ($\alpha = .96$)[1]. Further, individual items on the DBD appeared to have strong negative predictive power (NPP) for a full diagnosis of ADD (NPP rates per item all exceeded 0.95). In terms of positive predictive power (PPP), the items were not as strong (PPP rates ranged from 0.37 to 0.96), suggesting that items on the DBD were better at identifying students who did not meet the diagnostic criteria for ADD than it did in identifying those students who did meet the full criteria (Pelham et al., 1992).

The latest version of the DBD reflects the DSM-IV(-TR) criteria for ADHD, ODD, and CD virtually verbatim. For the purposes of the CHP-C study, researchers shortened the form to include only the 18

[1] Reliability descriptions based on recommendations by Kline (2005)

items relating to ADHD; nine items are associated with inattention and the other nine items are associated with hyperactivity-impulsivity. Consistent with the original DBD, raters respond to four-point scales, ranging from 0 to 3 representing *Not at all*, *Just a little*, *Pretty much*, and *Very much*, respectively, for each item. Separate scores can be computed for an inattention subscale, a hyperactivity-impulsivity subscale, and an overall total score. As such, the instrument maintains construct (and face) validity. The Inattention and Hyperactivity-Impulsivity subscales of the DBD range from 0 to 27, the total ADHD symptoms subscale of the DBD ranges from 0 to 54, and the two impairment items from the IRS range from 0 to 6. The ADHD subscale of the DBD is reproduced in Appendix B in the format used in the CHP-C study.

Although the CHP-C study version of the DBD (hereafter referred to simply as the DBD) lacks reliability, validity, or normative data, it clearly belongs to a class of narrow-band rating scales directly modeled from the DSM-IV criteria (Pelham, Fabiano, & Massetti, 2005). In this sense, the DBD is comparable to the ADHD Rating Scale, Fourth Edition (ADHD-RS-IV; DuPaul, Power, Anastopoulos, & Reid, 1998). For example, the ADHD-RS-IV also consists of 18 items, each based on

the DSM criteria for inattention and hyperactivity-impulsivity. Since both the DBD and ADHD-RS-IV are based on the DSM criteria, the items from both instruments are virtually identical. The response set is also very similar, with items from both instruments rated along four-point scales (0 to 3). However, the ADHD-RS-IV uses different anchors, which are *Never or rarely*, *Sometimes*, *Often*, and *Very often*. Both instruments are commonly used to assess children who are exhibiting symptoms consistent with ADHD.

Given the similarities of the two instruments, the psychometric properties between the DBD and ADHD-RS-IV may be comparable. DuPaul and colleagues (1998) report strong validity and reliability for the school version of the ADHD-RS-IV. Based on teacher ratings of 52 children, the DBD exhibited internal consistency of .94 for the total score, .96 for the inattention subscale, and .88 hyperactivity-impulsivity subscale. Test-retest reliability over a four-week interval is reported as .90 for total score, .89 for the inattention subscale, and .88 for the hyperactivity-impulsivity subscale. The ADHD-RS-IV was also found to correlate moderately well with direct observations of off-task classroom behaviors,

ranging from a low of .22 on the hyperactivity-impulsivity subscale, to .35 on the inattention subscale.

In terms of validity, the ADHD-RS-IV was found to differentiate between children with ADHD and non-impaired peers in a school setting. Teacher ratings alone accurately classified children with ADHD Predominately Inattentive Subtype versus the control group 75% of the time. Teacher responses to the ADHD-RS-IV were found to accurately classify students with ADHD Combined Subtype 78% of the time (DuPaul et al., 1998). Given its strong psychometric properties, the ADHD-RS-IV has been recommended for use as a narrow-band measures of ADHD for school psychologists (Demaray, Elting, & Shaefer, 2003). By virtue of their shared design, it appears likely that the ADHD-RS-IV and DBD have similar psychometric properties.

Impairment Rating Scale

The Impairment Rating Scale (IRS; Fabiano et al., 2006) is a broad-band measure of several functional impairments commonly exhibited among students with disruptive behavior disorders. There are separate parent and teacher version of this instrument. The teacher version of the IRS consists of six items: two items relate to social functioning (with peers and with the

teacher), two items relate to academic performance (progress and classroom functioning), and the last item asks the rater to consider if, overall, additional treatment or special services are required. Each item includes a line that the rater is asked to mark, with the left side of the line anchored by *No Problem, Definitely does not need treatment*, and the right side anchored by *Extreme Problem, Definitely needs treatment*. Items are scored by laying a transparent metric over top and then recording a score (1 to 6) based on the placement of the rater's mark. The IRS is reproduced is Appendix C.

According to preliminary studies of the IRS, the teacher version appears to have excellent internal consistency ($\alpha = .95$), adequate to excellent test-retest reliability (Pearson *r*s range from .74 to .96 with 3 to 4 month intervals between administrations), and good convergent and discriminant validity (Positive Predictive Power [PPP] = .90 and Negative Predictive Power [NPP] = .74) (Fabiano et al., 2006).

Recently, new norms for the parent and teacher version of the IRS were published based on the results of four separate studies of the instrument. Based on the results, items on the teacher version of the IRS appear to have less-than-adequate test-retest reliability

(Median *r* = .56), but overall interrater reliability between teachers and parents (*r* = .64) is adequate, and appears comparable to that of other measures of impairment. In terms of convergent validity, the IRS appears to correlate strongly with the DBD (*r*s ranging from .67 to .85). Further, the IRS was found to have moderate to high correlations when compared to other teacher instruments that measure impairment (Fabiano, Pelham, Waschbusch, Gnagy, Lahey, Chronis, et al., 2006). The present study examined teacher responses to the two items relating to academic impairment and overall impairment, respectively.

Teacher Questionnaire

The second aim of the present study was to examine potential sources of rater bias. While some information was known about the teachers in the CHP-C study (e.g., sex), data on other characteristics and experiences were specifically gathered for further analysis. To that end, teachers participating in the CHP-C study were administered a brief questionnaire, called the *Teacher Questionnaire*, which was designed by the researcher (see Appendix D).

Since the Teacher Questionnaire was designed specifically for the present study, the psychometric

properties are unknown; however, the individual items were designed to be easily understood and the data gathered are straightforward. Teachers may have been motivated to provide misleading data in some instances (e.g., age and highest degree attained), but it was not anticipated that misleading information would occur to a significant degree.

Items on the questionnaire inquired about demographic information, including teacher life experience, parenting experience, professional training, classroom experience, experience teaching children with disabilities, and workload. The items used to measure each construct are described in detail below, along with the rationale for their inclusion in this study.

Life experience. Life experience was defined as the breadth of each teacher rater's personal experiences, both within and outside of the school setting. To assess this construct, the researcher constructed an item on the Teacher Questionnaire related to age, that simply read: "What is your current age? Please check one." The response items included 6 categories: *20 to 28 Years*, *29 to 37 Years*, *38 to 46 Years*, *47 to 55 Years*, *56 to 64 Years*, and *65 Years or Older*. Each category was recoded in a database as the central year within each range; for

example, the 47 to 55 age category was recoded as 51. Age was treated in this fashion in an attempt to improve response rate and accuracy. The validity and reliability of the data collected by this item was anticipated to be good, given the wide ranges of response permitted by the item, but it is possible that some respondents provided incorrect data due to misinterpretation or purposeful deception.

Teacher age has been demonstrated to have a potentially indirect influence on rating bias in previous research, but this relationship may be mediated by other factors related to life experiences (e.g., Kokkinos et al., 1996). Regardless, teacher age is a straightforward demographic variable that appears to measure the construct of life experience. The researcher anticipated that this construct would correlate with other constructs measured by the Teacher Questionnaire. Specifically, it appeared that life experience would correlate with parenting experience, professional training, and classroom experience, all of which are somewhat related to age (refer to Figure 3).

Parenting experience. To address parental experiences, the Teacher Questionnaire asked about parental status. The first item simply read: "Are you a

parent?" to which raters circled either "Yes" or "No." Responses were recoded as 1 or 0, respectively. A second question read: "If yes, are any of your children middle school age or older?" Again, responses were recoded as 1 or 0, representing "Yes" and "No" respectively. The validity and reliability of the data collected by these items was anticipated to be excellent because the items were worded in a straightforward manner and it was not anticipated that teachers would be motivated to be deceptive in their responses.

Teacher parenthood and rating bias has not been examined in the existing literature; however, in informal discussions with the researcher, teachers in the present study conjectured that parenthood affected their perception of classroom behavior. Thus, this item was designed to gather the information needed to explore the relative ratings of teacher parents and nonparents.

Professional training. To assess professional training, teachers were asked, "What is your highest level of education? Please check one." Several response options were provided, including: *2 Year Junior College Degree (e.g., Associates)*, *4 Year College Degree (e.g., B.A.)*, *2 Year Graduate Degree (e.g., M.Ed.)*, *Graduate Degree + Certification (e.g., Ed.S.)*, *Doctorate (e.g.,*

Ed.D.), and *Other*. The "Other:" category includes a blank line for open-ended responses. For the purposes of analysis, the categories were recoded into the number of years traditionally associated with each level. For example, "2 Year Junior College" was recoded as 14; representing 12 years of grade school and 2 years of college. Following this pattern, the remaining categories were recoded as 16, 18, 19, and 20, respectively. Responses to the "Other" category were recoded based on the data provided, but credits earned toward an incomplete degree were not counted. Only earned degrees were entered into the database and recorded as the number of years that are traditionally associated with the degree.

A review of the literature uncovered no studies that have specifically examined teacher education and rater bias. However, it seems likely that teachers with advanced degrees have greater exposure to child mental health information when compared to Bachelor's level teachers who generally receive inadequate training in mental health issues (Weist, 2005). As a result, teachers with advanced degrees may perceive classroom behavior problems differently than their peers. Hence,

this item was included in the questionnaire based on literature related to teacher training issues.

Professional training was further assessed using items related to the primary classroom subject that each rater was trained to teach. Two related questions were asked on the Teacher Questionnaire, including "What subject(s) were you trained to teach?" and "What subject do you enjoy teaching most?" The first item was intended to be the primary source of this information and the second item was intended to clarify responses. The validity and reliability of these items was anticipated to be good because of the ease of interpretation, but some respondents may have been motivated to include incomplete degrees or continuing education experiences in their responses.

While there is no specific research examining differential teacher perceptions of behavior based on academic subject, it is clear that students with ADHD lag behind their peers in academics, particularly math and spelling (Barkley, 1990). It is conceivable that teachers in demanding academic settings are more likely to observe student behaviors consistent with inattention and hyperactivity-impulsivity. In turn, this may impact teacher ratings, especially in scenarios where teachers

are excited and invested in the topic. Hence, these items were included on the teacher questionnaire to explore the possible relationship between academic subject and teacher perception of problem behaviors.

Classroom experience. Classroom experience was measured as part of the teacher survey, based on responses to one item. This item asked, "How many years have you been teaching?" Responses to this item were interpreted literally and treated as continuous data. In instances where teachers provide a range (e.g., "11 or 12 years"), an average was computed. In the example given, 11.5 was used in the analysis. The validity and reliability of these data was anticipated to be good because of the straightforward nature of the information; however, some teacher respondents may have misread the item or have been motivated to falsify this information in an attempt to mask personal information, such as age or lack of classroom experience.

In the existing research, it has been shown that experienced teachers have differing perceptions of disruptive classroom behavior when compared to that of inexperienced teachers (e.g., Kokkinos et al., 2004). Based on this research, it appears likely that a relationship exists between rater bias and teaching

experience, so this item is included in the present study in an attempt to replicate these findings.

Experience with student disabilities. Respondents were asked about their experience with students with disabilities. This item read: "How much experience do you have teaching students with disabilities?" Immediately below this question is a 5-point scale, ranging from 1 ("Not at All") to 5 ("Almost Exclusively"). Raters were asked to circle one of the numbers along this scale. Responses that fell between the anchors were recoded as half-points (e.g., 2.5). The validity and reliability of this item was anticipated to be fair, as self-ratings of the nature used in the item are often misinterpreted or biased. For example, teacher respondents may have overestimate their experience based on a few instances of particularly difficult experiences with students with disabilities.

Similar to teaching experience and education, it seems likely that teachers with more experience teaching students with disabilities will have differing perceptions of classroom behavior issues as compared to less experienced peers. Again, a review of the literature did not uncover any studies that specifically

addressed this question, so this item was added to the questionnaire for exploratory purposes.

Workload. To assess teacher workload, respondents were asked about class size and the number of classes taught per day. In terms of class size, teachers were asked an item that read, "On average, how many students do you have in each of your classes?" This item included a blank line for open-ended responses. The data provided by teachers to this open-ended item were recorded as literally as possible. In instances where teachers provide a range (e.g., "24 to 25"), an average was computed and recorded. In addition, respondents were asked, "How many classes do you teach / co-teach per day?" This item was included to help interpret potentially unusable responses to the item pertaining to class size, but also for the purpose of combining the two items into one measure of teacher workload. Specifically, the researcher multiplied the two items to derive one single number that represents the estimated number of students each teacher encountered during the school day. The validity and reliability of these data were anticipated to be good for number of students per class, because it was an estimated average, and excellent for

number of classes because of the straightforward nature of the information.

Previous research suggests that a relationship exists between class size and teacher perception of classroom behavior problems (e.g., Havey et al., 2005; Glass & Wegar, 2000), but as discussed in the previous chapter, this relationship is unclear. Hence, these items are included in the scale to help clarify the impact of workload on teacher perceptions of problem behavior.

Other items not included. Research suggests a potential association between teacher burnout and rater bias (e.g., Kokkinos et al., 2005). In the present study, a measure of teacher burnout was infeasible, given the consultative relationship between the researcher and many of the participating teachers, and the potentially damaging impact such questions may have had on those relationships. Instead, the researcher chose to focus on straightforward demographic variables that required little personal disclosure. Similarly, there is research to suggest a relationship between race and teacher perception of student behavior (e.g., Sonuga-Barke et al., 1993); however, the issue of race is controversial and results from such studies can be easily

misinterpreted. Such issues require a careful examination of dyad-specific biases, but as mentioned previously, such an analysis is beyond the scope of the present study.

Procedures

This section describes the manner by which teacher ratings were collected in the CHP-C study, how information regarding the teachers was collected, and how the data were prepared for analysis. As explained previously, much of the present study utilized archived data that were collected in the CHP-C study; however, additional data were collected from teachers to address the second aim of the study. This section describes those procedures in detail.

Teacher Ratings

In the CHP-C study, rating scale data from the DBD and IRS were collected from each participant's four core course teachers (i.e., reading, math, social studies, and science) at the end of each month of the program (September through May, with the exception of December). The grant that supported the CHP-C research provided funds to remunerate teacher participants at the end of each school year for completing these ratings. Teachers earned $50 for each student they rated and ultimately

received anywhere from $50 to $600 at the end of each year, based on their level of involvement. While this reimbursement adequately recognized and rewarded the efforts of teacher raters, it was not considered enough to artificially influence response style or to motivate teachers who would otherwise be averse to providing this information.

In the first year of the CHP-C study (school year 2003-2004) there was high return rates for teacher ratings, with no less than 89% of all ratings returned each month. On most occasions, 100% of the monthly ratings were returned. Evans and colleagues (2007) attribute this high return rate to the reimbursement and a dual-system of data collection, which included an on-line version of the rating scales that are tied to a password- and identification-protected database. Each month, teacher participants received reminder emails when ratings were due and, when ratings were not completed online, this system was followed-up with a paper-and-pencil version of the rating scales that was delivered and collected at central locations at each site.

For both aims of the present study, a subset of the teacher ratings from the CHP-C study was selected for analysis. Specifically, the researcher selected teacher

ratings from the spring semester of 2005, collected at the end of each month from February to May. This timeframe was selected for three reasons. First, students often do not exhibit academic impairments until the spring semester of each academic school year (Fallah, Buvinger, Evans, Schultz, & Serpell, 2006). Thus, the time from the beginning of the school year to February is likely to allow teachers adequate opportunities to observe ADHD-related symptoms and impairments and to make accurate assessments of their students. Second, it appears that between-teacher reliability rates vary as a function of time, with large monthly fluctuations in the fall and stabilization of reliability rates in the spring. In a study of between-teacher reliability, Evans, Allen, Moore, and Strauss (2005) found that the highest agreement coefficients were observed very early in the school year (September; Intraclass Correlation [ICC] = .70), but then coefficients quickly dropped to their lowest levels in November and December (ICC = .26 and .28, respectively). By February, between-teacher reliability began to recover (ICC = .44) and to improve slightly until the end of the school year. Between-teacher reliability in February was closest to the average consistency rates over the entire school year

(*M* = .42). Thus, it appears that teacher ratings in February and thereafter are likely to include an average amount of error variance. And third, this timeframe represents the peak enrollment of student participants in the CHP-C study, at a time when both cohorts were actively involved, thereby providing more potential data for analysis and improving statistical power.

Teacher Characteristics and Experiences

For the second aim of the study (sources of rater bias), the researcher needed additional information regarding teacher characteristics and experiences. To gather these data, the researcher created the Teacher Questionnaire (described in detail above in the Instrument section). The researcher delivered these questionnaires in individual envelopes to the respective schools, using the interoffice mail systems at each site, beginning in December 2005. Each Teacher Questionnaire included a consent statement that made it clear to teachers that by returning the questionnaire, they were consenting to participate in the present study (see Appendix D for details). Teachers were asked to return the questionnaires either through special mailboxes set up at each site for the CHP-C study, or by direct mailing back to the researcher at James Madison University.

After six weeks, reminder emails were sent and a second follow-up interoffice mailing was delivered to all teachers who had not returned the questionnaires. The researcher continued to collect questionnaires from teachers until the spring of 2006, as several teachers returned the forms late and one teacher asked the researcher for another copy when she realized she had not returned the previous mailings.

Power and Sample Size

The sample size in this study was predetermined by the design of the larger CHP-C study. Seventy-nine students were enrolled in the CHP-C study, with each student receiving up to four monthly behavior ratings, from up to four teachers, depending on class schedule and returns. As a result, the total number of observations (ratings) that were potentially available was 1145. Using the participant sample size and the total number of observations, power analysis was conducted to determine the likelihood of discovering significant findings in both the analysis of the strength of interrater reliability and sources of rater bias, respectively.

To assess interrater reliability, average score ICCs were computed (see "Statistical Analysis" section for details). According to Cohen (1991), the power

attributable to correlations is based on the number of paired observations, significance criterion, and effect size. In the present analysis, ICCs were computed for teacher ratings of the 79 student participants in the CHP-C study. The significance criterion was set at .05 and effect size, which is equivalent to population r, was set to .50, based on the previous study conducted by Molina and colleagues (1998). Using these parameters, the power values exceeded .995. For smaller correlations, such as .40, .30, and .20, the power values were .96, .78, and .43, respectively (Cohen, 1991, p. 93). This meant that with a sample size of 79 students, the present study was very likely to detect correlations of .50 or greater, if they existed.

To assess sources of rater bias, hierarchical multiple regression analyses were conducted (See "Statistical Analysis" section for details). According to Cohen (1991), the power attributable to multiple regression models is based on the noncentral F distribution (λ) significance criterion, degrees of freedom of the numerator of the F ratio, and the degrees of freedom of the denominator of the F ratio. However, the researcher anticipated that the number of predictors would need to be adjusted to account for potential

instances of multicollinearity (i.e., bivariate correlations exceeding .90 among the IVs). Thus, power analysis was conducted for models with two, three, four, five, six, seven, and eight independent variables. With 234 potential observations and a significance criterion of .05, the power values all exceeded .99 in detecting a small (.15) effect sizes (Cohen, 1991, p. 421). Thus, it appears that the current analysis has acceptable power to reject the null hypothesis should small effects exist in the overall regression model.

When assessing the statistical power of a regression model it is also important to assess the relative power for each predictor variable. Given the lack of research regarding rater bias among teachers, it is difficult to anticipate the effect sizes of individual independent variables (IVs) in the present study; however, using Cohen's (1988) description of effect sizes as a guide, an effect of .10 represents a very small effect. With 234 teachers, an alpha of .05, an individual IV effect size of .10, and an overall model effect size of .20, the statistical power for each unique IV is .88. Thus, there is an acceptable probability that the null hypothesis would be rejected for each predictor should a very small effect be present in the data.

Statistical Analyses

Since the aims of this study were two-fold, two separate statistical analyses were conducted. Prior to the analyses, all data were screened for obvious data entry errors or outliers, and the basic assumptions of the statistical analyses were checked. Then, the two primary research questions were addressed. The research questions, as well as the hypotheses, variables, statistical procedures, and assumptions, are summarized in Table 4.

Data Screening

All data were screened prior to analysis to uncover possible data entry errors or outliers. Histograms and boxplots were examined for all ordinal to continuous variables using the Statistical Package for the Social Sciences (SPSS) Graduate Pack 15.0 for Windows®. In addition, descriptive statistics such as frequencies, skewness, and kurtosis were examined for all ordinal to continuous variables, and dichotomous variables were examined with the use of descriptive statistics and bar graphs (Field, 2005).

It is important to note that teacher behavior ratings collected within the CHP-C study technically produce ordinal data because the intervals between

Table 4

Research Questions, Hypotheses, Variables, Statistical Analyses, and Statistical Assumptions

Research Questions	Hypotheses	Variables	Statistic	Assumptions
1. To what extent do teachers agree on behavior ratings of adolescents with ADHD?	Teacher ratings will be inconsis-tent(ICC < .54)	Teacher ratings on three subscales of the DBD and two subscales of the IRS	Intraclass correlation (one-way model)	1.Continuous data 2.Normal distributions 3.Linear relationships
2. Are discrepancies between teacher ratings (error variance) associated with source characteristics?	Yes, but no specific hypothesis is tenable	IV: Teacher sex, life experience, training, classroom experience, disabilities experience, parental experience, and workload DV: Deviation scores on DBD and IRS subscales	Hierarchical Multiple Regression	1.Normal distributions 2.Linear relationships 3.Independent observations 4.Constant error variance (homoscedasticity)

response options on the Likert-type scales are not uniformly spaced. For example, the difference between *Often* and *Very often* on the DBD is subjective and likely to be interpreted differently among various raters. However, for the purposes of the present study, these data were treated as continuous variables rather than ordinal variables in all subsequent analyses. Despite this technical inaccuracy, such practices are common in research using rating scale data (e.g., Molina et al., 1998) because the assumption is that rater responses will generally fall along the continuum from low scores to high scores in continuous manner.

