<u>EDUCATION AND HUMAN DEVELOPMENT</u>

STABILITY OF EXTERNALIZING PROBLEM BEHAVIORS WITH ONSET IN EARLY CHILDHOOD: A META-ANALYTIC REVIEW

XINSHENG CAI

<u>Dissertation under the direction of Professors Ann P. Kaiser and Mark W. Lipsey</u>

A meta-analysis was conducted to examine the magnitude of stability of externalizing problem behaviors with onset before age 6 and the variables affecting the stability effect sizes. Gender difference in the stability was also investigated. Seventy empirical research reports, representing 12,111 non-referred children assessed before age 6 drawn from 72 independent aggregated samples and 27 pairs of matched gender samples, met inclusion criteria. Stability was coded as correlational effect sizes for the relationship between externalizing behaviors at Time 1 and Time 2. Results showed great variability in the weighted mean stability effect sizes ranging from.12 to .52 with most of the effect sizes around .30. Boys' externalizing behaviors were more enduring than girls' externalizing behaviors. The effects of informants and subtypes of externalizing behaviors were the most robust findings: the stability effect sizes were larger if Time 1 and Time 2 measured the same subtypes of externalizing behaviors and used the same type of informants. The stability of children's externalizing behaviors decreased as time intervals between measurement points increased. Children assessed before age 3 and from low socioeconomic status (SES) and Caucasian backgrounds had less stable externalizing behaviors. Low SES had differential effects on boys and girls: externalizing behaviors were less stable for boys from low SES families than girls. The findings suggest that externalizing behaviors in young children are not as stable as those in school age children and the information on externalizing behaviors in early childhood alone is insufficient to predict later antisocial behaviors accurately.

STABILITY OF EXTERNALIZING PROBLEM BEHAVIORS WITH ONSET IN EARLY CHILDHOOD: A META-ANALYTIC REVIEW

By

Xinsheng Cai

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Education and Human Development

December, 2004

Nashville, Tennessee

Approved:

Professor Ann P. Kaiser

Professor Mark W. Lipsey

Professor Mark Wolery

Professor Kathleen Lane

To my parents, Mr. and Mrs. Changhe and Yuzhen Ma Cai,

with much love and appreciation

ACKNOWLEDGEMENTS

This work would not have been possible without the kind support of my committee members. Professors Ann Kaiser, Kathleen Lane, Mark Lipsey, and Mark Wolery took time from their busy schedule to read my dissertation, provide insightful feedbacks, and attend the defenses. Professors Ann Kaiser and Mark Lipsey have been my academic advisors and have provided me with extensive professional guidance. They taught me a great deal about doing research and unselfishly invested their time in my professional life.

I am grateful to researchers and staff at the Center for Evaluation Research and Methodology of Vanderbilt Institutes for Public Policy Studies. They helped with literature research and shared their codebook with me. Professor Mark Lipsey is the director of the center and gave me the access to their resources.

I am grateful to Professor Teris Schery who has taken time to mentor me throughout my time at Vanderbilt. Her insight, kindness, generosity, and encouragement are invaluable in helping me finish this project.

I would like to thank Marjorie Boerema for proof reading the final draft of this dissertation.

Nobody has been more important to me in my educational endeavors than my parents, Mr. and Mrs. Changhe and Yuzhen Ma Cai. They have taught me to value education, to love and honor others, and to live a life of integrity and humility. My brother, Guangyuan, and my sisters, Bingqing and Junsheng, have also been important supporters of my educational pursuit.

I am thankful to many dear friends, particularly those from Belmont Church of Nashville, Tennessee. They have prayed for me, loved me, and given me a family to belong to.

Above all, I thank God in Jesus Christ for His divine guidance, love, grace, and strength.

TABLE OF CONTENTS

Page

Appendix

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Externalizing problem behaviors in young children have evolved into an important field of research in recent years. The prevalence of externalizing problem behaviors in young children is high, ranging from 3% to 21% in normative samples (Achenbach & Rescorla, 2000; Caruso & Corsini, 1994; Lavigne, Gibbons, Christoffel, Arend, Rosenbaum, Binns, Dawson, Sobel, & Issacs, 1996), and 33% to 62% in clinically referred samples (Achenbach & Rescorla, 2000; Keenan & Wakschlag, 2000). In addition, early externalizing problem behaviors negatively affect later academic, behavioral, and peer relationship outcomes (e.g., Hinshaw, 1992; Moffitt, 1993; Pettit, Clawson, Dodge, & Bates, 1996). Although early externalizing problem behaviors have high prevalence rates and severe consequences, it is not clear how stable these behaviors are in young children. It is important to determine if externalizing problem behaviors in early childhood are sufficiently stable to warrant early intervention and prevention efforts. The major purpose of this study was to examine the longitudinal consistency of externalizing problem behaviors in non-referred children before age 6 as indicated in extant empirical research reports.

The study of stability of early externalizing behaviors is an important topic, which has relevance in understanding of the origin, causes, and control of deviant behaviors in children. Over decades, researchers from diverse disciplines have conducted studies to determine to what extent later antisocial problem behaviors can be explained by early problem behaviors, and what accounts for the consistency and change in problem behaviors.

In the literature review section, externalizing behaviors will be classified first. Next, the theoretical perspectives for understanding the stability of externalizing problem behaviors will be discussed. Then, findings from previous meta-analytic studies on the stability of externalizing problem behaviors will be reviewed. Finally, research questions to be addressed in this study will be presented.

Classification of Externalizing Problem Behaviors

Externalizing problem behaviors in childhood refer to a wide variety of behavior symptoms such as attention deficit, hyperactivity, oppositional defiant behaviors, aggression, and conduct disorders (Campbell, 1990, 1995; Campbell, Shaw, & Gilliom, 2000; Keenan, Shaw, Delliquadri, Giovannelli, & Walsh, 1998). Although some researchers (e.g., Quay, 1979) posited that subtypes of externalizing behaviors were not distinct, empirical evidence has shown that subtypes of externalizing behaviors have differential etiologies and trajectories (e.g., Fergusson & Horwood, 1995; Hinshaw, 1987, 1992; Hinshaw & Anderson, 1996; Hinshaw, Lahey, & Hart, 1993). Hinshaw (1987) reviewed 60 factor analytic studies to determine the validity for classifying subcategories of externalizing behaviors. The majority of the studies Hinshaw reviewed (41 of 60) yielded two distinctive factorial dimensions: attention-deficit hyperactive disorders (ADHD) and conduct disorders/aggression (CD) across different sources of informants (e.g., parents, teachers), and different gender and age groups. Further examination of the associated features of ADHD and CD with external criterion variables provided more evidence that ADHD and CD were at least partially independent disorders. Aggression/CD disorders were more likely to be associated with environmental variables such as socioeconomic status, family risk variables (e.g., negative-family interaction, family adversity), and later delinquency

(Fergusson & Horwood, 1995; Fergusson, Horwood, & Lynskey, 1993; Hinshaw, 1989, 1992; Nadder, Rutter, Silberg, Maes, & Eaves, 2002). On the other hand, ADHD had much higher heritability, and was less influenced by environmental variables but was more strongly related to cognitive development and later academic failure (Fergusson & Horwood, 1995; Hinshaw, 1989, 1992; McGee, Willams, & Silva, 1985; Nadder et al., 2002; Thapar, Holmers, Poulton, & Harrington, 1999).

Although ample empirical evidence is available to demonstrate the independence of ADHD and conduct disorders as subcategories of externalizing problem behaviors, these two disorders often overlap. It is estimated that 30%-90% of children in one category also will be classified in the other category. Comorbidity rates are much higher for clinical than nonclinical samples (Hinshaw, 1987; Jensen, Martin, & Cantwell, 1997; McConaughy & Achenbach, 1994; McGee, Williams, & Feehan, 1992). Children with ADHD, conduct disorders, and comorbid ADHD and conduct disorders differ with respect to family risk variables, severity of antisocial behaviors, cognitive abilities, peer status, academic achievement, and prognosis (Barkley, Fischer, Edelbrock, & Smallish, 1990, 1991; Coie & Dodge, 1998; Hinshaw & Anderson, 1996; Hinshaw et al., 1993; Jensen et al., 1997; McGee et al., 1992). Therefore, researchers suggested that the construct for co-occurring ADHD and conduct disorders should be treated as a third distinct subclassification of externalizing behaviors (Biederman, Newcorn, & Sprich, 1991; Jensen et al., 1997).

Based on research related to the taxonomy of externalizing behaviors discussed above, in the current study externalizing problem behaviors were classified into three major categories: 1) attention deficit, hyperactive, and/or impulsive behaviors that conflict with age-appropriate expectations, 2) oppositional defiant, conduct or aggressive behaviors that violate the basic rights of other people, inflict harm or pain on other people, or conflict with major age-appropriate social rules (e.g., hostile to adult or peers; disobedient to parents; object or physical aggression), and 3) the combination of these two categories. ADHD is generally characterized by three major symptoms: attention deficit, hyperactive and impulsive behaviors. Different types of ADHD have been proposed. Behavior genetic studies have shown that at the phenotypic level, these three behaviors are derived from the same underlying behavioral construct (Nadder, Silber, Rutter, Maes, & Eaves, 2001). Therefore, ADHD refers to various symptoms characterizing ADHD in this study. Oppositional deviant disorders (ODD) are listed as a separate category of antisocial behaviors in the Diagnostics and Statistical Manual of Mental Disorders (DSM-IV-TR; American Psychiatric Association, 2000). However, ODD in essence represents a milder form of conduct disorders and taps the same underlying construct as CD (Hinshaw, 1987; Hinshaw, et al., 1993; Lahey, Waldman, & McBurnett, 1999; Nadder et al., 2002). Behavior genetic studies (Eaves, Rutter, Silberg, Shillady, Maes, & Pickles, 2000) have shown that CD and ODD share the same underlying genetic liability. In research reports, it is common to combine these two categories of problem behaviors (e.g., Nadder et al., 2002). In the current study, ODD and CD were combined and treated as the same subcategory of externalizing behaviors. Further, for individuals over 18 years old, antisocial personality, delinquent behaviors or crimes also are considered as externalizing behaviors in the same subcategory with CD.

Theoretical Perspectives on Stability of Externalizing Problem Behaviors

The longitudinal consistency of externalizing behaviors in this study is defined as the persistence of a single type of externalizing behavior at different times (e.g., aggression at both time 1 and time 2) and the heterotypic continuity of phenotypically different behaviors (Kagan, 1969; Moffitt, 1993; Pulkkinen, 2001; e.g., conduct disorders at T1 and the comorbidity of ADHD and CD at time 2) that presumably have the same genotypic process (Moffitt, 1993). Although researchers usually agree that the early deviant behaviors are a precursor of later antisocial behaviors in children (e.g., Farrington, 1995; Gottfredson & Hirschi, 1990; Robins, 1966, 1978; Nagin & Farrington, 1992; White, Moffitt, Earls, Robins, & Silva, 1990) and that problem behaviors achieve at least modest stability (e.g., Olweus, 1979, 1984; Zumkley, 1992, 1994), the cause and mechanism of the persistence of problem behaviors over time are much debated.

Classical psychological and developmental theories have emphasized continuity in development. For example, Freud's psychoanalytical theory posits that later behaviors have their origins in early childhood. Piaget's stage theory suggests the cohesiveness in development because each new stage of cognitive development is built on the foundation of previous stages of development. Although such theories have contributed to our understanding of child development, those theories have lost their prominent influences in the fields of special education, child development, developmental psychopathology, and criminology because they do not explain the mechanism of development and lack specific empirical support (Bird, 2001; Campbell, 1990). Empirically driven theories and models have begun to dominate the field (Bird, 2001; Campbell, 1990; Rutter & Sroufe, 2000). In recent years the fields of special education, child development, developmental psychopathology, and criminology have flourished with empirically driven theories and models. Two theoretical perspectives that are relevant in explaining the continuity of problem behaviors in children will be reviewed and later be tested using data from the current study. These theories are: propensity theory and the contextual theory.

Propensity theory

The basic premise of propensity theory is that stable antisocial behavior is caused by enduring individual characteristics. For example, one such individual characteristic is the criminal propensity. In summarizing findings from the Cambridge Study, a longitudinal study on male delinquency, Farrington (1995) stated, “There are individual differences between people in some general underlying theoretical construct which might be termed ‘antisocial tendency,’ which is relatively stable from childhood to adulthood” (p. 956). Therefore, criminal propensity is a latent characteristic rather than an indicator of fully measured behaviors. According to propensity theory, such criminal propensity or disposition in children remains stable over time. Findings from several empirical studies provide evidence supporting propensity theory because persistent criminal behaviors were usually preceded by a history of antisocial behaviors during early childhood and adolescence (Mottiff, 1993; Nagin & Farrington, 1992; Robins, 1966, 1978; White et al., 1990). In her classical study on sturdy childhood predictors of adult antisocial behaviors, Robins (1978) demonstrated that all types of childhood antisocial behaviors predicted a high level of adult antisocial behaviors. In addition, adult antisocial behaviors were predicted better by childhood behaviors than by family background and socioeconomic status variables. Robins’ major conclusion was replicated across four samples with different racial compositions in her study. White et al. (1990) used a New Zealand birth cohort of 1037 children to determine the predictive efficacy of preschool predictors of antisocial behaviors. They conducted discriminate analysis to identify early childhood variables that distinguished children with and without antisocial behaviors at age 11. Their results indicated parent-reported problem behaviors at age 5 were the single best predictor of antisocial behaviors determined by multiple informant reports at age 11. Thus, behavior problems in early childhood were the most reliable predictor for later behavior problems and thus supported the view of propensity theory.

Contextual theory

In contrast, the environmental or contextual theory (Kolvin, Miller, Scott, Gatzanis, & Fleeting, 1990; Lewis, 1990, 1999) postulates that the stability of antisocial behaviors is due to the continuing influences or consistency of the risk factors in the environment. Among the environmental or contextual risks, low socioeconomic status is one factor frequently associated with problem behaviors in children (e.g., Cottle, Lee, & Helbrun, 2001; Dodge, Pettit, & Bates, 1994; Duncan, Brooks-Gunn, & Klebanov, 1994; Gagnon, Craig, Tremblay, Zhou, & Vitrao, 1995; Greenberg, Lengua, Coie, Pinderhughes, & the Conduct Problems Prevention Research Group, 1999; Keenan, Shaw, Walsh, Delliquadri, & Giovannelli, 1996; Kolvin et al., 1990; Pagani, Boulerice, Vitaro, & Tremblay, 1999). The incidence of externalizing behaviors was much higher in children from low-income families than in children from community samples (Keenan et al., 1996). Socio-economic status has been shown to predict externalizing problem behaviors in children (Dodge et al., 1994; Greenberg et al., 1999). For example, Dodge et al. (1994) reported that children’s socioeconomic status assessed in preschool significantly predicted teacher-rated externalizing behaviors from kindergarten through third grade in a sample of 585 children. Similarly, Duncan et al. (1994) demonstrated that low-income status predicted children’s externalizing behaviors at age of 5 and the time when children became poor during early childhood had the same effect on later problem behaviors. In the Newcastle 1000 family study in England, Kolvin et al. (1990) demonstrated that total family deprivation was a stable phenomenon. The correlation of total deprivation scores between 1952 and 1957 when children were 5 and 10 years old was .61. During that time period, if the family moved into deprivation, the rate of offending by children at the age of 15 from those families increased by 50%. If the family moved out of deprivation, the rate of offending by children from those families decreased by 40%. Changes in children's behaviors related to family income. Therefore, their results provided evidence for their original hypothesis that "the total conditions under which a child lives influence his or her development and functioning in physical, social, emotional and intellectual terms" (Kolvin et al., 1990, p. 5).

Critique

Both propensity theory and contextual theory have been challenged because they do not explain fully the development of antisocial behaviors. Although early antisocial behavior was the best predictor for adult antisocial behaviors, the use of early antisocial behavior as the only predictor for later antisocial behaviors sometimes results in high false positive rates. For example, in White et al.'s study, 84.7% of the children who were predicted to have antisocial behaviors at age 11 from their early childhood problem behaviors failed to develop stable antisocial behaviors. The trait or propensity in children may be the necessary but not sufficient condition to develop antisocial behaviors later in life. Similarly, low-income status predicted children's externalizing behaviors; however, the effect was usually small. For example, in Duncan et al.'s study (1994), the inclusion of family income status in the regression model for predicting externalizing behaviors resulted in negligible change in the amount of variance explained. Therefore, low-income status of the family cannot be solely responsible for the development of externalizing behaviors in children.

Although these two theories alone could not explain fully the development and continuity of antisocial behaviors in children, they highlight the contribution of individual traits and environmental risks to the development and continuity of problem behaviors in children. Thus, findings based on these theories help intervention and prevention programs target individual trait factors and environmental risks to reduce antisocial behaviors. These theoretical perspectives are mainly based on the findings from school age children. The current study might provide empirical evidence for propensity and contextual theories to determine the magnitude of antisocial propensity in the form of externalizing behaviors in young children and the influence of environmental risk factors on the continuity of early externalizing problem behaviors in these children. If the propensity theory is true, the externalizing behavior problems with onset in early childhood should achieve high stability in the current study. Likewise, if contextual theory is correct, then a significant influence of socioeconomic status on the stability of externalizing behaviors should be evident in the data.

Previous Reviews on Stability of Externalizing Behaviors

Two types of quantitative literature reviews have been conducted to synthesize empirical findings on the stability of externalizing behaviors: (a) reviews with sole focus on longitudinal consistency of these behaviors (Bennett, Lipman, Racine & Offord, 1998; Olweus, 1979, 1984; Zumkley, 1992, 1994), and (b) reviews of various predictors of later antisocial behaviors (Cottle et al., 2001; Hubbard & Pratt, 2002; Lipsey & Derzon, 1998; Loeber & Dishion, 1983). The most frequently used index for effect sizes of the stability of externalizing behaviors is correlation coefficients (Cottle et al., 2001; Hubbard & Pratt, 2002; Lipsey & Derzon, 1998; Olweus, 1979, 1984; Zumkley, 1992, 1994). Although approaches to examining the literature on stability of externalizing behaviors have varied, findings from most reviews support the notion that childhood externalizing behaviors have considerable stability into adolescence and adulthood.

In a systematic review of stability of aggressive behaviors, Olweus (1979) reported an average stability effect size using correlation coefficients of .55 based on 16 independent samples of males in 14 publications. The ages of the subjects ranged from 2 to 18 years at the first time of measurement. The interval between time 1 (T1) and time 2 (T2) measures ranged from 6 months to 21 years with a mean of 5.7 years. Following the criteria formulated by Olweus (1979), Zumkley (1994) examined the longitudinal consistency of aggressive behaviors in 14 independent samples of males described in 12 research reports published after Olweus' article in 1979. Zumkley reported an average stability effect size using correlation coefficients of .49. The ages of the subjects in Zumkley's review varied from 2 to 19 years at the first time of measurement. The interval between T1 and T2 measures ranged from 1 to 22 years with a mean of 5.7 years.

