UNIVERSAL SCREENING FOR BEHAVIOR: CONSIDERATIONS IN THE USE OF BEHAVIOR RATING SCALES

A Dissertation

by

BENJAMIN ALLEN MASON

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Kimberly Vannest
Committee Members,	Jamilia Blake
	Jorge Gonzalez
	Paul Wellman
Head of Department,	Victor Willson

December 2012

Major Subject: School Psychology

ABSTRACT

Universal screening for behavior is the use of a measure of social, emotional or behavioral function across an entire population with a goal of preventing future difficulties by intervening with students identified by the screening protocol. Multiple screening procedures have been used, with most including behavior rating scales in the selection process. The purpose of the present research was to investigate two central questions related to the use of universal screeners for behavior in school settings: first, can scores on universal screeners be used as an outcome measure investigating program based interventions, and second, what evidence of teacher bias exists when an external criterion of behavior is included. The purpose of study one was to determine if differences in teacher-rated behavior could be detected between a sample of students that attended public preschool and a nonattending peer group matched for ethnicity, gender, and a gross measure of socioeconomic status (total n= 138). Results of Study One indicated no significant differences between preschool-attending and nonattending groups (p=.61) or between Hispanic and Caucasian participants. Limitations related to sampling and measurement were discussed. In study two, a best-evidence synthesis of peer-reviewed articles investigating teacher bias in behavior ratings of students was conducted. Strict inclusion criteria were chosen to allow for inferential judgment of teacher accuracy. Results of Study Two found a final total of 25 studies of teacher bias that suggested mixed evidence for bias due to student ethnicity or gender and stronger evidence for bias due to expectancies (disability label), teacher culture, unrelated

behaviors (halo effects), and teacher training and experience. Limitations, implications for practice and directions of future research were discussed.

DEDICATION

To my wife Rose, you have been my touchstone for the last (and best) fifteen years of my life. Who would have thought a college dropout from East Texas and a girl from rural Virginia would get this far? We turned some sharp corners, didn't we?

To Reese McKenna, you have taught me what it feels like to be relieved to see those who will take my place. You have your father's bones--you are streamlined for flight, for moving through the world.

To Sloane Parker, you have shown me how a human being can instinctively give to others. You have your mother's heart--a gift to the world that will outlast us both.

To Cohen Reilly, you remind me what it was like to live a dozen lives in a day. You have your father's eyes, so that when you bear witness, your report may be true.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Kimberly Vannest, for supporting me as a student and as a person for the last seven years. I cannot express how much it meant to have an advisor that cared what a second year grad student thought about project aims. It is a rare trait. I would also like to thank Dr. Blake for stepping into my committee and assisting me at a critical juncture. If there were an MVP award for timeliness and honesty, it would be yours. I would like to thank Dr. Gonzalez for being a source of positive support for me throughout my graduate program. I would also like to thank Dr. Wellman for being my sounding board in the last five years and winning the award for best stories in a graduate classroom.

For their support in this process and honest advice, Drs. Jan Hughes, Richard Parker, and Anita McCormick have all contributed to my graduation in real ways that cannot be expressed in such a finite space. For the help in talking out difficult steps in the process and offering his office space so that I could work evenings and weekends, John Davis was invaluable. I would like to thank our girls-- Lily, Amada, Emily, Lunda, Caitlin, Megan, Jordan and Margot for caring for my children at critical stages of my coursework and dissertation process. It is difficult to express how meaningful it was to know such wonderful and fun individuals were taking care of my babies so I could work without guilt.

Finally, thanks to my wife and children. This process has at times been brutal and I wanted very much to surrender and return to smaller ponds. Lao Tzu wrote that being

deeply loved by someone gives you strength, while loving someone deeply gives you courage. In this process, I have needed both and had both because of you four.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

Screening has been defined as a "brief assessment procedure designed to identify children who should receive more intensive diagnostic assessment" (Meisels & Provence, 1989, p. 13). In the medical world, screening is used within a developmental model to assess risk and identify the presence of a potential developmental delay or disability (American Academy of Pediatrics, 1999). Thus, the rationale for health screening is to identify the presence of disease or infirmity among individuals that otherwise appear well. Universal screening, in turn, is the use of a screening procedure across an entire population of interest and is typically used when the screening procedure is not costly and the targeted concern is not so common as to indicate treatment for the entire population is a more efficient means of service delivery (Wilson & Jungner, 1968).

While screening is traditionally associated with the field of medicine, the approach has been adopted in the educational setting. Briefly, screening in educational settings is used to identify students at high risk of developing difficulties in the domain screened as well as to develop targeted interventions to improve school functioning (Kamphaus, 1999; Volpe, Briesch, & Chafouleas, 2010). Consequently, a discerning, high-quality screening result should generate a referral to early intervention designed to improve a child's "developmental, behavioral, and/or school-readiness trajectory" (Marks et al., 2008, p. 866). Social and behavioral screening in schools is used for similar purposes, but the targeted intervention may result in a special education referral

if a student continues to exhibit behavioral difficulties despite intervention (Essex et al., 2009; Individuals with Disabilities in Education Act, IDEA, 2004).

Social and behavioral screening can also serve as a tool to recognize early signs of mental health concerns. In data collected by the MECA (Methodology for Epidemiology of Mental Disorders in Children and Adolescents; Schaffer et al., 1996) study, almost 21% of U.S. children ages 9 to 17 sampled had a diagnosable mental or addictive disorder associated with at least minimum impairment. Additionally, 11% of children and adolescents evidenced significant impairment, and 5% showed evidence of severe functional impairment at the time of testing. As studies use alternate cut score for impairment, different percentages of children evidencing mental health impairment are commonly identified. Additionally, screening results may differ due to the age of the child at the time of screening as well as disorder inclusion/exclusion criteria.

In a cohort of 1,420 children assessed annually for mental health impairment, a three-month prevalence rate of 13.3% and a "one point in time" prevalence rate of 36.3% were determined (Costello, Mustillo, Erkanli, Keeler, & Angold, 2003). Similarly, school and community population studies have shown 12-month prevalence of close to 30% (Kessler et al., 1994). Studies that have used "one point in time" versus current or trailing 3- to 12-month screening methods suggest that some concerns resolve without intervention. As an example, in the Costello et al. (2003) sample, enuresis, encopresis, and separation anxiety disorder had all but resolved by the age of 12. Conversely, other behavioral and mental health concerns may span and negatively impact entire educational careers.