Research Question One

The first research question examined interrater reliability in teacher ratings of ADHD symptoms and impairment. The researcher hypothesized that the consistency between teacher raters would be low to moderate, as measured by intraclass correlations (ICCs) within the range of .21 to .52, based on the findings of previous research (e.g., Molina et al., 1998).

Intraclass correlations utilize analysis of variance (ANOVA) methods to provide a ratio of between-group variance to total variance in instances where variables share common metrics and variances (Molina et al., 1998),

and can be thought of as an average correlation among a group of correlations. Using the terminology of the present study, between-group variance equated to between-*target* variance (i.e., disagreements between raters), and total variance was the between-target variance combined with within-target variance. Similar to correlations, ICCs range from 0.0 for perfect inconsistency (i.e., within-target variance equals between-target variance) to 1.0 for perfect consistency (all variance is attributable to between-target variance). In instances where all variance is attributable to within-target variance, ICCs will be -1.00.

While there are several types of ICCs, the basic approach (one-way) is found in Formula 1, where *MS* = Mean Squares derived from ANOVA and *k* = the number of targets. Formula 1 is the statistical method for computing a basic ICC.

Formula 1:
$$\frac{MS_{\text{Between Ss}} - MS_{\text{Within Ss}}}{MS_{\text{Between Ss}} + (k - 1)MS_{\text{Within Ss}}}$$

However, the complex design of the CHP-C study (see Figures 1 and 2) precluded such straightforward approaches to ICC (McGraw & Wong, 1996; Shrout & Fleiss, 1979). In their study of between-teacher reliability,

Molina and colleagues (1998) encountered similar complications and, in response, used a one-way random effects model, consistent with Case 1 as described by Shrout and Fleiss (1979), with a correction for varying number of raters per target provided by Bartko and Carpenter (1976). In Case 1 ICC, the effects due to targets, raters, and the interaction effect between targets and raters are inseparable because each target is rated by different groups of raters.

Case 1 ICCs also apply to instances where raters are selected from a larger group of raters (i.e., population) who are exchangeable with the raters under investigation. As a result, the model is considered random because the raters are, in effect, randomly selected from a larger population of equally acceptable raters. Other ICCs (e.g., Case 2 and Case 3) are used in instances where the underlying design is a fully-crossed, complete box design, or when the raters are not sampled from a larger population of raters (Shrout & Fleiss, 1979). The design of the CHP-C study was consistent with Case 1, as the raters in the sample are theoretically exchangeable with other teacher raters, and teacher observations of student targets were not fully crossed. Molina and colleagues (1998) came to the same conclusion using a similar

dataset of teacher ratings. However, the partially nested design of the CHP-C study dataset meant that, from an analysis standpoint, the variances due to rater biases and target behaviors were completely confounded (Brennan, 2001), as depicted in Figure 2.

Another complication arose due to missing data, which resulted in an unbalanced measurement design where targets were rated by two, three, or four raters in a single measurement occasion. The ANOVA correction provided by Bartko and Carpenter (1976) allows for varying numbers of raters by adjusting the degrees of freedom for the within subjects variance. The formula for this ICC is provided in Formula 2, where i = target, M = total number of ratings, and n = number of ratings per target (p. 316).

Formula 2:

$$\frac{MS_{\text{Between Ss}} - mMS_{\text{Within Ss}}}{[MS_{\text{Between Ss}} + m(R_o - 1)MS_{\text{Within Ss}}]}$$

where $R_o = [M - \sum_{i=1}^{N} n_i^2/M]/(N - 1)$

and $m = N(R_o - 1)/[N(R_o - 1)MS_{\text{Within Ss}}$

Using this formula, the researcher computed ICCs for each of the five measurements of interest for each of the

months in the timeframe. The variables used to test the hypotheses related to the first research question included the teacher responses to the inattention subscale of the DBD, the hyperactivity-impulsivity subscale of the DBD, the total score on the DBD, the raw score on the academic impairment item on the IRS, and the raw score on the overall impairment item of the IRS. The researcher selected the teacher responses to these items collected during the spring semester of 2005 in the archived CHP-C study dataset for the reasons described previously.

Research Question Two

The second research question addressed by this study involved the degree to which teacher characteristics predicted rater bias. Given the lack of research on rater bias in the existing literature, specific hypotheses regarding sources of rater bias were untenable. However, using the little research available specific to rater bias, related research, and experience-driven hypotheses, the researcher selected several potential variables to assess, including teacher sex, life experience (age), training, classroom experience, experience with student disabilities, parental experience, and workload.

Based on the recommendations of Hill and colleagues (1988) and Hoyt (2002), a multiple regression (MR) analysis was used to assess the strength of potential predictors for the variance between the raters. As mentioned in the introduction, the Hill study was largely unsuccessful using this technique, due partly to high interrater reliability prior to analysis. However, in the present study, it was anticipated that there would be low to moderate interrater reliability, thus providing ample variance for the regression analysis.

When comparing ratings from multiple informants, several options are available to quantify the discrepancies. For example, Reyes and Kazdin (2004) have recommended standardized difference scores. To compute the standardized differences between ratings, each individual rating is converted into a *z* score, relative to each rater's set of ratings. In this manner, ratings are assigned standardized scores relative to the set of ratings submitted by each rater. Converted ratings can then be subtracted from other converted ratings, producing standardized differences. However, this method effectively summarizes the data and, in the process, much of the original variance is lost. For example, if four ratings were offered, three of which were the same and a

fourth rating that was higher, the *z* scores for the first three ratings would be -0.5 and the high score would be 1.5, no matter the actual raw score difference. In the present analysis, standardized deviation scores were thought to potentially have the effect of ignoring the true impact of rater-specific biases, especially when comparing across raters, as was planned in the present study.

Given the potential problems with standardized deviation scores, the researcher opted to subtract each observed rating from the average rating, per target per occasion, to derive an unstandardized deviation score. In other words, between-rater variance was defined as the raw score differences between each rater's rating and the average rating. The unstandardized deviation scores were then used as the dependent variables in a regression analysis. An identical method was used by Hoyt (2002) and Hill and colleagues (1988) in their analyses of observer bias in ratings of psychotherapy sessions.

Five separate multiple regression analyses were planned; the first three examined teacher responses to the three subscales of the DBD and the remaining two examined teacher responses on the academic and overall impairment items of the IRS. The researcher sought to

enter predictors into the analyses hierarchically in two blocks, beginning with ancillary teacher characteristics (sex and age), followed by another block consisting of personal experience (parenting experience), professional experience (training, classroom experience), experience with student disabilities), and teacher workload (students per class multiplied by classes per day).

Summary

This chapter described the process by which the two research questions were addressed in the present study. Specifically, the researcher examined interrater reliability among middle school teachers' ratings of adolescents with ADHD and potential sources of teacher bias in those ratings. In both questions, the analyses were static group comparisons that speak to relationships that were occurring within this sample only. As a result, there can be no assumptions of cause and effect.

The participants in the study included 79 middle school students with ADHD and 234 middle school teachers, all of whom participated in the CHP-C study as it was conducted in five schools in Augusta and Rockingham Counties in Virginia. The student sample was comprised mostly of Caucasian children from families that, in most cases, earned less than $60,001. Diagnoses of ADHD were

confirmed through clinical assessment and diagnostic conferences between a university-based school psychologist and clinical child psychologist. At the start of the present study, less was known about the teacher participants than the child participants in the CHP-C study.

To address the first research question (between-teacher reliability), the researcher examined middle school teachers' consistency on five separate ratings of ADHD symptoms and impairment by selecting ratings collected from 108 teachers during the spring semester of 2005 in the CHP-C study. The researcher chose this timeframe because it allowed teachers enough time to observe student behavior, consistency between teacher ratings were likely to range near average for the year (based on previous research), and this was the point of the CHP-C study of highest enrollment among student participants. It was hypothesized that when these ratings were analyzed, the ICCs among teacher ratings on all measures would fall below .52, based on previous research.

To address the second research question (sources of rater bias), the researcher gathered information regarding the demographic and personal experiences of the

teachers who participated in the CHP-C study. Since there was little research on sources of rater bias to guide this process, the researcher chose teacher characteristics that were preliminarily explored as potential sources for bias in the existing literature, hypothesized to be related to bias in the existing literature, or based on the researcher's experience-based hypotheses. From this, the researcher created a questionnaire that was sent to teachers in the CHP-C study beginning in the fall of 2005. The items on the questionnaire were designed to assess several areas, including teacher life experience (i.e., age), training, classroom experience, experience with student disabilities, parental experience, and workload. Teachers' randomly-selected target ratings (subtracted from the student- and occasion-specific average teacher rating) were regressed onto teacher responses to the questionnaire. By subtracting the target rating from the average, an unstandardized deviation score was produced. Next, the researcher constructed and tested a hierarchical multiple regression model, whereby ancillary teacher characteristics (sex and age) were entered first, followed by experience (training, classroom experience, parenting experience, and experience with student

disabilities), and variables related to teacher workload (students per class multiplied by classes per day).

CHAPTER IV

RESULTS

Introduction

This chapter describes the results of the present study, which aimed to assess the consistency among teacher ratings of ADHD-related symptoms and impairment among middle school students, as well as potential sources of rater bias. Specifically, this chapter will focus on the complications that arose during the study, the computer programs used to analyze the data, and the results of the statistical analyses.

Complications

During the course of the present study, several complications were encountered. First, complications arose from the attrition rates in the CHP-C study, which was expected because attrition is endemic to longitudinal studies. In the CHP-C study, student participants started and left the program at varying times and, as a result, teacher ratings for 9 of the 79 student participants were not collected during the targeted timeframe between February and May 2005. Thus, the total number of student targets rated during the timeframe was 70. Within this subsample of 70 students, some were not actively enrolled in the program for all four months of

the timeframe. As a result, only one month of data were available for three students (4.29%), only two months of data were available for one student (1.43%), and only three months of data were available for eight students (11.43%). Data for all four months were available for the remaining 58 student participants (82.86%). However, there were imperfect return rates for teacher ratings. With 70 students actively enrolled for the entire targeted timeframe and adjusting for the 12 students who were not enrolled all four months, a maximum of 1044 teacher ratings were expected. However, only 930 teacher ratings were collected, for a return rate of 89.08%. This limited the number of observations available for analysis in the present study.

Second, only 108 of the 234 (46.15%) teachers participating in the CHP-C study provided ratings during the targeted timeframe, due to the shifting enrollment of student participants and missing data. Further, the researcher was unable to gather Teacher Questionnaire results from the full set of 108 teachers who provided target ratings during the spring 2005 timeframe. Despite two separate mailings and an extended collection period, the researcher received only 80 returned questionnaires. Of these, one respondent (1.25%) did not include

identifying information to match to the monthly target ratings in the CHP-C study, one respondent (1.25%) submitted a repeated return (i.e., a second return from a teacher who had already responded to the previous mailing), and two respondents (2.50%) had not submitted target ratings during the timeframe. After removing these four problematic responses, there were 76 usable Teacher Questionnaires, for a return rate of 70.37% of available teachers. Again, this limited the number of observations available for analysis in the present study.

In response to missing teacher data, the researcher recomputed the power analysis for the planned hierarchical multiple regression model and found that with 76 teacher participants, two blocks of predictors (sex and age in block one, the remaining predictors in block two), alpha at .05, and the incremental change to the squared multiple correlation (R^2) set at 0.10 for each block (i.e., small effect size), the estimated overall model power for four, five, six, seven, and eight predictor models were .94, .92, .89, .87, and .85, respectively. However, the power associated with each increment was substantially reduced from the original projections. For four predictors, the power estimates were .76 for each block. For five predictors, the

estimates were .76 and .70 for blocks one and two respectively. For six predictors, the estimates were .75 and .64, respectively. For seven predictors, the estimates were .74 and .60, respectively. For eight predictors, the estimates were .74 and .56, respectively. Based on the changes to the power estimates, the researcher decided to select no more than five predictors from the available list to create a parsimonious model that maintained power of .70 or greater for each block in the design. Details of the final design are provided below in the Analysis section.

Third, Teacher Questionnaires were not ready for dissemination until the fall of 2005, thus producing a lag between the timeframe targeted in the current study and the time that questionnaires were answered. While it is uncertain how many teacher participants in the CHP-C study were still available to respond to the Teacher Questionnaire in the fall (i.e., it was unclear how many teachers had retired or moved within their respective school districts), it was not anticipated that this delay would result in an appreciable drop in questionnaire returns. As mentioned above, the final return rate was 70.37%. Still, it is possible that teacher responses to the Teacher Questionnaire were less accurate as a result,

due to inaccurate memories of conditions during the targeted timeframe, such as average class size, number of classes per day, and how experienced the teacher was at that time with students with disabilities. Despite the potential complications, the researcher continued to accept late returns of the questionnaire until the spring of 2006 to ensure the highest return rate possible.

Computer Programs

Data checking and analysis were conducted with database, spreadsheet, and statistical analysis software. At the start, the teacher rating data from the CHP-C study was collected in a Microsoft Access® database. The researcher conducted initial data checks using electronic queries to scan for duplicate records, and then selected just the records that were returned during the spring of 2005 (February through May). The researcher then used Microsoft Access 2002® to count the number of records associated with teachers, targets, and measurement occasions, and to identify cases of missing data. Data from the Teacher Questionnaire was also entered in this same database, in a separate table.

The teacher ratings data were then exported to Microsoft Excel 2002® in preparation for the analysis to address the first aim of the study. Given the complex

design of the CHP-C study dataset, the analysis of between-teacher reliability could not be analyzed with straightforward statistical procedures commonly found in commercial statistical software. Rather, the researcher used the Data Analysis feature in Microsoft Excel 2002® to compute one-way analysis of variance (ANOVA) of teacher ratings and then used spreadsheet functions to compute the formula for an adjusted intraclass correlation (ICC) provided by Bartko and Carpenter (1976), described in the Methods chapter (see Formula 2).

For the second aim of the study, the researcher used Microsoft Access 2002® to compute the teacher rating deviation scores used as dependent variables prior to statistical analysis. To this end, the researcher programmed a crosstab query to store average monthly teacher ratings for each student in a table for each independent variable, including the inattention subscale of the CHP-C version of the Disruptive Behavior Disorders scale (DBD), the hyperactivity-impulsivity subscale of the DBD, the total score of the DBD, the academic impairment item of the Impairment Rating Scale (IRS), and the overall impairment item of the IRS. Next, unstandardized deviation scores were computed for every teacher rating electronically by another query that

referenced the monthly averages per child and then subtracting the raw score. These data were then stored alongside the raw teacher ratings in the database.

The researcher then used Microsoft Access 2002® to randomly select the teacher ratings that were used as dependent variables in the second aim of the present study. Specifically, the researcher composed an Access Visual Basic 2002® script that randomly selected one record per teacher, where the target and measurement occasion (combined) were not duplicated with any other selected record. By selecting records in this fashion, the researcher ensured independence of observations by avoiding the use of two or more records where deviation scores were computed using the same average. The script used to achieve this random selection is provided in Appendix E. Once the data table was constructed and an independent rating for each teacher who returned a usable Teacher Questionnaire was randomly selected, these data were also exported for analysis in statistical software. Multiple regression analyses were conducted using the SPSS Graduate Pack 15.0 for Windows®.

Analysis

There were two primary aims of the present study. The first aim was to assess interrater reliability among

teacher ratings of middle school students with ADHD. The researcher was interested in assessing the overall consistency of ratings among groups of up to four teachers rating the same target during a specific measurement occasion, as well as the correlation among teacher dyads within those groups. The second aim of the study was to assess potential sources of rater bias by regressing the discrepancies between teacher ratings onto indices of teacher characteristics and experiences. This section will focus on the outcomes of these analyses separately.

Between-Teacher Reliability

For the first aim of the study, the researcher assessed consistency among teacher ratings of middle school students with ADHD using intraclass correlations (ICCs), which measure "the proportion of a variance (variously defined) that is attributable to objects of measurement" (McGraw & Wong, 1996, p. 30). The researcher hypothesized that ICCs for interrater reliability in the CHP-C study would fall below .52, based on similar research in the existing literature. Previous research has focused on ADHD symptoms (e.g., Molina et al., 1998), but in the present study the researcher examined five separate ratings collected from

all of the teacher participants, including the inattention subscale of the DBD, the hyperactivity-impulsivity subscale of the DBD, the total score on the DBD, the academic impairment item of the IRS, and the overall impairment item of the IRS. Descriptive statistics for these variables are provided in Table 5.

Using the formula provided by Bartko and Carpenter (1976), the researcher computed ICCs for each of the measures examined in the present study, for each of the months in the targeted timeframe of spring 2005 where at least two ratings per target were available. The results of these analyses are summarized in Table 6. ICCs for all five measures across all four months were statistically significant ($p < .001$), based on the F-ratio from the underlying ANOVA, and ICC values ranged from 0.45 to 0.59. Consistent with previous research (e.g., Evans et al., 2005), the ICCs from February to May remained appreciably stable. Of the five measures, consistency among teacher ratings of inattention on the DBD appeared highly stable, with ICCs ranging from 0.51 to 0.55. The most variance over time was seen among teacher ratings of academic impairment on the IRS, where ICCs ranged from 0.52 to 0.59.

Table 5

Descriptive Statistics for the Five Teacher Ratings Provided in the CHP-C Study: February to May 2005

	M	SD	Min	Max	Skew	Kurt
DBD Subscales						
Inattention	12.99	7.72	0	27	0.1	-1.0
Hyper-Impulsivity	7.75	6.87	0	27	0.8	-0.1
Total Score	20.74	13.38	0	56	0.4	-0.6
IRS Items						
Academic Impairment	3.07	2.20	0	6	-0.1	-1.4
Overall Impairment	2.92	2.20	0	6	0.0	-1.5

Table 6

Intraclass Correlations for Teacher Ratings of ADHD Symptoms and Impairment

	Feb (*n*=62)	March (*n*=58)	April (*n*=66)	May (*n*=60)
DBD Subscales				
Inattention	0.52	0.55	0.55	0.55
Hyperactivity-Impulsivity	0.48	0.46	0.48	0.52
Total Score	0.51	0.53	0.55	0.56
IRS Items				
Academic Impairment	0.52	0.58	0.56	0.59
Overall Impairment	0.50	0.45	0.53	0.50

Note. *n* was limited to targets with two or more ratings in the respective month. All correlations were significant ($p < .001$).

The researcher also examined correlations between teacher pairs to assess the degree of consistency between any two ratings of the same target. To this end, the researcher randomly selected two ratings for each month for each target in the available dataset and computed Pearson correlations. Molina and colleagues (1998) used the same strategy in their study of between-teacher reliability to examine consistency between teacher dyads and found that the correlations were virtually identical to average correlations for all possible dyads, with results generally falling within the .40 to .50 range. Given the consistency of correlations between randomly selected dyads and all teacher combinations in the previous study, the researcher examined correlations from randomly selected teacher dyads only in the present study. The results of this analysis are summarized in Table 7.

All pairwise Pearson correlations were statistically significant ($ps < .01$) and ranged from .42 to .74. Among the measures examined, Pearson correlations appeared to vary most across measurement occasions for the hyperactivity-impulsivity subscale of the DBD. On this measure, correlations ranged from .42 to .74, which was the maximum range for any two correlations on any

Table 7

Pearson Correlations for Pairwise Ratings of ADHD Symptoms and Impairment

	Feb (*n*=62)	March (*n*=58)	April (*n*=66)	May (*n*=60)
DBD Subscales				
Inattention	.73	.65	.70	.65
Hyperactivity-Impulsivity	.55	.42	.66	.74
Total Score	.66	.55	.72	.71
IRS Items				
Academic Impairment	.52	.61	.57	.63
Overall Impairment	.53	.56	.65	.58

Note. *n* was limited to targets with two or more ratings in the respective month. All correlations were significant ($p < .01$).

measure. In contrast, the correlations for the inattention subscale of the DBD ranged from .65 to .73, suggesting that during the timeframe from February to May, consistency between pairs of teacher ratings were appreciably stable, relative to other measures.

Sources of Teacher Bias

In the second aim of the study, the researcher examined teacher bias. Teacher bias was assessed by regressing unstandardized deviation scores of teacher ratings (relative to the mean for all teachers) onto indices of teacher characteristics and experiences. Given the lack of research on sources of rater bias, specific hypotheses were untenable. The researcher used hierarchical multiple regression to address this component of the study.

As mentioned above in the Complications section, the researcher encountered issues of attrition in the archived CHP-C study dataset, as well as a 70.37% return rate on the Teacher Questionnaires that supplied the predictor variables used in this analysis. As a result, the statistical power was substantially reduced from initial projections and, in response, the researcher chose a smaller set of predictors than originally planned to create the final model. Based on power analysis, the

researcher targeted a model with no more than five predictors, entered hierarchically in no more than two blocks, to provide an estimated power of .70 or better for each increment (see Complications section above for details).

Construction of the Regression Model

To arrive at a more parsimonious model of five predictors or less, the researcher began by examining the raw data provided by the Teacher Questionnaire to look for obvious item misinterpretations or instrument errors. An overview of the continuous data collected by the Teacher Questionnaire is provided in Table 8 and an overview of the dichotomous and categorical data collected by the Teacher Questionnaire is provided in Table 9. It is clear from the descriptive statistics that the item designed to assess the subjects teachers were trained to teach resulted in problematic data, as the vast majority (63.2%) of respondents indicated a category that the researcher had not anticipated. For example, several teachers responded that their training was for "K-12," or "Secondary Education." Clearly, the responses to this item were unusable for the original intent (to contrast with the class the target was enrolled in), so this variable was removed as a candidate

Table 8

Descriptive Statistics for Continuous Data Items of the Teacher Questionnaire

Variable	M	SD	Min	Max	Skew	Kurt
Teacher Age	42.95	10.96	24	60	-0.4	-0.9
Highest Degree	16.64	1.04	16	19	*1.1	-0.4
Years Taught	15.16	10.09	1	32	0.2	**-1.5
Disabilities Exp.	3.31	0.76	1	5	-0.3	0.5
Students per Class	21.67	4.75	6	28	*-1.6	***2.8
Classes per Day	4.04	1.07	1	7	0.1	1.0

Note. * = significant positive skew; ** = significant platykurtosis; *** = significant leptokurtosis

Table 9

Descriptive Statistics for Dichotomous and Categorical Data Items of the Teacher Questionnaire

Variable	*n*	Percentage
Teacher Sex		
Male	19	25.0
Female	57	75.0
Parental Status		
Yes	53	69.7
No	23	30.3
Child Middle School Age		
Yes	41	53.9
Subject Trained to Teach		
English/Reading	7	9.2
Social Studies	8	10.5
Math	2	2.6
Science	7	9.2
Special Education	4	5.3
Other	48	63.2

Note. Teacher sex was determined based on the researcher's personal knowledge of the teachers and was not included as an item on the questionnaire.

for the final regression model.