The stability correlations reported in Olweus' studies (1979, 1984) and Zumkley's studies (1992, 1994) indicated externalizing behaviors in childhood achieved considerable stability over time. Using a different analytical approach, Bennett and colleagues (1998) reviewed studies on longitudinal stability of externalizing behavior by examining how accurately externalizing problem behaviors measured during kindergarten and first grade predicted later externalizing problem behaviors. In their review, which included 13 longitudinal studies, Bennett et al (1998) reported high false positive and false negative rates when early externalizing behaviors were used to predict later externalizing behaviors. For example, among the 15 estimates, two thirds of them had sensitivity at or below 50%, and only 2 of the 14 estimates had specificity over 90%. They concluded that externalizing problem behaviors first measured in kindergarten and first grade children were only modestly enduring and less stable than had been claimed in the literature. The samples in the review the Bennett et al. (1998) review were non-referred children and the interval between T1 and T2 measurements ranged from 1 to 7 years.

Reviews synthesizing research on predictors of antisocial behaviors in adolescence and adulthood have also provided evidence about the stability of externalizing problem behaviors. In general, the stability effect sizes in the format of correlation coefficients in these reviews were slightly lower than those reported in the first type of reviews. For example, in their meta-analytic review of 66 publications, Lipsey and Derzon (1998) found significant effect sizes ranging from .16 to .35 for various antisocial behaviors measured at 6-11 years of age and violent behaviors measured at 15-25 years of age. The effect sizes for various antisocial behaviors (e.g., aggression, general offenses) measured at 12-14 years of age and violent behaviors measured at 15-25 years of age were also significant but slightly lower, ranging from .07 to .27. Compared to other types of early predictors (i.e., personal characteristics, family characteristics, and social factors), several antisocial behaviors (e.g., aggression, general offenses) at an early age were among the top three strongest predictors of later violent behaviors. Similarly, early antisocial behaviors were found to be a strong predictor for juvenile recidivism in the Cottle et al. (2001) meta-analytic study of 23 publications on 22 independent samples. Early conduct problems correlated with juvenile recidivism at .26. The average age at the first measurement point in their study was 14.7 years ranging from 6 to 12 years. The average interval between T1 and T2 measures was 3.8 years, with a range from 1 month to 16 years.

In another review of early predictors of male delinquency, Loeber and Dishion (1983) used the relative improvement over chance (RIOC) as an index for the appraisal of the strength of various predictors for male delinquency in their review of 11 unique samples. Early problematic behaviors (including aggression) and reports of stealing and lying were the second and third best predictors of delinquency after the composite measures of parental family management techniques. These two predictors improved the prediction by 32% and 26 % respectively. The meta-analytic review by Hubbard and Pratt (2002) examined the association between early history of antisocial behaviors and later delinquency among girls, and provided some further evidence about the stability of externalizing behaviors. Hubbard and Pratt (2002) included 11 published research reports and found an average correlation between prior history of antisocial behavior and delinquency of .48.

The two types of quantitative literature reviews on stability of externalizing behaviors have provided consistent evidence that externalizing behavior problems are relatively stable over time. However, there are several limitations in the previous literature reviews. Critical examination of these limitations could help future research to address the areas needing improvement.

Limitations of Previous Reviews

Five issues might limit the findings from previous reviews of the stability of externalizing problem behaviors. First, most of the studies examined the stability of externalizing problem behaviors in school age children. Existing reviews included few studies with the first measurement occurring before children entered elementary school. For example, 7 out of 16 studies in Olweus' review (1979) used samples of children aged below 6 years old at T1 measurement point; the mean stability effect size for preschool children was not reported. In Zumkley's review, only 1 study examined the externalizing behaviors starting in early childhood. The only review focusing on younger children (Bennett et al., 1998) had a very limited age range at T1 between kindergarten to 1st grade, and included children who had already started formal schooling. The stability of externalizing behaviors with onset during early childhood is not clear from the previous reviews. Because of the unique developmental changes occurring before age 6, it is not appropriate to generalize findings about stability based on school age children to young children. In addition, different development stages (e.g., toddler and preschool periods) in early childhood may have effect on the stability of externalizing behaviors because children's development in the areas of language, self-control, and social development are different in toddler and preschool periods (e.g., Campbell, 1990; Erikson, 1963; Kopp, 1982).

Second, as discussed earlier, externalizing behaviors in early childhood refer to a wide variety of behavior symptoms with differential etiologies and trajectories (Hinshaw, 1987, 1992). Previous literature reviews on stability of problem behaviors have either grouped all the subtypes of externalizing behaviors together (Bennett et al., 1998) or examined the stability of a single type of externalizing behavior such as aggression (Olweus, 1979; Zumkley, 1994). It is not clear if the stability of externalizing behavior differs if measured as the same (i.e., homotypic stability) or different subtypes (i.e., heterotypic stability) of externalizing behaviors at T1 and T2.

Third, the contribution of informants to the assessment of stability has not been systematically analyzed. In previous reviews, the stability effect sizes were usually aggregated across measures by various informants with one exception (Bennett et al., 1998). Research studies have provided overwhelming evidence of strong informant effects in behavior ratings and in longitudinal stability that behavior ratings correlated much higher between the same type of informants than different types of informants (e.g., Achenbach, McConaughy, & Howell, 1987b; Fagot & Leve, 1998; Garrison & Earls, 1985; Schmitz & Fulker, 1995). For example, Fagot and Leve (1998) found very little evidence of continuity between parent reported externalizing behaviors of children at age 2 and teacher reported externalizing behaviors at age 5 in their longitudinal study. However, considerable stability was found for externalizing behaviors reported by parents at T1 and T2. Systematic examination of the effect of informants on longitudinal consistency of externalizing behaviors is needed.

Fourth, great variability was found in the stability correlations for externalizing problem behaviors in previous review studies. Only the effect of time on stability of externalizing behaviors has been examined systematically. Olweus (1979) and Zumkley (1994) reported similar findings: the stability effect sizes were affected by the length of the interval between T1 and T2 measurements. Aggressive behaviors became less stable as the time interval increased. Although Olweus (1979) attempted to characterize the effects of environmental and methodological variables on the stability of aggressive behaviors in some studies, his description was qualitative rather than empirical.

Fifth, gender differences in the longitudinal consistency of externalizing problem behaviors have not been adequately examined in previous reviews. Gender differences in externalizing behaviors emerge at about age four (Keenan & Shaw, 1997). For example, the prevalence of externalizing behaviors is much higher in boys than in girls during preschool (Earls, 1987; Keenan & Shaw, 1997). Most reviews of stability of externalizing behaviors have used male only samples or mixed gender samples. Only two reviews (Olweus, 1984; Zumkley, 1992) examined gender differences in the stability of externalizing behaviors using matched samples of males and females. Olweus (1984) found the average stability correlation for aggressive behaviors was .50 for males and .44 for females in six matched male and female samples. Gender differences in stability correlations for aggressive behaviors were greater in Zumkley's review of eight matched samples of males and females: .56 for males and .44 for females. Findings from previous reviews suggested that males' externalizing behaviors were more enduring than the externalizing behaviors of females, but that externalizing behaviors in females also had a high degree of stability. The small number of independent matched samples in previous reviews limits confidence in generalizing these results. The sample populations in which gender differences were explored were mainly school age children in the previous reviews. Gender might have different effects on the stability of externalizing behaviors in younger children.

Overview of the Current Study and Research Questions

The current study utilized a meta-analytic approach to examine the empirical findings from longitudinal studies to determine the magnitude of stability of externalizing problem behaviors with onset before the age of 6. The study explored factors that might account for the variability in stability effect sizes, and investigated gender differences in the stability of externalizing behaviors. Specifically, the study addressed the following research questions:

1) What is the magnitude of stability of externalizing problem behaviors with onset before age 6?
2) What variables (e.g., time, measurement, sample characteristics) account for the variability in stability effect sizes of externalizing problem behaviors?
3) What are the gender differences in the magnitude of stability effect sizes and variables accounting for the variability in stability effect sizes?

In sum, the present study filled the void in current literature by examining the stability of externalizing behaviors with onset before the age of 6 in general and by examining the stability of externalizing behaviors with onset during toddler and preschool periods in particular. To disentangle the effect of behavioral construct, the current study examined both homotypic and heterotypic stability of externalizing behaviors. To address the informant effect, the current study compared the stability of externalizing problem behaviors measured by the same and different informants at T1 and T2 measurement points. To disaggregate the effects of informant and construct types, the stability of externalizing behaviors was examined by forming more homogeneous subgroups of studies based on T1 and T2 informants and behavioral constructs. Unlike the previous reviews, the current study coded information of various aspects of the eligible reports to determine the effects of time, measurement features, and demographic characteristics of the sample (e.g., SES, race) on the stability of externalizing behaviors. Moreover, the current study used 27 pairs of matched gender samples to examine gender differences in the stability of externalizing behaviors in young children. The current study also examined if the effects of time, measurement, and sample characteristics on the stability of externalizing problem behaviors varied by gender.

CHAPTER II

METHOD

Inclusion Criteria

The inclusion criteria for this study are summarized in Table 1. These criteria were adapted from those used in a large meta-analysis underway by Lipsey and colleagues (Lipsey & Derzon, 1998; Derzon & Lipsey, 1999) regarding the predictors of antisocial behavior and substance abuse conducted at the Center for Evaluation Research and Methodology of Vanderbilt Institutes for Public Policy Studies (2001). To be eligible for the current meta-analysis, the study must have used a longitudinal panel design. Specifically, each study must have obtained at least

Table 1. Summary of eligibility criteria

Area	***Criteria for eligibility***	***Examples of ineligibility***
Design	Longitudinal and prospective design; children should not receive any interventions on problem behaviors	Cross-sectional and retrospective studies; children received intervention because of low birth weight
Construct	Externalizing behaviors	Internalizing and total problem behaviors
Effect size	Reporting *r* or sufficient data to calculate an effect size	Mean and standard deviation of externalizing behavior scores at time 1 and time 2
Age	Children below age 6 at time 1; sufficient details to make inference about age	School age children with no information regarding the grade level or age
Child subject	General population or at risk population; children in good physical health without mental disabilities	Children with developmental and psychiatric disorders; children referred for clinical treatment
Publication	Studies must be published after 1950 and the data collection finished by 1945; studies conducted in Western cultures, and published in English	Berkeley Guidance Growth Study with children born in the late 1920s; studies published in German only; studies conducted in India and published in English

two waves of measures on the same persons. The first wave of data must have occurred before children entered elementary school and their ages were not older than 6. However, the second wave of data collection could occur anytime after the first wave of data had been collected and children's age could be younger or older than 6. Only prospective longitudinal studies were included. Retrospective studies were excluded because behaviors measured in retrospective fashion may have been contaminated by distorted memory and/or the tendency to evaluate previous behaviors using current behaviors as reference.

Studies evaluating the effectiveness of interventions were excluded because the focus of the current study was the stability of externalizing behaviors in natural environment/conditions. For example, the study by Achenbach, Edelbrock, and Howell (1987a) was not included because the subjects in that study were involved in a treatment program for low birth weight infants.

The construct of the problem behaviors at T1 and T2 was externalizing behavior. Behavioral measures at either T1 or T2 using total problem behaviors scores based on internalizing and externalizing behaviors or including items of feeding, sleeping, physical problems of the children were not included in this meta-analysis (e.g., Richman, Stevenson, & Graham, 1982). Temperament measures were not included because temperament is widely used as a construct of natural disposition of an individual and is considered to be biologically rooted (e.g., Bates, Pettit, Dodge, & Ridge, 1998). Unlike externalizing behaviors, temperament is usually not a target for early intervention or prevention efforts. Only measures assessing observable and actually occurring externalizing behaviors were included. Social-cognitive processes indicating aggressive styles were not included. For example, correlations between aggressive solutions in solving problems at T1 and externalizing behaviors at T2 were excluded (e.g., Coy, Speltz, DeKlyen, & Jones, 2001).

To be included, the study must have provided at least one longitudinal effect size on the stability of externalizing behaviors. The study must have reported correlational effect sizes or provided sufficient quantitative data to compute an effect size in the form of Pearson correlation coefficient on the relationship between T1 and T2 externalizing problem behaviors.

The age of the sample must have been described in sufficient detail to allow reasonable inference about the age of the sample at each time of measurement. For example, a study may provide the age information at T1 measurement point and indicate that T2 measures were taken 6 months later. In such case, age information can be inferred for the study sample at T2.

Child participants in selected studies were from the general population and the population at-risk for antisocial behaviors due to disadvantaged SES and family environments. These children were in good physical health and without signs of gross brain damage, severely delayed or impaired language development, or severe developmental or psychiatric disorders (i.e., mental retardation or autistic-like behavior). Studies that enrolled children who were clinically referred for problem behaviors were not included in the current study.

Eligible studies must have been published after 1950 and the data collections must have been finished since1945 to reflect the research and life after World War II. Both published and nonpublished studies were included to minimize publication biases. In addition, eligible studies were conducted in a Western, economically developed culture, although studies may have included minority members of that culture.

Retrieval of Studies

Relevant studies were retrieved in four ways. First, PsyINFO, Education Abstracts, Exceptional Child Education Resources, Education Resources Information Center (ERIC), Sociological Abstracts, and PubMed were searched for potential eligible studies. The search terms used were modified according to the indices of the particular database. The general terms used were: problem behavior, externalizing behavior, disruptive behavior, impulsive, antisocial behavior, acting out, aggressive behavior, conduct disorder, attention deficit disorder, oppositional defiant disorder, psychopathology, behavior problems, fighting, and longitudinal. Second, previous meta-analysis and conventional reviews of the stability literature were searched for potential eligible studies. Third, the reference lists of coded reports were examined for additional studies. Fourth, the database of predictors of antisocial behaviors and substance abuse from the Center for Evaluation Research and Methodology (CERM) at Vanderbilt Institute for Public Policy Studies (VIPPS) was searched and eligible studies were included. The database at CREM included both published and unpublished research reports on antisocial behaviors and substance abuse obtained through various sources. References of all the eligible research reports for this meta-analytic study are included in the reference section with asterisks.

Coding of the Empirical Studies

Studies meeting the criteria specified above were coded by the author into a FileMaker database using the modified codebook originally developed by the Center for Evaluation Research and Methodology at VIPPS (2001) for the project on antecedent risk predictors of antisocial behavior (Lipsey & Derzon, 1998; Derzon & Lipsey, 1999). Information extracted from each eligible study falls into two categories: (a) information regarding the effect sizes, and (b) information regarding the study descriptors (Lipsey & Wilson, 2001). Information regarding the effect sizes is the outcome variable for the meta-analysis, while information regarding the study descriptors is similar to independent variables that may account for the variation in effect sizes across studies. Table 2 summarizes the information abstracted from each eligible study. The following are highlights on each section of the coding.

Table 2. Summary on coding and definitions

Area of coding	**Definition**
Study Level Information	
Country	Country where the study was conducted
Design	Type of longitudinal design: e.g., single cohort or multiple cohort designs
Population	Type of population from which the sample was drawn: e.g., general population, at-risk population for behavior problems
Wave Information	
Low year	The year when the measurement period began
High year	The year when the measurement period ended
Low age	Age of the subjects when the measurement period began
High age	Age of the subjects when the measurement period ended
Age-average	Average age of the subjects
Total sample size	Total number of subject measured at each wave
Check attrition effect	Researchers tested the effect of attrition: 1=yes, 2=no
Any attrition effect?	If tested, were there any attrition effects found: 1=yes, 2=no
Sample Information	
Type of sample:	Aggregated or subsamples
Total sample size	Total number of subject in the sample
Number of boys	Total number of boys in the sample
Number of girls	Total number of girls in the sample
SES –Poor/Low	Total number of low SES subjects in the sample
SES --Working	Total number of working class subjects in sample (unskilled laborers)
SES --Middle +	Total number of middle class or above subjects in sample (professionals, skilled laborers)
Rank SES breakdowns	If the exact number of subjects in each SES categories was not reported, rank SES breakdowns using a 3-point scale: 1=majority, 2=present, 3=clearly minority
Race-White	Total number of white
Race-Black	Total number of black
Race-Hispanic	Total number of Hispanic
Race-Other	Total number of other minority
Rank race breakdown	If the exact number of subjects in each race categories was not reported, rank race breakdowns using a 3-point scale: 1=majority, 2=present, 3=clearly minority

Table 2 continued

Construct Information	
Description	Describe in detail the behavior measures used: e.g., type of behaviors, name of the measure.
Type of behavior	Type of externalizing behaviors: 1=attention deficit and hyperactivity, 2=conduct disorder/aggression, 3=comorbidity
Method of data collections	How were the data collected: 1=questionnaire, 2=interview, 3=archival records, 4=observation, 5=physical tests, 6=more than more types of method.
Informant	Who provided the information: e.g., parents, teachers, observers, psychologist, etc.
Type of measures	Information on how the measure was derived: e.g., single item, unvalidated multiple items, factor scales, standardized instrument
Effect Size Information	
Type of effect sizes	How the original effect sizes were reported: e.g., correlation, *t*-test, chi-square, etc.
Type of scale at T1	The state of the data at point of analysis: dichotomized scale(i.e., measurements with only two categories), discrete scale (i.e., interval scales with 3 to 8 categories), and continuous scale (i.e., ratio or continuous scales, or interval scales with more than 8 categories).
Type of scale at T2	The state of the data at point of analysis: dichotomized scale(i.e., measurements with only two categories), discrete scale (i.e., interval scales with 3 to 8 categories), and continuous scale (i.e., ratio or continuous scales, or interval scales with more than 8 categories).
Sample size	Sample size on which the effect sizes was based
Effect sizes	Magnitude of the effect sizes
Significance	Indicate if the effect size was reported as statistically significant at the .05 level.

Coding effect sizes

Pearson correlation coefficient was used as the index of stability of externalizing behaviors in the present study because *r* statistic has been most frequently used in studies on stability of personality. If a correlation coefficient was not reported in a study but sufficient information was available, the correlation effect sizes were calculated. The formulas for converting various statistics, such as contingency tables, *F*-test statistics, *t*-test statistics, chi-square test statistics, and significant *p* values of correlation tests, into correlation coefficients can be found in Lipsey and Wilson (2001). It is common for a study to report more than one eligible effect size. For example, a study may report stability correlations for different gender and age groups on multiple measures of externalizing behaviors at different points of measurement. All eligible longitudinal effect sizes were coded.

Coding study descriptors

Four major categories of study descriptors were coded: study level, wave, sample, and construct information. The study level information section includes information on study design (e.g., how the sample was drawn, the type of sample, and the location of the study). Although multiple research reports were included for an independent study, there was only one record for coding study design information for each study. Previous meta-analytic studies indicated that study design variables affected effect sizes (Knight, Fabes, & Higgins, 1996; Lipsey & Derzon, 1998). For example, Lipsey and Derzon (1998) found that the country where studies were conducted and the type of samples used had significant influence on the effect sizes representing the relationship between predictors and violent or serious delinquency in adolescents and early adulthood. Therefore, information regarding the population and study design may help researchers decide to which population the study results may be generalized.