A follow-up report by the U.S. Public Health Service (2000) suggested that only 30% of the 20% percent identified by the MECA study would receive any services given current practices. Working from these data, a hypothetical school of 1,000 students would contain 200 students struggling with a diverse array of maladaptive internalizing (e.g., depression, anxiety) and externalizing (e.g., aggression, hyperactivity) behaviors that have reached levels deserving of special attention. Of these, only 60 would receive any school or community intervention, and approximately 10 would receive special education services for emotional disturbance, roughly one half the estimated prevalence of emotional/behavioral disturbance (EBD; Kauffman, 2001). Thus, traditionally applied identification strategies upon which schools rely are insufficient for identification of behavioral or mental health needs. The disparity between those with a need for temporary or long-term mental health services (20%), those ever receiving services in the school setting (6%), and those receiving special education (1%) for emotional and behavioral concerns is important, given the critical role schools play in initiating mental services for children and adolescents (and future adults).

In the Great Smoky Mountains longitudinal study of 4,500 youths, the educational setting was the first entry point for mental health services for more than 60% of students and was the *sole* provider of mental health services for more than three-fourths of students receiving services (Burns et al., 1995). This pattern of service use was later replicated in a targeted study of 1,420 youths from the same sample by Farmer, Burns, Phillips, Angold, and Costello (2003) as well as a National Institute of Mental Health study of 1,385 youths (Wu et al., 1999). Thus, mental health services within the

school system may be the primary entry point for the majority of children with mental health needs and often the only place in which those needs are met. Consequently, the school's role in connecting students with behavioral and emotional concerns with much needed services within the school milieu or by referral to outside agencies should not be underestimated.

While the Individuals with Disabilities in Education Act (IDEA) encourages districts to use universal screening to identify students in need of services, this recommendation is not accompanied by standardized selection methods (Maag & Katsiyannis, 2008). As discussed, traditional referral methods expanded to an entire school population are insufficient to identify students in need of monitoring and intervention. Consequently, firmer guidelines for best practice in screening must be developed.

Glover and Albers (2007) noted that universal screeners in schools should possess three critical attributes: developmental appropriateness, technical adequacy, and usability. Developmental appropriateness refers to the need for measures to capture behaviors across multiple ages. Technical adequacy consists of sensitivity and specificity as well as statistical support for its use with the population of interest. Statistical support includes normative data with a sample that is representative of the population of interest, evidence of reliability and validity, as well as acceptable correlation with longer behavioral measures such as the Behavior Assessment System for Children, Second Edition (BASC-2; Reynolds & Kamphaus, 2004) or the Achenbach System of Empirically Based Assessment (ASEBA; Achenbach & Rescorla, 2001). This evidence

should already be present in normed screening measures (Distefano & Kamphaus, 2007; Reynolds & Kamphaus, 2003; Smith, McCarthy, & Anderson, 2000).

Statistical support for a behavioral screener also requires a sufficiently broad item set to capture externalizing and internalizing behaviors as both predict future outcomes (Eklund, et al., 2009; Elliott, Huai, & Roach, 2007; Ervin, Schaughency, Goodman, McGlinchey, & Matthews, 2006; Hinshaw, 1992; Levitt, Saka, Romanelli, & Hoagwood, 2007). Mental health issues encompass externalizing and internalizing behaviors, and students with combined symptoms have shown poorer long-term outcomes in large-scale epidemiological studies of serious self-injury and completed suicides (Shinn et al., 2009; Sourander et al., 2009). Thus, behavioral screening should use instrumentation that captures both types of behavior reliably with sufficient distinction to make informed judgments about student needs (Caldarella, Young, Richardson, Young, & Young, 2008; Cook et al., 2011; Crocker & Algina, 1986; Essex et al., 2009).

Last, usability refers to resource constraints related to the screener's use, such as administration costs that may include forms costs, scoring software, as well as time spent completing and scoring forms. Ideally, screeners should provide sufficient data to allow for at least some degree of intervention selection (Caldarella et al., 2008; Cook, Volpe, & Livanis, 2010; Marchant, Brown, Caldarella, & Young, 2010; Marchant et al., 2009). The school setting is the sole access point for mental health services for many students and the most common mental health concerns are under identified (internalizing disorder). Therefore, screening is a high-stakes decision by default. Furthermore, the

importance of decision accuracy (and technical adequacy required to arrive at such accuracy) grows as the decision stakes are raised (Kahn & Baron, 1995).

Given the relatively new use of normed universal screeners in the research literature, this dissertation focused on two main questions. First, the use of universal screeners as an equitable measure of Hispanic and Caucasian student behavior was investigated. In study one, the relationship of preschool participation and behavioral risk at Kindergarten was investigated using data from the Behavior Assessment System for Children, Second Edition (BASC-2; Reynolds & Kamphaus, 2004). While preschool's effects on behavior for children entering school have been investigated in the past, this is the first article at the time of publication to use normed universal screening data as the outcome measure. In other words, are children who participated in public preschool perceived as having different levels of behavioral risk than those who did not.. The BASC-2 BESS yields a reliable score of behavioral risk that includes both externalizing and internalizing risk factors, and has shown evidence of validity for schoolwide screening, but its use as a secondary measure for intervention efficacy has not been addressed in the literature.

In study two, the evidence base for teacher bias in behavioral screeners was examined. A review of current publications indicated universal screening tools have not been sufficiently examined for bias in published research separate from test publisher data. As universal screener use in peer reviewed journals is relatively new, and normed universal screening tools are typically behavior rating scales, a best-evidence synthesis of the literature investigating teacher bias in behavior rating scales was undertaken to

inform the field and advance knowledge in this area. Strict guidelines for evidence were used, with articles included only if an external criterion of behavior was collected so as to allow for investigations of actual rather than potential bias.