Similarly, the item related to the highest degree attained (HD) resulted in a significantly positively skewed distribution because 54 of the respondents (71.1%) reported to have earned a Bachelor's Degree, 17 reported to have earned a Master's Degree (22.4%), and 5 respondents reported to have earned a Master's Degree plus certification (6.6%). Thus, this variable did not appear to include enough variability at the higher levels of the range because the sample did not include many teachers with advanced degrees. Hence, the researcher dropped this variable from the list of predictor candidates because there was no easily interpretable way to correct the skew or to combine these data with another predictor.

The researcher then constructed a correlation matrix of the remaining predictor candidates to assess the likelihood of multicollinearity (redundancy) in the eventual regression model. The correlation matrix is reproduced in Table 10. Several predictor variables were found to be significantly correlated with one another; however, for the purpose of selecting variables to avoid multicollinearity in the regression model, Tabachnick and

Table 10

Correlation Matrix for the Potential Independent (Predictor) and Dependent Variables

	S	TA	PS	OC	YT	DE	NS	NC	DV1	DV2	DV3	DV4	DV5
Sex (S)	1.00												
Teacher Age (TA)	.20	1.00											
Parental Status (PS)	.22	.56	1.00										
Older Child (OC)	.20	**.60	**.71	1.00									
Years Taught (YT)	*.27	**.73	**.41	**.51	1.00								
Disabilities Exp. (DE)	**.32	.17	.19	.22	**.31	1.00							
Number of Students (NS)	-.09	.12	.02	.09	.04	-.09	1.00						
Number of Classes (NC)	.02	**.33	.11	*.26	*.26	*.24	**.32	1.00					
Disc. on DBD-IA (DV1)	.17	.00	.14	.19	.01	.12	.19	.16	1.00				
Disc. on DBD-HI (DV2)	.19	-.15	.07	.04	-.19	.22	-.02	.20	**.53	1.00			
Disc. on DBD-Tot (DV3)	.21	-.08	.12	.13	-.10	.19	.10	.21	**.88	**.87	1.00		
Disc. on IRS Acad. (DV3)	.15	.05	.15	.21	-.01	.00	*.24	.12	**.66	**.40	**.61	1.00	
Disc. on IRS Total (DV4)	.14	-.02	.12	.20	-.01	.13	.20	.20	**.63	**.51	**.65	**.85	1.00

Note. For the IV Sex (S), 1 = man and 2 = woman. For yes/no items, 0 = no and 1 = yes.
* Correlation is significant at .05 level (two-tailed)
** Correlation is significant at .01 level (two-tailed)

Fidell (1996) suggest that correlations exceeding .70 should be carefully considered before entering both variables as separate predictors. The two items related to parental status exceeded this recommended threshold when examined using a Pearson correlation coefficient (r = .71, p < .01), which assumes an equal distribution, and a phi coefficient (Φ = .71, p < .001), which does not assume an equal distribution. Thus, the two variables would potentially result in multicollinearity if both were entered into the regression model. In response, the researcher chose to use only the first variable (i.e., parental status [PS] regardless of child age) because there was no way to logically combine these items into one easily interpretable predictor. Further, the PS variable seems more straightforward to interpret because the OC variable is dependent on the child's age, while the PS variable is dependent only on whether the teacher has children.

In addition, the variable for teacher age (TA) was correlated above the .70 with years taught (YT) (r = .73, p < .01). The researcher anticipated these variables would be correlated (see Figure 4), but the strength of this correlation presented potential problems related to multicollinearity in the regression model. In the

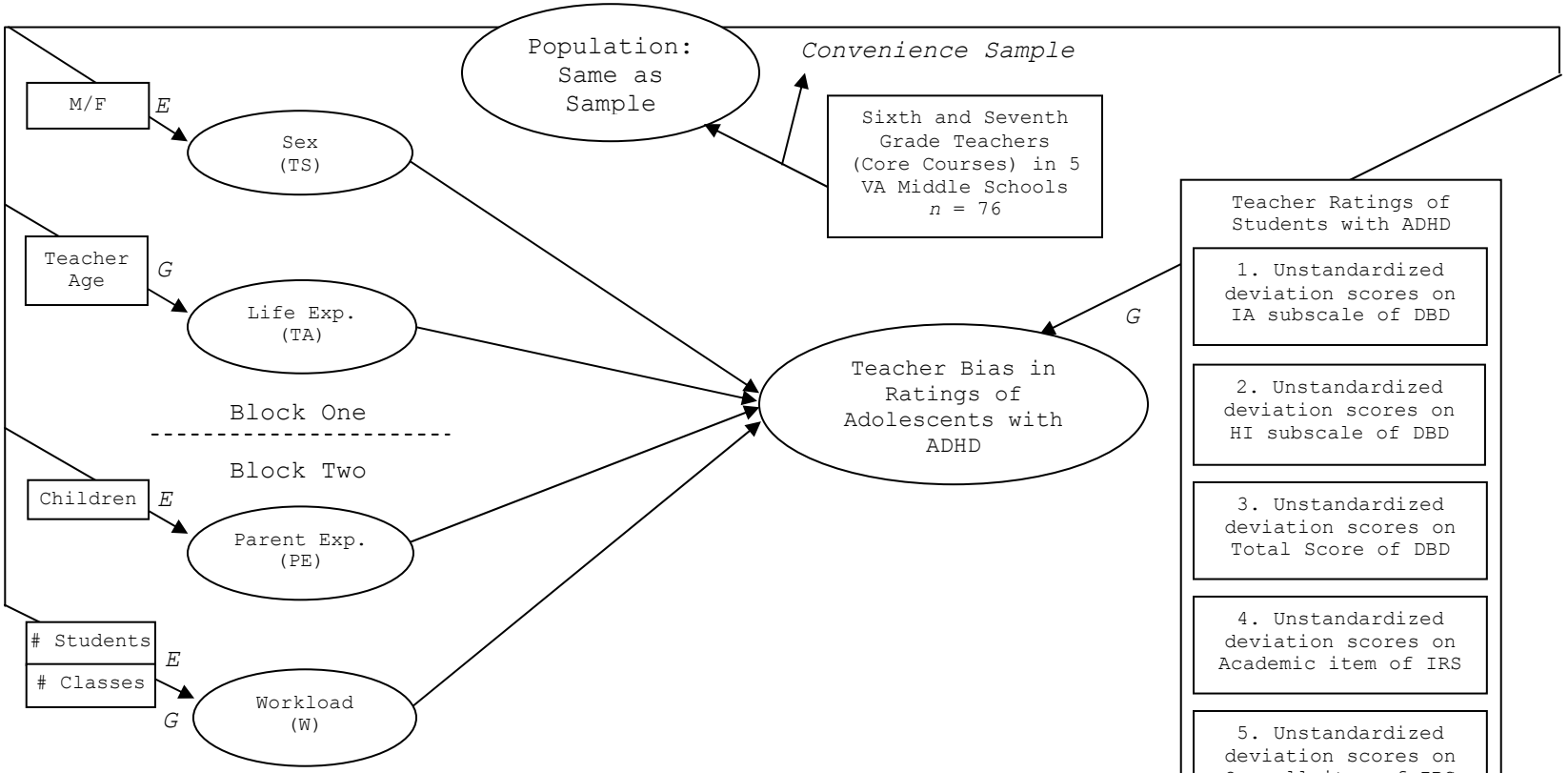
interest of creating a parsimonious model, the researcher chose to drop YT from the list of predictor candidates because YT would return identical data for a teacher who started a teaching career immediately after college and another who returned to teaching after a previous career. It is conceivable that teachers with such differing backgrounds would view childhood disruptive behaviors in different ways. Further, there was no logical way to combine both YT and TA into one easily interpretable predictor.

To further simplify the regression model, the researcher combined teacher responses to the average number of students per class (NS) and number of classes per day items (NC) by multiplying the variables. When multiplied, the data constituted a new variable, referred to as workload (W), as planned prior to the analysis. The workload variable (M = 89.24, SD = 35.21) was then assessed to see if correlations with the other predictor candidates exceeded the .70 threshold recommended by Tabachnick and Fidell (1996). Pearson correlations with the workload variable and all remaining predictor candidates ranged from r = -.16, p = .16 (HD) to r = .56, p < .001 (TA). No other correlations exceeded the recommended threshold.

To further assess potential problems with multicollinearity, the researcher regressed each remaining candidate predictor onto the other candidate predictors, as recommended by Kline (2005). As a rule of thumb, when the multiple R^2 exceeds .90 in any of these analyses potential problems with multicollinearity are likely. Although no multiple R^2 exceeded .90 (multiple R^2s ranged from .05 [TS] to .35 [TA]), it became apparent that teacher sex (TS) was a significant predictor (p = .01) of experience with student disabilities (DE) when the effects of other predictor candidates were held constant. Specifically, women reported significantly more experience with student disabilities than did men in the sample. The relationship between TS and DE was unanticipated and difficult to interpret without additional information. Hence, the researcher dropped the DE variable from the candidate predictors because the discrepancies between men and women respondents suggested either an instrument error, which resulted in underestimation among men and overestimation among women, or a possible sampling error, where women teachers with high levels of experience with student disabilities were overrepresented in the sample.

The changes made to the original regression model resulted in a four predictor model that included teacher sex (TS), teacher age (TA), parental status (PS), and workload (W). The researcher entered these predictors hierarchically in two blocks, with the ancillary variables, including teacher sex and age, in block one and the remaining two experience-based predictors, including parental status and workload, in block two. Figure 5 provides an overview of the final regression model.

The dependent variables included unstandardized deviation scores on the inattention subscales of the DBD, the hyperactivity-impulsivity subscale of the DBD, the total score on the DBD, the academic impairment item on the IRS, and the overall impairment item on the IRS. As described in the Methods section, the deviation scores were computed for each measurement occasion by subtracting individual teacher ratings from the average teacher rating. Then the researcher randomly selected one rating occasion for each teacher, without overlapping observations based on target and occasion (i.e., each observation was independent), and used these data as the dependent variable in the multiple regression analyses. Table 11 provides the descriptive statistics for the



Note. *E* = Excellent and *G* = Good estimated validity and reliability

Figure 5. Final regression model for the second aim of the present study: Teacher bias analysis.

Table 11

Descriptive Statistics for the Dependent Variables

Variable	M	SD	Min	Max	Skew	Kurt
DBD Subscales						
Inattention	0.4	4.3	-13.8	9.0	-0.5	0.5
Hyper.-Impulsivity	0.1	4.1	-10.3	9.3	-0.3	0.3
Total Score	0.5	7.4	-18.3	15.8	-0.3	0.0
IRS Items						
Academic Impair.	0.1	1.2	-3.3	2.8	-0.4	0.3
Overall Impair.	0.1	1.4	-3.3	2.8	-0.5	-0.3

Note. The academic impairment item of the IRS was missing from one of the randomly selected records; otherwise, n = 76.

dependent variables randomly selected from the dataset. A datum was missing from the academic impairment item of the IRS, thereby disallowing the calculation of a deviation score for that record. Otherwise, the data needed to prepare the dependent variables were available for all 76 teacher participants. Multiple regression analyses were conducted for each dependent variable using the final regression model.

Ratings of Inattention

In the first analysis, the researcher regressed unstandardized deviation scores from the inattention subscale of the DBD onto the indices of teacher characteristics and experiences in the regression model. The researcher assessed the robustness of this and all regression analyses in the present study by first examining the model fit statistics and casewise diagnostics. In terms of model parameters, the Durbin-Watson statistic (2.02) suggested that the assumption of independent errors was most likely met. Further, tolerance statistics, which ranged from .63 to .94, and the variance inflation factor (VIF), which ranged from 1.07 to 1.59, did not suggest any potential problems with multicollinearity among the predictors, based on rules of thumb provided in the literature (e.g., Field, 2005).

Specifically, tolerances did not fall below .20, no single VIF was greater than 10, and the average VIF was not substantially greater than 1.0. Thus, there were no apparent violations of the assumption of independent errors or multicollinearity.

In terms of casewise diagnostics, three standardized residuals fell outside of two standard deviations, with the greatest standard residual reaching -2.75. However, since it is reasonable to expect approximately 5% of any normally distributed sample to exceed these parameters, up to 3.8 cases would exceed these parameters for a sample of 76 teachers. Thus, the data appear to fit the model. In no instances did cases appear to have an undue influence on the model (Cook's distances < .14, Mahalanobis distances < 12.31), based on the rules of thumb provided by Pallant (2001), which state that Cook's distances should not exceed 1.0 and Mahalanobis distances should not exceed 18.47 for a four-predictor model. Similarly, the DFBeta statistics did not exceed the threshold 1.0 for any of the predictors (Field, 2005).

Finally, the researcher visually scanned the standardized residuals plotted against standardized predicted values and the distribution did not suggest problems with heteroscedasticity or non-linearity. A

histogram of the standardized residuals and the normal probability plot both suggested a normal distribution of residuals. None of the partial plots of the residuals for each predictor suggested the presence of outliers or problems related to heteroscedasticity. Given that the model fit statistics and casewise diagnostics suggested the data fit the model, the analysis appears to have met the statistical assumptions inherent in multiple regression analysis.

Table 12 summarizes the results of the hierarchical multiple regression analysis for teacher bias on the inattention subscale of the DBD. The second model (all predictors) explained an estimated 12.0% of the total variance in the dependent variable, based on the value of R^2 (.12), but the model did not represent a significant improvement in prediction over the mean alone (F = 2.42, p = .06). The adjusted R^2 adjusts for inflation in R^2 due to the number of predictors (Pallant, 2001). In the present analysis, the value of the adjusted R^2 for the second model (.07) was appreciably smaller than the value of R^2 and suggested that the full model only explained an estimated 7% of the error variance in teacher ratings. Thus, the model did not improve prediction and did not

Table 12

Hierarchical Multiple Regression Results for the Inattention Subscale of the DBD

	ΔR^2	*B*	*SE B*	*β*	*t*	*p*
Step 1	.03					
Constant		-2.16	2.58			
Teacher Sex		1.78	1.16	.18	1.54	.128
Teacher Age		-0.01	0.05	-.04	-0.29	.770
Step 2	*.09					
Constant		-3.37	2.69			
Teacher Sex		1.85	1.13	.19	1.63	.107
Teacher Age		-0.09	0.06	-.22	-1.58	.120
Parental Status		1.79	1.25	.19	1.43	.158
Workload		0.03	0.01	.27	2.34	.022

Note. Model R^2 = .12, F = 2.42, p = .06.

$^{*}p < .05$

Ratings of Hyperactivity-Impulsivity

Next, the researcher examined teacher bias on the hyperactivity-impulsivity subscale of the DBD using the same regression model. In terms of model parameters, the Durbin-Watson statistic (2.00) suggested that the assumption of independent errors was most likely met. Given that the same predictors were used in this analyses as used in the first analysis (see above), the tolerance statistics and the VIF were unchanged and did not suggest any potential problems with multicollinearity among the predictors, based on rules of thumb provided in the literature. Thus, there were no apparent violations of the assumption of independent errors or multicollinearity.

In terms of casewise diagnostics, four standardized residuals fell outside of two standard deviations, but the largest of these was -2.18. As explained above, it is reasonable to expect that up to 3.8 cases would exceed two standard deviations in a sample of 76 participants, so the residuals appear to fit the model. In no instances did cases appear to have an undue influence on the model (Cook's distances < .13, Mahalanobis distances < 12.32), and the DFBeta statistics did not exceed the threshold 1.0 for any of the predictors.

Visual scans of the standardized residuals plotted against standardized predicted values did not suggest problems with heteroscedasticity or non-linearity. Similarly, a histogram of the standardized residuals and the normal probability plot both suggested a normal distribution of residuals. All partial plots of the residuals of the outcome and each predictor did not suggest the presence of outliers or problems related to heteroscedasticity.

The results of this analysis are summarized in Table 13. The overall model appeared to significantly improve prediction of teacher bias above the mean alone (F = 2.85, p = .03), accounting for an estimated 13.8% of the total variance in teacher ratings. However, the adjusted R^2 was appreciably more conservative than the R^2 and suggested that the second model accounted for only 9.0% of the variance in teacher ratings.

The change in R^2 in each block did not represent significant incremental improvements in the prediction of teacher bias (ps > .05). In the first block, neither sex of the teacher (TS) or teacher age (TA) were statistically significant in predicting bias; however, when the experiential predictors were added in block two, both TS (β = .23, p =.048) and TA (β = -.37, p = .040)

Table 13

Hierarchical Multiple Regression Results for the Hyperactivity-Impulsivity Subscale of the DBD

	ΔR^2	*B*	*SE B*	*β*	*t*	*p*
Step 1	.07					
Constant		-0.53	2.41			
Teacher Sex		2.16	1.08	.23	2.00	.050
Teacher Age		-0.07	0.04	-.19	-1.69	.096
Step 2	.07					
Constant		-1.19	2.55			
Teacher Sex		2.15	1.07	.23	2.01	.048
Teacher Age		-0.14	0.05	-.37	-2.66	.010
Parental Status		1.83	1.19	.21	1.54	.127
Workload		0.03	0.01	.21	1.84	.070

Note. Model R^2 = .14, F = 2.85, p = .03.

were statistically significant. Specifically, it appeared that severe ratings were associated with younger teachers and women teachers, once the impact of the experiential predictors were taken into account. Based on the unstandardized beta values, the results suggest that women teachers provided more severe ratings with an average of 2.15 points higher than men on the hyperactivity-impulsivity subscale when the effect of other predictors were held constant. It also appeared that an increase of one standard deviation in teacher age (10.96 years) predicted .37 standard deviations, or 1.52 points, greater leniency on the hyperactivity-impulsivity subscale, when the effects of the other predictors were held constant.

The 95% confidence interval in *B* was 0.02 to 4.29 for TS, and -0.24 to -0.03 for TA, which suggests that in both instances the effect of these variables on teacher ratings of hyperactivity-impulsivity would most likely result in relationships trending toward rater severity among women and younger teachers, if the analysis were repeated and the experiential variables were included. However, the range of the confidence interval for TS suggests that this finding is less reliable because the

spread between the lower and upper limit is relatively wide.

Ratings of Overall ADHD

In the next analysis, the researcher examined teacher bias on total score ratings of ADHD symptoms. In terms of model parameters, the Durbin-Watson statistic (2.00) suggested that the assumption of independent errors was most likely met, and again, the tolerance statistics and the VIF did not suggest any potential problems with multicollinearity as in the previous analyses because the same predictors were used. Thus, there were no apparent violations of the assumption of independent errors or multicollinearity.

In terms of casewise diagnostics, three standardized residuals fell outside of two standard deviations. The largest of these was 2.59. However, as with the models for inattention and hyperactivity-impulsivity, a model with three residuals of this magnitude in a sample of 76 participants suggests that the data fit the model. In no instances did cases appear to have an undue influence (Cook's distances < .17, Mahalanobis distances < 12.32, DFBeta statistics < 1.0). As in the previous analyses, visual scans of plots, histograms, and partial plots suggested no problems with heteroscedasticity, non-

linearity, non-normal distribution of residuals, or outliers.

The results of the analysis are summarized in Table 14. When compared the accuracy of prediction using just the means, the overall model provided a significant improvement in prediction of teacher bias (F = 3.35, p = .01). It appeared that the addition of the experiential variables in block two significantly improved the predictive power of the model over the ancillary variables alone, and the total model was estimated to account for 15.9% of the variance in teacher ratings of ADHD. Again, however, the adjusted R^2 provided a more conservative estimate and suggested that approximately 11.1% of the variance in teacher ratings was explained by the predictors.

Several variables appeared to significantly contribute in the second block, including teacher age (TA; β = -.34, p = .02), workload (W; β = .28, p = .02), and teacher sex (TS; β = .24, p = .02). Specific to teacher age, the results suggest that for every increase of 10.96 years of teacher age, total score ADHD ratings were about 2.50 points more lenient on average, after the effects of all other predictors were removed. In terms of workload, the results suggest that for each increase

Table 14

Hierarchical Multiple Regression Results for the Total Score Subscale of the DBD

	ΔR^2	*B*	*SE B*	*β*	*t*	*p*
Step 1	.06					
Constant		-2.70	4.35			
Teacher Sex		3.94	1.95	.23	2.02	.047
Teacher Age		-0.09	0.08	-.13	-1.11	.271
Step 2	*.10					
Constant		-4.56	4.51			
Teacher Sex		3.99	1.90	.24	2.11	.039
Teacher Age		-0.23	0.09	-.34	-2.44	.017
Parental Status		3.62	2.10	.23	1.72	.089
Workload		0.06	0.02	.28	2.44	.017

Note. Model R^2 = .16, F = 3.35, p = .01.

*$p < .05$

of 35 students, total score ADHD ratings increased by 2.06 points, after the effects of all other predictors were removed. In terms of teacher sex, the unstandardized beta weights suggests that women rated the targets more severely than did men, providing ratings that were on average 3.99 points higher, after the effects from all other predictors were removed.

Taken together, it appears that young teachers, high workloads, and women teachers were likely to be associated with overall ADHD rating severity, after the effects of all other predictors were removed. Parental status did not significantly add to the prediction. The 95% confidence interval in *B* for TS ranged from 0.22 to 7.78, suggesting that the trend toward severity among women teachers would likely be repeated in other samples, but the spread between the upper and lower limit suggest that there would be great variability and overlap in ratings from men and women. The 95% confidence intervals for TA (ranged from -0.04 to -0.41) and W (ranged from .01 to .11) were smaller than that for TS, suggesting that the findings related to these variables were more reliable.

Ratings of Academic Impairment

In the next analysis, teacher bias in ratings of academic impairment on the IRS was examined. As mentioned previously, one teacher in the randomly selected sample failed to provide a response to this item, so this case was excluded from the analysis listwise, leaving a sample of 75 teachers rather than 76. The Durbin-Watson statistic (1.90) suggested that the assumption of independent errors was likely met. In this analysis, the same predictors were used as in previous analyses, which were found to have no apparent problems with multicollinearity. The listwise deletion of the one missing case did not have a meaningful impact on the tolerance statistics (ranged from .63 to .96) or the VIF (ranged from 1.04 to 1.60). Thus, the assumptions of independent errors and lack of multicollinearity appear to have been met.