The wave section includes information on when the measurements were taken, sample sizes at each measurement point, and attrition information. The number of wave records corresponds to the number of measurement points for a particular study. Information on all eligible waves was coded. Each study had at least two Wave records corresponding to T1 and T2 measurement points. Longitudinal studies may have more than two waves.

Olweus (1979) and Zumkley (1994) found the time interval influenced the stability correlations of externalizing behavior problems in their analyses: the stability correlations decreased as the time interval between the two measurement points increased. Wave information coded in this study will be used in the final analyses to determine how various predictors regarding time effects may influence the stability correlations of externalizing behaviors in young children.

The sample information section includes information on the demographic characteristics of the sample: gender, age, race, and socioeconomic status of the subjects. When a study reported stability data on more than one sample, for example, the whole sample, and boy and girl subsamples, one record was created for each sample and subsample. The number of sample records for one study corresponds to the number for the total sample plus the number for all eligible subsamples.

Previous research has shown that children's externalizing behaviors are influenced by various demographic and environmental risk factors such as gender, race, and SES (e.g., Kolvin et al., 1990; Olweus, 1979; Zumkley, 1994; Feld, 1999). Regarding the gender effect, the level of externalizing behavior problems has been found to be much higher for boys than for girls (Keenan & Shaw, 1997) and the continuity of externalizing behaviors is greater for boys than for girls (Olweus, 1979; Zumkley, 1994). Regarding the race effect, children from African American and other minority backgrounds have been found to be associated with higher crime rates and race predicted antisocial behaviors (Cottle et al., 2001; Feld, 1999). Regarding the effect of SES, findings from previous research indicated children from low-income families were more likely to develop problem behaviors than children from middle class families and problem behaviors of children from low-income families were more likely to persist (Kolvin et al., 1990). In this study, the effect of demographic and environmental risk factors on the stability of externalizing behaviors will be examined.

The construct section includes information on eligible measures assessing the construct of interest (e.g., the description of a measure, how a measure was taken, the type of the measure, and the type of behaviors measured). If a study reported more than one eligible measure (e.g., different measures assessing conduct disorders), one record was created for each eligible measure for each independent study. Construct information helps examine the effects of different subtypes of externalizing behaviors and informants on the stability of externalizing behaviors.

Coding Reliability

The author of the study was the primary coder for this meta-analysis project and coded all the eligible studies. Ten percent of the studies were coded independently by another trained coder from the Center for Evaluation and Research Methodology at VIPPS to determine the reliability of the coding. Coding reliability was calculated using the point-by-point agreement formula (Kazdin, 1982):

$$\text{point-by-point agreement} = \frac{A}{A+D}\times 100$$

where A = agreement for the coding and D= disagreement for the coding.

Table 3 summarizes the coding reliability. In general, the averaged reliability for each

Table 3. Summary on coding reliability

Variable	**Mean**	**Range**
Study features	95%	67% - 100%
Wave information	97%	89% - 100%
Sample characteristics	91%	75% - 100%
Construct features	91%	63% - 100%
Effect sizes features	99%	93% - 100%
Overall reliability	96%	88% - 100%

section of the coding was above 90%. The reliability for the overall coding ranged from 88% to 100% with a mean of 96%. In addition, the coding of all eligible reports was checked by the primary coder a second time to ensure accuracy particularly in the areas indicated by the coding reliability data that errors were more likely to occur.

CHAPTER III

DATA ANALYSIS

Data analyses in meta-analytic review studies include analyzing both the effect sizes and the study descriptors. Descriptive statistical analyses are used to describe study design, wave, sample, and construct characteristics. Inferential statistical analyses are conducted to estimate the magnitude of effect sizes and to determine the sources of variability in the effect sizes. Because the descriptive statistical analysis for study descriptors is quite straightforward, this section will be devoted to the discussion of inferential statistical tests for meta-analysis in relation to how they might be applied to the current study.

Inferential statistical techniques in meta-analysis can be grouped into two categories: single effect-size and multivariate effect-size data analyses depending on whether one or multiple effect sizes will be used from an independent study for a single analysis. If one effect size is chosen from an independent study, single effect-size data analytic approaches should be applied. If more than one effect size is chosen per study for a single analysis and the dependence of these effect sizes is modeled, multivariate effect-size data analytic approaches should be applied (Becker, 2000; Becker & Schram, 1994). The two categories of inferential statistical approaches to meta-analysis depend largely on the data structure.

Single effect-size data analytic techniques were used to analyze the data in this study because most of the studies failed to provide information on the correlations among the eligible dependent variables, which is critical to estimate the dependence of multiple effect sizes from a single study in the multivariate meta-analysis. Single effect-size data analytic techniques involve a sequence of statistical testing procedures including: preparing data for analysis, applying adjustments to effect sizes, testing homogeneity of the effect sizes, and modeling variability in effect sizes. These procedures apply only in situations in which one effect size for an independent study is selected for any single analysis.

Data Set Construction

The initial step of data analysis in univariate meta-analysis is to prepare sets of data with only one effect size from each study for different analyses. One effect size can be obtained by (a) selecting randomly one effect size from all the eligible effect sizes within a study, (b) selecting an effect size according to specified criteria (e.g., quality of construct), or (c) averaging all the eligible effect sizes within a study (Becker, 2000; Lipsey & Wilson, 2001). In addition to using the study as the unit of effect size selection, researchers could also use variables of interest, such as constructs and measurement points. For example, if more than one effect size is provided for more than one eligible construct, then one effect size can be selected for each eligible construct using the strategies noted above, and separate analysis can be conducted for each construct of interest.

In this study, the strategies of averaging effect sizes using certain criteria and selecting effect sizes randomly were both used to create data sets with independent effect sizes. The effect sizes were first averaged by samples. Independent samples rather than independent studies were used to select effect sizes for each analysis. There are two types of samples in the current meta-analysis: aggregated samples and matched gender samples. An aggregated sample includes all child participants from one independent study. An matched gender sample includes children of the same sex. Effect sizes for boy and girl samples are matched on sample characteristics (e.g., SES, age), time variables (e.g., time interval), construct features (e.g., informant and construct types), and effect size characteristics (e.g., scale type)

Because the stability effect sizes might vary as a function of subtypes of externalizing behaviors and informants used at T1 and T2, averaging effect sizes by samples across different constructs and informants might conceal meaningful differences related to different constructs and informants. In addition, the stability might be different for externalizing behaviors with onset in toddlerhood from those with onset in preschool. It is inappropriate to average effect sizes measured in these two different developmental periods. Therefore, in addition to samples, effect sizes were also averaged T1 age categories (i.e., before and after age 3), informants (i.e., if T1 and T2 used the same or different informants), and subtypes of externalizing behaviors. For example, if one study provided three effect sizes on aggressive behaviors using three different behavior measures rated by parents at both T1 and T2 for children who were first assessed at age 4, these three effect sizes were averaged to form a single effect size. In another study, if one effect size was provided on correlation between aggression rated by parents at T1 and ADHD rated by teachers at T2 when children were age 5, and one effect size was provided on correlation between aggression rated by parents at T1 and ADHD rated by parents at T1 when children were age 5, these two effect sizes could not be averaged. Averaging effect sizes according to these criteria within a study still resulted in more than one effect size from a single study. In the analyses of estimating weighted mean effect sizes by informants and constructs, such strategy of averaging effect sizes was sufficient. However, in the weighted multiple regression analyses, random numbers were created to select one effect size among multiple effect sizes from each study. Similar strategies for creating data sets of independent effect sizes were used for all samples. Criteria for aggregating effect sizes changed slightly because of the different purposes of different analyses. A brief description on how the data set was created for a particular analysis is given in the results section.

Transformation and Calculation

After the data sets of effect sizes are prepared, the next step is to apply effect size adjustment. The sample size on which an effect size is based affects how precisely the effect size represents the relationship of interest in the population. Effect sizes based on small sample sizes are less precise estimates of the relationship in the population than effect sizes from large sample sizes. In this study, the inverse variance was used as the weight to represent the reliability of information associated with each effect size (Lipsey & Wilson, 2001). First, the correlation effect sizes were transformed using Fisher Z transformation (Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Lipsey & Wilson, 2001). The formula is: $z = .05 \ln (1 + r)/(1 - r)$, where z is Fisher z-transformed correlation and r is the Pearson correlation coefficient. Next, the inverse sampling error variance weight was applied, which is $w = n - 3$ (n is the sample size), to each z-transformed effect sizes for estimating the weighted mean effect size using the formula:

$$\bar{z} = \Sigma (w_i z_i) / \Sigma w_i \tag{1}$$

where $i = 1, 2, \ldots, k$ independent studies. The weighted z-transformed mean effect size was converted back to correlation effect sizes for interpretation using the formula: $r = [\exp(2z - 1)/(\exp(2z + 1)]$.

Homogeneity Test

The homogeneity test (i.e., Q statistic) was performed to determine if effect sizes from different studies share a common population effect size (Hedges & Olkin, 1985). Nonsignificant

Q statistic suggests the effect sizes may be selected from the same population of effect sizes. Significant *Q* statistic suggests the variability in the effect sizes is not likely to be the result of sampling error alone.

The test of homogeneity for correlation effect sizes based on Fisher's *z*-transformed effect sizes is the following formula:

$$\mathrm{Q} = \sum_{i=1}^{k} (n_i - 3)(z_i - \bar{z})^2 \qquad (2)$$

where z_i is z-transformed effect size for study i = 1, 2, … k independent studies, and $\bar{z}$ is the weighted mean effect size. Under homogeneity, *Q*s are asymptotically distributed as a Chi square distribution with *k*-1 degrees of freedom. The homogeneity test has relatively low power to detect heterogeneity in samples with small sample sizes (Lipsey & Wilson, 2001).

Model Testing

Model testing is an important issue in estimating weighted mean effect sizes because the estimates might differ when different types of models are used. Two different classes of models can be used to estimate weighted mean effect sizes: fixed and random effects models (Hedges & Olkin, 1985; Hedges, 1994; Overton, 1998; Raudenbush, 1994). These two types of models differ in the weights used in the analysis and in their statistical and theoretical assumptions. The fixed effects model assumes that all effect sizes are used to estimate the same population effect size and the variability in the effect sizes can be totally explained by subject level sampling error. Such a fixed effects model is appropriate when empirical studies in the meta-analysis are selected through an exhaustive search, the findings of the meta-analysis are generalized to studies similar to those under investigation, and effect sizes are assumed not to vary randomly from each other (Hedges, 1994; Overton, 1998). The weight used in the fixed effects model contains only the inverse sampling error variance. In contrast, the random effects model assumes effect sizes in the sample may come from a population of effect sizes that truly differ across studies (Becker, 2000). The weight for random effects models consists of the sampling error variance and the random variance component, which can be derived from the total *Q* statistic estimated by the fixed effects model. A random effects model is recommended when the homogeneity hypothesis is rejected in the fixed effects model (Lipsey & Wilson, 2001) and the studies selected for the meta-analysis are assumed to be a random sample of a hypothetical universe of possible eligible studies (Raudenbush, 1994; Overton, 1998). Data analysts usually disagree about which assumptions best fit in various research situations. Results from both fixed and random effects models were reported and interpreted in relation to their generalizability. Reporting findings from both types of models might inform research about the magnitude of stability of externalizing behaviors under different theoretical assumptions and help theory building in the field.

Similarly, there are two types of models to explain the variability in the effect sizes using weighted regression analyses: fixed effects models with predictors and mixed effects models (Hedges, 1994; Lipsey & Wilson, 2001; Overton, 1998; Raudenbush, 1994). A fixed effects model assumes that the variability in the effect sizes can be explained by the sampling error within subjects and systematic differences between studies. Under the assumption of the fixed effects model with predictors, researchers can conduct weighted multiple regression analyses using coded information such as sample characteristics, time variables, and constructs as independent variables and the sampling error variance matrix as the weight. In contrast, a mixed effect model assumes that the variability in the effect sizes can be partitioned into sampling error, systematic differences between studies, and random variation across studies (Lipsey & Wilson, 2001). Therefore, the weights incorporate both the sampling variance and the random error variance, which is the residual estimated by the fixed effects model with predictors. Although more conservative than the fixed effects model, the mixed effects model results have greater generalizability because the estimation of mixed effects models take into consideration the random errors across studies. In the current study, results from both the fixed and mixed effects models for explaining the variability in the stability effect sizes are reported and interpreted in relation to their generalizability.

Analysis for Gender Differences

The procedures described above applied to the data analyses for the aggregated samples as well as matched gender samples. The weighted mean effect sizes were calculated for the boy and girl samples separately using fixed and random effects models. The variability in the effect sizes was examined using weighted multiple regression analysis in fixed and mixed effects models including both effect sizes from the boy and girl samples. One exception in the weighted multiple regression analyses for matched gender samples was that interaction effects between gender and other predictors as well as the main effects of predictors were tested to determine if certain predictors had differential effects on the stability effect sizes for boys and girls. A Q-test of between-group difference was conducted to determine if the stability of externalizing behaviors was different for boys and girls. The Q-test of between-group difference is a variation of fixed effects model using ANOVA procedures to test group differences in weighted mean effect sizes (Hedges, 1996; Lipsey & Wilson, 2001). The Q-test of between-group difference partitions the variability in the effect sizes into variability between groups and within groups. If the variability between groups is significant, the differences in the weighted mean effect sizes are significantly different between groups.

CHAPTER IV

RESULTS

The results were developed through five major sets of analyses. First, the distributions of the effect sizes and sample sizes were examined to identify outliners. Second, characteristics of the eligible empirical research studies included in this meta-analysis were described. Third, weighted mean effect sizes were calculated aggregated by samples, informants, and constructs for the aggregated and matched gender groups. Fourth, weighted multiple regression analyses were conducted to account for the variability in the effect sizes for the aggregated sample. Fifth, the issue of gender differences was investigated regarding the variables accounting for the variability in the effect sizes for matched gender samples. Because the sample sizes in this study were relatively small, the α level was set at .10 as the significance level to increase statistical power and to minimize Type II error rates (Lipsey, 1998).

Preliminary Analysis to Identify Outliers

Before proceeding with the major data analyses, the distribution of effect sizes was examined for extreme values. Because the main goal of meta-analytical reviews is to obtain a general estimate of the effect sizes in a body of empirical research studies, extreme values may be misleading and distort the analysis of the effect sizes (Hunter & Schmidt, 1990; Lipsey & Wilson, 2001). No extreme values were found in the distribution of the aggregated effect sizes from the aggregated and matched gender samples. All the effect sizes were within three standard deviations from the mean. Because effect sizes were weighted by the inverse variance weight, which is influenced greatly by sample size, extreme values of the sample sizes could have significant influence on the effect sizes. Therefore, sample sizes were also examined for outliers. A few studies had extremely large sample sizes. Sample sizes over 500 were adjusted and were recoded into 500, a procedure known as Winsorizing (Lipsey & Wilson, 2001). A sensitivity analysis was conducted to determine the effect of Winsorizing sample sizes on the estimation of weighted mean effect sizes. Sensitivity analysis is a systematic approach to determine how sensitive the conclusion is to the methods of analysis (Greenhouse & Iyengar, 1996). The results showed that the weighted mean effect sizes using the original sample sizes were similar to those using Winsorized sample sizes (see Table 10). Thus, Winsorized sample sizes were used for all the inferential statistical analyses in this study.

Descriptive Analysis

Tables 4 through 9 present the information on the characteristics of the eligible research reports included in this meta-analysis. Seventy research reports (see Table 4) met eligibility

Table 4. Characteristics of eligible reports (N=70)

Variable	**N**	**%**
Publication Type		
Journal articles	56	80.0%
Book or book chapters	5	7.0%
Conference proceedings	1	1.0%
Dissertations	8	11.0%
Year of Publication		
1960-1969	2	3.0%
1970-1979	2	3.0%
1980-1989	13	19.0%
1990-1999	35	50.0%
2000-2003	18	26.0%

criteria for the study. The majority of the research reports were journal articles (80%). Nineteen percent of the reports were published in the 1980s, 50% were published in 1990s, and 26% were published between 2000 and 2003. Two reports were published in the 1960s and two were published in the 1970s.

Table 5 summarizes the characteristics of study design features. The frequency counts in

Table 5. Study design features. The sample size for the aggregated sample is 72. The sample size for the matched gender sample is 54.

	Aggregated sample		**Matched gender sample**	
Variable	**N**	**%**	**N**	**%**
Country where study was conducted				
USA	60	83.3%	44	81.5%
Great Britain	1	1.4%	0	0%
Canada	4	5.6%	4	7.4%
Scandinavia	3	4.2%	2	3.7%
Australia New Zealand	1	1.4%	2	3.7%
Other Western European Countries	3	4.2%	2	3.7%
Study Design Features				
Single cohort follow-up	49	68.1%	36	66.7%
Multiple cohort follow-up	23	31.9%	18	33.3%
Attrition Rate				
0% to 5%	31	43.1%	18	33.3%
5 % to 10%	6	8.3%	5	9.3%
10% to 20%	11	15.3%	11	20.4%
20% to 30%	11	15.3%	7	13.0%
30% to 40%	4	5.6%	1	1.9%
40% to 51%	9	12.5%	12	22.2%
Attrition Effect				
Tested attrition effect	22	30.6%	20	37.0%
Did not test attrition effect	50	69.4%	34	63.0%
Significant attrition effect found	5	22.7%	2	10.0%
Significant attrition effect not found	17	77.3%	18	90.0%

the table indicate the characteristics of independent samples or groups, which are considered as independent studies in this meta-analysis. The aggregated (N=72) and matched gender (N=54) studies were very similar in study design features and other characteristics. Over 80% of the aggregated and matched gender studies were conducted in the United States. Two-thirds of the studies used single cohort follow-up designs and the rest used multiple age group designs. Over 60% of the studies had attrition rates less than 20%. One-third of the studies (22 aggregated samples and 20 matched gender samples) tested whether participants who stayed and those who dropped out of the studies were significantly different on important demographic and behavioral variables. For aggregated samples, 5 of the 22 studies that examined attrition found significant attrition effects (23%). Attrition effects were not consistent because some research reports indicated that children with higher risks remained in the longitudinal studies while others indicated that children with lower risks remained in the studies. For example, children who participated in the later waves of measurements were reported to have more behavior problems (Schmitz, Fulker, & Mrazek, 1995; Verhulst & Althaus, 1988), consist of fewer boys with fewer behavior problems (Garrison & Earls, 1985), have lower ratings in mother and child communication, neighborhood violence, and neighborhood disorders (Ingoldsby, 2002), or come from families with lower socioeconomic status (Verhulst & Althaus, 1988). For matched gender studies, two studies from one report found significant attrition effects (Verhulst & Althaus, 1988). The mean SES at T1 for non-responders was significantly lower than the mean SES for the responders. In addition, the mean of total problem scores by parent reports was slightly but significantly higher for nonresponders than responders.