CHAPTER II

PRESCHOOL PARTICIPATION AND LATER TEACHER-REPORTED MEASURES OF BEHAVIORAL RISK

Hispanics are the most rapidly growing ethnic group in United States schools. According to the U.S. Census Bureau the population claiming Hispanic or Latino descent more than doubled between 1990 and 2010. The number of non-white Hispanics increased from 22 million to over 50 million and has passed African Americans as the largest minority group during the same period (16.3% Hispanic vs. 12.9% AA; U.S. Census Bureau, 2010). Moreover, fifty percent of the Hispanic population growth in the United States during the 1990s was due to new immigration (Aloise-Young & Chavez, 2002; Velez & Saenz, 2001), and recent data indicate that trend has continued (U.S. Census Bureau, 2006). Hispanic families report larger numbers of children per household, with more than 11% of Hispanic households reporting three or more children compared to 5.4% for whites alone. These patterns suggest Hispanic children will constitute 25% of all public school students by the year 2030 (Kindler, 2002).

Unfortunately, local educational authorities (LEAs) struggle to meet Hispanic student needs upon school entry and thus Hispanic students are more likely to begin school with fewer academic language and literacy skills (Magnuson & Waldfogel, 2005) and are more likely to be retained (NCES, 2003), with grade retention typically serving as the strongest predictor of school dropout (Close & Solberg, 2008; Goldenring Fine, & Davis, 2003; McCoy & Reynolds, 1999; Rumberger, 1995). Additionally, there is evidence to suggest that Hispanic students may be perceived as having fewer emotional

and behavioral needs than Caucasian students even when exhibiting similar behaviors and thus may not receive much-needed assistance in the educational setting (Massa, 2011; Prieto & Zucker, 1981). Supporting this concern is lower identification rates in the category of emotional and behavioral disorders for Hispanic students despite relatively high rates of anxiety and depression (Polo & Lopez, 2005). In a study of 4500 youths, Burns and colleagues found that the educational setting was the primary entry point for mental health services for approximately 60% of subjects and the sole provider for three-fourths of those that received any services at all. This critical point of contact may be eliminated if Hispanic student symptomology is underestimated for reasons related to race.

Screening

Screening has been defined as a "brief assessment procedure designed to identify children who should receive more intensive diagnostic assessment" (Meisels & Provence, 1989, p.13). While better developed in the areas of academic achievement, screening for emotional and behavioral concerns at the school level offers promise in identification of those in need of emotional and behavioral supports, particularly those that may be missed by traditional referral practice (Lane et al., 2008). If Hispanic students are missed by traditional teacher referral processes due to underestimation of social and behavioral need, then an alternative process that identifies those students at a rate that is commensurate with needs-occurrence is preferable to standard school practice.

While emotional and behavioral screeners are typically used in an assessment capacity with sufficient reliability and validity to meet that purpose (DiStefano & Kamphaus, 2007), their potential utility as outcome measures for assessment and progress monitoring has been explored less successfully (Cronbach, 1970; Greenwood et al, 1979). The secondary use of screeners in either capacity, if supported, would lend credence to LEA adoption of schoolwide screening for emotional and social behaviors. Unfortunately, screeners used by schools suffer from difficulties inherent to both social and emotional measurement and short forms in particular (Smith, McCarthy, & Anderson, 2000). Consequently, any appraisal of screener validity as an outcome measure would require an intervention potent enough to change teachers' global perceptions of social and emotional competence relative to peers that had not received the intervention

Preschool Programs

One intervention shown to improve student's behavioral academic and behavioral functioning is preschool participation. Preschool is a broad term describing a continuum of structured educational experiences that begin before the age of traditional school attendance (typically age 5 in the United States). Preschool programs originally focused on serving children of poverty with the programmatic aim of remedying experiential gaps that contributed to academic deficits (Belsky & MacKinnon, 1994; Ramey & Ramey, 1998). While the most frequently researched preschool program is Head Start, preschool programs that share similar goals but differ in intensity and scope have also been investigated for child outcomes. The Abecedarian Project and the Perry

Preschool Program are examples of intense, full-day programs with small teacher-student ratios, but half and full-day public preschool programs that mirror Kindergarten have also been implemented and assessed for child benefits (Belfield, 2006; Love et al., 2007; Nores & Barnett, 2009).

The current evidence base suggests preschool participation results in consistently positive but small effects on academic achievement at school entry (Barnett & Hustedt, 2005). Magnuson, Ruhm, and Waldfogel (2004) analyzed a large, meta-analytic dataset comprised of dozens of studies and found an effect size improvement of .14 standard deviations (SD) for preschool participation on academic skills at school entry, with differences shrinking each successive year. Yet as gains in measures of achievement fade, life-outcome benefits persist for preschool participants. These include a reduction of 40% to 60% in special education placement (Reynolds & Temple, 2008; Schweinhart & Weikart, 1997) as well as reduced odds of grade retention by age 15 compared to matched controls. Only 54% of Abecedarian program participants had been retained compared to 88% of matched controls (Campbell & Ramey, 1995). The search for a mechanism of effect that explain these differences in the absence of persistent academic differences-- the "black box" of preschool (Currie & Neidell, 2003), has remained elusive. As mechanisms of these benefits do not appear to function through differences in achievement between participants and nonparticipants, differences in socioemotional and behavioral outcomes have been explored.

Behavioral benefits have been shown to exist for preschool participants for over four decades. In the first large scale study of Head Start, the Coleman Report (Coleman

et al., 1966) examined multiple outcomes for Head Start participation. In the report including 13,326 Head Start attenders and matched non-attenders, students were rated on a behavioral measure by teachers in subsequent school years. Head Start attenders were rated more favorably than non-attenders by teachers on a measure of classroom behavior, with the most significant differences found for those children reared in the most deprived home environments. On average, Head Start attenders were found to receive one more favorable rating (of a possible eight) than matched controls in the study-designed measure of classroom behavior. In a predominately African American sample, Schweinhart and Weikart (1979) found teachers rated 22 Perry Preschool participants more positively on measures of classroom behavior and competence than 46 nonparticipants that received either a typical nursery school or an academic intervention alone. In a large sample of more than 3000 families, Love et al. (2005) found modest but positive effects of Early Head Start on behavior upon follow-up prior to formal school entry.