In terms of casewise diagnostics, three standardized residuals fell outside of two standard deviations and the largest of these was -2.64. As with the models for ADHD symptoms, a model with three residuals of this magnitude with 76 participants is consistent with expectations for accuracy. In no instances did cases appear to have an undue influence (Cook’s distances < .12, Mahalanobis

distances < 12.40, DFBeta statistics < 1.0). As in the previous analyses, visual scans of plots, histograms, and partial plots suggested no problems with heteroscedasticity, non-linearity, non-normal distribution of residuals, or outliers.

The results of the subsequent analysis are summarized in Table 15. Neither block provided significant improvement in prediction, and the overall model only accounted for an estimated 8.0% of the variance in teacher ratings based on R^2. Based on the adjusted R^2, the model appeared to only account for 2.8% of the variance in teacher ratings. Not surprisingly, this does not suggest statistical improvement in prediction over prediction based on the mean alone (F = 1.53, p = .20). For all predictors, the lower bound of the 95% confidence interval was below zero and the upper bound was above zero, suggesting that if the model were retested with different samples, there would be no clear trends in either direction for the predictors. In sum, the predictors did not appear to be related to the observed error variance on the academic impairment scale.

Ratings of Overall Impairment

For the final analysis, the researcher regressed the deviation scores on the overall impairment item of the

Table 15

Hierarchical Multiple Regression Results for the Academic Impairment Item of the IRS

	ΔR^2	*B*	*SE B*	*β*	*t*	*p*
Step 1	.02					
Constant		-0.71	0.74			
Teacher Sex		0.42	0.33	.15	1.26	.213
Teacher Age		0.00	0.01	.02	0.15	.883
Step 2	.06					
Constant		-0.96	0.78			
Teacher Sex		0.42	0.32	.15	1.28	.204
Teacher Age		-0.02	0.02	-.14	-0.96	.341
Parental Status		0.42	0.36	.16	1.16	.251
Workload		0.01	0.00	.22	1.78	.079

Note. Model R^2 = .08, F = 1.53, p = .20.

IRS onto the four predictors in the regression model. In terms of model parameters, the Durbin Watson statistic (1.90) suggested that the assumption of independent errors was most likely met. Also, since the same predictors were used in this model as in the previous models, VIF statistics and tolerances continued to suggest that multicollinearity was not a problem in the model. Thus, the assumptions of independent errors and lack of multicollinearity appear to have been met.

In terms of casewise diagnostics, only one standardized residual fell outside two standard deviations, with a score of -2.45. Again, as in the previous analyses, this suggests that the data fit the model. A visual scan of the standardized residuals plotted against the predicted values did not suggest any problems with heteroscedasticity, curvilinear relationships, or outliers, and a histogram of the standardized residuals appeared to fit a normal distribution. Likewise, the partial plots for all of the predictors appeared consistent with the statistical assumptions of multiple regression.

The results of the regression analysis are summarized in Table 16. While the first block (teacher sex and age) did not significantly improve prediction,

Table 16

Hierarchical Multiple Regression Results for the Overall Impairment Item of the IRS

	ΔR^2	*B*	*SE B*	*β*	*t*	*p*
Step 1	.02					
Constant		-0.51	0.85			
Teacher Sex		0.47	0.38	.15	1.24	.221
Teacher Age		-0.01	0.02	-.04	-0.37	.711
Step 2	*.10					
Constant		-0.97	0.89			
Teacher Sex		0.51	0.37	.16	1.35	.180
Teacher Age		-0.03	0.02	-.23	-1.65	.104
Parental Status		0.56	0.41	.18	1.35	.181
Workload		0.01	0.01	.30	2.52	.014

Note. Model R^2 = .12, F = 2.36, p = .06.

*$p < .05$

the second block appeared to provide a significant incremental improvement over block one (change in R^2 = .10, p = .03); however, the final model did not appear to offer an improvement over predictions based on means alone (F = 2.36, p = .06). The second model (all predictors) are estimated to account for approximately 9.7% if the variance in ratings, based on the R^2, and approximately 6.8% of the variance based on the conservative adjusted R^2.

The relative improvement in prediction afforded by the second model appeared to be largely attributable to unique contribution of teacher workload (W; β = .30, p = .01), once the effects of the other predictors were removed statistically. The direction of this relationship was positive, suggesting that as teacher workload increased, ratings of overall impairment increased. Specifically, it appeared that when workload increased by 35 students, teacher ratings of overall target impairment increased by 0.42 points on average, when the effects of the other predictors were held constant.

Summary

This chapter described the results of the analyses conducted in the present study. It is important to note

that several complications were encountered in the study that ultimately affected the projected statistical power for the analyses, leading to unavoidable adjustments to the original methods. Specifically, the usable sample size was significantly lower than anticipated as a result of missing data and limits created by the selection of a specific timeframe within the longitudinal CHP-C study archived data. In response, the researcher assessed the consistency of teacher ratings using a modified version of the ICC and a smaller, more parsimonious regression model than originally planned. Thus, the results of the present analyses should be interpreted with caution.

The results suggest that consistency among teacher ratings as measured by ICCs fell within the range of 0.45 to 0.59. All ICCs were significant ($ps < .001$), based on the underlying ANOVA F-test. Pairwise consistency between randomly selected teacher dyads in the dataset were also significant ($ps < .01$), and ranged from .42 to .74.

In preparation for the analysis of potential sources of rater bias, the researcher selected predictors from the original model where data appeared accurate, intercorrelations were unlikely to result in multicollinearity, and adequate variability was available

based on teacher responses to the Teacher Questionnaire. Other predictors were combined to make single predictors. In the end, the researcher constructed a four-predictor model, consisting of teacher sex, teacher age, parental status, and workload. The dependent variables (unstandardized deviation scores on five separate teacher ratings) were regressed onto the predictors in the model, one at a time, and the results are presented in this chapter.

In general, the regression model significantly improved prediction of teacher bias in the case of ADHD symptoms and in ratings of student overall impairment. On the measure of inattention symptoms, teachers with greater workloads provided relatively more severe ratings than their peers with lesser workloads. On the measure of hyperactivity-impulsivity, the ancillary variables of teacher sex and age appeared to improve prediction of teacher bias, with woman teachers and older teachers associated with severe ratings. However, these relationships were only apparent after the experiential predictors were added to the model, suggesting possible interaction effects. On overall ADHD symptoms, workload, woman teachers, and older teachers were all associated with relatively severe ratings. On the measure of

academic impairment, the model did not improve prediction of teacher bias. Finally, on the measure of overall impairment, teacher workload appeared to uniquely improve prediction of teacher bias, with higher workloads associated with relatively severe ratings.

CHAPTER V

DISCUSSION

Introduction

This chapter addresses the results of the present study. Specifically, the researcher will interpret the results, based on the generalizability and robustness of the analyses, and provide conclusions that take into consideration the many limitations of the present study. Finally, several recommendations are offered for future research to improve school psychologists' understanding of between-teacher reliability in behavior ratings and potential sources of rater bias in ratings of adolescents with ADHD.

Interpretation

The present study consisted of two aims; first, to examine the rates of interrater reliability among teacher behavior ratings of middle school students with ADHD, and second, to examine potential sources of rater bias. The existing literature on interrater reliability among teachers suggests that school psychologists can expect modest reliability rates among teachers (Evans, Allen, et al., 2005), and that ratings collected in secondary school environments are potentially less consistent than those collected at the elementary level (Achenbach, et

al., 1987). Thus, in the present study, it was hypothesized that middle school teacher ratings of students with ADHD would be congruent with that of similar research at the secondary level, which found that ICCs generally fall below .52 (Molina et al., 1998). However, the present study resulted in ICCs that ranged from 0.45 to 0.59, with all ICCs reaching statistical significance (*p*s < .001). On a measure of DSM-IV(-TR) defined inattention, all ICCs matched or exceeded .52, and on the measure of DSM-IV(-TR) defined hyperactivity-impulsivity, all ICCs matched or fell below .52. This finding suggests that interrater reliability on the measure of inattention in the present study exceeded the highest level of interrater reliability found by previous researchers using virtually identical instruments and statistical methods (Molina et al., 1998). Indeed, the measure of ADHD symptoms used in the previous research informed the hypotheses specific to the researcher's first aim of the study, and the approach to computing ICCs followed the same statistical formula.

On a measure of impairment (IRS), between-teacher reliability on an item relating specifically to academic impairment appeared to match or exceed the level predicted by the researcher (ICCs ≥ .52). Indeed,

teacher consistency on academic impairment resulted in the highest ICCs found in the present study, with ICCs reaching .59 in May of 2005. On the item relating to overall impairment, ICCs fell below the predicted threshold in all but one month.

To assess the consistency between two randomly selected teachers, Pearson correlations within teacher dyads were computed. The results of this analysis resulted in correlations that ranged from .42 to .74, suggesting that school psychologists collecting ratings from two randomly selected teachers within this sample could reasonably expect moderate consistency rates. In comparison to previous research that found correlations that ranged from .37 to .53 (Molina et al., 1998), the correlations in the present study appeared appreciably stronger. Hence, when the results of the ICCs and the Pearson correlations are taken into account, the hypothesis related to the first aim of this study appeared largely incorrect; expressly, between-teacher reliability exceeded the anticipated levels in many instances.

There are several potential explanations for why the correlations in the present study generally exceeded those of previous research. First, teachers in the

present study were participants in the CHP-C study, which provided five hours of training prior to the school year and—for teachers in the treatment condition—there was ongoing behavior consultation with a certified school psychologist (author). In contrast, most previous studies, such as the study that informed the researcher's first hypothesis (Molina et al., 1998), did not provide teachers with training related to ADHD beyond that provided by their schools. Further, research has often been based on single measurements only and not on longitudinal data where participants were aware of a previous diagnosis (e.g., Amador-Campos et al., 2006; Mitsis et al., 2000; Molina et al., 1998).

To assess the potential impact of the school consultation in the treatment condition of the CHP-C study on between-teacher consistency, the researcher performed post hoc ICCs for all measures, using only the teachers in the treatment condition. The results of this analysis suggested that the treatment condition did not exhibit appreciably stronger ICCs than did all teachers combined. For example, on Total Score subscale of the DBD for May, treatment group teachers exhibited virtually identical interrater reliability (ICC = .55) as the overall sample of teachers (ICC = .55). Hence, it seems

reasonable that the relatively stronger ICCs observed in the current study as compared to previous studies may be related to the teacher trainings provided by the CHP-C study. It also seems likely that teachers in both conditions of the CHP-C discussed the student participants more frequently than would have occurred otherwise. Teachers were aware of the ongoing monitoring for the students in the CHP-C study and it seems likely that, over time, they were cued by the monitoring system (i.e., monthly rating scale) to closely observe student behavior and to discuss these students often. Frequent cues and discussions relevant to the targets may have altered teacher perceptions, resulting in consensus judgments.

Second, it is conceivable that teachers are generally more familiar with ADHD in recent years as opposed to previous years, especially as it relates to the DSM-IV diagnostic criteria. For example, repeated exposure to rating scales, which likely occurs for teachers over time, may sensitize raters to the behaviors of interest and perhaps provide prompts for careful observance of ADHD-related behaviors. Further, as school professionals become increasingly familiar with best practice assessment techniques for ADHD (Demaray,

Schaefer, & Delong, 2003), it is likely that teachers have access to high quality workshops and in-service trainings on the disorder and its related impairments. The researcher's original hypothesis regarding the anticipated level of consistency between teacher ratings was based on a study conducted nearly ten years ago (just years after the publication of the DSM-IV) and those results may not apply to contemporary rater performances.

Third, while the measure of ADHD symptoms was virtually identical to that of Molina and colleagues (1998), the measure of impairment in the present study was unique. Interestingly, the strongest correlations occurred on the measure of academic impairment, which exceeded the anticipated range by the greatest amount. One possible interpretation of this finding is that, when compared to ratings of ADHD symptoms, teachers are more likely to come to similar conclusions regarding a student's need for additional academic assistance. Since teacher training primarily focuses on classroom curriculum and the creation of effective lesson plans, it seems reasonable to conclude that teachers are better prepared to identify academic needs versus behavioral and mental health needs. Researchers in the field of school mental health have come to similar conclusions, in that

teachers appear inadequately prepared to identify student mental health needs (Weist, 2005) and schools generally ignore mental health issues until academics are clearly impacted (Adelman & Taylor, 2004).

Fourth, the teacher ratings in the present study were collected during the second half of the school year. Previous research suggests that interrater reliability fluctuates over the course of the school year, and that average rates of consistency are observed in February and then slowly climb from that point (Evans, Allen, et al., 2005). It seems likely that, by focusing only on the second half of the school year, the sample examined in this study represents average to high average consistency. Ratings collected from the beginning of the school year may have returned far different results. These findings, to the extent they reflect the "time effect" (Evans, Allen, et al., 2005, p. 702), have practical implications for school psychologists. Specifically, teacher ratings of ADHD symptoms collected during the first semester in secondary schools are likely to be less reliable than those collected during the second semester. As a result, replication of these measures may be advisable, thus allowing teachers

adequate time to observe the behaviors and impairments of interest.

Taken together, the present analysis of between-teacher reliability appears credible, given the congruence of these findings and those of previous research in the context of potential explanations for the appreciably stronger correlations found within this sample. However, design issues stemming from the measurement strategy of the CHP-C study preclude strong conclusions about the external validity (generalizability) of the results to a larger middle school teacher population, and how other teachers rate students with ADHD. It is possible, for example, that due to missing data, some teacher teams were underrepresented or overrepresented in the CHP-C study data, thus potentially affecting overall interrater reliability, and it cannot be assumed that the missing data in the CHP-C study dataset occurred randomly. Hence, the findings of the present study must be interpreted with caution.

In the second aim of the present study, the researcher investigated potential sources of rater bias. With interrater reliability exceeding that of previous research in many instances, the ICCs and most of the

Pearson correlations attained in the present study suggested that a substantial proportion of rater variance was still unexplained. For example, even the strongest ICC (.59), which occurred on the academic impairment item of the IRS in May, 2005, suggested that an estimated 41% of the variance was unexplained. Similarly, the strongest Pearson correlation, which occurred on the hyperactivity-impulsivity subscale of the DBD in May, suggested that 45% of the variance was unexplained. Thus, the consistency of teacher ratings in the present study can be considered moderate, based on guidelines provided in the literature (e.g., Kline, 2005). Further, the level of rater consistency in the CHP-C study was below the average between-teacher reliability (ICC = .62) found in Achenbach and colleague's (1987) landmark meta-analysis of interrater reliability across many childhood disorders.

The error variance in within-target ratings was necessary for the second aim of the study, which was to assess potential sources of rater bias. With moderate interrater reliability rates found in the first aim of the study, the researcher anticipated that the analysis of rater bias would be based on ample within-target variance. However, several complications, including

imperfect return rates of rating scales and questionnaires, limited the available sample size and lowered the statistical power of the design. As a result, the original regression model was simplified, but the estimated power for the final regression model was .76, which was still below the level of .80 recommended by Cohen (1988). Further, the incomplete and unbalanced block design of the CHP-C study measurement strategy precluded clear partialing of variance components specific to targets and raters. In other words, the effects of rater biases and the effects of student behavior were inseparable (refer to Figure 2). As a result, rater bias in the present study cannot be interpreted as rater "error," per se, as differences between raters may be due to true target behavioral shifts across environments. Instead, the results of the analysis of rater bias relate to teacher *perception* of student behavior and impairment, which included both accurate and inaccurate interpretations. An examination of rater error would require a complete block design and complete overlap in observations, which is virtually impossible to achieve in field-based settings such as secondary schools.

Based on the results of hierarchical multiple regression, it appears that teacher characteristics and experiences predicted rater bias differentially across types of ratings. Among ratings of ADHD-related symptoms, teacher workload (defined as the number of students taught per day) appeared to explain a significant, albeit small, proportion of the error variance. Specifically, as workload increased, teacher ratings of target symptoms of inattention became more severe. This suggests that experiential factors, such as work-related stress, affect teacher perception of inattentive symptoms to some degree. However, it should be noted that when workload was taken into consideration with the effects of teacher age, teacher sex, and parenting status, the model only explained 7.0% of the variance in the ratings, using the most conservative estimate. While workload represented a statistically significant proportion of the explained variance, the predicted effects on ratings were not clinically meaningful. For example, the regression equation suggests that as workload increased by a little more than one class, or about 35 students per day, the predicted increase in inattention ratings increased by only 1.16 points on the inattention subscale. Thus, it appears

that workload may have a minimal impact on teacher perception of inattention, but this impact is not likely to be clinically meaningful, given the 27 point range of the inattention subscale.

Among ratings of hyperactivity-impulsivity symptoms, it appeared that teacher characteristics, including sex and age, significantly predicted variance in teacher ratings, but only when the effects of teacher workload and parental status were removed statistically. When combined, the predictors appeared to explain an estimated 9.0% of the variance in hyperactivity-impulsivity ratings, based on the most conservative estimate. It appeared that women and younger teachers were more likely than men and older teachers to provide relatively severe target ratings, once the effect of workload and parenting experiences were removed. Specifically, it appeared that women provided ratings that were 2.15 points more severe than men. Given the similarity of this finding with that commonly found between mothers and fathers, where mothers generally provide more severe ratings (e.g., Reynolds & Kamphaus, 1992), the present finding appears robustly consistent with similar research. However, on a scale with a range of 27 points, two point differences between women and men are unlikely to be clinically meaningful

unless examiners adhere too tightly to specific cutpoints and ignore "close calls."

In terms of teacher age, an increase in teacher age of 10.96 years predicted about 1.52 points greater rating leniency on the hyperactivity-impulsivity subscale. Unlike the finding regarding teacher sex, the finding for teacher age is potentially meaningful in cases of large discrepancies between teachers. For example, the regression equation predicts that the rating provided by a 58-year old teacher would on average be 4.56 points more lenient than that provided by a 25-year old teacher. A difference of this magnitude can clearly affect diagnostic and treatment decisions. It should be noted however, that the impact of teacher sex and age appeared statistically significant only after the effect of teacher parental status and workload were held constant.

Taken together, the complex results for the hyperactivity-impulsivity subscale may suggest that an interaction existed between the experiential and demographic variables, such as a moderated or mediated relationship (Frazier, Tix, & Barron, 2004). However, further analysis of this possibility was clearly beyond the scope of the present study, given the limited statistical power and the other complications encountered

(described above). One possible explanation is that there is an interaction between teacher workload and teacher characteristics (sex and age) that affects perception of student hyperactivity-impulsivity. For example, it may be that older teachers who have developed effective classroom management strategies observe fewer disruptive behaviors in their classrooms than do younger teachers, but this relationship is mediated or moderated to some degree by experiences, such as workload. It may also be that men perceive hyperactive-impulsive symptoms as less disruptive than do women teachers, perhaps due to gender match between the rater and the target, as hyperactive-impulsive children are more likely to be boys (Gaub & Carlson, 1997). Similar effects have been noted in cases of rater-target racial match, suggesting dyad-specific rater bias as noted in the literature review (e.g., Downey & Pribesh, 2004; Sonuga-Barke et al., 1993). However, in the present study this finding appears to be mediated or moderated by teacher experiences, including workload. Unfortunately, the present study lacks the statistical power to examine complex interactions such as these due to missing data and imperfect return rates on rating scales and questionnaires.

In overall ratings of ADHD, the researcher found that both teacher characteristics, including sex and age, as well as experiences, specifically workload, appeared to explain a significant proportion of variance in teacher ratings. It appears that the statistical significance of multiple predictors in this analysis is perhaps an artifact of the combination of the findings for inattentive and hyperactive-impulsive symptoms. It seems likely that teacher perception of inattention is predicted by workload and teacher perception of hyperactivity-impulsivity is predicted by a more complex interaction between teacher demographics (age and sex) and experiences, including workload. The overall regression model used in the present study appeared to explain an estimated 11.1% of the rating variance, using the most conservative estimate. However, the practical implications are questionable, as the predictors were associated with small changes in the dependent variable. Thus, from a school psychology standpoint, the results are interesting but not robust enough to support strong conclusions regarding the interpretation of inconsistent teacher ratings. One possible exception is teacher age. Again, as with the findings for the hyperactivity-impulsivity subscale, the regression equation predicts

that large differences in teacher ages may result in meaningfully different ratings of the same target. For example, the results predict that a rating provided by a 58-year old teacher would on average be 7.50 points more lenient than that provided by a 25-year old teacher. On a scale with a range of 52 points, this may result in meaningful discrepancies, depending on how strictly examiners interpret the findings relative to potential diagnostic or treatment thresholds.

Among measures of impairment, it appeared that the regression model was unsuccessful in explaining a significant or meaningful amount of variance in teacher ratings of academic impairment. As discussed earlier, interrater reliability on this item was generally higher than all other measures examined in this study, perhaps because teachers are more reliable raters of academic performance. Thus, it is not surprising that teacher age, teacher sex, teacher parental status, or workload predicted teacher bias. Teachers have access to objective measures of academic performance, in the form of tests, quizzes, classwork, and homework, thus informing perceptions of student achievement. In contrast, fewer objective measures of ADHD-related symptoms are available to teachers on a daily basis,

thereby leading to more between-teacher inconsistencies that can be partly predicted by teacher characteristics and experiences.

In the final analysis, the researcher examined bias on an overall measure of impairment. The results of this analysis suggested that the regression model explained 6.8% of the variance in teacher ratings, using the most conservative estimate. Similar to the first regression analysis of teacher ratings of inattention, this analysis suggested that prediction was improved by the unique contribution of teacher workload. Again, it appeared that as workload increased, teacher ratings became significantly more severe. However, from a practical standpoint, the change predicted in the dependent variable was relatively small, as an increase of 35 students per day was associated with only a 0.42 point increase in rating severity on the IRS item. Given the range of the item (6 points), the change associated with workload would only be meaningful when comparing the ratings of teachers with large discrepancies in workload.