For both aggregated and matched gender studies (see Table 6), the majority of the samples were drawn from the normal population of children, 74% and 82% for the aggregated

Table 6. Sample characteristics (N=72 for aggregated sample and N=54 for matched gender Sample)

	Aggregated sample		Matched gender sample	
Variable	**N**	**%**	**N**	**%**
Population Risk				
Normal population	53	73.6%	44	81.5%
Risk population	19	26.4%	10	18.5%
Gender				
All males (>95%)	18	25.0%	27	50.0%
60% -95% males	2	2.8%	0	0%
50% - 60% males	30	41.7%	0	0%
<50% males	7	9.7%	0	0%
No males (<5%)	15	20.8%	27	50.0%
Ethnicity				
All white (>95%)	34	47.2%	26	48.1%
60% -95% white	17	23.6%	14	25.9%
50% - 60% white	4	5.6%	4	7.4%
<50% white	13	18.1%	7	13.0%
No white (<5%)	4	5.6%	3	5.6%
Middle and above SES				
All middle and above middle SES (>95%)	17	23.6%	14	25.9%
60% -95% middle and above middle SES	16	22.2%	6	11.1%
50% - 60% middle and above middle SES	3	4.2%	0	0%
<50% middle and above middle SES	17	23.6%	22	40.7%
No middle and above middle SES (<5%)	19	26.4%	10	22.2%
Low SES				
All low SES (>95%)	15	20.8%	10	18.5%
60% -95% low SES	3	4.2%	2	3.7%
50% - 60% low SES	2	2.8%	0	0.0%
<50% low SES	22	30.6%	24	44.4%
No low SES (<5%)	30	41.7%	18	33.3%
Mean number of subjects				
Less than 100	32	44.4%	32	59.2%
101-200	19	26.4%	13	24.1%
201-500	13	18.1%	7	13.0%
501-1000	5	6.9%	2	3.7%

and matched gender samples respectively. The majority of the children in these samples were Caucasian. Children in the eligible studies were from diverse SES backgrounds with over one third of the samples consisting of predominantly middle or above middle class families. Most samples (42%) had an equal number of boy and girl participants in the aggregated sample. For the aggregated samples, the sample sizes ranged from 18 to 1244 with a median of 122; 71% of these studies had sample sizes with less than 200. For matched gender samples, the sample sizes ranged from 24 to 580 with a median of 80; 80% of these studies had sample sizes with less than 200.

The wave characteristics were similar for the aggregated and matched gender samples (see Table 7). For the aggregated samples, the age of the child participants at T1 ranged from 1.1 to 6.2 years with a median of 4 years. For the matched gender samples, the age of the child participants at T1 ranged from 1.5 to 6 years with a median of 4 years. The average age at T2 ranged from 2 to 24 years with a median of 6 years for aggregated group and ranged from 2 to 15.5 years with a median of 6.9 years for matched gender groups. At T1, most of the children were between 3 and 6 years old and at T2 most of the children were between 3 and 12 years old across different samples. A majority of the T1 and T2 measurements were taken after the 1980s. For the aggregated samples, the time intervals between T1 and T2 measures ranged from .04 to 20.5 years with a median of 2 years. For the matched gender samples, the time intervals between T1 and T2 measures ranged from .3 to 13.3 years with a median of 2.8 years.

For both the aggregated and matched gender samples, a majority of the measures assessed the construct of oppositional defiant behaviors, aggression, and conduct disorders at T1 and T2 (see Table 8). Close to 90% of the measures used at T1 and T2 were behavior scales consisting of multiple items derived through factor analysis techniques or validated through empirical research. Very few studies used measures with single items (e.g., "can't sit still") at either time points. Questionnaires and clinical interviews constituted the most frequent methods

Table 7. Wave characteristics. The number of unique wave pairs for the aggregated sample is 163 for the aggregated sample and 132 for the matched gender sample.

	Aggregated sample		**Matched gender sample**	
Variable	**N**	**%**	**N**	**%**
Mean Age at T1				
T1 age 0-3	36	22.1%	42	31.8%
T1 age 3-6	127	77.9%	90	68.2%
Mean Age at T2				
T1 age 0-3	3	1.8%	4	3.0%
T1 age 3-6	73	44.8%	32	24.2%
T2 age 6-12 yrs	77	47.2%	76	57.6%
T2 adolescent age 12-18 yrs	9	5.5%	20	15.2%
T2 age >18 yrs	1	0.6%	0	0.0%
Mean year T1 measure was taken				
1945-1959	6	3.7%	0	0.0%
1960-1969	31	19.0%	28	21.2%
1970-1979	25	15.3%	46	34.8%
1980-1989	39	23.9%	40	30.3%
1990-2002	62	38.0%	18	13.6%
Mean year T2 measure was taken				
1945-1959	6	3.7%	0	0.0%
1960-1969	25	15.3%	16	12.1%
1970-1979	21	12.9%	34	25.8%
1980-1989	42	25.8%	64	48.5%
1990-2001	69	42.3%	18	13.6%
Mean interval between measures				
Less than 1 year	39	23.9%	12	9.1%
1-3 years	60	36.8%	54	40.9%
3-5 years	32	19.6%	24	18.2%
5-10 years	28	17.2%	34	25.8%
10 -15 years	3	1.8%	8	6.1%
Over 15 years	1	0.6%	0	0.0%

Table 8. Measurement characteristics. A total of 305 measures were used in the aggregated sample and 216 measures were used in the matched gender sample.

	Aggregated sample		**Matched gender sample**	
Variable	**N**	**%**	**N**	**%**
Type of behavior at T1				
ADHD	38	12.5%	26	12.0%
Oppositional/aggressive/conduct	200	65.6%	156	72.2%
Comorbidity	67	22.0%	34	15.8%
Type of behavior at T2				
ADHD	30	9.8%	26	12.0%
Oppositional/aggressive/conduct	215	70.5%	144	66.7%
Comorbidity	60	19.7%	46	21.3%
Type of T1 measures				
Single item	33	10.8%	6	2.8%
Multiple items	272	89.2%	210	97.2%
Type of T2 measures				
Single item	5	1.6%	0	0.0%
Multiple items	300	98.4%	216	100.0%
How T1 Measure was collected				
Questionnaire	165	54.1%	104	48.1%
Interview	44	14.4%	26	12.0%
Observation	94	30.8%	86	39.8%
More than one type of measure	2	7.0%	0	0.0%
How T2 Measure was collected				
Questionnaire	182	59.7%	110	50.9%
Interview	38	12.5%	26	12.0%
Archival record	13	4.3%	6	2.8%
Observation	72	23.6%	74	34.3%
Source of T1 measure				
Parent	128	42.0%	92	42.6%
Teacher	67	22.0%	38	17.6%
Peer	13	4.3%	0	0.0%
Researcher/observer	94	30.8%	86	39.8%
Multiple sources	3	1.0%	0	0.0%
Source of T2 measure				
Self-report	15	4.9%	2	0.9%
Parent	96	31.5%	90	41.7%
Teacher	86	28.2%	42	19.4%
Peers	18	5.9%	4	1.9%

Table 8, continued

Police	12	3.9%	4	1.9%
Researcher/observers	72	23.6%	74	34.3%
Multiple sources	6	2.0%	0	0.0%

of behavior assessments at both T1 and T2. Parents, teachers, and observers/researchers were the top three most frequent informants for assessing children's externalizing behavior problems at both T1 and T2.

A total of 607 effect sizes was used for data analysis for the aggregated sample and 325 effect sizes were used for the matched gender samples (see Table 9). Over 90% of the effect sizes were calculated on continuous measures for both aggregated and matched gender samples at both T1 and T2. Very few effect sizes were calculated on dichotomized or discrete measures.

Table 9. Effect size characteristics (N=607 for the aggregated sample and N=325 for the matched gender sample).

	Aggregated sample		**Matched gender sample**	
Variable	**N**	**%**	**N**	**%**
Scaling of T1 measure				
Dichotomized	27	4.4%	6	1.8%
Discrete (3-8 categories)	48	8.0%	29	8.9%
Continuous (>8 categories)	532	87.6%	290	89.2%
Scaling of T2 measure				
Dichotomized	25	4.1%	6	1.8%
Discrete (3-8 categories)	19	3.1%	21	6.5%
Continuous (>8 categories)	563	92.8%	298	91.7%
Original effect size statistics				
Correlation	577	95.1%	317	97.5%
Crosstab frequency table	16	2.6%	6	1.8%
Chi square analysis	6	1.0%	0	0.0%
Mean and Standard Deviation	6	1.0%	0	0.0%
Imputed from significance level	2	0.3%	2	0.6%

Over 95% of the effect sizes were reported in the form of Pearson correlation coefficients. Less than 10% of the effect sizes were calculated by using information in frequency tables, Chi square test results, mean and standard deviation, or the significance level of the correlation coefficients provided in the reports across samples.

Magnitude of Weighted Mean Effect Sizes

Table 10 presents the weighted mean effect sizes for different samples using both fixed and random effects models. In this analysis effect sizes were aggregated by independent samples. For the aggregated sample, the weighted mean effect size estimated using a fixed effects model was .40, significantly greater than zero, with a 95% confidence interval of .39 to .42. For matched gender samples, the weighted mean effect size was .44 for boys (95% confidence interval: .41 to .47), and .38 for girls (95% confidence interval: .35 to .42). The weighted mean effect sizes estimated by random effects models were slightly higher than those estimated by fixed effects models with wider 95% confident intervals. The Q-tests for between-group difference were conducted to determine whether boy and girl effect sizes were significantly different using both fixed and random effects models. The Q-test for between-group difference using the fixed effects model showed significant between differences with Hedges' Q between = 4.70 and *p*=.03 indicating the stability effect sizes were larger for boys than for girls. However, the Q-test for between group-difference using the random effects model did not show significant group differences.

Table 10. Aggregated effect sizes by sample. ES refers to effect sizes. 95% CI refers to 95% confidence interval. Winsor refers to weighted mean effect sizes calculated using winsorized sample sizes. Original refers to weighted mean effect sizes calculated using original sample sizes. *Q* refers to homogeneity test used to determine heterogeneity in the sample. *p* refers to the *s*ignificant level for homogeneity tests. *N* refers to the number of effect sizes used in estimating the weighted mean effect sizes.

Sample	**Type of ES**	**Weighted mean ES**	**-95% CI**	**+95% CI**	***Q***	***P***	***N***
		Fixed effects model					
Aggregated sample	Winsor	0.40	0.39	0.42	392.81	0.00	72
	Original	0.39	0.38	0.41	451.08	0.00	72
Boy sample	Winsor	0.44	0.41	0.47	80.38	0.00	27
	Original	0.44	0.41	0.47	80.38	0.00	27
Girl sample	Winsor	0.38	0.35	0.42	104.95	0.00	27
	Original	0.38	0.35	0.42	104.95	0.00	27
		Random effects model					
Aggregated sample	Winsor	0.43	0.39	0.47	69.95	0.51	72
	Original	0.43	0.39	0.47	69.99	0.51	72
Boy sample	Winsor	0.45	0.39	0.51	24.74	0.53	27
	Original	0.45	0.39	0.51	24.81	0.53	27
Girl sample	Winsor	0.39	0.32	0.47	27.52	0.38	27
	Original	0.39	0.32	0.47	27.52	0.38	27

Next, weighted mean effect sizes were examined by sample, informant categories (if T1 and T2 used the same or different types of informant), T1 construct categories (i.e., ADHD, Aggression/CD, and Comorbidity), and T2 construct categories (i.e., ADHD, Aggression/CD, and Comorbidity) for the aggregated sample using fixed effects models (see Table 11) and random effects models (see Table 12). A clear pattern of informant and construct effects was seen using both fixed and random effects models. The weighted mean effect sizes measuring the same subtype of externalizing behaviors rated by the same type of informant at T1 and T2 were much higher than effect sizes measuring different subtypes of externalizing behaviors rated by different types of informant. For example, in the fixed effects model (Table 11), the weighted mean effect size was .50 if T1 and T2 measured aggression/CD rated by the same informant.

However, the weighted mean effect size was only .23 if T1 and T2 measured aggression/CD but used different informants. In another example, the weighted mean effect size was .55 if T1 and T2 measured the same subtypes of externalizing behaviors—comorbidity of ADHD and CD by the same type of informant. But the weighted mean effect size was only .32 if T1 measured comorbidity and T2 measured ADHD by different types of informant.

Table 11. Weighted mean effect sizes by sample, construct, and informant for the aggregated sample using fixed effects models. ES refers to effect sizes. 95% CI refers to 95% confidence interval. *Q* refers to homogeneity test used to determine heterogeneity in the sample. *p* refers to the *s*ignificant level for homogeneity tests. *N* refers to the number of effect sizes used in estimating the weighted mean effect sizes. n/a refers to Not Available.

Time 1	**Time 2**	**Weighted**	**95% CI**				
Construct	**Construct**	**Mean ES**	**Lower**	**Upper**	***Q***	***p***	***N***
		Same informant at time 1 and time 2					
	ADHD	0.41	0.35	0.48	22.097	0.01	11
ADHD	CD	0.41	0.29	0.51	0	n/a	1
	Comorbidity	n/a	n/a	n/a	n/a	n/a	n/a
	ADHD	0.30	0.10	0.47	0.48	0.49	2
CD	CD	0.50	0.48	0.52	249.80	0.00	49
	Comorbidity	0.42	0.38	0.47	23.66	0.01	11
	ADHD	0.32	0.19	0.44	1.86	0.39	3
Comorbidity	CD	0.39	0.27	0.49	1.95	0.58	4
	Comorbidity	0.55	0.51	0.58	34.93	0.00	17
		Different informants at time 1 and time 2					
	ADHD	0.23	0.15	0.31	0.92	0.82	4
ADHD	CD	0.15	0.11	0.20	2.25	0.90	7
	Comorbidity	0.20	-0.03	0.41	0.00	n/a	1
	ADHD	0.19	0.09	0.29	0.18	0.67	2
CD	CD	0.23	0.19	0.26	24.03	0.09	17
	Comorbidity	0.25	0.14	0.36	4.77	0.57	7
	ADHD	-0.01	-0.40	0.39	0.00	n/a	1
Comorbidity	CD	0.25	0.15	0.34	5.19	0.52	7
	Comorbidity	0.38	0.31	0.44	5.50	0.48	7

Table 12. Weighted mean effect sizes by construct and informant for the aggregated sample using random effects models. ES refers to effect sizes. 95% CI refers to 95% confidence interval. *Q* refers to homogeneity test used to determine heterogeneity in the sample. *p* refers to the significant level for homogeneity tests. *N* refers to the number of effect sizes used in estimating the weighted mean effect sizes. n/a means Not Available.

Time 1	**Time 2**	**Weighted**	**95% CI**				
Construct	**Construct**	**Mean ES**	**Lower**	**Upper**	***Q***	***p***	***N***
			Same informant at time 1 and time 2				
ADHD	ADHD	0.40	0.29	0.49	10.34	0.41	11
	CD	n/a	n/a	n/a	n/a	n/a	n/a
	Comorbidity	n/a	n/a	n/a	n/a	n/a	n/a
CD	ADHD	0.30	0.10	0.47	0.48	0.49	2
	CD	0.49	0.45	0.54	50.57	0.37	49
	Comorbidity	0.41	0.32	0.48	5.58	0.85	11
Comorbidity	ADHD	0.32	0.19	0.44	1.86	0.39	3
	CD	0.39	0.27	0.49	1.95	0.58	4
	Comorbidity	0.55	0.50	0.61	16.33	0.43	17
			Different informants at time 1 and time 2				
ADHD	ADHD	0.23	0.15	0.31	0.92	0.82	4
	CD	0.15	0.11	0.20	2.25	0.90	7
	Comorbidity	n/a	n/a	n/a	n/a	n/a	n/a
CD	ADHD	0.19	0.09	0.29	0.18	0.67	2
	CD	0.22	0.18	0.26	15.36	0.50	17
	Comorbidity	0.25	0.14	0.36	4.77	0.57	7
Comorbidity	ADHD	n/a	n/a	n/a	n/a	n/a	n/a
	CD	0.25	0.15	0.34	5.19	0.52	7
	Comorbidity	0.38	0.31	0.44	5.50	0.48	7

For matched gender groups, weighted mean effect sizes were examined by sample, and T1 and T2 informant and construct categories using fixed (see Table 13) and random effects models (see Table 14). Similar informant and construct effects were found for the boy and girl samples as were found for the aggregated sample using both fixed and random effects models. The weighted mean effect sizes were much higher if T1 and T2 used the same type of informant and measured the same subtype of externalizing behaviors. The Q-tests of between-group difference were conducted for samples with at least 10 matched gender effect sizes using both fixed and random effects models. The stability of externalizing behaviors was similar for boys and girls. Only one Q-test of between-group difference using the fixed effects model approached significance level with Hedges' Q = 2.56 and *p* = .11. The stability of externalizing behaviors rated by the same type of informant and measuring aggression/CD at both T1 and T2 were slightly higher for boys than for girls. Results from random effects models found no significant gender differences.

Table 13. Weighted mean effect sizes by sample, construct, and informant for the matched gender samples using fixed effects models. ES refers to effect sizes. 95% CI refers to 95% confidence interval. *Q* refers to homogeneity test used to determine heterogeneity in the sample. *p* refers to the significant level for homogeneity tests. *N* refers to the number of effect sizes used in estimating the weighted mean effect sizes. n/a means Not Available.