More recently, differences in outcomes between preschool participants have focused on executive function (EF), also known as cognitive control. Executive function skills critical for classroom success include inhibitory control, working memory, and cognitive flexibility in adjusting to change (Diamond, Barnett, Thomas, & Munro, 2007). These executive skills contribute to academic success (Merz & McCall, 2010), school readiness (Duncan et al., 2007) and predict later measures of teacher-reported externalizing and internalizing behaviors (Hughes & Ensor, 2011; Riggs, 2003; Schachar & Logan, 1990). While specific curricula have been assessed in the attempt to modify

preschool-age children's executive skills (Wilson, 2012), it has not been established if preschool attendance in general contributes to executive skill variance.

However, Hispanics constituted negligible percentages of the Coleman or the Schweinhart and Weikart studies. While the study conducted by Love and colleagues (2002) used a diverse sample, investigated effects were one to two years *prior to* school entry with no information regarding behavior or emotional status at follow-up and included intense parent-training components in addition to the preschool/child care component. To date, no studies that included significant numbers of Hispanics within the sample have investigated teacher-ratings of behavior for preschool participants in the following academic year. Additionally, normed behavioral measures that capture measures of behavioral risk that are not disruptive such as anxiety were either not available or eschewed in favor of researcher-developed measures in prior work. Last, with the exception of the first studies of Head Start, no measures of behavior have addressed these proximal measures of emotional and behavioral function within a school program that may be more typical of school settings. This study investigates the relationship between attendance of a typical preschool program on teacher perceptions of Hispanic and Caucasian students' emotional and behavioral function in the first semester of Kindergarten.

Hypotheses

Hypothesis One: Differences in teacher-rated behavioral risk will be reduced when a normed measure is used to assess risk in a typical school sample.

Hypothesis Two: Preschool participation will be associated with reduced scores on a measure of behavioral risk in the following year for both Hispanic and Caucasian Kindergartners.

Method

Participants

Seventeen elementary teachers completed behavioral screeners (the BASC 2-Behavioral and Emotional Screening System, known as the BASC 2-BESS; Kamphaus & Reynolds, 2007) on all Kindergarten students (N=315). Instructional personnel were typically early career, with 54.2% of classroom teachers reporting fewer than five years of experience and all rating teachers were identified as Caucasian. The sample was drawn from a small Southwestern city (population approximately 11,000). The school reported an enrollment of 766 students for the most recent school year, and the school composition for the year of analysis was a reported 15.1% Hispanic and 83.3% White non/Hispanic. Additionally, the school population consisted of 36.5% economically disadvantaged students.

Instruments and Measures

The child/adolescent form of the BASC 2-BESS (Kamphaus & Reynolds, 2008) was used to assess teacher perception of students' behavioral and emotional status. The BASC-2 BESS is a 27-item behavioral screener designed to assess a broad array of emotional and behavioral strengths and weaknesses based on the Behavior Assessment System for Children, Second Edition (BASC-2, Reynolds & Kamphaus, 2004).The measure yields a single score that has been shown to predict a number of behavioral

concerns (including difficulties with both internalizing and externalizing behaviors) as well as achievement test scores up to at least 5 years later. The BASC-2 BESS has shown acceptable correlations with the full BASC-2 (Behavior Assessment System for Children, Second Edition; Reynolds & Kamphaus, 2004) and the ASEBA (Achenbach System of Empirically Based Assessment; Achenbach & Rescorla, 2001). The Teacher Form, Level Child/Adolescent (CA) was used in this study. The BESS classifies students into three broad risk categories for the development of future emotional and behavioral problems (normal risk, elevated risk, and extremely elevated risk) based on a combined risk score that includes internalizing and externalizing behaviors. Reported test retest reliability of the BASC-2 BESS is .91 and split-half reliability is .96. Correlations between the BASC-2 BESS and the BASC-2 teacher rating scales are .80 for externalizing problems, .64 for internalizing problems, .89 for school problems, -.85 for adaptive skills, and .91 for the behavioral symptoms index analysis.

Procedures

After a training session explaining screener protocol presented by program staff, seventeen elementary teachers completed behavioral screeners BASC 2-BESS on all Kindergarten students (N=315) in the fall of the academic year. All raters included a gross measure of economic disadvantage (free or reduced lunch) and district-reported ethnicity for rated students. Screeners were collected by project personnel in the following week and were scored using the BASC-2 BESS Scoring software. District criteria for preschool admission and a list of all students attending district preschool in the previous academic year were then obtained.

Sampling procedures. The preschool-attending sample consisted of 69 students (55 Caucasian, 14 Hispanic). The comparison group of preschool nonattenders was selected as follows: first, all nonattending students were grouped according to socioeconomic status (SES), ethnicity, and gender. Next, a student with identical SES, ethnicity and gender was randomly sampled without replacement from the nonparticipant group. This process continued until a group of 69 comparison students, matched on all three categorical variables had been drawn from the larger, nonparticipating pool. These 69 students served as the control group.

Data analysis. Descriptive statistics were generated for all independent and independent variables. Levene's Test of Homogeneity of Variances was then used to determine if the residual variances were invariant across groups. A univariate one way analysis of variance (ANOVA) was then conducted to determine if there were differences in mean risk scores between the two groups across condition (preschool attendance and nonattendance). In the analysis, combined group, age-corrected deviation scores of behavioral risk were the dependent variable of interest. Student ethnicity, SES, and gender were independent variables included in the full model. Student age was not included in the model, as first, student age varied only slightly and second, age was already included as a component of the T-scores themselves, and inclusion of age in the model would have allowed age to contribute unique variance at two levels of analysis. Accordingly, gender was a variable of interest, and thus combined gender T-scores on the dependent measure were used rather than gender specific T-scores, reducing the

impact of gender to one level of contribution rather than the two possible if gender-adjusted T scores were used.

Results

Table 1.

Background characteristics of the participant sample.