In summarizing the results of the five regression analyses, it appears that teacher workload most consistently predicted teacher bias. As discussed in the literature review, similar findings were reported by

Havey and colleagues (2005), as overidentification of ADHD was associated with larger class sizes; however, other research has suggested opposite results in private school settings (e.g., Glass & Wegar, 2000). The present study appears to support a positive relationship between class size (a component of workload) and rating severity across several measures, where increased workload predicted increased rating severity. In this study, the relationship between rater bias and workload appeared to occur across multiple teacher ratings, including inattentive symptoms, overall ADHD symptoms, and overall impairment. Interestingly, the partial plots generated by the regression analysis suggested that a positive relationship between workload and rating severity occurred for every examined measure, although not to a significant degree on ratings of hyperactivity-impulsivity or academic impairment. When the effect of workload was statistically significant, the predicted effects on the dependent variable were small and unlikely to be clinically meaningful. For example, on the total score subscale of the DBD, which has a range of 52 points, an increase of 35 students per day predicted only an estimated 2 point increase in severity on teacher ratings.

One possible explanation for the recurring and positive effect of workload on teacher ratings may be related to the experiential differences between general education and special education teachers. On the Teacher Questionnaire, four teachers indicated that they were trained to be special education teachers. While this item proved problematic and was removed from the regression analyses (see Complications section), the researcher performed a post hoc analysis to compare the reported workload of the self-identified special education teachers versus all others. The differences in workload between the groups were statistically significant (t = 2.44, p = .02), as the average teacher workload for special education teachers (M = 48.75) and general education teachers (M = 91.22) differed substantially. Even though no outliers were observed in the regression analyses relative to workload, it is clear that special education teachers in this sample interacted with fewer students per day than did general education teachers.

Conceivably, special education teachers rate students with ADHD more leniently than their general education counterparts because judgments are made relative to other students in their classrooms. For

general education teachers, the comparisons are mostly to students without identified special needs, and for special education teachers the comparisons are made to students with clearly defined needs. In comparison all students receiving special education services, students with ADHD represent a relatively mild disability population. In most instances, students with ADHD who receive special education services are identified with specific learning disabilities (U.S. Department of Education, 2005), which are clearly less impairing on average than other special education categories, such as autism or mental retardation.

Conclusions

The present study encountered several complications that challenged the internal and external validity of the results. For example, poor return rates and missing data substantially reduced the statistical power of the planned analysis of teacher bias, which required the researcher to adjust the regression model. Thus, the results of the present study must be interpreted carefully.

Based on the results of the analyses, it appears that between-teacher reliability on ratings of ADHD symptoms and impairments are modest and generally less

reliable than ratings among teachers observed in previous research at the elementary level (Achenbach et al., 1987). Further, the analyses of teacher bias suggested that teacher workload significantly predicts rater bias, whereby increased workloads were associated with rating severity. In addition, teacher sex and age may play a role in teacher perception of hyperactivity-impulsivity symptoms, whereby younger teachers and women teachers generally provided more severe ratings. However, the effect sizes observed in the teacher bias analyses suggested that there are few if any practical implications for school psychologists.

At the outset, I hoped to provide general guidelines for school psychologists faced with inconsistent teacher ratings; however, the results suggested that the variance associated with teacher characteristics and experiences are unlikely to meaningfully change teacher ratings. The only potential exception to this general conclusion is in the case of extreme differences in teacher age, as older teachers were found to provide more lenient ratings of hyperactivity-impulsivity on average. Once teacher age discrepancies go beyond 20 to 30 years of age, for example, the results of the present study suggest that

school psychologists can expect meaningful discrepancies in ratings of hyperactivity-impulsivity.

Teacher workload was also found to have consistently positive relationships with rating severity, but again the effect of workload on teacher ratings would only be meaningful in cases where teacher raters vary widely on this variable, based on the results of the present study. It appears that such discrepancies are most likely to occur when ratings are provided by both special and general education teachers because the discrepancies in workload between these environments are often quite large, as found in this sample. Further, special education and general education teachers may compare the symptoms and impairments of students with ADHD to differing student populations. In other words, special educators and general educators have access to different local norms.

Given the complications encountered in the present study, more research is needed to clarify the findings. Specifically, future research is needed to cross-validate the present results using other samples. Further, future research on issues of interrater reliability among secondary teachers and sources of rater bias would benefit from larger samples and measurement designs that

approximate fully-crossed complete block designs, with overlap in rater observations. As mentioned previously, this design is virtually impossible to achieve in naturalistic secondary school settings. As such, future research may benefit from combining both naturalistic observations of teacher ratings and analogue assessment of rater biases using experimentally controlled stimuli (e.g., videotaped scenarios) and settings.

References

Abikoff, H., Courtney, M., Pelham, W. E., & Koplewicz, H. S. (1993). Teachers' ratings of disruptive behaviors: The influence of halo effects. *Journal of Abnormal Child Psychology*, *21*, 519-533.

Abramowitz, A.J., & O'Leary, S.G. (1991). Behavioral interventions for the classroom: Implications for students with ADHD. *School Psychology Review, 20,* 220-234.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*, 213-232.

Adelman, H.S., Barker, L.A., & Nelson, P. (1993). A study of a school-based clinic: Who uses it and who doesn't? *Journal of Clinical Psychology, 22,* 52-59.

Adelman, H.S., & Taylor, L. (2004). Mental health in schools: A shared agenda. *Report on Behavioral Disorders in Youth*, *4*, 59-78.

Amador-Campos, J. A., Forns-Santacana, M., Guàrdia-Olmos, J., & Peró-Cebollero, M. (2006). DSM-IV Attention deficit hyperactivity disorder symptoms: Agreement between informants in prevalence and factor structures at different ages. *Journal of Psychopathology and Behavioral Assessment*, *28*, 23-32.

American Academy of Child and Adolescent Psychiatry (1997). Practice parameters for the assessment and treatment of children, adolescents, and adults with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, *36*, 85S-121S.

American Academy of Pediatrics (2000). Clinical practice guidelines: Diagnosis and evaluation of the child with attention-deficit/hyperactivity disorder. *Pediatrics*, *105*, 1158-1170.

American Psychiatric Association. (1994). *Diagnostic and Statistical Manual of Mental Disorders (4th Edition)*. Washington DC: American Psychiatric Association.

American Psychiatric Association (2000). *Diagnostic and Statistical Manual of Mental Disorders (4th Edition), Text Revision*. Washington DC: American Psychiatric Association.

American Psychological Association (2003). *Senate introduces bill to provide incentives to increase the ranks of child mental health providers*. Public Policy Office. Available at http://www.apa.org/ppo/issues/es12230603.html.

Anastopoulos, A. D., & Shelton, T. L. (2001). *Assessing attention-deficit/hyperactivity disorder.* New York: Kluwer Academic/Plenum Publishers.

Angold, A., Erkanli, A., Egger, H.I., & Costello, E.J. (2000). Stimulant treatment for children: A community perspective. *Journal of the American Academy of Child and Adolescent Psychiatry, 39,* 975-984.

Antrop, I. , Roeyers, H., Van Hoost, P., & Buysse, A. (2000). Stimulation seeking and hyperactivity in children with ADHD. *Journal of Child Psychology and Psychiatry*, *41*, 225-231.

Applegate, B., Lahey, B. B., Hart, E. L., Biderman, J., Hynd, G. W., Barkley, R. A., Ollendick, T., Frick, P. J., Greenhill, L., McBurnett, K., Newcorn, J. H., Kerdyk, L., Garfinkel, B., Waldman, I., & Shaffer, D. (1997). Validity of the age-of-onset criterion for ADHD: A report from the DSM-IV field trials. *Journal of the American Academy of Child and Adolescent Psychiatry*, *37*, 657-664.

Archer, J., & Coyne, S. M. (2005). An integrated review of indirect, relational, and social aggression. *Personality and Social Psychology Review*, *9*, 212-230.

Arnold, L. E., Chuang, S., Davies, M., Abikoff, H. B., Conners, C. K., Elliott, G. R., Greenhill, L. L., Hechtman, L., Hinshaw, S. P., Hoza, B., Jenson, P.

S., Kraemer, H. C., Langworthy-Lam, K. S., March, J. S., Newcord, J. H., Pelham, W. E., Severe, J. B., Swanson, J. M., Vitiello, B., Wells, K. C., & Wigal, T. (2004). None months of multicomponent behavioral treatment for ADHD and effectiveness of MTA fading procedures. *Journal of Abnormal Child Psychology*, *32*, 39-51.

Bagwell, C., Molina, B. S. G., Pelham, W. E., & Hoza, B. (2001). Attention-Deficit Hyperactivity Disorder and problems in peer relations: Predictions from childhood to adolescence. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*, 1285-1292.

Barkley, R.A. (1990). *Attention Deficit Hyperactivity Disorder: A Handbook for Diagnosis and Treatment*. New York: Guilford Press.

Barkley, R. A. (1998). *Attention Deficit Hyperactivity Disorder: A Handbook for Diagnosis and Treatment*. New York: Guilford Press.

Barkley, R. A. (2002). International consensus statement on ADHD. *Clinical Child and Family Psychology Review*, *5*, 89-111.

Barkley, R. A. (2006). *Attention-Deficit Hyperactivity Disorder* (3rd Ed.). New York: The Guilford Press.

Barkley, R. A., Anastopoulos, A. D., Guevremont, D. C., & Fletcher, K. E. (1991). Adolescents with ADHD: Patterns of behavioral adjustment, academic functioning, and treatment utilization. *Journal of the American Academy of Child and Adolescent Psychiatry, 30*, 752-761.

Barkley, R. A., & Biederman, J. (1997). Toward a broader definition of the age-of-onset criterion for Attention-Deficit Hyperactivity Disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, *36*, 1204-1210.

Barkley, R. A., DuPaul, G. J., & Connor, D. F. (1999). Stimulants. In J. S. Werry & M. G. Aman (Eds.), Practitioner's guide to psychoactive drugs for

children and adolescents (2nd ed., pp. 213-247). New York: Plenum Medical Book.

Barkley, R. A., DuPaul, G. J., & McMurray, M. B. (1990). A comprehensive evaluation of attention deficit disorder with and without hyperactivity. *Journal of Consulting and Clinical Psychology*, *58*, 775-789.

Barkley, R. A., Edwards, G., Laneri, M., Fletcher, K., & Metevia, L. (2001). The efficacy of problem-solving communication training alone, behavior management training alone, and their combination for parent-adolescent conflict in teenagers with ADHD and ODD. *Journal of Consulting and Clinical Psychology*, *69*(6), 926-941.

Barkley, R. A., Fischer, M., & Edelbrock, C. S. (1990). The adolescent outcome of hyperactive children diagnosed by research criteria: An 8-year prospective follow-up study. *Journal of the American Academy of Child and Adolescent Psychiatry*, *29*, 546-557.

Barkley, R. A., Fischer, M., Smallish, L., & Fletcher, K. (2002). The persistence of attention-deficit/hyperactivity disorder into young adulthood as a function of reporting source and definition of disorder. *Journal of Abnormal Psychology*, *111*, 279-289.

Barkley, R. A., Guevremont, D. C., Anastopoulos, A. D., & Fletcher, K. E. (1992). A comparison of three family therapy programs for treating family conflicts in adolescents with Attention-Deficit Hyperactivity Disorder. *Journal of Consulting and Clinical Psychology, 60,* 450-462.

Barron, K. E., Evans, S. W., Baranick, L. E., Serpell, Z. N., & Buvinger, E. (2006). Achievement goals of students with ADHD. *Learning Disability Quarterly*, *29*, 137-158.

Bartko, J. J., & Carpenter, W. T. (1976). On the methods and theory of reliability. *The Journal of Nervous and Mental Disease*, *163*, 307-317.

Becker, K. B., & McCloskey, L. A. (2002). Attention and conduct problems in children exposed to family violence. *American Journal of Orthopsychiatry*, *72*, 83-91.

Börger, N., & van der Meere, J. (2000). Visual behaviour of ADHD children during an attention test: An almost forgotten variable. *Journal of Child Psychology and Psychiatry*, *41*, 525-532.

Breggin, P. R. (2001). *Talking back to Ritalin: What doctors aren't telling you about stimulants and ADHD* (Rev. Ed.). Massachusetts: Perseus Publishing.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Bridgett, D. J., & Walker, M. E. (2006). Intellectual functioning in adults with ADHD: A meta-analytic examination of full scale IQ differences between adults with and without ADHD. *Psychological Assessment*, *18*, 1-14.

Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical linear models.* Newbury Park, CA: SAGE.

Cairns, E., & Green, J. A. (1979). How to assess personality and social patterns: Observations or ratings? In R. B. Cairns (Ed.), *The analysis of social interactions* (pp. 209-226). Hillsdale, NJ: Lawrence Erlbaum.

Carpenter, S. (2001). Stimulants boost achievement in ADHD teens. *Monitor on Psychology*, *32*(5), 26-27.

Carroll, K. M. & Nuro, K. F. (2002). One size cannot fit all: A stage model for psychotherapy manual development. *Clinical Psychology: Science and Practice, 9,* 396-406.

Chambers, J.G., Shkolnik, J., Perez, M. (2003). *Total expenditures for students with disabilities, 1999-2000: Spending variation by disability*. Report published by the American Institutes for Research and available at http://csef.air.org/.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66, 7-18.*

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.

Center for Mental Health in Schools. (2003, April). *Working collaboratively: From school-based teams to school-community-higher education connections. An introductory packet.* Los Angeles: Author.

Cluett, S. E., Forness, S. R., Ramey, S. L., Ramey, C. T., Hsu, C., Kavale, K. A., & Gresham, F. M. (1998). Consequences of differential diagnostic criteria on identification rates of children with emotional or behavioral disorders. *Journal of Emotional and Behavioral Disorders, 6*, 130-141.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, P., Velez, N., Kohn, M., Schwab-Stone, M., & Johnson, J. (1987). Child psychiatric diagnosis by computer algorithm: Theoretical issues and empirical tests. *Journal of the American Academy of Child and Adolescent Psychiatry, 26*, 631-638.

Conners, C. K. (2000). *Conners' Continuous Performance Test II user's manual*. Toronto: MHS.

Conners, C. K., Epstein, J. N., March, J. S., Angold, A., Wells, K. C., Klaric, J., et al. (2001). Multimodal treatment of ADHD in the MTA: An alternative outcome analysis. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*, 159-167.

Coolidge,F. L., Thede, L. L., & Young, S. E. (2000). Heritability and the comorbidity of attention deficit hyperactivity disorder with behavioral disorders and executive function deficits: A preliminary investigation. *Developmental Neuropsychology, 17*, 273-287.

Cox, E.R., Motheral, B.R., Henderson, R.R., & Mager, D. (2003). Geographic variation in the prevalence of stimulant medication use among children 5 to 14 years old: Results from a commercially insured US sample. *Pediatrics*, *111*, 237-243.

Cuffe, S. P., McKeown, R. E., Jackson, K. L., Addy, C. L., Abramson, R., & Garrison, C. Z. (2001). Prevalence of Attention-Deficit/Hyperactivity Disorder in a community sample of older adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry*, *40*, 1037-1044.

Danforth, J. S., & DuPaul, G. J. (1996). Interrater reliability of teacher rating scales for children with Attention-Deficit Hyperactivity Disorder. *Journal of Psychopathology and Behavioral Assessment*, *18*, 227-237.

De Los Reyes, A., & Kazdin, A.E. (2004). Measuring informant discrepancies in clinical child research. *Psychological Assessment*, *16*, 330-334.

Demaray, M. K., Elting, J., & Shaefer, K. (2003). Assessment of attention-deficit/hyperactivity disorder (ADHD): A comparative evaluation of five, commonly used, published rating scales. *Psychology in the Schools*, *40*, 341-361.

Demaray, M. K., Schaefer, K., & Delong, L. K. (2003). Attention-deficit hyperactivity disorder (ADHD): A national survey of training and current assessment practices in the schools. *Psychology in the Schools*, *40*, 583-597.

Deshler, D.D. & Schumaker, J.B. (1990). Learning strategies: An instructional alternative for low-achieving adolescents. In S.B. Sigmon (Ed.) *Critical Voices in Special Education: Problems and Progress Concerning the Mildly Handicapped.* (pp. 155-166). Albany, NY: State University of New York Press.

Di Nocera, F., Ferlazzo, F., & Borghi, V. (2001). G theory and the reliability of psychophysiological measures: A tutorial. *Psychophysiology*, *38*, 796-806.

Diener, M. B., & Milich, R. (1997). Effects of positive feedback on the social interactions of boys with attention deficit hyperactivity disorder: A test of the self-protective hypothesis. *Journal of Clincal Child Psychology*, *26*, 256-265.

DiScala, C., Lescoheir, I., Barthel, M., & Li, G. (1998). Injuries to children with attention deficit hyperactivity disorder. *Pediatrics*, *102*(6), 1415-1421.

Dishion, T. J., Nelson, S. E., Kavanagh, K., Beidel, D., Brown, T. A., Lochman, J., & Haaga, D. A. (2003). The family check-up with high-risk young adolescents: Preventing early-onset substance use by parent monitoring. *Behavior Therapy*, *34,* 553-572.

Dishion, T. J., & Kavanagh, K. (1999). *Adolescent Transitions Program; assessment and intervention sourcebook* . Eugene, OR: OSLC.

Dishion, T. E., & Kavanagh, K. (2003). *Intervening in adolescent problem behavior: A family centered approach*. New York: Guilford Press.

Doherty, S.L., Frankenberger, W., Fuhrer, R., & Snider, V. (2000). Children's self reported effects of stimulant medications. *International Journal of Disability, Development & Education*, *47*, 39-54.

Douglas, V. I. (1999). Cognitive control processes in attention-deficit/hyperactivity disorder. In H.C. Quay, A.E. Hogan et al. (Eds.) *Handbook of Disruptive Behavior Disorders*. (pp. 105-138). New York, NY: Kluwer Academic/ Plenum Publishers.

dosReis, S., Zito, J.M., Safer, D.J., Gardner, J. F., Puccia, K. B., & Owens, P. L. (2005). Multiple psychotropic medication use for youths: A two-state comparison. *Journal of Child and Adolescent Psychopharmacology*, *15*, 68-77.

dosReis, S., Zito, J.M., Safer, D.J., Soeken, K.L., Mitchell, J.W., & Ellwood, L.C. (2003). Parental perceptions and satisfaction with stimulant medication for attention-deficit hyperactivity

disorder. *Journal of Developmental & Behavioral Pediatrics*, *24*, 155-162.

Downey, D. B., & Pribesh, S. (2004). When race matters: Teachers' evaluations of students' classroom behaviors. *Sociology of Education*, 77, 267-282.

Dulcan, M., Dunne, J.E., Ayres, W., Arnold, V., Benson, R.S., Bernet, W., Bukstein, O., Kinlan, J., Leonard, H., Licamele, W., McClellan, J., Sloan, L.E., & Miles, C.M. (1997). Practice parameters for the assessment and treatment of children, adolescents, and adults with Attention-Deficit/Hyperactivity Disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 36,* 85-121.

Dumas, J. E., Lynch, A. M., & Laughlin, J. E. (2001). Promoting intervention fidelity: Conceptual issues, methods, and preliminary results from the EARLY ALLIANCE prevention trial. *American Journal of Preventive Medicine*, *20*, 38-47.

Durston, S. (2003). A review of the biological bases of ADHD: What have we learned from imaging studies? *Mental Retardation and Developmental Disabilities Research Reviews*, *9*, 184-195.

DuPaul, G. J., Power, T. J., Anastopoulos, A. C., Reid, R., McGoey, K. A., & Ikeda, M. J. (1997). Teacher ratings of Attention-Deficit Hyperactivity Disorder Symptoms: Factor structure and normative data. *Psychological Assessment, 9*(4), 436-444.

DuPaul, G.J., Power, T.J., Anastopoulos, A.D., & Reid, R. (1998). *ADHD Rating Scale – IV: Checklist norms and clinical interpretation.* New York: The Guilford Press.

Dupaul, G.J., & Stoner, G. (2002). Interventions for attention problems. In M.R. Shinn, H.M. Walker, & G. Stoner (Eds.) *Interventions for Academic and Behavior Problems II: Preventive and Remedial Approaches*. (pp. 913-938). Washington, DC: National Association of School Psychologists.

DuPaul, G. J., & Stoner, G. (1994). *ADHD in the schools: Assessment and intervention strategies*. New York: The Guilford Press.

Epstein, J. N., Conners, C. K., Erhardt, D., Arnold, L. E., Hechtman, L., Hinshaw, S. P., Hoza, B., Newcorn, J. H., Swanson, J. M., & Vitiello, B. (2000). Family aggregation of ADHD characterisics. *Journal of Abnormal Child Psychology*, *28*, 585-594.

Epstein, J. N., Willoughby, M., Valencia, E. Y., Tonev, S. T., Abikoff, H. B., Arnold, L. E., & Hinshaw, S. P. (2005). The role of children's ethnicity in the relationship between teacher ratings of attention-deficit/hyperactivity disorder and observed classroom behavior. *Journal of Consulting and Clinical Psychology*, *73*, 424-434.

Erhardt, D., & Hinshaw, S. P. (1994). Initial sociometric impressions of attention-deficit hyperactivity disorder and comparison boys: Predictions from social behaviors and from nonbehavioral variables. *Journal of Consulting and Clinical Psychology*, *62*, 833-842.

Ervin, R.A., DuPaul, G.J., Kern, L., & Friman, P.C. (1998). Classroom-based functional and adjunctive assessments: Proactive approaches to intervention selection for adolescents with Attention Deficit Hyperactivity Disorder. *Journal of Applied Behavior Analysis, 31,* 65-78.

Evans, S.W. (1999). Mental health services in schools: Utilization, effectiveness, and consent. *Clinical Psychology Review, 19,* 165-178.

Evans, S. W., Allen, J., Moore, S., & Strauss, V. (2005). Measuring symptoms and functioning of youth with ADHD in middle schools. *Journal of Abnormal Child Psychology*, *33*, 695-706.

Evans, S.W., Allen, J., Moore, S., & Timmins, B. (2004, July). *Reliability and validity of secondary school teachers' ratings of behavior.* Poster presented at annual meeting of the American Psychological Association, Honolulu, HI.

Evans, S.W., Axelrod, J.L., & Langberg, J. (2004). Efficacy of a school-based treatment program for middle school youth with ADHD: Pilot data. *Behavior Modification, 4,528-547*.

Evans, S. W., Langberg, J., Axelrod, J., Achey, S., McAllister, M., Rogers, R., & Cole, W. (2001). *School-based psychosocial treatment for middle school youth with ADHD*. Poster presented at the biennial meeting of the International Society for Research on Child and Adolescent Psychopathology, Vancouver, Canada.