	Time 1	**Time 2**	**Weighted**	**95% CI**				
Gender	**Construct**	**Construct**	**Mean ES**	**Lower**	**Upper**	***Q***	***p***	***N***
		Same informant at time 1 and time 2						
		ADHD	0.48	0.38	0.57	9.557	0.05	5
	ADHD	CD	0.50	0.22	0.70	0	n/a	1
		Comorbidity	n/a	n/a	n/a	n/a	n/a	n/a
		ADHD	0.41	0.22	0.58	0.29	0.59	2
Boys	CD	CD	0.46	0.42	0.49	60.94	0.00	19
		Comorbidity	0.36	0.26	0.45	1.04	0.90	5
		ADHD	n/a	n/a	n/a	n/a	n/a	n/a
	Comorbidity	CD	n/a	n/a	n/a	n/a	n/a	n/a
		Comorbidity	0.56	0.49	0.62	1.76	0.88	6
		ADHD	0.39	0.28	0.49	9.096	0.06	5
	ADHD	CD	0.33	-0.06	0.63	0	n/a	1
		Comorbidity	n/a	n/a	n/a	n/a	n/a	n/a
		ADHD	0.28	0.05	0.48	0.32	0.57	2
Girls	CD	CD	0.41	0.37	0.45	90.74	0.00	19
		Comorbidity	0.28	0.17	0.38	4.55	0.34	5
		ADHD	n/a	n/a	n/a	n/a	n/a	n/a
	Comorbidity	CD	n/a	n/a	n/a	n/a	n/a	n/a
		Comorbidity	0.55	0.48	0.62	5.13	0.40	6

Table 13 continued

		Different informants at time 1 and time 2						
Boys	ADHD	ADHD	0.27	0.10	0.41	0.42	0.52	2
		CD	0.32	0.18	0.45	3.00	0.08	2
		Comorbidity	n/a	n/a	n/a	n/a	n/a	n/a
	CD	ADHD	n/a	n/a	n/a	n/a	n/a	n/a
		CD	0.19	0.09	0.29	3.04	0.55	5
		Comorbidity	0.32	0.11	0.51	0.68	0.41	2
	Comorbidity	ADHD	n/a	n/a	n/a	n/a	n/a	n/a
		CD	0.39	0.14	0.60	0.02	0.88	2
		Comorbidity	0.43	0.17	0.63	0.09	0.76	2
Girls	ADHD	ADHD	0.27	0.10	0.42	0.581	0.45	2
		CD	0.34	0.21	0.46	2.022	0.155	2
		Comorbidity	n/a	n/a	n/a	n/a	n/a	n/a
	CD	ADHD	n/a	n/a	n/a	n/a	n/a	n/a
		CD	0.16	0.06	0.26	3.87	0.42	5
		Comorbidity	0.16	-0.12	0.41	2.77	0.10	2
	Comorbidity	ADHD	n/a	n/a	n/a	n/a	n/a	n/a
		CD	0.39	0.12	0.60	0.06	0.81	2
		Comorbidity	0.38	0.11	0.60	0.04	0.85	2

Table 14. Weighted mean effect sizes by sample, construct, and informant for the matched gender samples using random effects models. ES refers to effect sizes. 95% CI refers to 95% confidence interval. *Q* refers to homogeneity test used to determine heterogeneity in the sample. *p* refers to the significant level for homogeneity tests. *N* refers to the number of effect sizes used in estimating the weighted mean effect sizes. n/a means not available.

	Time 1	**Time 2**	**Weighted**	**95% CI**				
Gender	**Construct**	**Construct**	**Mean ES**	**Lower**	**Upper**	***Q***	***p***	***N***
		Same informant at time 1 and time 2						
Boys	ADHD	ADHD	0.47	0.30	0.61	4.961	0.29	5
		CD	n/a	n/a	n/a	n/a	n/a	n/a
		Comorbidity	n/a	n/a	n/a	n/a	n/a	n/a
	CD	ADHD	0.41	0.22	0.58	0.29	0.59	2
		CD	0.47	0.40	0.54	18.64	0.41	19
		Comorbidity	0.36	0.26	0.45	1.04	0.90	5
	Comorbidity	ADHD	n/a	n/a	n/a	n/a	n/a	n/a
		CD	n/a	n/a	n/a	n/a	n/a	n/a
		Comorbidity	0.56	0.49	0.62	1.76	0.88	6.0

Table 14 Continued

Girls	ADHD	ADHD	0.37	0.18	0.53	3.581	0.47	5
		CD	n/a	n/a	n/a	n/a	n/a	n/a
		Comorbidity	n/a	n/a	n/a	n/a	n/a	n/a
	CD	ADHD	0.28	0.05	0.48	0.32	0.57	2
		CD	0.44	0.34	0.53	21.73	0.24	19
		Comorbidity	0.30	0.17	0.42	3.55	0.47	5
	Comorbidity	ADHD	n/a	n/a	n/a	n/a	n/a	n/a
		CD	n/a	n/a	n/a	n/a	n/a	n/a
		Comorbidity	0.55	0.48	0.62	5.13	0.40	6
		Different informants at time 1 and time 2						
Boys	ADHD	ADHD	0.27	0.10	0.41	0.42	0.52	2
		CD	0.28	-0.007	0.52	1.00	0.32	2
		Comorbidity	n/a	n/a	n/a	n/a	n/a	n/a
	CD	ADHD	n/a	n/a	n/a	n/a	n/a	n/a
		CD	0.19	0.09	0.29	3.04	0.55	5
		Comorbidity	0.32	0.11	0.51	0.68	0.41	2
	Comorbidity	ADHD	n/a	n/a	n/a	n/a	n/a	n/a
		CD	0.39	0.14	0.60	0.02	0.88	2
		Comorbidity	0.43	0.17	0.63	0.09	0.76	2
Girls	ADHD	ADHD	0.27	0.10	0.42	0.581	0.45	2
		CD	0.32	0.11	0.50	1	0.317	2
		Comorbidity	n/a	n/a	n/a	n/a	n/a	n/a
	CD	ADHD	n/a	n/a	n/a	n/a	n/a	n/a
		CD	0.16	0.06	0.26	3.87	0.42	5
		Comorbidity	0.19	-0.27	0.58	1.00	0.32	2
	Comorbidity	ADHD	n/a	n/a	n/a	n/a	n/a	n/a
		CD	0.39	0.12	0.60	0.06	0.81	2
		Comorbidity	0.38	0.11	0.60	0.04	0.85	2

Practical Significance of the Effect Sizes

To understand the practical significance of the correlational effect sizes, the stability effect sizes were translated into various indicators of predictive accuracy in terms of positive predictive value, sensitivity, negative predictive value, and specificity using different prevalence rates of externalizing behaviors at T1 and T2 measurement points. *Positive predictive value* is the proportion of children with externalizing behavior problems at T1 who continue to have externalizing behavior problems at T2. *Sensitivity* is the proportion of children with externalizing behavior problems at T2 who also have externalizing behavior problems at T1. *Negative predictive value* is the proportion of children without externalizing behavior problems at T1 who did not develop externalizing behavior problems at T2. *Specificity* is the proportion of children without externalizing behavior problems at T2 who do not have externalizing behavior problems at T1 either. To simplify interpretation, the prevalence rates of T1 and T2 externalizing behavior problems were set at the same values: 5%, 10%, and 20% because the prevalence rates of externalizing behaviors were reported in the range of less than 5%, between 5-10%, and between 10-20% for preschoolers as well as school age children from the normal population (American Psychiatric Association, 2000; Achenbach, 1991; Achenbach & Rescorla, 2000; Bird, 1996; Caruso & Corsini, 1994; Lavigne et al., 1996; Offord, Boyle, Racine, Szatmari, Fleming, Sanford, & Lipman, 1996; Rietveld, Hudziak, Bartels, Beijsterveldt, & Boomsma, 2004). The prevalence rates of externalizing behaviors reported in the literature were based on different measures and thresholds for defining clinical behaviors, and the prevalence rates that were set for translating correlational effect sizes into predictive accuracy were rather arbitrary. However, the indicators derived for predictive accuracy might, to certain extent, represent the reality of research on externalizing behaviors problems in children.

Table 15 displays the indicators of predictive accuracy when the prevalence rates of externalizing behaviors at T1 and T2 were set at 5%, 10%, and 20% across different stability effect sizes. The values in Table 15 were calculated through two steps. First, Taylor-Russell Tables (Taylor & Russell, 1939) were consulted to determine the true positives in a 2x2 frequency table of prediction for a certain correlation effect size at given prevalence rates of

Table 15. Predictive accuracy

Effect size	**Positive predictive values**	**Sensitivity**	**Negative predictive values**	**Specificity**
Prevalence rates of externalizing behaviors at T1 and T2 = .05				
0.10	0.070	0.070	0.951	0.951
0.15	0.090	0.090	0.952	0.952
0.20	0.110	0.110	0.953	0.953
0.25	0.120	0.120	0.954	0.954
0.30	0.140	0.140	0.955	0.955
0.35	0.170	0.170	0.956	0.956
0.40	0.190	0.190	0.957	0.957
0.45	0.220	0.220	0.959	0.959
0.50	0.240	0.240	0.960	0.960
Prevalence rates of externalizing behaviors at T1 and T2 = .10				
0.10	0.130	0.130	0.903	0.903
0.15	0.150	0.150	0.906	0.906
0.20	0.170	0.170	0.908	0.908
0.25	0.190	0.190	0.910	0.910
0.30	0.220	0.220	0.913	0.913
0.35	0.240	0.240	0.916	0.916
0.40	0.270	0.270	0.919	0.919
0.45	0.290	0.290	0.921	0.921
0.50	0.320	0.320	0.924	0.924
Prevalence rates of externalizing behaviors at T1 and T2 = .20				
0.10	0.240	0.240	0.810	0.810
0.15	0.260	0.260	0.815	0.815
0.20	0.280	0.280	0.820	0.820
0.25	0.310	0.310	0.828	0.828
0.30	0.330	0.330	0.833	0.833
0.35	0.360	0.360	0.840	0.840
0.40	0.380	0.380	0.845	0.845
0.45	0.410	0.410	0.853	0.853
0.50	0.440	0.440	0.860	0.860

externalizing behaviors at T1 and T2. Second, when the true positives were determined, values in other cells of the prediction table were imputed for given marginal prevalence rates and the prediction statistics in terms of positive predictive value, negative predictive value, sensitivity, and specificity were computed. Because the prevalence rate of T1 externalizing behavior problems was set to be the same as that of T2 externalizing behavior problems, the values of positive predictive power equaled the values of sensitivity and the values of negative predictive power equaled the values of specificity.

The findings were consistent across different prevalence rates and various effect size values. If T1 and T2 externalizing behaviors correlated between .10 and .50, the positive predictive values and sensitivity ranged from 7% to 24% for a prevalence rate of externalizing behaviors of 5%. The positive predictive values and sensitivity ranged from 13% to 32% for a prevalence rate of externalizing behaviors of 10%. The positive predictive values and sensitivity ranged from 24% to 44% for a prevalence rate of externalizing behaviors of 20%. The positive predictive values and sensitivity were higher if the prevalence rates of externalizing behaviors and stability effect sizes were larger. Nevertheless, these indicators were all below 50% under the three prevalence conditions (5%, 10%, and 20%) and across different values of the correlational effect sizes. Less than 55% of the children with externalizing problem behaviors at T1 would be expected to show externalizing behavior problems at T2 in all these prevalence scenarios. Likewise, 7% to 54% of the children with externalizing behaviors at T2 (46% to 93% false negative rates) would be expected to have externalizing behavior problems at T1. The negative predictive values and specificity were in the range of 90% if the prevalence rates of externalizing behaviors were 5% and 10%, and in the range of 80% if the prevalence rate was 20%. Thus, children without externalizing behavior problems at T1 are not likely to develop them at T2 and children without externalizing behavior problems at T2 were not likely to have them at T1 (with 10% to 20% false positive rates). In sum, using Time 1 externalizing problem behaviors to predict Time 2 externalizing problem behaviors may result high error rates.

Accounting for Effect Size Variability in the Aggregated Sample

To account for the variability in the effect sizes for the aggregated sample, weighted multiple regression analyses were conducted using both fixed and mixed effects models. Effect sizes were first averaged to uniquely represent the combination of following criteria: (a) independent samples, (b) T1 age categories (i.e., before and after age 3), (c) informant categories (i.e., whether T1 and T2 used the same or different types of informants), and (d) construct categories (i.e., whether T1 and T2 measured the same or different subtypes of externalizing behaviors). For studies that contributed more than one effect size according to these selection criteria, random numbers were created and one effect size was randomly chosen for each independent sample to form the data set for the analysis. Random numbers were used to select one effect size for 27 (of 70) aggregated samples that provided more than one effect size after effect sizes were averaged by the criteria discussed above.

Predictor variables in the weighted multiple regression analyses were chosen based on the strength of their correlations with the dependent variables, the intercorrelations of the predictor variables, and the theoretical importance of the predictor variables in accounting for the stability of externalizing behavior problems. A correlation matrix is included in Appendix A on the relationship among effect size and other study variables. The following predictors were used in the regression models: T1 age categories (1=Before age 3 at T1 and 0=After age 3 at T1), informant effect (1=T1 and T2 used the same type of informants and 0=T1 and T2 used different types of informants), construct effect (1=T1 and T2 measured the same subtypes of externalizing behaviors and 0=T1 and T2 measured the same subtypes of externalizing behaviors), the time interval in years between T1 and T2 measurement points, the proportion of children from low

SES backgrounds, and the proportion of Caucasian children in the sample. The results of the fixed and mixed effects models are summarized in Table 16.

Table 16. Summary of the weighted regression analysis for the aggregated sample

Variable	***β***	***B***	**-95% *CI***	**+95% *CI***	***Q***	***p***	***df***
Fixed effects model							
T1 age categories	-0.08	-0.05	-0.10	0.01	2.68	0.10	
Informant effects	0.51	0.24	0.19	0.29	87.24	0.00	
Construct effects	0.19	0.10	0.04	0.15	12.09	0.00	
Time interval	-0.16	-0.01	-0.01	0.00	7.76	0.01	
Low SES	-0.15	-0.10	-0.18	-0.02	6.77	0.01	
Race	-0.12	-0.09	-0.18	-0.01	4.52	0.03	
Constant	0.00	0.34	0.21	0.46	29.42	0.00	
Overall model						282.72	6
Residual						227.14	65
R-square: .56							
Random effects model							
T1 age categories	-0.06	-0.03	-0.13	0.07	0.36	0.55	
Informant effects	0.43	0.23	0.12	0.35	16.91	0.00	
Construct effects	0.19	0.10	-0.01	0.22	3.26	0.07	
Time interval	-0.18	-0.01	-0.03	0.00	2.62	0.11	
Low SES	-0.07	-0.04	-0.21	0.12	0.26	0.61	
Race	-0.09	-0.06	-0.25	0.13	0.41	0.52	
Constant	0.00	0.32	0.06	0.57	5.76	0.02	
Overall model					49.05	0.00	6
Residual					68.70	0.35	65
R-square: .417							

Results from the fixed effects model indicated all predictors had significant independent effects on the magnitude of effect sizes. The externalizing behaviors tended to be more stable if T1 measures were taken when children were after than before 3 years old. Effect sizes using the same type of informants at T1 and T2 were much larger than the ones using different informants. Similarly, effect sizes measured on the same subtype of externalizing behaviors at T1 and T2 were much larger than the ones measured on different subtypes of externalizing behaviors at the two measurement points. As the time interval between T1 and T2 measurements increased, the stability of externalizing behaviors decreased. Externalizing behaviors were less stable in samples of children from low SES and Caucasian backgrounds. The fixed effects model was significant and accounted for 56% of the variation in the effect sizes.

Result from the mixed effects model showed that only informant and construct predictors had significant independent effects on the effect sizes. The effect of time interval between T1 and T2 measurements on effect sizes was approaching significance (p =.11). The standardized beta weights were similar in the mixed and fixed effects models. The direction of effects of these significant predictors in the mixed effects model was the same as in the fixed effects model. The mixed effects model was significant and accounted for 42% of the variation in the effect sizes.

Accounting for Effect Size Variability in the Matched Gender Studies

Because of the relatively small number of the aggregated effect sizes in matched gender groups, there was not enough statistical power to test all the predictors and their interaction effects with gender simultaneously in a single weighted multiple regression analysis. Therefore, the examination of variability in the effect sizes for matched gender samples was done in three steps. First, weighted multiple regression analyses were conducted to test the effects of gender, T1 age categories, informant effects, and construct effects. Second, if any main effects were found for these predictors, the significant predictors were entered in a second weighted multiple regression model along with the interaction terms between gender and those significant predictors. Third, a weighted multiple regression analysis was performed by choosing the aggregated effect sizes measured on the same subtype of externalizing behaviors and using the same type of informant at both T1 and T2 to determine the contribution of demographic variables on the variability of effect sizes while keeping the informant and construct effects constant. A correlation matrix is included in Appendix B on the relationship among effect size and other study variables for the matched gender sample.

Step 1

A weighted multiple regression analysis was conducted to determine the main effect of gender, T1 age categories, informant effects, and construct effects. Effect sizes for this analysis were first averaged to uniquely represent the combination of the following criteria: (a) independent matched gender samples, (b) T1 age categories (i.e., before and after age 3), c) informant effect (whether T1 and T2 used the same or different types of informants), and (d) construct effect (i.e., whether T1 and T2 measured the same or different subtypes of externalizing behaviors). For studies that contributed more than one effect size on these selection criteria, random numbers were created and one effect size was randomly chosen for each independent sample to form the data set for the analysis.

The weighted multiple regression analyses using fixed and mixed effects models were conducted with the following predictors: gender (1=Girls and 2=Boys), T1 age categories (1=Before age 3 and 0=After age 3), informant effect (1=T1 and T2 used the same type of informants and 0= T1 and T2 used different types of informants), and construct effect (1=T1 and T2 measured the same subtypes of externalizing behaviors and 0=T1 and T2 measured different subtypes of externalizing behaviors). The results of the weighted regression analyses (see Table 17) indicated the significant main effect of informant and subtypes of externalizing problem behaviors using both fixed and random effects models. The effect sizes measured on the same

Table 17. Summary of weighted regression analysis for the matched gender sample—Step I

Variable	***β***	***B***	**-95% *CI***	**+95% *CI***	***Q***	***p***	***df***
				Fixed effects model			
Gender	-0.16	-0.66	-1.25	-0.07	4.81	0.03	
T1 age categories	-0.04	-0.20	-0.94	0.55	0.27	0.60	
Informant effects	0.31	1.88	1.00	2.77	17.44	0.00	
Construct effects	0.21	0.96	0.30	1.61	8.16	0.00	
Constant	0.00	0.36	0.30	0.41	165.38	0.00	
Overall model					40.44	0.00	4
Residual					152.84	0.00	49
R-square: .21							
				Random effects model			
Gender	-0.14	-0.62	-1.75	0.51	1.17	0.28	
T1 age categories	0.02	0.12	-1.19	1.43	0.03	0.86	
Informant effects	0.24	1.65	-0.14	3.44	3.25	0.07	
Construct effects	0.24	1.11	-0.08	2.30	3.33	0.07	
Constant	0.00	0.37	0.27	0.48	53.31	0.00	
Overall model					9.29	0.05	4
Residual					50.24	0.42	49
R-square: .16							

subtype of externalizing behaviors rated by the same type of informant at T1 and T2 were significantly larger than those measured on different subtypes of externalizing behaviors rated by different informants. This result on the effects of informant and construct replicated the findings in the aggregated sample. Gender had significant independent effect on effect sizes in the fixed effects model: externalizing behaviors were more stable for boys than for girls. Gender had limited generalizability because its effect was not significant in the mixed effects model. T1 age categories had negligible effects in either the fixed or mixed effects models. Both fixed and mixed effects models were significant, accounting for 21% and 16% variances respectively.

Step 2

The second weighted multiple regression analysis was conducted to determine if there were any interaction effects between gender and the variables that had significant main effects in the first regression models. Because the stability effect sizes were not influenced by children's age at T1 as indicated in the results from the first regression analysis, the variable of T1 age category was not entered in the second regression model. The predictors tested in the weighted multiple regression models were: gender, informant effect, construct effect, the interaction between gender and informant effect, and the interaction between gender and construct effect (see Table 18).