Ind.Variable	**Number**	**Percent of Sample**
Ethnicity	Caucasian (n=110)	79.7
	Hispanic (n=28)	20.3
Gender	Male (n=84)	60.9
	Female (n=54)	39.1
Socioeconomic Status	Meeting Federal Guidelines for Free and Reduced Lunch[1] (n =120)	86.9
	Not Meeting Federal Guidelines for Free and Reduced Lunch[2] (n= 18)	13.1

[1] Annual income below $39,220 (family of four) for the year of data collection
[2] Annual income above $39,220 (family of four) for the year of data collection

Descriptive statistics for the full sample (n=138) are presented in Table 1. As previously described, the comparison group was randomly selected from a pool of students matched for background characteristics and there were no students with missing data, thus no t-tests were necessary to assess for differences in the independent variables

(SES, gender, ethnicity). The variance in outcome variable (T-score of behavioral risk) was compared across the control and comparison groups to assess for equivalence across groups. Levene's Test resulted in statistically nonsignificant results (p= .610), suggesting variances were similar between groups.

Table 2.

T-score descriptive statistics.

	Preschool	Comparison
Mean	48.362	49.101
Median	47.000	47.000
Mode	44	34[1]
Range	40	41
Standard Deviation	10.242	10.898
Skewness	19.64	13.629
Std. Error of Skewness	.289	. 289
Kurtosis	-.422	-.770
Std. Error of Kurtosis	.570	.570

[1]Multiple modes exist. The smallest value is shown

Distribution statistics for the preschool attending, nonattending, and full sample are presented in Table 2. The range of scores included multiple students with a T score of 33, the lowest possible raw score of zero signifying a student rated by the teacher as having never shown any behavioral concerns in any area. The overweighting of students

rated as having no problems at all resulted in the negative kurtosis, positive skew, and mean that was slightly below the normative sample's standardized T score of 50. However, the normative sample's T score was captured by the grand mean when 95% confidence intervals were calculated (47.049-50.415).

Score distributions for the preschool attending and nonattending groups are presented in Figures 1 and 2. The slightly positive skew is evident through visual inspection of both figures.

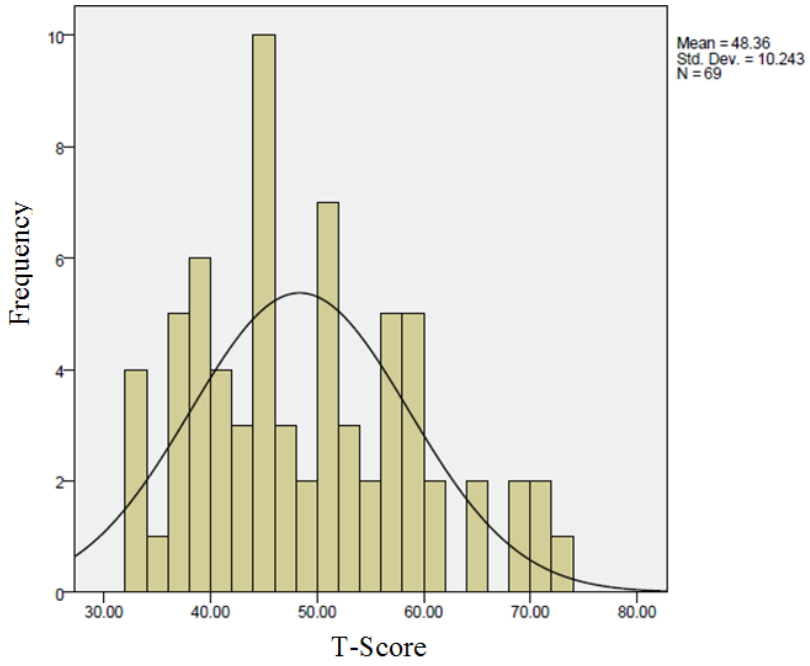

Figure 1. T-score distribution for preschool attenders

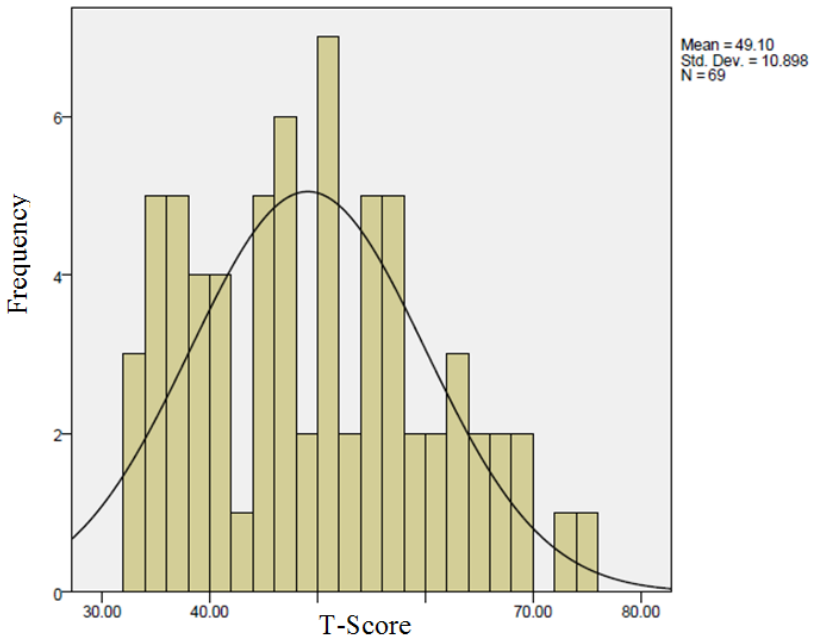

Figure 2. T-score distributions for preschool non-attenders

A univariate analysis of variance was then examined to determine if statistically significant variance in outcome risk scores could be attributed to program participation, with ethnicity, gender, program participation, and SES included as factors in the full model. The results are presented in Table 3. As can be seen, the full model did not yield statistically significant results.

Table 3.