Evans, S. W., Langberg, J., Raggi, V., Allen, J., & Buvinger, L. (2005). Development of a school-based treatment program for middle school youth with ADHD. *Journal of Attention Disorders*, *9*, 343-353.

Evans, S. W., Langberg, J., & Williams, J. (2003). Treatment generalization in school-based mental health. In M. D. Weist, S. W. Evans, & N. Tashman (Eds.), *Handbook of School Mental Health: Advancing Practice and Research*. (pp. 335-348). New York, NY: Kluwer Academic/Plenum Publishers.

Evans, S. W., & Pelham, W. E. (1991). Psychostimulant effects on academic and behavioral measures for ADHD junior high school students in a lecture format classroom. *Journal of Abnormal Child Psychology, 19,* 537-552.

Evans, S. W., Pelham, W. E., Gnagy, E., Smith, B., & Molina, B. (1999, November). Behavioral and educational interventions to improve academic performance in youth with ADHD. In S. Evans (Chair), *Effective Strategies for Behavior Therapists in Schools*. Symposium conducted at the meeting of the Association for Advancement of Behavior Therapy, Toronto, Canada.

Evans, S. W., Pelham, W., & Grudberg, M. V. (1995). The efficacy of note-taking to improve behavior and comprehension of adolescents with Attention Deficit Hyperactivity Disorder. *Exceptionality, 5,* 1-17.

Evans, S.W., Pelham, W.E., Smith, B.H., Bukstein, O., Gnagy, E.M., Greiner, A.R., Altenderfer, L., &

Baron-Myak, C. (2001). Dose-response effects of methylphenidate on ecologically-valid measures of academic performance and classroom behavior in adolescents with ADHD. *Experimental and Clinical Psychopharmacology*, *9*(2), 163-175.

Evans, S. W., Sapia, J. L., Axelrod, J. L., & Glomb N. (2002). Practical issues when establishing a school based mental health program: A case example. In H.S. Ghuman, M.D. Weist, & R.M. Saries (Eds.), *Providing Mental Health Services to Youth Where They Are: School- and Community Based Approaches*. New York: Brunner/Mazel.

Evans, S.W., Serpell, Z.N., Schultz, B. & Pastor, D. (2007). Cumulative benefits of secondary school-based treatment of students with ADHD. *School Psychology Review*, *36*, 256-273.

Evans, S. W., Serpell, Z., Williams, A., Gearing, F., Swensson, K. & Ingram, R. (2003, June). *Using Community Development Teams to Transport Science to Practice*. Poster presented at ISRCAP: 11th Scientific Meeting, Sydney, Australia.

Evans, S. W., Vallano, G., & Pelham, W. E. (1995). Attention-Deficit Hyperactivity Disorder. In V.B. Van Hasselt & M. Hersen (Eds.), *Handbook of Adolescent Psychopathology. A Guide to Diagnosis and Treatment.* (pp. 589-617). New York: Lexington Books.

Evans, S. W. & Weist, M. D. (2004). Implementing empirically supported treatments in the schools: What are we asking? *Clinical Child and Family Psychology Review*, 7, 263-267.

Evans, S. W., Williams, A., Schultz, B., & Weist, M. (2004). Behavioral assessment in schools. In *Encyclopedia of Applied Psychology*. Oxford, UK: Elsevier Inc.

Evans, S. W., & Youngstrom, E. (2006). Evidence-based assessment of attention-deficit/hyperactivity disorder: Measuring outcomes. *Journal of the American Academy of Child and Adolescent Psychiatry*, *45*, 1132-1137.

Fabiano, G. A., Pelham, W. E., Waschbusch, D. A., Gnagy, E. M., Lahey, B. B., Chronis, A. M., et al. (2006). A practical measure of impairment: Psychometric properties of the impairment rating scale in samples of children with attention deficit hyperactivity disorder and two school-based samples. *Journal of Clinical Child and Adolescent Psychology*, *35*, 369-385.

Fallah, N., Buvinger, E., Evans, S. W., Schultz, B., & Serpell, Z. (2006, August). *Outcomes of a consultation model school-based psychosocial intervention for middle School Students with ADHD.* Poster presented at the Annual Meeting of the American Psychological Association, New Orleans, LA.

Field, A. (2005). *Discovering statistics using SPSS* (2nd Ed.). London: Sage Publications.

Findling, R. L., Short, E. J., & Manos, M. J. (2001). Developmental aspects of psychostimulant treatment in children and adolescents with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*, 1441-1447.

Flaherty, L. T., & Osher, D. (2003). History of school-based mental health services. In M. D. Weist, S. W. Evans, N. A. Lever (Eds.), *Handbook of school mental health: Advancing practice and research* (pp. 11-22). New York: Kluwer Academic/Plenum Publishers.

Frazier, P.A., Tix, A.P., & Barron, K.E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology*, *51*, 115-134.

Freidman, D., Vaughan, H., & Erlenmeyer-Kimling, L. (1978). Stimulus and response related components of the late positive complex in visual discrimination tasks. *Electroencephalography and Clinical Neurophysiology*, *45*, 319-330.

Frick, P. J., Lahey, B. B., Applegate, B., Kerdyck, L., Ollendick, T., Hynd, G. W., Garfinkel, B., Greenhill, L., Biederman, J., Barkley, R. A.,

McBurnett, K., Newcorn, J., & Waldman, I. (1994). DSM-IV field trials for the disruptive behavior disorders: Symptom utility estimates. *Journal of the American Academy of Child and Adolesent Psychiatry*, *33*, 529-539.

Furman, L. (2005). What is attention-deficit hyperactivity disorder (ADHD)? *Journal of Child Neurology*, *20*, 994-1002.

Gaub, M., & Carlson, C. L. (1997). Gender differences in ADHD: A meta-analysis and critical review. *Journal of the American Academy of Child and Adolescent Psychiatry*, *36*, 1036-1045.

Gadow, K. D., Drabick, D. A., Loney, J., Sprafkin, J., Salisbury, H., Azizian, A., & Schwartz, J. (2004). Comparison of ADHD symptom subtypes as source-specific syndromes. *Journal of Child Psychology and Psychiatry*, *45*, 135-1149.

Gibbons, R.D., Hedeker, D.R., Elkin, I., Waternaux, C., Kraemer, H., Greenhouse, J., Shea, M.T., Imber, S.D., Sotsky, S.M., & Watkins, J.T., (1993). Some statistical and conceptual issues in the analysis of longitudinal psychiatry data. *Archives of General Psychiatry, 50*, 739-750.

Glass, C. S., & Wegar, K. (2000). Teacher perceptions of the incidence and management of attention deficit hyperactivity disorder. *Education*, *121*, 412-420.

Greene, R. W., Beszterczey, S. K., Katzenstein, T., Park, K., & Goring, J. (2002). Are students with ADHD more stressful to teach? *Journal of Emotional and Behavioral Disorders*, *10*, 79-90.

Greenhill, L. L. Abikoff, H. B., Arnold, L. E., Cantwell, D. P., Conners, C. K., Elliott, G., Hechtman, L., Hinshaw, S. P., Hoza, B., Jensen, P. S., March, J. S., Newcorn, J., Pelham, W. E., Severe, J. B., Swanson, J. M., Vitiello, B., & Wells, K. (1996). Medication treatment strategies in the MTA study: Relevance to clinicians and researchers. *Journal of the American Academy of Child and Adolescent Psychiatry*, *34*, 1304-1313.

Gresham, F.M. (1998). Social skills training: Should we raze, remodel, or rebuild. *Behavioral Disorders*, *24*, 19-25.

Gresham, F.M. & Elliott, S.N. (1990). *Social Skills Rating System*. Circle Pines, MN: American Guidance Service, Inc.

Grove, A.B., Evans, S.W., Thompson, J., & Barnett, L. (2004, July). *Barriers to children's mental health care.* Poster presented at annual meeting of the American Psychological Association, Honolulu, HI.

Guevara, J., Lozano, P., Wickizer, T., Mell, L., & Gephart, H. (2002). Psychotropic medication use in a population of children who have attention-deficit/hyperactivity disorder. *Pediatrics*, *109*, 733-739.

Guevara, J., Lozano, P., Wickizer, T., Mell, L., & Gephart, H. (2001). Utilization and cost of health care services for children with attention-deficit/hyperactivity disorder. *Pediatrics*, *108*, 71-78.

Hart, E. L., Lahey, B. B., Loeber, R., & Hanson, K. S. (1994). Criterion validity of informants in the diagnosis of disruptive behavior disorders in children: A preliminary study. *Journal of Consulting and Clinical Psychology*, *62*, 410-414.

Harter, S. (1982). The Perceived Competence Scale for Children. *Child Development, 53*, 87-97.

Harter, S. (1988). *Manual for the Self-Perception Profile for Adolescents*. Denver, CO: University of Denver.

Hartman, C. A., Willcutt, E. G., Rhee, S. H., & Pennington, B. F. (2004). The relation between sluggish cognitive tempo and DSM-IV ADHD. *Journal of Abnormal Child Psychology*, *32*, 491-503.

Hartung, C. M., Willcutt, E. G., Lahey, B. B., Pelham, W. E., Loney, J., Stein, M. A., & Keenan, K. (2002). Sex differences in young children who meet criteria for attention deficit hyperactivity disorder.

Journal of Clinical Child and Adolescent Psychology, *31*, 453-464.

Harris, M. J., Milich, R., Johnston, E. M., & Hoover, D. W. (1990). Effects of expectancies on children's social interactions. *Journal of Experimental Social Psychology*, *26*, 1-12.

Havey, J. M., Olson, J. M., McCormick, C., & Cates, G. L. (2005). Teachers' perceptions of the incidence and management of attention-deficit hyperactivity disorder. *Applied Neuropsychology*, *12*, 120-127.

Hechtman, L., & Greenfield, B. (2003). Long-term use of stimulants in children with attention deficit hyperactivity disorder. *Pediatric Drugs*, *5*, 787-794.

Hedeker, D.R. & Gibbons, R.D. (1996) MIXREG:a computer program for mixed-effects linear regression analysis with autocorrelated errors. *Computer Methods and Programs in Biomedicine, 49*, 229-252.

Hedeker, D., Gibbons R.D., and Waternaux, C. (1999). Sample size estimation for longitudinal designs with attrition: comparing time-related contrasts between two groups. *Journal of Educational and Behavioral Statistics*, *24*(1), 70-93.

Hendren, R. L., De Backer, I., & Pandina, G. J. (2000). Review of neuroimaging studies of children and adolescent psychiatric disorders from the past 10 years. *Journal of the American Academy of Child and Adolescent Psychiatry*, *39*, 815-827.

Henggeler, S. W. (2002). Transporting efficacious treatments to field settings: The link between supervisory practices and therapist fidelity in MST programs. *Journal of Clinical Child and Adolescent Psychology*, *31*, 155-167.

Hibbs, E. D., Clarke, G., Hechtman, L., Abikoff, H. B., Greenhill, L. L., & Jensen, P. S. (1997). Manual development for the treatment of child and adolescent disorders. *Psychopharmacology Bulletin, 33*, 619-629.

Hill, C.E., O'Grady, K.E., Price, P. (1988). A method for investigating sources of rater bias. *Journal of Counseling Psychology*, *35*, 346-350.

Hinshaw, S. P., Zupan, B. A., Simmel, C., Nigg. J. T., & Melnick, S. (1997). Peer status in boys with and without attention-deficit hyperactivity disorder: Predictions from over and covert antisocial behavior, social isolation, and athoritative parenting beliefs. *Child Development*, *68*, 880-896.

Hoagwood, K., Jensen, P., Roper, M., Arnold, L. Eugene, Odbert, C., Severe, J. et al. (in progress). *The reliability of the Services for Children and Adolescents Parent Interview (SCAPI).* Manuscript in preparation.

Hodgens, J. B., Cole, J., & Boldizar, J. (2000). Peer-based differences among boys with ADHD. *Journal of Clinical Child Psychology*, *29*, 443-452.

Homack, S. R., & Reynolds, C. R. (2005). Continuous performance testing in the differential diagnosis of ADHD. *The ADHD Report*, *13*(5), 5-9.

Homack, S., & Riccio, C. A. (2004). A meta-analysis of the specificity and sensitivity of the Stroop Color and Word Test for children. *Archives of Clinical Neuropsychology*, *19*, 725-743.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it. *Psychological Methods*, *5*, 64-86.

Hoyt, W. T. (2002). Bias in participant ratings of psychotherapy process: An initial generalizability study. *Journal of Counseling Psychology*, *49*, 35-46.

Hoyt, W. T., & Melby, J. N. (1999). Dependability of measurement in counseling psychology: An introduction to generalizability theory. *The Counseling Psychologist*, *27*, 325-352.

Hoza, B., Gerdes, A. C., Mrug, S., Hinshaw, S P., Bukowski, W. M., Gold, J. A., Arnold, L. E., Abikoff, H. B., Conners, C. K., Elliott, G. R., Greenhill, L. L., Hechtman, L., Jensen, P. S.,

Kraemer, H. C., March, J. S., Newcorn, J. H., Severe, J. B., Swanson, J. M., Vitiello, B., Wells, K. C., & Wigal, T. (2005). Peer-assessed outcomes in the multipmodal treatment study of children with Attention Deficit Hyperactivity Disorder. *Journal of Clinical Child and Adolescent Psychiatry*, *34*, 74-86.

Hsieh, P., Acee, T., Chung, W., Hsieh, Y., Kim, H., Thomas, G. D., You, J., & Levin, J. R. (2005). Is education intervention research on the decline? *Journal of Educational Psychology*, *97*, 523-529.

Hunter, L. (2003). School psychology: a public health framework III. Managing disruptive behavior in schools: the value of public health and evidenced-based perspective. *Journal of School Psychology, 41,* 39-59.

Hynd, G. W., Voeller, K. K., Hern, K. L., & Marshall, R. M. (1991). Neurobiological basis of attention-deficit hyperactivity disorder (ADHD). *School Psychology Review*, *20*, 174-186.

Jackson, D. A., & King, A. R. (2004). Gender differences in the effects of oppositional behavior on teacher ratings of ADHD symptoms. *Journal of Abnormal Child Psychology*, *32*, 215-224.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12-19.

Jenson, P. S., Hinshaw, S. P., Kraemer, H. C., Lenora, N., Newcorn, J. H., Abikoff, H. B., March, J. S., Arnold, L. E., Cantwell, D. P., Conners, C. K., Elliott, G. R., Greenhill, L. L., Hechtman, L., Hoza, B., Pelham, W. E., Severe, J. B., Swanson, J. M., Wells, K. C., Wigal, T., & Vitiello, B. (2001). ADHD comorbidity findings from the MTA study: Comparing comorbid subgroups. *Journal of the American Academy of Child and Adolescent Psychiatry*, *40*, 147-158.

Jensen, P. S., Kettle, L., Roper, M.T., Sloan, M.T., Dulcan, M., Hoven, C., Bird, H.R., Baurmeister, J.J., & Payne.J.D. (1999). Are stimulants

overprescribed? Treatment of ADHD in four U.S. communities. *Journal of the American Academy of Child and Adolescent Psychiatry, 38,* 797-804.

Jensen, P. S., Hoagwood, K., Roper, M., Arnold, L. Eugene, Odbert, C., Crowe, M., Molina, B.S.G., Hechtman, L., Hinshaw, S.P., Hoza, B., Newcorn, J., Swanson, J., & Wells, K. (in press). The Services for Children and Adolescents Parent Interview (SCAPI): Development and Performance Characteristics. *Journal of the American Academy of Child and Adolescent Psychiatry*.

Kamphaus, R. W., & Frick, P. J. (1996). *Clinical assessment of child and adolescent personality and behavior*. Boston: Allyn & Bacon, Inc.

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., Williamson, D., & Ryan, N. (1997). Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): Initial Reliability and Validity Data. *Journal of the American Academy of Child and Adolescent Psychiatry*, *36*(7), 980-988.

Kaufman, A.S., & Kaufman, N.L. (1990). *Kaufman Brief Intelligence Test Manual*. Circle Pines, MN: American Guidance Service.

Kazdin, A. E. (1990). Premature termination from treatment among children referred for antisocial behavior. *Journal of Child Psychology & Psychiatry & Allied Disciplines*, *31*, 415-425.

Kazdin, A.E. (1993). Psychotherapy for children and adolescents: Current progress and future research directions. *American Psychologist, 48,* 644-657.

Kazdin, A. E. & Weisz, J. R. (1998). Identifying and developing empirically supported child and adolescent treatments. *Journal of Consulting and Clinical Psychology, 66*, 19-36.

Kendall, P.C. (1998). Directing misperceptions: Researching the issues facing manual-based treatments. *Clinical Psychology: Science and Practice*, *5*, 396-399.

Klassen, A. F., Miller, A., & Fine, S. (2004). Health-related quality of life in children and adolescents who have a diagnosis of attention-deficit/hyperactivity disorder. *Pediatrics*, *114*, 541-547.

Kline, R. B. (2005). *Principles and practices of structural equation modeling* (2nd edition). New York: The Guilford Press.

Kokkinos, C. M., Panayiotou, G., & Davazoglou, A. M. (2005). Correlates of teacher appraisals of student behaviors. *Psychology in the Schools*, *42*, 79-89.

Kokkinos, C. M., Panayiotou, G., & Davazoglou, A. M. (2004). Perceived seriousness of pupils' undesirable behaviours: The student teacher's perspective. *Educational Psychology*, *24*, 109-120.

Kos, J. M., Richdale, A. L., & Jackson, M. S. (2004). Knowledge about attention-deficit/hyperactivity disorder: A comparison of in-service and preservice teachers. *Psychology in the Schools*, *41*, 517-526.

Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *American Journal of Psychiatry*, *160*, 1566-1577.

Krain, A. L., & Costellanos, F. X. (2006). Brain development and ADHD. *Clinical Psychology Review*, *26*, 433-444.

Lahey, B. B., Applegate, B., McBurnett, K., Biederman, J., Greenhill, L., Hynd, G. W., Barkley, R. A., Newcorn, J., Jensen, P., Richters, J., Garfinkel, B., Kerdyk, L., Frick, P. J., Ollendick, T., Perez, D., Hart, E. L., Waldman, I., & Shaffer, D. (1994). DSM-IV field trials for attention deficit hyperactivity disorder in children and adolescents. *American Journal of Psychiatry*, *151*, 1673-1685.

Laird, N.M. (1988) Missing data in longitudinal studies. *Statistics in Medicine,* 7, 305-315.

Landau, S., & Moore, L. A. (1991). Social skill deficits in children with attention-deficit hyperactivity disorder. *School Psychology Review*, *20*, 235-251.

Lee, S. S., & Hinshaw, S. P. (2004). Severity of adolescent delinquency among boys with and without attention deficit hyperactivity disorder: Predictions from early antisocial behavior and peer status. *Journal of Clinical Child and Adolescent Psychology*, *33*, 705-716.

Lambert, N. M., & Hartsough, C. S. (1998). Prospective study of tobacco smoking and substance dependencies among samples of ADHD and non-ADHD participants. *Journal of Learning Disabilities*, *31*, 533-544.

Landau, S., & Moore, L.A. (1991). Social skill deficits in children with attention-deficit hyperactivity disorder. *School Psychology Review*, *20*, 235-251.

Leckman, J.F., Sholomskas, D., Thompson, W.D., Belanger, A., & Weissman, M.M. (1982). Best estimate of lifetime psychiatric diagnosis: A methodological study. *Archives of General Psychiatry*, *39*, 879-883.

Levine, M. (2002). *A mind at a time: America's top learning expert shows how every child can succeed.* New York : Simon & Schuster.

Lipsey, M. W. & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*, 1181-1209.

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

Loo, S. K., & Barkley, R. A. (2005). Clinical utility of EEG in attention deficit hyperactivity disorder. *Applied Neuropsychology*, *12*, 64-76.

Marshal, M. P., Molina, B. S., & Pelham, W. E. (2003). Childhood ADHD and adolescent substance use: An examination of deviant peer group affiliation as a

risk factor. *Psychology of Addictive Behaviors*, *17*, 293-302.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30-46.

Meichenbaum, D., Gnagy, B., Flammer, L., Molina, B., & Pelham, W. E., Jr. (2001, April). *Why stop success? Exploration of long-term use of medication in a clinical ADHD sample from childhood through young adulthood*. Paper presented at the Eighth Florida Conference on Child Health Psychology, Gainesville, Florida.

Meichenbaum, D.L., Pelham, W., Gnagy, E., Smith, B.H., & Bukstein, O. (1999). *Effects of methylphenidate on parent-adolescent interactions in families with an ADHD teenager*. Unpublished manuscript.

Michael, R. I., Klorman, R., Salzman, L. F., Borgstedt, A. D., & Dainer, K. B. (1981). Normalizing effects of methylphenidate on hyperactive children's vigilance performance and evoked potentials. *Psychophysiology*, *18*, 665-677.

Mick, E., Biederman, J., & Faraone, S. V. (1998). Comment on Lambert and Hartsough (1998). *Journal of Learning Disabilities*, *33*, 314.

Mitsis, E. M., McKay, K. E., Schulz, K. P., Newcorn, J. H., & Halperin, J. M. (2000). Parent-teacher concordance for DSM-IV attention-deficit disorder in a clinic-referred sample. *Journal of the American Academy of Child and Adolescent Psychiatry*, *39*, 308-313.

Molina, B. & Pelham, W. E. (1999). Alcohol and other substance use and abuse in ADHD adolescents: Patterns of use compared to controls and prediction from childhood. In W. E. Pelham (Chair), *Adolescent substance use and abuse: Prediction from childhood psychopathology and personality and mediating pathways*. Annual meeting of the American Psychological Association, Boston, August.

Molina, B., & Pelham, W. E. (1999). Explaining ADHD risk for adolescent substance use and abuse: Exploring

the intrapersonal domain of personality, attitudes, and beliefs. In B. Molina (Chair), *Longitudinal studies of substance use and abuse into adolescence: Childhood psychopathology and mediating pathways.* Symposium at the Biennial Meeting of the International Society for Research on Child and Adolescent Psychopathology, Barcelona, Spain, June.

Molina, B., & Pelham, W. E. (2003). Childhood predictors of adolescent substance use in a longitudinal study of children with ADHD. *Journal of Abnormal Psychology*, *112*, 497-507.