Table 18. Summary of weighted regression analysis for the matched gender sample—Step II

Variable	***β***	***B***	**-95% *CI***	**+95% *CI***	***Q***	***p***	***df***
				Fixed effects model			
Gender	-0.19	-0.75	-1.85	0.34	1.81	0.18	
Informant effects	0.31	1.88	0.83	2.93	12.31	0.00	
Construct effects	0.21	0.97	0.31	1.62	8.40	0.00	
Gender and informant interaction	0.02	1.69	-19.28	22.66	0.02	0.87	
Gender and construct interaction	0.01	0.78	-12.32	13.88	0.01	0.91	
Constant	0.00	0.35	0.30	0.41	158.74	0.00	
Overall model					40.21	0.00	5
Residual					153.07	0.00	48
R-square: .21							
				Mixed effects model			
Gender	-0.15	-0.66	-2.60	1.29	0.44	0.51	
Informant effects	0.24	1.62	-0.29	3.54	2.76	0.10	
Construct effects	0.24	1.08	-0.11	2.27	3.17	0.07	
Gender and informant interaction	-0.01	-1.09	-39.38	37.20	0.00	0.96	
Gender and construct interaction	0.05	4.74	-19.08	28.57	0.15	0.70	
Constant	0.00	0.38	0.28	0.48	58.79	0.00	
Overall model					9.115	0.11	5
Residual					48.77	0.44	48
R-square: .16							

Similar findings were seen using the fixed and mixed effects models. Gender had no significant main effect; however, both informants and constructs had significant main effects in the model. The effect sizes were significantly higher if they were measured on the same subtype of externalizing behaviors rated by the same type of informant at T1 and T2. The effects of informant and construct were the same for the boy and girl samples because the interaction effects between gender and informant and between gender and construct were not statistically significant. The fixed effect model was significant, accounting for 21% of the variance. The mixed effects model approached the significance level (*p*=.11), accounting for 16% of the variance.

Step 3

Because informants and constructs had significant effects on the magnitude of effect sizes, their effects needed to be kept constant when testing the effects of other important variables. To accomplish this, effect sizes were aggregated by samples, informant (i.e., whether T1 and T2 used the same or different types of informants), and construct (i.e., whether T1 and T2 measured the same or different subtypes of externalizing behaviors). Only those effect sizes measured on the same subtype of externalizing behaviors rated by the same type of informants at T1 and T2 were chosen for the third weighted multiple regression analysis to keep the informant and construct effect constant. Because T1 age category was not a significant predictor for effect sizes for the matched gender studies, effect sizes were aggregated across different T1 age categories.

The following predictors were entered in the third weighted multiple regression model simultaneously: gender (0=Girls and 1=Boys), the time interval in years between T1 and T2 measurement points, the proportion of children from low SES in the sample, the interaction between gender and time interval, and the interaction between gender and low SES. The fixed effects model was significant and accounted for 29% of the variance (see Table 19). In the fixed effects model, three variables showed significant main effects: gender, time interval, and low SES. Externalizing behaviors were much more stable for boys than for girls. The stability of externalizing behaviors decreased when the time interval between T1 and T2 measurement points increased. Children from low socioeconomic backgrounds had more stable externalizing behaviors than children from high SES backgrounds. The significant interaction effect between

Table 19. Summary of weighted regression analysis for the matched gender sample—Step III

Variable	***β***	***B***	**-95% *CI***	**+95% *CI***	***Q***	***p***	***df***
				Fixed effects model			
Gender	0.28	0.12	-0.02	0.25	2.90	0.09	
Time interval	-0.49	-0.04	-0.06	-0.02	19.23	0.00	
Low SES	0.26	0.17	0.02	0.32	5.22	0.02	
Gender and time interval interaction	0.05	0.00	-0.02	0.03	0.08	0.77	
Gender and SES interaction	-0.38	-0.31	-0.52	-0.11	8.77	0.00	
Constant	0.00	0.58	0.49	0.68	140.37	0.00	
Overall model					48.31	0.00	5
Residual					119.02	0.00	36
R-square: .29							
				Mixed effects model			
Gender	0.33	0.15	-0.12	0.42	1.22	0.27	
Time interval	-0.35	-0.04	-0.08	0.01	2.87	0.09	
Low SES	0.20	0.12	-0.13	0.36	0.88	0.35	
Gender and time interval interaction	-0.05	-0.01	-0.06	0.05	0.03	0.86	
Gender and SES interaction	-0.38	-0.28	-0.62	0.06	2.53	0.11	
Constant	0.00	0.56	0.37	0.75	33.88	0.00	
Overall model					9.97	0.08	5
Residual					39.21	0.33	36
R-square: .20							

gender and low SES indicated the differential effect of low SES on boys and girls. For girls, coming from low SES families resulted in more stable externalizing behaviors while for boys, coming from low SES families resulted in less stable externalizing behaviors. The interaction effect between gender and time interval was not statistically significant.

To graphically examine the relationship between effect sizes and low SES for matched gender groups, the effect sizes were plotted against the percentage of children from low SES backgrounds in the sample in Figure 1. Separate prediction lines were drawn for boys and girls. The prediction lines pointed to different directions indicating the differential effects of low SES for boys and girls. Externalizing problem behaviors were more stable for girls from low SES families than for boys.

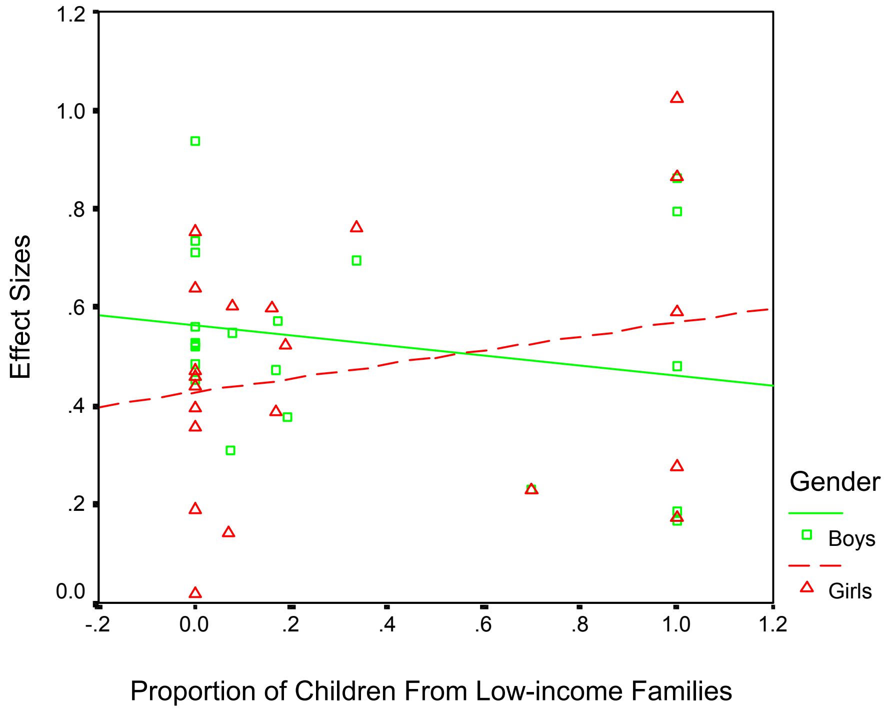

Figure 1. Interaction effect between low SES and children's gender. Regression lines indicating the relationship between stability effect sizes and proportion of children from low-income families in the boys' and girls' samples.

The relationship between effect sizes and the time interval between T1 and T2 measurements in years was plotted in Figure 2. Although the prediction line for the girl sample was steeper than that of the boy sample indicating that boys' externalizing behaviors were more stable than girls' over time, there was no significant difference between the slopes for the two groups. Effect sizes decreased significantly with increasing time intervals between measurement points and the time interval effect was similar for both the boy and girl samples.

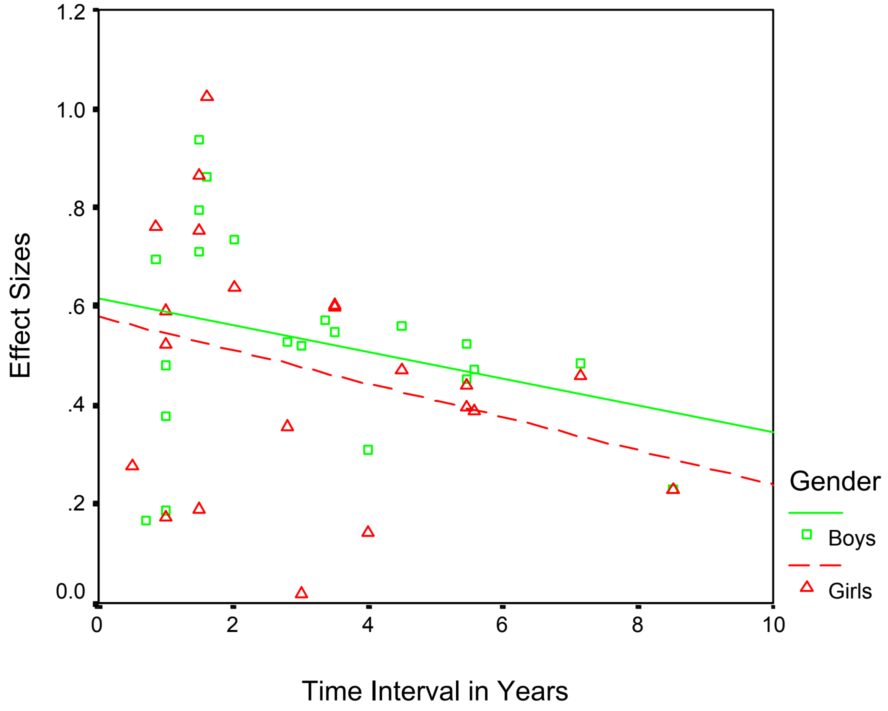

Figure 2. Interaction effect between time interval in years and children's gender. Regression line indicating the relationship between stability coefficients and time interval in years between T1 and T2 measurement points for the boys' and girls' samples.

Results from the random effects model showed that only the time interval between T1 and T2 measurements had a significant independent effect. The longer the interval between measurement points, the smaller the effect sizes. The interaction effect between gender and low

SES approached significance (*β*=-.38, *p*=.11). The random effects model was significant, accounting for 20% of the variance in the effect sizes.

CHAPTER V

DISCUSSION

This meta-analysis study synthesized the empirical research on the issues of longitudinal continuity of externalizing behaviors in children before they entered elementary school. The present investigation is the first study that systematically examined the empirical evidence on stability of externalizing behaviors with onset in early childhood. Although the sample aggregated mean effect sizes for stability of externalizing behaviors in young children were in the rage of .40, the magnitude of mean effect sizes aggregated by sample, informant types, subtypes of externalizing behaviors varied considerably. There were two types of findings with different levels of generalizability. Findings from the fixed effects models of the weighted regression analyses may be generalized to studies similar to the studies included in this meta-analysis. Findings from the mixed effects models may have a greater generalizability. Several variables including measurement features, time effects, and sample characteristics were found to account for the variability in the stability effect sizes in the fixed effects models. Those effects were replicated across aggregated and matched gender samples in this study. Findings related to informant and construct effects were robust and were supported by results from both the fixed and mixed effects models. In this study, boys' externalizing behaviors were more enduring than girls, and different sample characteristics (i.e., SES) had differential effects on the stability of externalizing behaviors in boys and girls.

Limitations of the Current Study

Several major weaknesses of the current study should be noted before discussing implications of the findings. First, the weighted mean effect sizes, such as the weighted mean effect sizes measured on different constructs using different types of informants at T1 and T2 for the matched gender samples, were calculated from a very small number of aggregated effect sizes. Because homogeneity tests and effect sizes estimates are less reliable for small samples, the confidence in these results is limited. Second, because the meta-analytic techniques for calculating effect sizes are limited, and the information that was reported in some research reports was not adequate to derive a correlational effect size, many studies were excluded from the present meta-analysis. For example, effect sizes cannot be calculated for many multivariate statistical tests, such as growth curve analysis, if descriptive statistics are not provided in the research reports. As a result, the range of the studies included in this study is limited. However, a thorough and exhaustive search had been done to include as many relevant reports as possible. The current study should reflect the reality of empirical research on the stability of externalizing behaviors in young children. Third, the stability effect sizes in the correlation format only imply whether the relative position of individuals on externalizing problem behaviors changed over time. These correlations do not indicate how externalizing problem behaviors change over time, or which participant continues to show high levels of externalizing behaviors. The magnitude of the stability effect sizes does not indicate the severity of externalizing problem behaviors. Further, stability effect sizes might be under- or over-estimated because of the time of measurement (Loeber, 1990). Some children might be in remission from severe externalizing behaviors at the time of measurement but continue to show previous pattern of high level externalizing behaviors when behavior assessments are not taken. It is also possible that some children demonstrate escalated externalizing behaviors at the time of measurement and desist later in life. Therefore, the correlational effect sizes are not an absolute index of the true magnitude of stability of externalizing problem behaviors in children.

Effects of Informants and Behavior Constructs

Despite these limitations, the present study extended the research on preschool problem behaviors in several important ways. Results from the mixed effects model indicated a strong effect of informants on the stability of externalizing problem behaviors in longitudinal studies of young children. Empirical evidence regarding informant effects on behavioral assessment in cross-sectional studies is very strong. For example, in their meta-analysis of 119 studies Achenbach et al. (1987) reported a mean correlation of .60 between similar informants, and .28 between different informants on children's behavioral/emotional functioning. The present study indicated an informant effect in the longitudinal consistency of externalizing behaviors in young children. The stability effect sizes on externalizing behaviors rated by the same informants at Time 1 and Time 2 were generally bigger than by different informants. Both the fixed and random effects models of the weighted multiple regression analyses showed significant informant effects for the stability of effect sizes across different samples suggesting that informant effects might be found in other longitudinal studies on children's problem behaviors. The consistent large informant effect suggests that a significant amount of the longitudinal continuity of externalizing behaviors may be an artifact of informants and externalizing behaviors with onset in early childhood might not be as stable as previously thought.

Another robust finding from the mixed effects model was the effect of constructs in the stability of externalizing behavior problems in young children. Different subtypes of externalizing behaviors were not identical and externalizing behaviors in general were not unidemensional. Attention deficits/hyperactivity, aggression/construct disorders, and the combined behaviors (i.e., attention/hyperactivity plus aggression/conduct disorders) measured at T1 correlated much higher with the same construct counterparts at T2 than with the different construct counterparts. Attention deficits/hyperactivity, aggression/conduct disorders, and the combined behaviors apparently are at least partially distinct types of externalizing behaviors (Campbell et al., 2000; Hinshaw, 1987; Hinshaw et al., 1993). This finding is consistent with other research studies (Hinshaw, 1992; Moffitt, 1990) indicating that different subtypes of externalizing behaviors might have different etiologies and developmental trajectories.

Effects of Time Interval, Time 1 Age, and Sample Characteristics

In addition to informant and behavior construct variables, time intervals, T1 age, and sample characteristics were found to account for the variability in the stability effect sizes. The interval between T1 and T2 measurement points was a significant predictor of stability effect sizes. This effect was replicated across aggregated and matched gender samples. The stability of externalizing problem behaviors decreased over time. Similar time interval effects were reported by Olweus (1979) and Zumkley (1994) although both Olweus and Zumkley used male samples mainly in the school age. The declining trend in the stability effect sizes over time found in this study was not as strong as informant and construct effects.

Findings from the fixed effects model for the aggregated sample suggested that externalizing behavior problems were less enduring for children assessed before age 3 than after age 3. Several lines of research in developmental psychology may shed light on why toddler's externalizing behaviors are less stable than preschoolers' externalizing behaviors. First, children before the age of 3 experience a period of intense exploration and independent seeking, therefore, they are more likely to be noncompliant with parents and other adults (Campbell, 1990; Erikson, 1963). Second, during toddlerhood, children's self-control is just emerging and they are less capable of controlling their behaviors as preschoolers (Kopp, 1982). Third, not until age 3 do children master complex language structure, and empirical findings indicate that children's noncompliance behaviors in toddlers are mainly due to their failure to understand the directions (Kaler & Kopp, 1990). Due to these developmental factors, children's externalizing behavior problems might be more temporary in the toddlerhood. As children's develop social, language, and self-control skills, their behavior pattern may become more stable over time.

Low socioeconomic status also contributed to the variability in the stability effect sizes in the aggregated and in the matched gender samples in this study. Children from low SES backgrounds had less stable externalizing problem behaviors. The finding on low SES as a predictor for the stability of externalizing behaviors might be consistent with the contextual theory that environmental factors have a significant impact on children's problem behaviors. Children's SES status was measured only at the T1 measurement point in empirical reports included in the current study. It is possible that children's SES status changed over time. Children's problem behaviors may improve or deteriorate if they move out of poverty or move into a deeper level of poverty (Kolvin et al., 1990). As a result, externalizing problem behaviors were less stable in children from low SES backgrounds. Future research is warranted to determine the relationship between the trajectories of SES and children's problem behaviors to further test the contextual theory. Results from this study also demonstrated the effect of minority backgrounds on the stability of externalizing behavior problems: externalizing behavior problems were more stable in children from minority backgrounds. Further research is needed to identify specific risk variables in minority children and in their environment that may cause these children's externalizing behaviors to persist.

Gender Differences

The effects of gender on the stability of externalizing behaviors were examined in the current study. Boys' externalizing behaviors were more enduring than girls' externalizing behaviors. Girls' externalizing behaviors also achieved a degree of stability. In addition, great variability exists in the stable effect sizes in both boy and girl samples. A portion of the heterogeneity in their stability effect sizes could be accounted for by informant, construct, time, and sample characteristic variables. Most of the predictors, such as informant, construct, and time effects, had similar impact on boys' and girls' stability effect sizes. For example, for both the boy and girl samples, the stability of externalizing behaviors declined as the time between measurement occasions increased; no gender differences were found in the rate of decline. Only low SES had differential effects on the stability of externalizing behaviors in boys and girls. Externalizing problem behaviors in boys from low SES backgrounds were less stable. In contrast, externalizing problem behaviors in girls from low SES backgrounds were more stable. Such finding extends the contextual theory and provides some evidence that developmental trajectories of externalizing problem behaviors might be different for boys and girls from low SES backgrounds.

Implications

There are several implications for research and practice related to early externalizing behavior problems in young children. First, the magnitude of stability of externalizing behaviors in young children is lower than that reported for school age children. In his review of stability of aggressive behaviors in boys, Olweus (1979) concluded that the degree of stability of aggressive behaviors was not much lower than the stability of IQ. However, in the current study the stability effect sizes for externalizing behaviors with onset before age 6 were much lower than the stability of IQ first tested in early childhood. For example, the stability effect sizes for externalizing behaviors across aggregated and matched gender samples by informant and construct ranged from -.01 to .56 with most of the effect sizes around .30. In contrast, the stability effect sizes of children's IQ first tested between 12 months to 6 years are in the range of .31 to .87 with majority of the effect sizes around .50 and .60 (Bartels, Rietveld, Van Baal, & Boomsma, 2002; Schuerger & Witt, 1989; Wilson, 1983). Young children's externalizing behaviors are not as stable as externalizing behaviors in school age children.