Test of between subjects effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	1945.940[a]	4	486.485	4.871	.001	.128
Intercept	6657.598	1	6657.598	66.660	.000	.334
Ethnicity	8.776	1	8.776	.088	.767	.001
Gender	860.745	1	860.745	8.618	**.004**	.061
SES	1432.004	1	1432.004	14.338	**.000**	.097
Pre-K	26.988	1	26.988	.270	**.604**	.002
Error	13283.140	133	99.873			
Total	342951.000	138				
Corrected Total	15229.080	137				

[a] R Squared = .128 (Adjusted R Squared = .002)

An R squared for preschool participation signified a behavioral risk score that was nonsignificant for both Hispanic and Caucasian preschool participants. Consequently, the first research hypothesis that Hispanic and Caucasian students' behavioral risk scores would be similar if a normed measure were used to screen was supported. The second hypothesis, that preschool participation would be associated with a reduction in behavioral risk scores was not supported (p=.604). Significant differences in obtained scores were found for two covariates, gender and socioeconomic status. While females and participants that did not qualify for free and reduced lunch were the

minority of the rated students, their risk scores were lower and contributed the only significant variance to the model.

Discussion

This study investigated teachers' ratings of behavioral and emotional risk for a sample of Hispanic and Caucasian preschool attenders and non-attenders. The hypothesis that the use of a normed measure would result in similar scores across Hispanic and Caucasian participants was supported in the current dataset. Scores for Hispanic and Caucasian students on the behavioral risk screener were not significantly different. This is important for schoolwide screening implementation in schools with Hispanic students, as otherwise, screening would not be supportable and would suffer from the same biases as teacher referral. While extended validity studies using Hispanic data should be forthcoming, as a preliminary finding, similarity in scores is encouraging for screening in diverse settings.

The second hypothesis, that preschool attenders would be perceived as evidencing less risk and more acceptable behaviors, was not supported in the study. This finding conflicts with prior studies of both Head Start and the Perry Preschool Project that obtained positive behavioral effects as measured by teachers on less reliable measures. However, there were two key differences that render comparisons of effects difficult. First, Perry Preschool was a very intensive program with a low teacher to student ratio and used a population at very high risk compared to this study. Additionally, like the Head Start investigations, the most salient results for preschool effects on behavior were conducted in the 1960s and 1970s, a period in which structured

childcare was not as common as it is currently. If many of the same academic benefits of public preschool and Head Start are similar for structured childcare (as found by Loeb et al., 2006), the same behavioral benefits may exist as well.

At its face, the use of a normed screener with evidence of reliability and validity should be more sensitive to perceived differences in behavior than past studies that used tailored measures designed by researchers. Instead, both groups were virtually indistinguishable from one another. Specific outcomes may require targeted measures to detect and the use of a broadband measure, even if such a measure quantifies risk reliably, may not qualify as sufficiently targeted. The study found no significant differences in teacher-rated behavior for preschool participants but also found no significant score differences between Hispanic and Caucasian participants. This suggests teachers were not over or underrating students based on group membership. Hispanic students have historically been scored more favorably on behavioral measures by teachers than Caucasian and African American peers. If scores of both sample and control are restricted in range even slightly, then measurement may be compromised by this limited variability for an extended discussion of this topic, see Sackett & Yang, 2000). While this possibility may be explored in the future, the similar distributions displayed by the entire sample, and corresponding null effects displayed by Caucasian students did not support this hypothesis. It is also of note in tests of various assumptions that variance was equally distributed in both groups, but that normality of kurtosis was not. This would be expected in a measure of behavior, as behavior measured by rating scales may not be normally distributed in the general population. The finding that gender

was a significant contributor to ratings differences is unsurprising, particularly when using scales that are heavily influenced by externalizing items. It is noted that early work on Head Start found larger behavioral benefits on long term outcomes for female than male participants, and it is unknown if this is a Matthew Effect in which females "behave" better but also benefit more from programs broadly meant to improve learning behaviors/executive skills.

Limitations and Future Research

A fundamental limitation to the current study is the lack of information available on the control group. While the sample group received the standard preschool offered by the district, the control group likely received "something else." That "something else" may range from babysitting, unstructured child care, or high-quality center-based care. As a counterexample, the first reference to preschool behavioral effects referenced in this article stems from Head Start attendance cited in the Coleman Report. It is almost certain that few of the nonparticipating low-income families were able to enroll nonparticipating children in a quality, education-centered formal setting in the early 1960s.

Second, while "sleeper effects" of behavioral differences in preschool attenders that might not appear initially but may surface at some point in later schooling are possible, they were not the purpose of this study and data were not collected to explore this hypothesis. Future work into program components of school preparation may shed light on this topic and other long-term mechanisms of preschool participation. The relatively early timing of behavioral rating (October) may have influenced ratings in a

positive or neutral manner. Students were rated more positively than the normative sample, with an average score below the norm average, and rating this early in the year may have influenced teachers to rate in this manner due to concerns that it may be too early in the year to draw attention to even moderate misbehavior. Both are empirical questions that could not be answered with the current dataset.

Third, the preschool program used by the participating district did not follow any specific implementation model. A common complaint of early intervention researchers is poor treatment fidelity by program implementers. In this instance, the program was essentially Kindergarten offered a year prior to the Kindergarten year, with identical curricula and no social or emotional components. This limitation may actually have improved external validity, as programs may be more likely to use a "home-grown" package than one with stringent requirements and suggests early childhood programs with no predetermined social or behavioral focus may not yield desirable results in those areas. Again, this is an empirical question as to how little or how much time and school resources should be required to yield results, and may be answered in longer term studies of structured programs designed to enhance executive skills.

Last, while students from low SES homes received preferential access, some students that did not meet free or reduced lunch were admitted and program staff related this is was the result of a "first come-first served" policy. However, roughly 65 percent of the nonattending sample qualified as free or reduced lunch but for unknown reasons, did not attend. Given the high cost of child care for families earning wages below federal poverty guidelines, this raises questions of unequal access for qualifying families. It is

unknown if this is an idiosyncratic finding or if there are similar issues of equity in access for other studies. Future research that investigates gating procedures used by schools when there are fewer spots than potential applicants may help to resolve this discrepancy.