Molina, B., Pelham, W.E., Blumenthal, J., & Galiszewaski, E. (1998). Agreement among teachers' behavior ratings of adolescents with a childhood history of attention deficit hyperactivity disorder. *Journal of Clinical Child Psychology*, *27*, 330-339.

Molina, B., Smith, B., & Pelham, W. E. (2001). Factor structure and criterion validity of secondary school teacher ratings of ADHD and ODD. *Journal Abnormal Child Psychology, 29*(1), 71-82.

Moline, S., & Frankenberger, W. (2001). Use of stimulant medication for treatment of attention-deficit/hyperactivity disorder: A survey of middle and high school students' attitudes. *Psychology in the Schools*, *38*, 569-584.

Myers, J.K. & Bean, L.L. (1968). *A Decade Later: A Follow-up of Social Class and Mental Illness.* New York: Wiley.

National Institute of Mental Health Consensus Forming Panel. (2000). National Institutes of Health Consensus Development Conference Statement: Diagnosis and treatment of attention-deficit/hyperactivity disorder (ADHD). *Journal of the American Academy of Child and Adolescent Psychiatry, 39,* 182-193.

O'Brien, N., O'Brien, S., Packman, A., & Onslow, M. (2003). Generalizability theory I: Assessing reliability of observational data in the communication sciences. *Journal of Speech, Language, and Hearing Research*, *46*, 711-717.

Ohan, J. L., & Johnson, C. (2002). Are the performance overestimates given by boys with ADHD self-protective? *Journal of Clinical Child Psychology*, *31*, 230-241.

Olfson, M., Marcus, S.C., Weissman, M.M., & Jenson, P.S. (2002). National trends in the use of psychotropic medications by children. *Journal of the American Academy of Child & Adolescent Psychiatry*, *41*, 514-521.

Olympia, D., & Larsen, J. (2005). Functional behavioral assessment: An emerging component of best school practices for ADHD. *The ADHD Report*, *13*(5), 1-5.

Osborne, J., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation*, 8(2). Retrieved November 15, 2004 from http://PAREonline.net/getvn.asp?v=8&n=2

Ostrander, R., Weinfurt, K.P., Yarnold, P.R., & August, G.J. (1998). Diagnosing attention deficit disorders with the Behavioral Assessment System for Children and the Child Behavior Checklist: Test and construct validity analyses using optimal discriminant classification trees. *Journal of Consulting and Clinical Psychology, 66,* 660-672.

Pallant, J. (2001). *SPSS survival manual*. Philadelphia: Open University Press.

Peele, P.B., Lave, J.R., & Kelleher, K. J. (2002). Exclusions and limitations in children's behavioral health care coverage. *Psychiatric Services*, *53*(5), 591-594.

Pelham, W. E., Fabiano, G. A., & Massetti, G. M. (2005). Evidence-based assessment of attention deficit hyperactivity disorder in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, *34*, 449-476.

Pelham, W.E., Gnagy, E., Breiner, A.R., Hoza, B., Hinshaw, S.P. Swanson, J.M., et al. (2000). Behavioral versus behavioral and pharmacological treatment in ADHD children attending a summer

treatment program. *Journal of Abnormal Child Psychology*, *28*, 507-525.

Pelham, W.E., Gnagy, E.M., Greenslade, K.E., & Milich, R. (1992). Teacher ratings of DSM-III symptoms for the disruptive behavior disorders. *Journal of the American Academy of Child and Adolescent Psychiatry*, *31*, 210-218.

Pelham, W.E.; Gnagy, E.; Waschbusch, D.A.; Willoughby, M.; Palmer, A.; Whichard, M.; Hall, S.; Shaffer, S.; Myers, D.; Billheimer, S. (1996, January). *A practical impairment scale for childhood disorders: Normative data and an application to ADHD.* Poster session presented at the annual meeting of the Society for Research in Child and Adolescent Psychopathology.

Pelham, W.E., Jr., & Hoza, B. (1996). Psychosocial treatments for child and adolescent disorders: Empirically based strategies for clinical practice. In E.D. Hibbs & P.S. Jensen (Eds.), *Intensive Treatment: A Summer Treatment Program for Children with ADHD.* (pp. 311-340). Washington: American Psychological Association.

Pelham, W. E., Jr., Molina, B. S. G., Meichenbaum, D., Gnagy, E., & Greenhouse, J. (2003). *Stimulant effects on longterm outcomes in ADHD individuals.* Presentation at NCDEU, Boca Raton, May 27, 2003.

Pelham, W., Wheeler, T., & Chronis, A. (1998). Empirically supported psychosocial treatments for attention deficit hyperactivity disorder. *Journal of Clinical Child Psychology*, *27*, 190-205.

Pfiffner, L.J. & McBurnett, K. (1997). Social skills training with parent generalization: Treatment effects for children with attention deficit disorder. *Journal of Consulting and Clinical Psychology, 65*, 749-757.

Power, T. J., Andrews, T. J., Eiraldi, R. B., Doherty, B. J., Ikeda, M. J., DuPaul, G. J., & Landau, S. (1998). Evaluating attention deficit hyperactivity disorder using multiple informants: The incremental utility of combining teacher with parent reports. *Psychological Assessment*, *10*, 250-260.

Power, T. J., & DuPaul, G. J. (1996). Attention-defiit hyperactivity disorder: The reemergence of subtypes. *School Psychology Review*, *25*, 284-296.

President's New Freedom Commission on Mental Health.(2003). *Achieving the promise: Transforming mental health care in America. Final report* (U.S. DHHS Pub. No. SMA-03-3832). Rockville, MD: U.S. Department of Health and Human Services.

Preston, A. S., Fennell, E. B., Bussing, R. (2005). Utility of a CPT in diagnosing ADHD among a representative sample of high risk children: A cautionary study. *Child Neuropsychology, 11,* 459-469.

Prinz, R.J., Foster, S.L., Kent, R.N, & O'Leary, K.D. (1979). Multivariate assessment of conflict in distressed and nondistressed mother-adolescent dyads. *Journal of Applied Behavior Analysis, 12*, 691-700.

Pronin, E., Luin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*, 369-381.

Prout, H.T., (1999). Counseling and psychotherpay with children and adolescents: An overview. In H.T. Prout & D.T. Brown (Eds.). *Counseling and psychotherapy with children and adolescents: Theory and practice for school and clinical settings* (3rd ed.). New York, NY: John Wiley & Sons, Inc.

Quinn, M.M., Kavale, K.A., Mathur, S.R., Rutherford, R.B., & Forness, S.R. (1999). A meta-analysis of social skill interventions for students with emotional or behavioral disorders. *Journal of Emotional and Behavioral Disorders, 7,* 54-64.

Raudenbush, S. W., & Liu, X.-F. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, *6*, 387-401.

Reid, R., DuPaul, G. J., Power, T. J., Anastopoulos, A. D., Rogers-Adkinson, D., Noll, M., & Riccio, C.

(1998). Assessing culturally different students for attention deficit hyperactivity disorder using behavior rating scales. *Journal of Abnormal Child Psychology*, *26*, 187-198.

Reid, R., Maag, J. W. (1994). How many fidgets in a pretty much: A critique of behavior rating scales for identifying students with ADHD. *Journal of School Psychology*, *32*, 339-354.

Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavior assessment system for children (BASC)*. Circle Pines, MN: American Guidance Services.

Rieppi, R., Greenhill, L.L., Ford, R.E., Chuang, S., Wu, M., Davies, M., et al. (2002). Socioeconomic status as a moderator of ADHD treatment outcomes. *Journal of the American Academy of Child and Adolescent Psychiatry*, *41*, 269-277.

Robin, A.L. (1990). Training families with ADHD adolescents. In R.A. Barkley (Ed.), *Attention Deficit Hyperactivity Disorder. A Handbook for Diagnosis and Treatment.* (pp. 462-497). New York: The Guilford Press.

Robin, A.L. (1998). *ADHD in Adolescents: Diagnosis and Treatment.* New York: The Guilford Press.

Robin, A. L., & Foster, S. L. (1989). *Negotiating parent-adolescent conflict: A behavioral-family systems approach*. NY: Guilford Press.

Robison, L. M., Skaer, T. L., Sclar, D. A., and Galin, R. S. (2002). Is attention deficit hyperactivity disorder increasing among girls in the US? *CNS Drugs*, *16*, 129-137.

Romano, E., Tremblay, Vitrano, Zoccolillo, M., & Pagani, L., (2001). Prevalence of psychiatric diagnoses and the role of perceived impairment: Findings from an adolescent community sample. *Journal of Clinical Psychology and Psychiatry, 42*, 451-461.

Romine, C. B., Lee, D., Wolfe, M. E., Homack, S., George, C., & Riccio, C. A. (2004). Wisconsin Card Sorting Test with children: A meta-analytic study of

sensitivity and specificity. *Archives of Clinical Neuropsychology*, *19*, 1027-1041.

Rounsaville, B. J., Carroll, K. M., & Onken, L. S. (2001). Methodological diversity and theory in the stage model: Reply to Kazdin. *Clinical Psychology: Science and Practice, 8,* 152-154.

Sandoval, J., (1994). [Review of the Behavior Assessment System for Children]. *Mental Measurements Yearbook*. Circle Pines, MN: American Guidance Service, Inc.

Schamer, L. A., & Jackson, M. J. (1996). Coping with stress: Common sense about teacher burnout. *Education Canada*, *36*, 28-31.

Schultz, B. K., & Cobb, H. (2005). Behavioral consultation for adolescents with ADHD: Lessons learned in the Challenging Horizons Program. *Report on Emotional & Behavioral Disorders in Youth*, *5*, 91-99.

Schultz, B. K., Reisweber, J., & Cobb, H. (in press). Mental health consultation in secondary schools. In S. Evans, M. Weist, & Z. Serpell (Eds.), *Advances in school-based mental health interventions* (Vol. 2). New York: Civic Research Institute.

Sciutto, M. J., Nolfi, C. J., & Bluhm, C. (2004). Effects of child gender and symptom type on referrals for ADHD by elementary school teachers. *Journal of Emotional and Behavioral Disorders*, *12*, 247-253.

Seidman, L. J. (2006). Neuropsychological functioning in people with ADHD across the lifespan. *Clinical Psychology Review*, *26*, 466-485.

Seidman, L. J., Valera, E. M., & Makris, N. (2005). Structural brain imaging of attention-deficit/hyperactivity disorder. *Biological Psychiatry*, *57*, 1263-1272.

Sergeant, J. (2004). EUNETHYDIS – Searching for valid aetiological candidates of attention-deficit hyperactivity disorder or hyperkinetic disorder. *European Child & Adolescent Psychiatry*, *13*, i43-i49.

Sergeant, J. A., Geurts, H., & Oosterlaan, J. (2002). How specific is a deficit of executive functioning for Attention-Deficit/Hyperactivity Disorder? *Behavioural Brain Research*, *130*, 3-28.

Shapiro, E. S., & Heick, P. F. (2004). School psychologist assessment practices in the evaluation of students referred for social/behavioral/emotional problems. *Psychology in the Schools*, *41*, 551-561.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A primer*. Thousand Oaks, California: SAGE Publications.

Sherman, M., & Hertzig, M.E. (1991). Prescribing practices of Ritalin: The Suffolk County, New York study. In L.L. Greenhill & B.B. Osman (Eds.), *Ritalin: Theory and Patient Management* (pp. 187-193). New York: Mary Ann Liebert Inc.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420-428.

Simonoff, E., Pickles, A., Hewitt, J., Silberg, J., Rutter, M., Loeber, R., Meyer, J., Neale, M., & Eaves, L. (1995). Multiple raters of disruptive child behavior: Using a genetic strategy to examine shared views and bias. *Behavior Genetics*, *25*, 311-326.

Smith, B.H., Molina, B.S. & Pelham, W.E. (2002). The clinically meaningful link between alcohol use and attention deficit hyperactivity disorder. *Alcohol Research and Health, 26*, 122–129.

Smith, B.H., Pelham, W.E., Evans, S.W., Gnagy, E., Molina, B.S.G., Bukstein, O., Greiner, A.R., Myak, C., Presnell, M., & Willoughby, M. (1998). Dosage effects of methylphenidate on the social behavior of adolescents diagnosed with attention-deficit hyperactivity disorder. *Experimental and Clinical Pscyhopharmacology, 6,* 187-204.

Smith, B. H., Pelham, W. E., Gnagy, E., Molina, B. S., & Evans, S. W. (2000). The reliability, validity, and unique contributions of self-report by adolescents receiving treatment for attention-

deficit/hyperactivity disorder. *Journal of Consulting and Clinical Psychology*, *68*(3), 489-499.

Smith, B. H., Pelham, W. E., Gnagy, E., & Yudell, R. S. (1998). Equivalent effects of stimulant treatment for Attention-Deficit Hyperactivity Disorder during childhood and adolescence. *Journal of the American Academy of Child and Adolescent Psychiatry, 37,* 314-321.

Smith, B. H., Waschbusch, D. A., Willoughby, M. T., & Evans, S. W. (2000). Treatments for Adolescents with Attention-Deficit Hyperactivity Disorder (ADHD): A Review of Efficacy, Safety, and Practicality. *Clinical Child and Family Psychology Review, 3,* 243-268.

Sonuga-Barke, E. J., Minocha, K., Taylor, E. A., & Sandberg, S. (1993). Inter-ethnic bias in teachers' ratings of childhood hyperactivity. *British Journal of Developmental Psychology*, *11*, 187-200.

Spencer, T., Biederman, J., & Wilens, T. (1998). Growth deficits in children with attention deficit hyperactivity disorder. *Pediatrics*, 102, 501-506.

Spencer, T., Biederman, J., Wilens, T., Harding, M., O'Donnell, D., & Griffin, S. (1996). Pharmacotherapy of Attention-Deficit Hyperactivity Disorder across the life cycle. *Journal of the American Academy of Child and Adolescent Psychiatry, 35,* 409-432.

Staller, J. A., (2006). Diagnostic profiles in outpatient child psychiatry. *American Journal of Orthopsychiatry*, *76*, 98-102.

Stevens, J., Quittner, A. L., Abikoff, H. (1998). Factors influencing elementary school teachers' ratings of ADHD and ODD Behaviors. *Journal of Clinical Child Psychology*, *27*, 406-414.

Stewart, K. G., & McLaughlin, T. F. (1992). Self-recording: Effects on reducing off-task behavior with a high school student with attention-deficit hyperactivity disorder. *Child & Family Behavior Therapy, 14*(3), 53-59

Stokes, T.F., & Baer, D.M. (1977). An implicit technology of generalization. *Journal of Applied Behavior Analysis, 10,* 349-367.

Stokes, T.F., & Osnes, P.G. (1989). An operant pursuit of generalization. *Behavior Therapy, 20,* 337-355.

Stormont, M. (2001). Social outcomes of children with AD/HD: Contributing factors and implications for practice. *Psychology in the Schools*, *38*, 521-531.

Strauss, M. E., Thompson, P., Adams, N. L., Redline, S., Burant, C. (2000). Evaluation of a model of attention with confirmatory factor analysis. *Neuropsychology*, *14*, 201-208.

Swanson, J. (2003). Compliance with stimulants for attention deficit/hyperactivity disorder: Issues and approaches for improvement. *CNS Drugs*, *17*, 117-131.

Swanson, J., Greenhill, L., Wigal, T., Kollins, S., Stehli, A., Davies, M., Chuang, S., Vitiello, B., Skrobala, A., Posner, K., McGough, J., Riddle, M., Ghuman, J., Cunningham, C., & Wigal, S. (2006). Stimulant-related reductions of growth rates in the PATS. *Journal of the American Academy of Child and Adolescent Psychiatry*, *45*, 1304-1313.

Swanson, J., McBurnett, K., Christian, D.L. & Wigal, T. (1995). Stimulant medications and the treatment of children with ADHD. In T.H. Ollendick & R.J. Prinz (Eds.), *Advances in Clinical Child Psychology*, Vol. 17 (pp. 265-322). New York: Plenum Press.

Swanson, J. M., Kraemer, H. C., Hinshaw, S. P., Arnold, L. E., Conners, C. K., Abikoff, H. B., Clevenger, W., Davies, M., Elliott, G. R., Greenhill, L. L., Hechtman, L., Hoza, B., Jensen, P. S., March, J. S., Newcorn, J. H., Owens, E. B., Pelham, W. E., Schiller, E., Severe, J. B., Simpson. S., Vitiello, B., Wells, K., Wigal, T., & Wu, M. (2001). Clinical relevance of the primary findings of the MTA: Success rates based on severity of ADHD and ODD symptoms at the end of treatment. *Journal of the American Academy of Child and Adolescent Psychiatry*, *40*, 168-179.

Swensen, A.R., Birnbaum, H.G., Secnik, K., Marynchenko,

M., Greenberg, P., & Claxton, A. (2003). Attention-deficit/hyperactivity disorder: Increased costs for patients and their families. *Journal of the American Academy of Child & Adolescent Psychiatry*, *42*, 1415-1423.

Szatmari, P., Offord, D., & Boyle, M. (1989). Correlates, associated impairments and patterns of service utilization of children with attention deficit disorder: Findings from the Ontario Child Health Study. *Journal of Child and Adolescent Psychiatry*, *30*, 205-217.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th edition). Boston: Allyn & Bacon.

Taffel, R. (2001). *The Second Family: How Adolescent Power is Challenging the American Family*. New York: St. Martin's Press.

Taffel, R. (2005). *Breaking Through to Teens: A New Psychotherapy for the New Adolescence*. New York: The Guilford Press.

Taylor, E. (1999). Development of clinical services for attention-deficit/hyperactivity disorder. *Archives of General Psychiatry*, *56*, 1097-1099.

Taylor, T.K., Eddy, J.M., & Biglan, A. (1999). Interpersonal skills training to reduce aggressive and delinquent behavior: Limited evidence and the need for an evidence-based system of care. *Clinical Child and Family Psychology Review, 2,* 169-182.

Teeter, P. A. (1998). *Interventions for ADHD: Treatment in developmental context*. New York, NY: The Guilford Press.

Teicher, M. H., Ito, Y., Glod, C., & Barber, N. I. (1996). Objective measurement of hyperactivity and attentional problems in ADHD. *Journal of the American Academy of Child and Adolescent Psychiatry*, *35*, 334-343.

Thiruchelvam, D., Charach, A., & Schachar, R. J. (2001). Moderators and mediators of long-term adherence to stimulant treatment in children with ADHD. *Journal*

of the American Academy of Child and Adolescent Psychiatry, 40, 922-928.

The MTA Cooperative Group (1999a). A 14-month randomized clinical trial of treatment strategies for attention-deficit/hyperactivity disorder. *Archives of General Psychiatry*, *56*, 1073-1086.

The MTA Cooperative Group (1999b). Moderators and mediators of treatment response for children with attention-deficit/hyperactivity disorder. *Archives of General Psychiatry*, *56*, 1088-1096.

The MTA Cooperative Group (2004a). National Institute of Mental Health Multimodal Treatment Study of ADHD follow-up: 24-month outcomes of treatment strategies for attention-deficit/hyperactivity disorder. *Pediatrics, 113,* 754-761.

The MTA Cooperative Group (2004b). National Institute of Mental Health Multimodal Treatment Study of ADHD follow-up: Changes in effectiveness and growth after the end of treatment. *Pediatrics, 113,* 762-769.

The Psychological Corporation. (2002). *Wechsler Individual Achievement Test – Second Edition: Examiner's Manual.* San Antonio: Harcourt Brace & Company.

Tucker, P. (1999). Attention-Deficit/Hyperactivity Disorder in the drug and alcohol clinic. *Drug and Alcohol Review, 18*, 337-344.

United States Department of Education, Office of Special Education and Rehabilitation Services, Office of Special Education Programs. (2005). *25th Annual Report to Congress on the Implementation of the Individuals with Disabilities Education Act* (Vol. 1). Washington, DC: Author.

United States Drug Enforcement Agency (1999). *Yearly Aggregate Production Quotas (1990-1999).* Washington, DC: Office of Public Affairs, Drug Enforcement Administration.

Vitelli, R. (1998). Childhood disruptive behavior disorders and adult psychopathy. *American Journal of Forensic Psychology*, *16*, 29-37.

Waldman, I. D., & Gizer, I. R. (2006). The genetics of attention deficit hyperactivity disorder. *Clinical Psychology Review*, *26*, 396-432.

Waldron, H.B., Slesnick, N., Brody, J.L., Turner, C.W. & Peterson, T.R. (2001). Treatment outcomes for adolescent substance abuse at 4- and 7-month assessments. *Journal of Consulting and Clinical Psychology*, *69*, 802-813.

Waschbusch, D. A. (2002). A meta-analytic examination of comorbid hyperactive-impulsive-attention problems and conduct problems. *Psychological Bulletin*, *128*, 118-150.

Waschbusch, D. A., & King, S. (2006). Should sex-specific norms be used to assess attention-deficit/hyperactivity disorder or oppositional defiant disorder? *Journal of Consulting and Clinical Psychology*, *74*, 179-185.

Webb. J. T., Amend, E. R., Webb, N. E., Goerss, J., Beljan, P., & Olenchak, F. R. (2005). *Misdiagnosis and dual diagnosis of gifted children and adults: ADHD, Bipolar, OCD, Asperger's, Depression, and other disorders.* Scottsdale, AZ: Great Potential Press, Inc.

Webster-Stratton, C. (1993). Strategies for helping early school-aged children with oppositional defiant and conduct disorders: The importance of home-school partnerships. *School Psychology Review*, *22*, 437-457.

Wechsler, D. (2003). *WISC-IV Technical and Interpretive Manual.* New York: The Psychological Corporation.

Weis, R., & Totten, S. J. (2004). Ecological validity of the Conners' Continuous Performance Test II in a school-based sample. *Journal of Psychoeducational Assessment*, *22*, 47-61.

Weist, M.D. (2005). Fulfilling the promise of school-based mental health: Moving toward a public mental health promotion approach. *Journal of Abnormal Child Psychology*, *33*, 735-741.

Weist, M.D. & Evans, S.W. (2005). Expanded school mental health: Challenges and opportunities in an emerging field. *Journal of Youth and Adolescence*, *34*, 3-6.

Weisz, J. R. (2000). Lab-clinic differences and what we can do about them: I. The clinic-based treatment development model. *Clinical Child Psychology Newsletter, 15*, 1-10.

Weisz, J. R. & Hawley, K. M. (2002). Developmental factors in the treatment of adolescents. *Journal of Consulting and Clinical Psychology, 70,* 21-43.