Second, because of the low stability of externalizing behaviors, high levels of prediction errors are inevitable when early externalizing behavior status is used to predict later antisocial behaviors. To illustrate this point, we translated the correlational effect sizes into predictive accuracy for different prevalence rates of externalizing behaviors at different measurement points. The overall pattern of findings suggested that T1 externalizing behavior status was not an accurate predictor of T2 externalizing behaviors. The current study challenges propensity theory. In early childhood, antisocial tendency as indicated by early externalizing behavior problems may not be a stable trait that would necessarily lead to later antisocial behaviors. This finding has important implications for intervention and prevention programs targeting externalizing behaviors problems in young children. Successful intervention and prevention programs begin with accurate identification of children with externalizing behavior problems. Early externalizing problem behavior alone is not sufficient to accurately predict later antisocial behaviors in nonreferred children (Bennett et al., 1998). Information about children’s externalizing behavior problems along with assessment of impairments in other areas of development (e.g., internalizing behaviors, language, cognitive development) and risks (e.g., parenting, family environment) should be considered in identifying children for intervention and prevention services (Bennett, Lipman, Brown, Racine, Boyle, & Offord, 1999; White et al., 1990).

Third, attention deficits/hyperactivity, aggression/conduct disorders, and the combined behaviors of attention/hyperactivity and aggression/conduct disorders appeared to be somewhat distinct subtypes of externalizing behaviors. At the same time, a longitudinal correlational relationship also existed among these constructs suggesting some common variance these subtypes of externalizing behaviors all share. Future research studies should examine whether there are common or different mechanisms underlying these common types of childhood externalizing behaviors and differential treatment for these behaviors.

Fourth, although the current study examined the construct effect on the stability of externalizing behaviors using the classification of three broad subtypes of externalizing problem behaviors, there are still subcategories of behaviors within each of these categories. For example, longitudinal continuity of physical vs. verbal aggression or proactive vs. reactive aggression might be quite distinctive rather than similar. In this meta-analysis, there were not enough studies to allow further examination of these important differences. With increased research on early childhood behavior problems, future meta-analytical reviews may be able to fill this void in the literature.

Fifth, although much current research has focused attention on boys’ externalizing behaviors, girls’ externalizing behaviors should not be neglected. Findings from the present study showed that girls from low SES background were more likely to have enduring externalizing behavior problems. Therefore, prevention and intervention programs should include girls from low-income background who exhibit clinical range of externalizing behavior problems. Early intervention and prevention might reduce later antisocial behaviors in girls since their early externalizing behaviors are more like to continue.

Finally, it might be inappropriate for intervention and prevention programs to target externalizing behavior problems alone in toddlers because some of the externalizing problem behaviors might be considered normal rather than psychopathological. With improvement in the areas of cognitive, language, self-control, and social development, toddlers might outgrow some of their problem behaviors without intervention or prevention efforts.

APPENDIX A

Correlation Matrix for the Aggregated Sample

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1 Effect size	1.00	-0.21	-0.34	0.18	-0.29	0.63	0.35	-0.06	-0.15	-0.09	0.26	0.25	-0.12	-0.02	0.06	-0.10
2 Time 1 age category		1.00	-0.01	-0.80	-0.26	0.08	0.06	-0.13	0.22	-0.04	0.15	0.21	-0.05	-0.07	0.17	-0.25
3 Time interval in years			1.00	-0.23	0.94	-0.22	-0.36	-0.13	0.15	-0.04	-0.26	-0.41	0.15	-0.10	-0.06	0.33
4 Time 1 age				1.00	0.10	-0.06	0.05	0.16	-0.21	0.10	-0.13	-0.16	0.09	0.09	-0.15	0.18
5 Time 2 age					1.00	-0.25	-0.36	-0.08	0.10	-0.01	-0.30	-0.47	0.19	-0.07	-0.11	0.41
6 Informant effect						1.00	0.40	0.02	-0.12	-0.05	0.28	0.27	-0.06	0.05	0.21	-0.28
7 Construct effect							1.00	0.18	-0.32	-0.05	0.22	0.22	0.04	0.08	0.00	-0.15
8 Percent of low SES								1.00	-0.62	0.44	0.19	0.11	0.31	0.41	-0.05	-0.11
9 Percent White									1.00	-0.07	0.05	0.05	-0.05	-0.23	0.02	0.04
10 Percent male										1.00	0.18	0.14	0.16	0.20	0.04	-0.16
11 Time 1 scale type											1.00	0.88	-0.12	0.05	0.11	-0.29
12 Time 2 scale type												1.00	-0.15	0.04	0.11	-0.45
13 Publication year													1.00	0.16	-0.24	0.25
14 Publication type														1.00	0.13	0.00
15 Attrition rate															1.00	-0.09
16 Sample size																1.00

APPENDIX B

Correlation Matrix for the Matched Gender Sample

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1 Effect size	1	-0.21	-0.31	0.24	-0.18	0.33	0.12	-0.26	0.23	0.19	0.11	0.10	0.04	0.01	0.03	-0.03
2 Time 1 age category		1.00	-0.11	-0.92	-0.48	0.03	-0.01	0.02	0.00	0.02	0.05	0.07	-0.26	0.18	-0.04	-0.44
3 Time interval in years			1.00	0.08	0.91	0.08	0.12	-0.13	0.14	0.00	0.06	0.06	0.34	-0.45	-0.03	0.34
4 Time 1 age				1.00	0.48	-0.02	0.05	-0.07	-0.01	-0.02	0.02	0.00	0.23	-0.15	-0.04	0.45
5 Time 2 age					1.00	0.05	0.13	-0.13	0.11	0.00	0.06	0.05	0.39	-0.45	-0.06	0.48
6 Informant effect						1.00	0.28	-0.47	0.47	0.00	-0.09	-0.08	0.22	-0.30	0.04	0.19
7 Construct effect							1.00	-0.03	0.01	-0.01	0.17	0.17	0.24	-0.29	-0.29	0.27
8 Percent of low SES								1.00	-0.82	0.01	0.03	0.03	0.02	0.19	-0.12	-0.02
9 Percent White									1.00	0.00	0.08	0.08	0.26	-0.33	0.07	0.17
10 Gender										1.00	0.01	0.01	0.02	0.00	-0.03	0.02
11 Time 1 scale type											1.00	0.96	-0.08	-0.18	-0.15	0.11
12 Time 2 scale type												1.00	-0.09	-0.17	-0.14	0.10
13 Publication year													1.00	-0.76	-0.29	0.68
14 Publication type														1.00	0.33	-0.79
15 Attrition rate															1.00	-0.24
16 Sample size																1.00

REFERENCES

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4-18 and 1991 Profile.* University of Vermont: Department of Psychiatry.

Achenbach, T. M., Edelbrock, C., & Howell, C. T. (1987a). Empirically based assessment of the behavioral/emotional problems of 2- and 3-year-old children. *Journal of Abnormal Child Psychology, 15*, 629-650.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987b). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 41,* 213-232.

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the ASEBA Preschool Forms & Profiles*. Burlington, VT: University of Vermont, Department of Psychiatry.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed.--Text Revision). Washington, DC: Author.

*Anderson, D. R. (1984). Stability of behavioral and emotional disturbance in a sample of disadvantaged preschool-aged children. *Child Psychiatry and Human Development, 14*, 249-260.

*Baker, B. L., Blacher, J., & Crnic, K. (2003). *Once a problem, always a problem?: Parenting and the continuity of child disorders from ages 3 to 5.* Paper presented at the 36th Annual Gatlinburg Conference, Annapolis, MD.

Barkley, R. A., Fischer, M., Edelbrock, C. S., & Smallish, L. (1990). The Adolescent outcome of hyperactive children diagnosed by research criteria: I. 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., Edelbrock, C., & Smallish, L. (1991). The adolescent outcome of hyperactive children diagnosed by research criteria: III. Mother-child interactions, family conflicts and maternal psychopathology. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 32*, 233-255.

Bartels, M., Rietveld, M. J. H., Van Baal, G. C. M., & Boomsma, D. I. (2002). Genetic and environmental influences on the development of intelligence. *Behavior Genetics, 32*, 237-249.

*Bates, J. E., Bayles, K., Bennett, D. S., Ridge, B., & Brown, M. M. (1989). Origins of externalizing behavior problems at eight years of age. In D. J. Pepler & K. H. Rubin (Eds.), *The development and treatment of childhood aggression* (pp. 93-120). Hillsdale, NJ: Lawrence Erlbaum.

Bates, J. E., Pettit, G. S., Dodge, K. A., & Ridge, B. (1998). Interaction of temperamental resistance to control and restrictive parenting in the development of externalizing behaviors. *Developmental Psychology, 34,* 982-995.

Becker, B. J. (2000). Multivariate meta-analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 499-526). San Diego: Academic Press.

Becker, B. J., & Schram, C. M. (1994). Examining explanatory models through research synthesis. In J. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 357-382). New York: Russell Sage Foundation.

Bennett, K. J., Lipman, E. L., Brown, S., Racine, Y., Boyle, M. H., & Offord, D. R. (1999). Predicting conduct problems: Can high-risk children be identified in kindergarten and grade I? *Journal of Consulting and Clinical Psychology, 67*, 470-480.

Bennett, K. J., Lipman, E. L., Racine, Y., Offord, D. R. (1998). Annotation: Do measures of externalizing behavior in normal populations predict later outcome?: Implications for targeted interventions to prevent conduct disorder. *Journal of Child Psychology and Psychiatry, 39*, 1059-1070.

Biederman, J., Newcorn, J., & Sprich, S. (1991). Comorbidity of attention deficit hyperactivity disorder with conduct, depressive, anxiety, and other disorders. *American Journal of Psychiatry, 148*, 564-577.

Bird, H. R. (1996). Epidemiology of childhood disorders in a cross cultural context. *Journal of Child Psychology and Psychiatry, 37,* 35-49.

Bird, H. R. (2001). Psychoanalytic perspectives on theories regarding the development of antisocial behavior. *Journal of the American Academy of Psychoanalysis, 29*, 57-71.

*Bhandari, R. P. (2003). Parenting and socio-emotional development: A longitudinal study of African American and Caucasian families (Doctoral dissertation, Wayne State University, 2003). *Dissertation Abstracts International, 63,* 5547.

*Bradley, M. E. (1999). Internalization of maternal descriptors in children at risk for the development of disruptive behavior disorder (Doctoral dissertation, University of Maryland, 1999). *Dissertation Abstracts International, 60,* 0853.

Campbell, S.B. (1990). *Behavioral problems in preschool children: Clinical and developmental issues.* New York: Guilford.

Campbell, S.B. (1995). Behavior problems in preschool children: A review of recent research. *Journal of Child Psychology and Psychiatry, 36,* 113-149.

Campbell, S. B., Shaw, D., & Gilliom, M. (2000). Early externalizing behavior problems: Toddlers and preschoolers at risk for later maladjustment. *Development and Psychopathology, 12*, 467-488.

*Carlson, E. A., Jacobvitz, D., & Sroufe, L. A. (1995). A developmental investigation of inattentiveness and hyperactivity. *Child Development, 66*, 37-54.

Caruso, G., & Corsini, D.A. (1994). The prevalence of behavior problems among toddlers in child care. *Early Education and Development, 5,* 27-40.

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

Center for Evaluation Research and Methodology. (2001). *Coding manual for the meta-analysis of antecedent risk predictors of antisocial behavior.* Unpublished Manuscript.

*Chamberlin, R. W. (1976). Can we identify a group of children at age 2 who are at high risk fro the development of behavior or emotional problems in kindergarten and first grade? *Pediatrics, 59*, 971-981.

Coie, J. D., & Dodge, K. A. (1998). Aggression and antisocial behavior. In W. Damon & N. Eisenberg (Eds.), *Handbook of child psychology* (5th ed., Vol. 3, pp. 779-862). New York: NY: John Wiley & Sons, Inc.

Cottle, C. C., Lee, R. J., & Heilbrun, K. (2001). The prediction of criminal recidivism in juveniles: A meta-analysis. *Criminal Justice and Behavior, 28*, 367-394.

Coy, K., Speltz, M. L., DeKlyen, M., & Jones, K. (2001). Social-cognitive processes in preschool boys with and without oppositional defiant disorder. *Journal of Abnormal Child Psychology, 29*, 107-119.

*Cummings, E. M., Iannotti, R. J., & Zahn-Waxler, C. (1989). Aggression between peers in early childhood: Individual Continuity and development change. *Child Development, 60*, 887-895.

*Deater-Deakard, K., Pinkerton, R., & Scarr, S. (1996). Child care quality and children's behavioral adjustment: A four-year longitudinal study. *Journal of Child Psychology and Psychiatry, 37*, 937-948.

*Denham, S. A., Caverly, S., Schmidt, M., Blair, K., DeMulder, E., Caal, S., Hamada, H., & Mason, E. (2002). Preschool understanding of emotions: Contributions to classroom anger and aggression. *Journal of Child Psychology and Psychiatry, 43*, 901-916.

Derzon, J. H., & Lipsey, M. W. (1999). A synthesis of the relationship of marijuana use with delinquent and problem behaviors. *School Psychology International, 20*, 57-68.

Dodge, K. A., Pettit, G. S., & Bates, J. E. (1994). Socialization mediators of the relation between socioeconomic status and child conduct problems. *Child Development, 65*, 649-665.

*Dodge, K. A., Pettit, G. S., Bates, J. E., & Valente, E. (1995). Social information-processing patterns mediate the effect of early physical abuse on later conduct problems. *Journal of Abnormal Psychology, 104*, 632-643.

Duncan, G. J., Brooks-Gunn, J., & Klebanov, P. K. (1994). Economic deprivation and early childhood development. *Child Development, 64*, 296-318.

Earls, F. (1987). Sex differences in psychiatric disorders: Origins and developmental influences. *Psychiatric Developments, 1*, 1-23.

Eaves, L., Rutter, M., Silberg, J. L., Shillady, L., Maes, H. H., & Pickles, A. (2000). Genetic and environmental causes of covariation in interview assessments of disruptive behavior in child and adolescent twins. *Behavior Genetics, 30*, 321-334.

*Emmerich, W. (1966). Continuity and stability in early social development: II. Teacher ratings. *Child Development, 37*, 17-27.

Erikson, E. H. (1963). *Childhood and society* (2nd ed.). New York: W. W. Norton & Company, Inc.

*Fagan, J., & Iglesias, A. (2000). The relationship between fathers' and children's communication skills and children's behavior problems: A study of Head Start children. *Early Education and Development, 11*, 307-320.

*Fagot, B. I. (1984). The consequents of problem behavior in toddler children. *Journal of Abnormal Child Psychology, 12*, 385-396.

*Fagot, B. I., & Kavanagh, K. (1990). The prediction of antisocial behavior from avoidant attachment classifications. *Child Development, 61*, 864-873.

*Fagot, B. I., & Leve, L. D. (1998). Teacher ratings of externalizing behaviors at school entry for boys and girls: similar predictors and different correlates. *Journal of Child Psychology and Psychiatry, 39*, 555-566.

Feld, B. C. (1999). *Bad kids: Race and the transformation of the juvenile court*. New York: Oxford University Press.

Farrington, D. P. (1995). The development of offending and antisocial behavior from childhood: Key findings from the Cambridge Study in Delinquent Development. *Journal of Child Psychology and Psychiatry, 360*, 929-964.

Fergusson, D. M., & Horwood, L. J. (1995). Early disruptive behavior, IQ, and later school achievement and delinquent behavior. *Journal of Abnormal Child Psychology, 23*, 183-200.

Fergusson, D. M., Horwood, L. J., & Lynskey, M. T. (1993). The effects of conduct disorder and attention deficit in middle childhood on offending and scholastic ability at age 13. *Journal of Child Psychology and Psychiatry, 34*, 899-916.

*Fischer, M., Rolf, J. E., Hasazi, J. E., & Cummings, L. (1984). Follow-up of a preschool epidemiological sample: Cross-age continuities and predictions of later adjustment with internalizing and externalizing dimensions of behavior. *Child Development, 55*, 137-150.

Gagnon, C., Craig, W. M., Tremblay, R. E., Zhou, R. M., & Vitaro, F. (1995). Kindergarten predictors of boys' stable behavior problems at the end of elementary school. *Journal of Abnormal Child Psychology, 23*, 751-766.

*Garcia, M. M., Shaw, D. S., Winslow, E. B., & Yaggi, K. E. (2000). Destructive sibling conflict and the development of conduct problem in young boys. *Developmental Psychology, 36*, 44-56.

Garrison, W. T., & Earls, F. (1985). The Child Behavior Checklist as a screening instrument for young children. *Journal of the American Academy of Child and Adolescent Psychiatry, 24*, 76-80.

*Garrison, W., & Earls, F. (1985). Change and continuity in behavior problems from the pre-school period through school entry: An analysis of mothers' reports. In J. E. Stevenson (Ed.), *Recent research in developmental psychopathology* (Vol. 4, pp. 51-65). Oxford: Pergamon Press.

*Gilliam, W. S. (1996). Developmental correlates and predictors of teacher-rated behavior problems in preschool children from low-income families: A longitudinal analysis (Doctoral dissertation, University of Kentucky, 1996). *Dissertation Abstracts International, 57*, 4055.

Gottfredson, M. R., & Hirschi, T. (1990). *A general theory of crime.* Stanford, California: Stanford University Press.

Greenberg, M. T., Lengua, L. J., Coie, J. D., Pinderhughes, E. E., & the Conduct Problems Prevention Research Group (1999). Predicting developmental outcomes at school entry using a multiple-risk model: Four American communities. *Developmental Psychology, 35*, 403-417.

Greenhouse, J. B., & Iyengar, S. (1996). Sensitivity analysis and diagnostics. In J. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 383-398). New York: Russell Sage Foundation.

*Haapasalo, J., Tremblay, R. E., Boulerice, B., & Vitaro, F. (2000). Relative advantages of person- and variable-based approaches for predicting problem behaviors from kindergarten assessments. *Journal of Quantitative Criminology, 16*, 145-168.

*Hagekull, B., & Bohlin, G. (1995). Day care quality, family, and child characteristics and socioemotional development. *Early Childhood Research Quarterly, 10*, 505-526.

*Harnish, J. D., Dodge, K. A., Valence, E., & Group, C. P. P. R. (1995). Mother-child interaction quality as a partial mediator of the roles of maternal depressive symptomatology and socioeconomic status in the development of child behavior problems. *Child Development, 66*, 769-753.

Hedges, L. V. (1994). Fixed effects models. In J. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 285-300). New York: Russell Sage Foundation.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

*Hetherington, E. M., Cox, M., & Cox, R. (1985). Long-term effects of divorce and remarriage on the adjustment of children. *Journal of the American Academy of Child and Adolescent Psychiatry, 24*, 518-530.