CHAPTER III

BIAS IN TEACHER RATINGS OF BEHAVIOR: A RESEARCH SYNTHESIS

Behavior rating scales are indirect ratings of behaviors, symptoms, and functioning gathered for assessment purposes (Beaver & Busse, 2000; Fennerty, Lambert, & Majsterek, 2000; Hosp, Howell, & Hosp, 2003). Typically presented in a standardized format, behavior rating scales have now entered the sixth decade of use in the psychological assessment of children and adolescents (Achenbach, 1966; Reynolds & Kamphaus, 2003; Sattler, 1992). While behavior rating scales differ in length and presentation, they typically require raters that know the subject well enough to form a summative judgment as to the subject's behavior to respond to questions or statements about subject function in a Likert-style format (Merrell, 1994; Reid, 1995; Reid & Maag, 1994). Behavior rating scales may be used to measure specific concerns such as attention-deficit hyperactivity disorders (ADHD) or depression or may be broadband in nature and used to capture a large number of adaptive and maladaptive behaviors.

Behavior rating scales are now considered standard assessment practice for both externalizing and internalizing disorders (Merrell, 2000; Sattler, 1992), as they allow for efficient information-gathering and comparison of information between informants. Additionally, shortened behavior ratings scales are used in screening (both universal and selected) as well as in behavioral progress monitoring (Gresham & Elliott, 2008; Kamphaus & Reynolds, 2007; Walker & Severson, 1992). This expansion suggests behavior rating scales may be used to provide information in virtually every step of service provision for special education: initial screening, progress monitoring,

assessment of intervention response, and full and individual evaluation for placement in special education.

However, behavior rating scales are measures, and all measures produce scores that contain measurement error (Haynes, Smith, & Hunsley, 2011). Measurement error strictly defined is variability in results when measuring the same variable in the same individual (Bland & Altman, 1996). Measurement error negatively affects interpretation of behavior rating scale scores by impacting their primary functions: classifying subjects and predicting outcomes for those subjects. Measurement error may be random or systematic in nature. Random error is not correctible, as it should theoretically have a self-cancelling mean of zero but is detectible through repeated measurement of the same subject (DeVellis, 2006). Conversely, systematic error moves scores consistently in one direction, is typically undetectable through repeated administration, but may be corrected once it has been identified (Campbell & Fiske, 1959). Any systematic error that affects the validity of scores is bias (Hunter & Schmidt, 1990). The three most commonly assessed types of bias associated with behavior rating scales are construct bias, predictive bias, and rater bias.

Bias

Construct Bias

Construct bias occurs when the factor structure of the latent construct (e.g. aggression, depression) differs across measured groups and thus compromises the ability to make valid score comparisons across those groups (French & Finch, 2006; Meredith, 1993). Construct bias may be addressed by answering a series of related questions.

First, is the same number of factors present across measured groups? Second, do the same items load on the same factors across groups? Third, do the common factors have the same meanings across groups? If these three questions are answered affirmatively, meaningful and defensible comparisons of scores for different groups may be made (Gregorich, 2006).

Predictive Bias

Predictive bias (also known as differential prediction) was defined by Cleary (1968) as consistent nonzero errors of prediction between members of a subgroup. Predictive bias is measured by assessing differences in the prediction of the criterion based on slope and intercept between groups (Nunnally & Bernstein, 1994). If predictive bias is present, then the predicted criterion will differ across groups despite similar scores on the measure of study. Examples of outcomes predicted by behavior rating scales and subscales include academic achievement (Kamphaus, Thorpe, Winsor, Kroncke, & Dowde, et al., 2007), independent measures of school adjustment (Hoge & McKay, 1986), peer victimization (Hanish & Guerra, 2000), and referral for intervention or placement in special education (Harris, Tyre & Wilkinson, 1993). Together, investigation of predictive and construct bias can be expressed simply as—do the tests scores predict outcomes equally well between groups and are the scores measuring the same thing for those groups. Both types of bias are typically addressed at the level of test and item development through statistical analysis and proper matching of normative sample with the population of interest.

Rater Bias

Rater bias is the presence of substantial and systematic error in ratings of performance or behavior caused by rater attitudes, beliefs, or experiences (Hoyt, 2000). First investigated empirically by Thorndike (1920), rater bias can lead to systematic overestimation or underestimation of true behavior (Saudino, 2005). Rater bias negatively impacts behavior rating scales' primary functions—to predict outcomes for individuals and to classify individuals as needing intervention and/or service provision. While not directly linked to rater bias, studies of inter-rater reliability are suggestive of large disparities in behavior ratings across informants**.** In a study of 60 articles offering sufficient data to calculate paternal-maternal agreement, Duhig, Renk, Epstein, and Phares (2000) found mean Pearson *r* values of .46 for internalizing problems, .66 for externalizing problems, and .61 for total problems. These values were similar to those found in Achenbach, McConaughey, and Howell's (1987) meta-analytic review of 119 studies in which parent-parent agreement for externalizing (.62) and internalizing (.59) symptoms were found. The resulting variance is large and must be attributed at least in part to rater differences due to shared observation setting.

As may be expected, teacher-parent agreement in rating scales is substantially lower than parent-parent agreement, with an obtained Pearson *r* of .31 for externalizing and .21 for internalizing symptoms found within the same review (Achenbach et al., 1987). The larger correlations found in ratings of externalizing rather than internalizing problems is robust and has been consistent across multiple studies (Diamond & Squires, 1993; Verhulst & Akkerhuis, 1989; Verhulst & van der Ende, 1991; Stanger & Lewis,

1993; Winsler & Wallace, 2002). The lower correlations found on behavior rating scale scores for teachers and parents may represent true variation in behaviors across settings and task demands or may be the result of some bias associated with characteristics or beliefs of the raters (Nunnally, 1978; Phares, Compas, & Howell, 1989). Best practice for assessment calls for multiple informants with teacher ratings considered an essential component (Loeber, Green, Lahey, & Stouthamer-Loeber, 1989), yet teacher bias is rarely discussed as a potential threat to score validity. As systematic error in teacher ratings may unduly influence service provision and empirical study of teacher bias beyond simple agreement has not yet been investigated in a structured literature review, it is the focus of the current paper.