Weisz, J.R., Jensen, A.L., & McLeod, B.D. (in press). Development and Dissemination of Child and Adolescent Psychotherapies: Milestones, Methods, and a New Deployment-Focused Model. In E.D. Hibbs & P.S. Jensen (Eds.). *Psychosocial Treatments for child and adolescent disorders: Empirically-based approaches*, 2nd edition. Washington, DC: American Psychological Association.

Wells, K. C., Pelham, W. E., Kotkin, R. A., Hoza, B., Abikoff, H. B., Abramowitz, A., Arnold, L. E., Cantwell, D. P., Conners, C. K., Carmen, R. D., Elliott, G., Greenhill, L. L., Hechtman, L., Hibbs, E., Hinshaw, S. P., Jensen, P. S., March, J. S., Swanson, J. M., & Schiller, E. (2000). Psychosocial treatment strategies in the MTA study: Rationale, methods, and critical issues in design and implementation. *Journal of Abnormal Child Psychology*, *28*, 483-505.

Wheeler, J., & Carlson, C. L. (1994). The social functioning of children with ADD with hyperactivity and ADD without hyperactivity: A comparison of their peer relations and social deficits. *Journal of Emotional and Behavioral Disorders*, *2*, 2-13.

White, L. C., Barbour, K., Schill, T., Vodra, A., Garrett, A., Schultz, B. K., & Evans, S. W. (2005, August). *Academic Underachievement and Symptom Severity in Adolescents with ADHD.* Poster presented at the 113th Annual American Psychological Association Conference, Washington, D.C.

White, J. L., Moffitt, T. E., Caspi, A., Bartusch, D. J., Needles, D. J., & Stouthamer-Loeber, M. (1994).

Measuring impulsivity and examining its relationship to delinquency. *Journal of Abnormal Psychology*, *103*, 192-205.

Wilens, T. E., Faraone, S. V., Biederman, J., Gunawardene, S. (2003). Does stimulant therapy of Attention-Deficit/Hyperactivity Disorder beget later substance abuse? A meta-analytic review of the literature. *Pediatrics*, *111*, 179-186.

Wilens, T., Pelham, W., Stein, M., Conners, C.K., Abikoff, H., Atkins, M., August, G., Greenhill, L., McBurnett, K., Palumbo, D.,; Swanson, J., & Wolraich, M. (2003). ADHD treatment with once-daily OROS methylphenidate: Interim 12-month results from a long-term open-label study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *42*, 424-434.

Wilson, G.T. (1998). Manual-based treatment and clinical practice. *Clinical Psychology: Science and Practice*, *5*, 396-399.

Wolraich, M. L., Lambert, E. W., Baumgaertel, A., Garcia-Tornel, S., Feurer, I. D. , Bickman, L., & Doffing, M. A. (2003). Teachers' screening for attention deficit/hyperactivity disorder: Comparing multinational samples on teacher ratings of ADHD. *Journal of Abnormal Child Psychology*, *31*, 445-455.

Wolraich, M. L., Lambert, E. W., Bickman, L., Simmons, T., Doffing, M. A., & Worley, K. A. (2004). Assessing the impact of parent and teacher agreement on diagnosing Attention-Deficit Hyperactivity Disorder. *Developmental and Behavioral Pediatrics*, *25*, 41-47.

APPENDIXES

Appendix A

Site Coordinator Letters of Permission

Site Coordinator Letter of Permission

Institutional Review Board
Office of Sponsored Programs,
MSC 5728, Medical Arts West, Suite 26
James Madison University
Harrisonburg, VA 22807

Dear Institutional Review Board,

I hereby agree to allow Dr. Steve Evans James Madison University to continue, per our verbal face to face agreement February 2004, to conduct his research at Elkton Middle School. I understand that this is a six-year research project funded by the Virginia Tobacco Settlement Foundation (VTSF) to examine whether improved quality and coordination of care for adolescents with ADHD reduces cigarette smoking and improves academic and social outcomes in this youth population. The overall project goal is to assess the extent to which the Challenging Horizons Program model of care significantly improves outcomes among adolescents with ADHD over and above what would be expected with the type of care they typically receive.

I am aware that five middle schools are involved-- three in Rockingham County and two in Augusta County and each has been randomly assigned to be either a control or treatment school by way of a coin toss. I am aware that Elkton Middle school was recruited as a result of difficulties the project was having meeting its recruitment goals for control school participants. As such, my staff and I will continue to aid recruitment efforts to fill control participant spaces that you have for students entering 8th grade in the Fall (first cohort) and student entering 7th grade in the Fall (second cohort) by providing mailing labels for students that you can use to mail information about the study to their parents. We are also glad to assist by putting notices of this opportunity for families in newsletters and other communications sent from the school.

In addition, I understand that as a control school, we will continue to aid your efforts to evaluate the Challenging Horizons Program by having our teachers and other school staff complete brief rating scales on a regular basis for the students involved in the project. I am aware that training sessions may be held for teachers that have not had it previously to ensure the accurate completion of rating scales. I understand further that, as you have in the past, you will continue to compensate teachers and other school staff for attending training sessions and that at the end of the year you will compensate them for their efforts completing the aforementioned rating scales.

I look forward to our continued collaboration.

Sincerely,

Laura Evy, *Principal*
Elkton Middle School

Site Coordinator Letter of Permission

Institutional Review Board
Office of Sponsored Programs,
MSC 5728, Medical Arts West, Suite 26
James Madison University
Harrisonburg, VA 22807

Dear Institutional Review Board,

I hereby agree to allow Dr. Steve Evans James Madison University to continue, per the verbal face to face agreement made with the previous principal, Mr. Fred Babbitt in April of 2003, to conduct his research at Hillyard Middle School. I understand that this is a six-year research project funded by the Virginia Tobacco Settlement Foundation (VTSF) to examine whether improved quality and coordination of care for adolescents with ADHD reduces cigarette smoking and improves academic and social outcomes in this youth population. The overall project goal is to assess the extent to which the Challenging Horizons Program model of care significantly improves outcomes among adolescents with ADHD over and above what would be expected with the type of care they typically receive.

I am aware that five middle schools are involved-- three in Rockingham County and two in Augusta County and each has been randomly assigned to be either a control or treatment school by way of a coin toss. I am aware that Hillyard Middle school was randomly assigned to be a control school and as such we will continue to aid recruitment efforts to reach the goal of 10 students for the first cohort (entering 8th grade in the Fall) and 10 students for the second cohort (entering 7th grade in the Fall) students by providing mailing labels for students that you can use to mail information about the study to their parents. We are also glad to assist, by putting notices of this opportunity for families in newsletters and other communications sent from the school.

I understand that as a control school, we will continue to aid your efforts to evaluate the Challenging Horizons Program by having our teachers and other school staff complete brief rating scales on a regular basis for the students involved in the project. I am aware that training sessions may be held for teachers that have not had it previously to ensure the accurate completion of rating scales. I understand further that, as you have in the past, you will continue to compensate teachers and other school staff for attending training sessions and that at the end of the year you will compensate them for their efforts completing the aforementioned rating scales.

I look forward to our continued collaboration.

Sincerely,

Doug Aldefer, *Principal*
Hillyard Middle School

Site Coordinator Letter of Permission

Institutional Review Board
Office of Sponsored Programs,
MSC 5728, Medical Arts West, Suite 26
James Madison University
Harrisonburg, VA 22807

Dear Institutional Review Board,

I hereby agree to allow Dr. Steve Evans James Madison University to continue, per our verbal face to face agreement April 2003, to conduct his research at Pence Middle School. I understand that this is a six-year research project funded by the Virginia Tobacco Settlement Foundation (VTSF) to examine whether improved quality and coordination of care for adolescents with ADHD reduces cigarette smoking and improves academic and social outcomes in this youth population. The overall project goal is to assess the extent to which the Challenging Horizons Program model of care significantly improves outcomes among adolescents with ADHD over and above what would be expected with the type of care they typically receive.

I am aware that five middle schools are involved-- three in Rockingham County and two in Augusta County and each has been randomly assigned to be either a control or treatment school by way of a coin toss. I know that Pence Middle school was randomly assigned to be a treatment school and as such we will continue to aid recruitment efforts to reach the goal of 10 students for the first cohort (entering 8th grade in the Fall) and 10 students for the second cohort (entering 7th grade in the Fall) students by providing mailing labels for students that you can use to mail information about the study to their parents. We are also glad to assist, by putting notices of this opportunity for families in newsletters and other communications sent from the school.

I understand that as a treatment school, we will continue to collaboratively implement the Challenging Horizons Program model of care using the school-based psychosocial intervention manual and associated training you have provided our teachers and other school staff over the course of the last two years. To aid your efforts to evaluate this program, I am aware that our staff will continue to complete brief rating scales on a regular basis for the students involved in the project and that staff involved in the provision of interventions may also occasionally complete a brief questionnaire about their implementation of these interventions. I am aware that training sessions may be held for teachers that have not had it previously to assist with the implementation of interventions and completion of rating scales, and that my staff will continue to work with a CHP school liaison for on-going training, support, and problem solving during the school year. I also understand that as you have in the past, you will continue to pay teachers for attending training sessions and at the end of the year will compensate them for completing the rating scales.

I look forward to our continued collaboration.

Sincerely,

Mary Schifflet, *Principal*
Pence Middle School

Site Coordinator Letter of Permission

Institutional Review Board
Office of Sponsored Programs,
MSC 5728, Medical Arts West, Suite 26
James Madison University
Harrisonburg, VA 22807

Dear Institutional Review Board,

I hereby agree to allow Dr. Steve Evans James Madison University to continue, per our verbal face to face agreement April 2003, to conduct his research at Stewart Middle School. I understand that this is a six-year research project funded by the Virginia Tobacco Settlement Foundation (VTSF) to examine whether improved quality and coordination of care for adolescents with ADHD reduces cigarette smoking and improves academic and social outcomes in this youth population. The overall project goal is to assess the extent to which the Challenging Horizons Program model of care significantly improves outcomes among adolescents with ADHD over and above what would be expected with the type of care they typically receive.

I am aware that five middle schools are involved-- three in Rockingham County and two in Augusta County and each has been randomly assigned to be either a control or treatment school by way of a coin toss. I know that Stewart Middle school was randomly assigned to be a control school and as such we will continue to aid recruitment efforts to reach the goal of 10 students for the first cohort (entering 8th grade in the Fall) and 10 students for the second cohort (entering 7th grade in the Fall) students by providing mailing labels for students that you can use to mail information about the study to their parents. We are also glad to assist, by putting notices of this opportunity for families in newsletters and other communications sent from the school.

I understand that as a control school, we will continue to aid your efforts to evaluate the Challenging Horizons Program by having our teachers and other school staff complete brief rating scales on a regular basis for the students involved in the project. I am aware that training sessions may be held for teachers that have not had it previously to ensure the accurate completion of rating scales. I understand further that, as you have in the past, you will continue to compensate teachers and other school staff for attending training sessions and that at the end of the year you will compensate them for their efforts completing the aforementioned rating scales.

I look forward to our continued collaboration.

Sincerely,

Donald Curtis, *Principal*
Stewart Middle School

Site Coordinator Letter of Permission

Institutional Review Board
Office of Sponsored Programs,
MSC 5728, Medical Arts West, Suite 26
James Madison University
Harrisonburg, VA 22807

Dear Institutional Review Board,

I hereby agree to allow Dr. Steve Evans James Madison University to continue, per our verbal face to face agreement April 2003, to conduct his research at Stuarts Draft Middle School. I understand that this is a six-year research project funded by the Virginia Tobacco Settlement Foundation (VTSF) to examine whether improved quality and coordination of care for adolescents with ADHD reduces cigarette smoking and improves academic and social outcomes in this youth population. The overall project goal is to assess the extent to which the Challenging Horizons Program model of care significantly improves outcomes among adolescents with ADHD over and above what would be expected with the type of care they typically receive.

I am aware that five middle schools are involved-- three in Rockingham County and two in Augusta County and each has been randomly assigned to be either a control or treatment school by way of a coin toss. I know that Stuarts Draft Middle was randomly assigned to be a treatment school and as such we will continue to aid recruitment efforts to reach the goal of 10 students for the first cohort (entering 8th grade in the Fall) and 10 students for the second cohort (entering 7th grade in the Fall) students by providing mailing labels for students that you can use to mail information about the study to their parents. We are also glad to assist, by putting notices of this opportunity for families in newsletters and other communications sent from the school.

I understand that as a treatment school, we will continue to collaboratively implement the Challenging Horizons Program model of care using the school-based psychosocial intervention manual and associated training you have provided our teachers and other school staff over the course of the last two years. To aid your efforts to evaluate this program, I am aware that our staff will continue to complete brief rating scales on a regular basis for the students involved in the project and that staff involved in the provision of interventions may also occasionally complete a brief questionnaire about their implementation of these interventions. I am aware that training sessions may be held for teachers that have not had it previously to assist with the implementation of interventions and completion of rating scales, and that my staff will continue to work with a CHP school liaison for on-going training, support, and problem solving during the school year. I also understand that as you have in the past, you will continue to pay teachers for attending training sessions and at the end of the year will compensate them for completing the rating scales.

I look forward to our continued collaboration.

Sincerely,

Betsy Agee, *Principal*
Stuarts Draft Middle School

Appendix B

DBD Rating Scale – CHP-C Study Teacher Version

Adolescent's name: _____

Teacher's name: _____ Date: _____

Completed by (*circle one*): (1) Science teacher (2) Math teacher (3) Social Studies teacher (4) Reading teacher

Please circle the number that *best describes* the child's school behavior over the past month. Please circle only one number for every question. Due to the confidential nature of these rating scales, please return them promptly to the CHP mailbox in a sealed envelope.

	Not at all	Just a little	Pretty much	Very much
1. Often interrupts or intrudes on others (e.g., butts into conversations or games)	0	1	2	3
2. Often talks excessively	0	1	2	3
3. Is often easily distracted by extraneous stimuli	0	1	2	3
4. Often fidgets with hands or feet or squirms in seat	0	1	2	3
5. Often does not seem to listen when spoken to directly	0	1	2	3
6. Often blurts out answers before questions have been completed	0	1	2	3
7. Often has difficulty playing or engaging in leisure activities quietly	0	1	2	3
8. Often fails to give close attention to details or makes careless mistakes in schoolwork, work, or other activities	0	1	2	3
9. Often leaves seat in classroom or in other situations in which remaining seated is expected	0	1	2	3
10. Often does not follow through on instructions and fails to finish schoolwork, chores, or duties in the workplace (not due to oppositional behavior or failure to understand instructions)	0	1	2	3
11. Often has difficulty sustaining attention in tasks or play activities	0	1	2	3
12. Often has difficulty awaiting turn	0	1	2	3
13. Is often "on the go" or often acts as if "driven by a motor"	0	1	2	3
14. Often loses things necessary for tasks or activities (e.g., toys, school assignments, pencils, books, or tools)	0	1	2	3
15. Often runs about or climbs excessively in situations in which it is inappropriate (in adolescents or adults, may be limited to subjective feelings of restlessness)	0	1	2	3
16. Often avoids, dislikes, or is reluctant to engage in tasks that require sustained mental effort (such as schoolwork or homework)	0	1	2	3
17. Often has difficulty organizing tasks and activities	0	1	2	3
18. Is often forgetful in daily activities	0	1	2	3

Appendix C

IRS Rating Scale

Please mark an "X" on the lines at the point that you believe reflects the severity of the child's problems in this area and whether he or she needs treatment or special services for the problems, beyond the treatment in place this month. Please consider behavior during the last **month** when making your ratings. **Please do not leave any of the items blank.**

(1) How this child's problems affect his or her relationship with other children

No Problem |_____| Extreme Problem
Definitely does not need treatment Definitely needs treatment

(2) How this child's problems affect his or her relationship with you the teacher

No Problem |_____| Extreme Problem
Definitely does not need treatment Definitely needs treatment

(3) How this child's problems affect his or her academic progress

No Problem |_____| Extreme Problem
Definitely does not need treatment Definitely needs treatment

(4) How this child's problem affects your classroom functioning

No Problem |_____| Extreme Problem
Definitely does not need treatment Definitely needs treatment

(5) Overall, does this child require additional treatment and special services?

No Problem |_____| Extreme Problem
Definitely does not need treatment Definitely needs treatment

Have there been any changes in your approach to working with this child during the past month (e.g., moved his/her seat; began a reward program)? YES NO

If YES, please describe.

Appendix D

Teacher Questionnaire

December 6, 2005

Dear Teacher:

I would like to invite you to participate in a small study using student ratings that you have previously provided as part of the **Challenging Horizons Program (CHP).** Specifically, this study will look at trends in teacher ratings and investigate whether those trends are related to teacher characteristics. The findings will be used as part of my doctoral dissertation.

To conduct this study, I will need to collect information about you and your teaching experiences. Although your participation is solicited, it is strictly voluntary. Whether you complete the questionnaire or not, your status with the CHP will not be impacted in any way.

If you do complete and return the questionnaire, all information will be kept confidential and incorporated into group data only. Your name will never be associated with any of the findings. Further, my research assistant will manage the information and enter it into a database without identifying information, so that only she can match your responses to your name.

If you are interested in participating, please complete and return the enclosed questionnaire and return it in the attached envelope by **December 21, 2005.** It is estimated that this will take less than 5 minutes to complete. Your return of a completed questionnaire implies consent. If you choose not to participate, please return the incomplete questionnaire in the attached envelope.

If you have any questions or require additional information, please feel free to contact either me or my dissertation committee chair (contact information listed below).

Thank you for your help!

Sincerely,

Brandon K. Schultz, Ed.S.
James Madison University
Alvin V. Baird Attention & Learning Disabilities Center
220 Blue Ridge Hall
Harrisonburg, VA 22801
(540) 568-7383
schultbk@jmu.edu

Dr. Joe Kovaleski, Professor
Indiana University of Pennsylvania
Educational & School Psychology
246B Stouffer Hall
Indiana, PA 15705
(724) 357-3785
jkov@iup.edu

Note: My doctoral program at Indiana University of Pennsylvania supports the practice of protection of human subjects participating in research. This project has been approved by the Indiana University of Pennsylvania Institutional Review Board for the Protection of Human Subjects (Phone:724/357-2223). There are no known risks or discomforts associated with this research.

Follow-up Post Card

January 16, 2006

Last month you should have received a brief questionnaire seeking information about you and your teaching experience. The questionnaire was sent to a selection of teachers who participated in the **Challenging Horizons Program (CHP)** last school year.

If you have already completed and returned the questionnaire, thank you. If not, please do so at your earliest possible convenience and return to the CHP mailbox in the attached envelope. Your input is vital for an investigation that will be used as part of my doctoral dissertation. Although your participation is solicited, it is strictly voluntary. Whether you decide to respond to the questionnaire or not, your status with the CHP will not be impacted in any way.

If by some chance you did not receive the questionnaire, or it was misplaced, please call me at (540) 568-7383 or email me at schultbk@jmu.edu and I will immediately send you another copy.

Sincerely,

Brandon K. Schultz, Ed.S.
James Madison University
Alvin V. Baird Attention & Learning Disabilities Center
220 Blue Ridge Hall
Harrisonburg, VA 22801

Note: My doctoral program at Indiana University of Pennsylvania supports the practice of protection of human subjects participating in research. This project has been approved by the Indiana University of Pennsylvania Institutional Review Board for the Protection of Human Subjects (Phone:724/357-2223). There are no known risks or discomforts associated with this research.

Teacher Questionnaire

Please answer the following questions and return this form to our mailbox at your earliest possible convenience. The information that you provide will be **confidential**; identifying information will not be shared with anyone outside of our research staff. If you have any questions, please contact me (Brandon Schultz) at schultbk@jmu.edu. Please be sure to complete **BOTH SIDES** of this questionnaire.

Your Name: _____ Date: _____

School: _____

1. **What is your highest level of education?**
   Please check one:

   _____ 2 Year Junior College Degree (e.g., Associates)
   _____ 4 Year College Degree (e.g., B.A.)
   _____ 2 Year Graduate Degree (e.g., M.Ed.)
   _____ Graduate Degree + Certification (e.g., Ed.S.)
   _____ Doctorate (e.g., Ed.D.)
   _____ Other: _____
   (Describe)

2. **How many years have you been teaching?** _____

3. **How many more years do you anticipate that you will be teaching?**

   _____
   (Please estimate in years)

4. **What subject(s) were you trained to teach?** _____
   (e.g., English, Special Ed.)

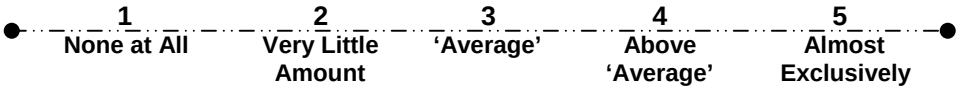
5. **What subject do you enjoy teaching most?** _____
   (e.g., Science, Resource Room)

6. **Are you a parent?** Circle: **Yes** **No**

   a. **If yes, are any of your children middle school age or older?**

      Yes **No**

7. How much experience do you have teaching students with disabilities?



8. How many classes do you teach / co-teach each day?

_____

a. On average, how many students do you have in each of your classes?

_____

9. What is your current age?

Please check one:

_____ 20 to 28 Years
_____ 29 to 37 Years
_____ 38 to 46 Years
_____ 47 to 55 Years
_____ 56 to 64 Years
_____ 65 Years or Older

Appendix E

Microsoft Access Visual Basic® Module to Select Random Records without Repeating Target and Occasion

```
Function InsertRandom()
Dim SQL_Insert As String, RndRecords As String
CurrentDb.Execute "DELETE * FROM TempTable"
RndRecords = " SELECT TOP 76 User_Name,UniqueKidMonth
FROM [Correct Discrepancy Scores] " & _
    "ORDER BY GetRnd(ID)"
SQL_Insert = "INSERT INTO
 TempTable(User_name,UniqueKidMonth) " & RndRecords
CurrentDb.Execute SQL_Insert
If GetRecordCount < 76 Then
    Do
       SQL_Insert = "INSERT INTO
        TempTable(User_name,UniqueKidMonth) " & _" SELECT
        TOP 1 User_name,UniqueKidMonth " & _" FROM
        [Correct Discrepancy Scores] " & _" ORDER BY
        GetRnd(ID)"
       CurrentDb.Execute SQL_Insert
       DoEvents
    Loop Until GetRecordCount = 76
End If
End Function
```