Hinshaw, S. P. (1987). On the distinction between attentional deficits/hyperactivity and conduct problems/aggression in child psychopathology. *Psychological Bulletin, 101,* 443-463.

Hinshaw, S. P. (1992). Externalizing behavior problems and academic underachievement in childhood and adolescence: Causal relationships and underlying mechanisms. *Psychological Bulletin, 111*, 127-135.

Hinshaw, S. P., & Anderson, C. A. (1996). Conduct and oppositional defiant disorders. In E. J. Mash & R. A. Barkley (Eds.), *Child psychopathology* (pp. 113-149). New York: NY: The Guilford Press.

Hinshaw, S. P., Lahey, B. B., & Hart, E. L. (1993). Issues of taxonomy and comorbidity in the development of conduct disorder. *Development and Psychopathology, 5*, 31-49.

Hubbard, D. J., & Pratt, T. C. (2002). A meta-analysis of the predictors of delinquency among girls. *Journal of Offender Rehabilitation, 34*, 1-13.

Hunter, J. E., & Schmidt, F. L. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Newbury Park: Sage Publications.

*Ingoldsby, E. M. (2002). Neighborhood contextual factors and early-starting antisocial behavior (Doctoral dissertation, University of Pittsburgh, 2002). *Dissertation Abstracts International, 63,* 4906.

*Johnson, D. R. (2003). The relationship between relational aggression in preschool children and friendship stability, mutuality, and popularity (Doctoral dissertation, Alliant International University, 2003). *Dissertation Abstracts International, 63,* 3958.

Jensen, P. S., Martin, D., & Cantwell, D. P. (1997). Comobidiy in ADHD: Implications for research, practice. and DSM-V. *Journal of the American Academy of Child* and Adolescent Psychiatry, 36, 1065-1079.

Kagan, J. (1969). The three faces of continuity in human development. In D. A. Goshin (Ed.), *Handbook of socialization theory and research* (pp. 983-1002). Chicago: Rand McNally.

Kaler, S. R., & Kopp, C. B. (1990). Compliance and comprehension in very young toddlers. *Child Development, 61*, 1997-2003.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.

*Keenan, K., & Shaw, D. S. (1994). The development of aggression in toddlers: A study of low-income families. *Journal of Abnormal Child Psychology, 22*, 53-78.

Keenan, K., & Shaw, D. S. (1997). Developmental and social influences on young girls' early problem behavior. *Psychological Bulletin, 121*, 95-113.

*Keenan, K., Shaw, D. S., Delliquadri, E., Giovannelli, J., & Walsh, B. (1998). Evidence for the continuity of early problem behaviors: Application of a developmental model. *Journal of Abnormal Child Psychology, 26*, 441-454.

Keenan, K., Shaw, D.S., Walsh, B., Delliquadri, E., & Giovannelli, E (1996). DSM-II-R disorders in preschool children from low-income families. *Journal of the American Academy of Child and Adolescent Psychiatry, 36*, 620-627.

Keenan, K., & Wakschlag, L. S. (2000). More than the terrible twos: The nature and severity of behavior problems in clinic-referred preschool children. *Journal of Abnormal Child Psychology, 28*, 33-46.

*Kilgore, K., Snyder, J., & Lentz, C. (2000). The contribution of parental discipline, parental monitoring, and school risk to early-onset conduct problems in African American boys and girls. *Developmental Psychology, 36*, 835-845.

*Kohn, M., & Rosman, B. L. (1972). A social competence scale and symptom checklist for the preschool children: Factor dimensions, their cross-instrument generality, and longitudinal persistence. *Developmental Psychology, 6*, 430-444.

Knight, G. P., Fabes, R. A., & Higgins, D. A. (1996). Concerns about drawing causal inferences from meta-analysis: An example in the study of gender differences in aggression. *Psychological Bulletin, 119*, 410-421.

Kolvin, I., Miller, F., Scott, D. M., Gatzanis, S., & Fleeting, M. (1990). *Continuities of deprivation? The Newcastle 1000 family study*. Aldershot: Avebury.

Kopp, C. B. (1982). antecedents of self-regulation: A developmental perspective. *Developmental Psychology, 18*, 199-214.

*Ladd, G. W., & Burgess, K. B. (1999). Charting the relationship trajectories of aggressive, withdrawn, and aggressive/withdrawn children during early grade school. *Child Development, 70*, 910-929.

*Ladd, G. W., & Troop-Gordon, W. (2003). The role of chronic peer difficulties in the development of children's psychological adjustment problems. *Child Development, 74*, 1344-1367.

Lahey, B. B., Waldman, I. D., & McBurnett, K. (1999). Annotation: The development of antisocial behavior: An integrative causal model. *Journal of Child Psychology and Psychiatry, 40*, 669-682.

*Laucht, M., Esser, G., & Schmidt, M. H. (2001). Differential development of infants at risk for psychopathology: The moderating role of early maternal responsivity. *Developmental Medicine and Child Neurology, 43*, 292-300.

Lavigne, J. V., Gibbons, R. D., Christoffel, K. K., Arend, R., Rosenbaum, D., Binns, H., Dawson, N., Sobel, H., & Issacs, C. (1996). Prevalence rates and correlates of psychiatric disorders among preschool children. *Journal of the American Academy of Child and Adolescent Psychiatry, 35*, 204-214.

Lewis, M. (1990). Models of developmental psychopathology. In M. Lewis & S. M. Miller (Eds.), *Handbook of developmental psychopathology* (pp. 15-28). New York: Plenum Press.

Lewis, M. (1999). Contextualism and the issue of continuity. *Infant Behavior and Development, 22*, 431-444.

*Lipman, E. L., Bennett, K. J., Racine, Y. A., Mazumdar, R., & Offord, D. R. (1998). What does early antisocial behaviour predict? A follow-up of 4- and 5-year-olds from the Ontario Child Health Study. *Canadian Journal of Psychiatry, 43*, 605-613.

Lipsey, M. W. (1998). Design sensitivity: Statistical power for applied experimental research. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods*. Thousand Oaks, CA: Sage Publications, Inc.

Lipsey, M. W., & Derzon, J. H. (1998). Predictors of violent or serious delinquency in adolescence and early adulthood: A synthesis of longitudinal research. In R. Loeber & D. P. Farrington (Eds.), *Serious and violent juvenile offenders: Risk factors and successful interventions.* Thousand Oaks: CA: SAGE Publications.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: SAGE Publications.

*Lochman, J. E., & Group, T. C. P. P. R. (1995). Screening of child behavior problems for prevention programs at school entry. *Journal of Consulting and Clinical Psychology, 63*, 549-559.

Loeber, R., & Dishion, T. (1983). Early predictors of male delinquency: A review. *Psychological Bulletin, 94,* 68-99.

Loeber, R. (1990). Antisocial behavior: More enduring than changeable? *Journal of the American Academy of Child and Adolescent Psychiatry, 30*, 393-397.

*Loughran, S. B. (1998). Assessing attention-deficit/hyperactivity disorder in preschool children: A longitudinal study (Doctoral dissertation, Fordham University, 1998) *Dissertation Abstracts International, 59,* 1068.

*Loukas, A., Fitzgerald, H. E., Zucker, R. A., & von Eye, A. (2001). Parental alcoholism and co-occurring antisocial behavior: Prospective relationships to externalizing behavior problems in their young sons. *Journal of Abnormal Child Psychology, 29*, 91-106.

*Marchand, J. F., Hock, E., & Widaman, K. (2002). Mutual relations between mothers' depressive symptoms and hostile-controlling behavior and young children externalizing and internalizing behavior problems. *Parenting: Science and Practice, 2*, 335-353.

McConaughy, S. H., & Achenbach, T. M. (1994). Comorbidity of empirically based syndromes in matched general population and clinical samples. *Journal of Child Psychology and Psychiatry, 35*, 1141-1157.

*McElwain, N. L., Olson, S. L., & Volling, B. L. (2002). Concurrent and longitudinal associations among preschool boys' conflict management, disruptive behavior, and peer rejection. *Early Education and Development, 13*, 245-263.

McGee, R., Williams, S., & Silva, P. A. (1985). Factor structure and correlates of ratings of inattention, hyperactivity, and antisocial behavior in a large sample of 9-year-old children from the general population. *Journal of Consulting and Clinical Psychology, 53*, 480-490.

*Miller, E. M. (1992). Impulsivity in middle childhood: Components, stability, antecedents, and relation to overall behavior problems (Doctoral dissertation, Indiana University, 1992). *Dissertation Abstracts International, 53,* 569.

Moffitt, T. E. (1990). Juvenile delinquency and attention-deficit disorder: Boys' developmental trajectories from age 3 to 15. *Child Development, 61,* 893-910.

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review, 100*, 674-701.

*Moffitt, T. E., Caspi, A., Rutter, M., & Silva, P. A. (2001). *Sex differences in antisocial behaviour: Conduct disorder, delinquency, and violence in the Dunedin Longitudinal Study*. Cambridge: UK: Cambridge University Press.

Nadder, T. S., Rutter, M., Silberg, J. L., Maes, H. H., & Eaves, L. J. (2002). Genetic effects on the variation and covariation of attention deficit-hyperactivity disorder (ADHD) and oppositional-defiant disorder/conduct disorder (ODD/CD) symptomatologies across informant and occasion of measurement. *Psychological Medicine, 32*, 39-53.

Nadder, T. S., Silberg, J. L., Rutter, M., Maes, H. H., & Eaves, L. J. (2001). Comparison of multiple measures of ADHD Symptomatology: A multivariate genetic analysis. *Journal of Child Psychology and Psychiatry, 42*, 475-486.

Nagin, D. S., & Farrington, D. P. (1992). The stability of criminal potential from childhood to adulthood. *Criminology, 30*, 235-260.

Offord, D. R., Boyle, M. H., Racine, Y., Szatmari, P., Fleming, J. E., Sanford, M., & Lipman, E. L. (1996). Integrating assessment data from multiple informants. *Journal of the American Academy of Child and Adolescent Psychiatry, 35*, 1078-1085.

*O'Leary, S. G., Slep, A. M. S., & Reid, M. J. (1999). A longitudinal study of mothers' overreactive discipline and toddlers' externalizing behavior. *Journal of Abnormal Child Psychology, 27*, 331-341.

*Olson, S. L. (1992). Development of conduct problem and peer rejection in preschool children: A social system analysis. *Journal of Abnormal Child Psychology, 20*, 327-350.

*Olson, S. L., & Brodfeld, P. L. (1991). Assessment of peer rejection and externalizing behavior problems in preschool boys: A short-term longitudinal study. *Journal of Abnormal Child Psychology, 19*, 493-503.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3*, 354-379.

Pagani, L., Boulerice, B., Vitaro, F., & Tremblay, R. E. (1999). Effects of poverty on academic failure and delinquency in boys: A change and process model approach. *Journal of Child Psychology and Psychiatry, 40*, 1209-1219.

*Patterson, G. R., Littman, R. A., & Bricker, W. (1967). Assertive behavior in children: A step toward a theory of aggression. *Monographs of the Society for Research in Child Development, 32*(5, Serial No. 113).

Pettit, G. S., Clawson, M., Dodge, K. A.; Bates, J. E. (1996). Stability and change in peer-rejected status: The role of child behavior, parenting, and family ecology. *Merrill-Palmer Quarterly, 42*, 267-294.

*Pianta, R. C., & Caldwell, C. B. (1990). Stability of externalizing symptoms from kindergarten to first grade and factors related to instability. *Development and Psychopathology, 2,* 247-258.

Pulkkinen, L. (2001). Reveller or striver? How childhood self-control predicts adult behavior. In A. C. Bohart & D. J. Stipek (Ed.), *Constructive and destructive behavior: Implications for family, school, and society* (pp. 167-185). Washington, DC: American Psychological Association.

Quay, H. C. (1979). Classification. In H. C. Quay & J. S. Werry (Eds.), *Psychopathological disorders of childhood.* (2nd ed., pp. 1-42). New York: Wiley.

Olweus, D. (1979). Stabilities of aggressive reaction patterns in males: A review. *Psychological Bulletin, 86,* 852-875.

Olweus, D. (1984). Stability of aggressive and withdrawn, inhibited behavior patterns. In R. M. Kaplan & V. J. Konecni & R. W. Novaco (Eds.), *Aggression in Children and Youth* (pp. 104-137). Boston: The Hague.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3*, 354-379.

Raudenbush, S. W. (1994). Random effects models. In J. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301-322). New York: Russell Sage Foundation.

Richman, N., Stevenson, J., & Graham, P. (1982). *Preschool to school: A behavioral study*. London: Academic Press.

Rietveld, M. J. H., Hudziak, J. J., Bartels, M., van Beijsterveldt, C. E. M., & Boomsma, D. I. (2004). Heritability of attention problems in children: Longitudinal results from a study of twins, age 3 to 12. *Journal of Child Psychology and Psychiatry, 45*, 577-588.

Robins, L. N. (1978). Sturdy childhood predictors of adult antisocial behavior: Replications from longitudinal studies. *Psychological Medicine, 8*, 611-622.

Robins, L. N. (1966). *Deviant children grown up*. Baltimore: Williams and Wilkins.

*Rose, S. L., Rose, S. A., & Feldman, J. F. (1989). Stability of behavior problems in very young children. *Development and Psychopathology*, 5-19.

*Rose, S. A., Feldman, J. F., Rose, S. L., Wallace, I. F., & McCarton, C. (1992). Behavior problems at 3 and 6 years: Prevalence and continuity in full-terms and preterms. *Development and Psychopathology, 4*, 361-374.

*Roseblum, K. L., & Olson, S. L. (1997). Assessment of peer neglect in the preschool years: A short-term longitudinal study. *Journal of Clinical Child Psychology, 26*, 424-432.

Rutter, M., & Sroufe, A. A. (2000). Developmental psychopathology: concepts and challenges. *Development and Psychopathology, 12*, 265-296.

Schmitz, S., & Fulker, D. W. (1995). Continuity due to which factors? An extension to the rater bias model. *Behavior Genetics, 25*, 287.

*Schmitz, S., Fulker, D. W., & Mrazek, D. A. (1995). Problem behavior in early and middle childhood: An initial behavior genetic analysis. *Journal of Child Psychology and Psychiatry, 36*, 1443-1458.

*Shaw, D. S., Keena, K., & Vondra, J. I. (1994). Developmental precursors of externalizing behavior: Ages 1 to 3. *Developmental Psychology, 30*, 355-364.

*Shaw, D. S., Winslow, E. B., Owens, E. B., Vondra, J. I., Cohn, J. F., & Bell, R. Q. (1998). The development of early externalizing problems among children from low-income families: A transformational perspective. *Journal of Abnormal Child Psychology, 26*, 95-107.

Schuerger, J. M., & Witt, A. C. (1989). The temporal stability of individually tested intelligence. *Journal of Clinical Psychology, 45*, 294-302.

*Spieker, S. J., Larson, N. C., Lewis, S. M., Keller, T. E., & Gilchrist, L. (1999). Developmental trajectories of disruptive behavior problems in preschool children of adolescent mothers. *Child Development, 70*, 443-458.

*Spivack, G., Arcus, J., & Swift, M. (1986). Early classroom behaviors and later misconduct. *Developmental Psychology, 22*, 124-131.

*Spivack, G., & Cianci, N. (1987). High-risk early behavior pattern and later delinquency. In J. D. Burchard & S. N. Burchard (Eds.), *Prevention of delinquent behavior* (pp. 44-74). Beverly Hills, CA: SAGE Publications Inc.

*Stanger, C. (1990). The developmental impact of stress and social network size on psychopathology in adolescence (Doctoral dissertation, Rutgers, The State University of New Jersey, 1990). *Dissertation Abstracts International, 51,* 1173.

*Stattin, H., & Trost, K. (2000). When do preschool conduct problems link to future social adjustment problems and when do they not? In L. R. Bergman & R. B. Cairns (Eds.), *Developmental science and the holistic approach* (pp. 349-375). Mahwah, NJ:: Lawrence Erlbaum Associates, Inc.

*Stevenson, J., & Goodman, R. (2001). Association between behaviour at age 3 years and adult criminality. *The British Journal of Psychiatry, 179*,197-202.

*Stillwell, R., & Dunn, J. (1985). Continuities in sibling relationships: Patterns of aggression and friendliness. *Journal of Child Psychology and Psychiatry, 26*, 627-637.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* 565-578.

Thapar, A., Holmers, J., Poulton, K., & Harrington, R. (1999). Genetic basis of attention deficit and hyperactivity. *The British Journal of Psychiatry, 174*, 105-111.

*Verhulst, F. C., & Althaus, M. (1988). Persistence and change in behavioral/emotional problems reported by parents of children aged 4-14: An epidemiological study. *Acta Psychiatrica Scandinavica, 77*, 1-28.

*Verhulst, F. C., Koot, H. M., & Berden, G. F. M. G. (1989). Four-year follow-up of an epidemiological sample. *Journal of the American Academy of Child and Adolescent Psychiatry, 29*, 440-448.

*Verhulst, F. C., & Van Der Ende, J. (1991). Four-year follow-up of teacher-reported problem behaviours. *Psychological Medicine, 21*, 965-977.

*Vitaro, F., Gagnon, C., & Tremblay, R. E. (1991). Teachers' and mothers' assessment of children's behaviors from kindergarten to grade two: Stability and change within and across informants. *Journal of Psychopathology and Behavioral Assessment, 13*, 325-343.

*Vitaro, F., Tremblay, R. E., Gagnon, C., & Pelletier, D. (1994). Predictive accuracy of behavioral and sociometric assessment of high-risk kindergarten children. *Journal of Clinical Child Psychology, 23*, 272-282.

*Wakschlag, L. S., & Hans, S., L. (1999). Relationship of maternal responsiveness during infancy to the development of behavior problems in high-risk youths. *Developmental Psychology, 35*, 569-579.

*White, J. L., Moffitt, T. E., Earls, F., Robins, L., & Silva, P. A. (1990). How early can we tell?: Predictors of childhood conduct disorder and adolescent delinquency. *Criminology, 28*, 507-527.

Wilson, R. S. (1983). The Louisville Twin Study: Developmental synchronies in behavior. *Child Development, 54*, 298-316.

*Zahn-Waxler, C., Iannotti, R. J., Cummings, E. M., & Denham, S. (1990). Antecedents of problem behaviors in children of depressed mothers. *Development and Psychopathology, 2*, 271-291.

*Zahn-Waxler, C., Schmitz, S., Fulker, D. W., Robinson, J., & Emde, R. (1996). Behavior problems in 5-year-old monozygotic and dizygotic twins: Genetic and environmental influences, patterns of regulation, and internalization of control. *Development and Psychopathology, 8,* 103-122.

Zumkley, H. (1992). Stability of individual differences in aggression. In A. Fraczek & H. Zumkley (Eds.), *Socialization and Aggression* (pp. 45-57). New York: Springer-Verlag.

Zumkley, H. (1994). The stability of aggressive behavior: A meta-analysis. *German Journal of Psychology, 18*, 273-281.