Types of Rater Bias

Rater bias is a broad construct that encompasses a large number of personal and situational factors that may impact teacher ratings of behavior. Examples of specific rater biases include the following common examples: halo bias, response-style bias, projection bias, expectancy bias, situational bias, and language bias. Halo bias occurs when raters tend to rate subjects as "generally good" or "generally bad" contaminating ratings of unrelated behaviors (Epkins & Myers, 1994; Thorndike, 1920). Response-style bias includes the following types: leniency bias in which raters favor positive responses (also known as acquiescent response style), stringency bias, in which raters favor negative responses (also known as disacquiescent response style), midpoint response style, in which raters favor middle responses, and extreme response style, in which raters tend to choose extremely positive or negative ratings in general (Alliger &

Williams, 1992; Weijters, Schillewaert, & Geuens, 2008). Projection bias may occur when raters score others higher or lower on the latent construct in accordance with the rater's own level of the construct (Hooman, 1982; Kroes, Veerman, & De Bruyn, 2005). Expectancy bias (also referred to as examiner or observer bias) is the result of examiners assessing subjects differently based on expectation or notions regarding expected behavior rather than behavioral differences. Expectancy bias is most commonly investigated in studies using vignettes that assign labels related to disability or socioeconomic status. Situational bias refers to the effect of environmental variables that are specific to the testing environment (Cole & Bruner, 1971). Last, language bias is the result of examiners using a language or register that the subject cannot understand sufficiently (Reynolds, Lowe, & Saenz, 1999).

Additional rater variables not typically considered in studies of bias but that may contribute systematic variance to teacher ratings include practice effects, regression effects, and order effects. Practice effects occur when raters evaluate subjects more positively after initial rating without evidence of actual behavioral improvement due to increased familiarity with rating content (Sandoval, 1977). Similarly, regression effects occur when extreme, initial ratings (both lenient and harsh ratings) are less extreme in subsequent ratings (Milich, Loney, Roberts, & Caputo, 1980). Order effects occur when the order of ratings impact rating scores (e.g, students with extreme misbehavior, if rated first, positively impact behavior ratings of subsequent students). Categorization of bias types can be difficult due to definitional overlap, but all are similar in that they systematically impact scores and thus compromise score validity.

Measurement of Rater Bias

While each of the aforementioned biases represents independent, systematic threats to the validity of rating interpretation, presence of bias cannot be determined by simple comparison of rater scores and requires a reasonable criterion serving as a measure of "true" current or predicted behavior (Kingstrom & Bass, 1981). Simple differences between raters is insufficient to answer this question and does not meet the "best evidence" criteria as described by Slavin (1986) due to the lack of internal validity inherent to any study of behavior rating without directly measured or controlled behavior. Direct measurement of rating bias is desirable, as it allows for the study of unambiguous relationships between the rater and the latent construct and has been investigated through the use of vignettes and direct observation. With an accurate metric of child behavior serving as criterion, ratings may be compared and shown to deviate from actual behavior (Guilford, 1954; Podsakoff, MacKenzie, & Podsakoff, 2012).

The use of vignettes (both scripted and video) to alter variables of interest in a systematic fashion while holding other student characteristics constant has been used to investigate multiple areas of teacher bias in behavior ratings**.** In these studies, the researchers have typically manipulated: disability category (Foster & Ysseldyke, 1976), ethnicity (Chang & Sue, 2003), and gender (Kelter & Pope, 2011) while leaving other information identical. This means of holding constant all student characteristics apart from the variable of interest is efficient, but suffers from loss of external validity due to the absence of real-world variables that may impact teacher ratings of students in actual settings. For example, the quality of teacher-student relationships may impact teacher

ratings of student behavior. While no direct research has been completed on this issue, studies of maternal rating behaviors are suggestive of unique rater variance attributable to dyadic relationship variables. Seifer, Sameroff, Barrett, and Krafchuk (1994) trained mothers to rate infant temperament and found high correspondence between ratings by mothers and those of trained observers when rating unrelated infants but negligible agreement with trained observers when mothers rated their own children. Accordingly, relationship variables may account for unique variance in teacher ratings of actual students that may not be discoverable in research that uses vignettes due to some combination of past interactions and relationship quality. This source of specific variance is unique to the relationship between rater and subject and is rarely, if ever, explored (Hoyt, 2000).

Direct observation data collected by a trained observer allows for study of real students in real settings and may capture elements of student-teacher interactions. Direct observation data is considered the gold standard of behavioral data (Chafouleas, Riley-Tillman, & McDougal, 2002; Hintze & Matthews, 2004; Intille et al., 2003) and may be used as the criterion against which behavior ratings are compared. Collection of direct observation data also allows for contextual information-gathering such as behavior of same-class peers. However, direct observation data is costly to collect and may lead to subject reactivity in which observed students or peers behave differently due to observer presence (Harris & Lahey, 1982). Additionally, low-frequency, high intensity events as discussed by Gresham, MacMillan, and Bocian (1996) may be missed by direct observation, are important contributors to understanding of child functioning, and may

be impossible to discount in a retrospective rating of child function. While neither vignettes nor direct observation is without drawback for determining rater bias, each offers useful information regarding systematic variance that is likely to go unexplored during standard development of behavior rating scales.

Rationale

The search for articles addressing teacher bias yields a large number of results, yet the resulting articles are often conceptual in nature and do not offer quantifiable evidence of mean differences directly attributable to teacher characteristics or beliefs. Also, studies that do offer mean differences that would allow for effect size calculation address very different bias types, rendering a meta-analysis impractical. The absence of a structured literature review or best-evidence synthesis (Slavin, 1986) that examines what bias is present in teacher ratings is surprising given the significant contribution of teacher ratings of behavior to the assessment and intervention process and the low correspondence in agreement between teacher and parent ratings. This gap in the literature supports the need for a comprehensive review and synthesis related to the following research question:

What evidence of teacher bias in ratings of student behavior exists when an unbiased or third-party criterion measure of behavior is collected for comparison?

Method

The study examined evidence of teacher bias in behavior ratings and was conducted in two phases. The first phase was a comprehensive literature review using: an electronic database search, a hand search of the past ten years for journals frequently

publishing articles on teacher bias, and a search of reference lists for all identified articles. In the second phase, the articles were organized and presented in a best-evidence synthesis format as described by Slavin (1986). A best-evidence synthesis should include a thoughtful extraction of the *best* evidence in the topic of study rather than a broader, quantitative assessment of relatively weak evidence.

Comprehensive Literature Review

A comprehensive review of the literature should be methodical and exhibit clear criteria necessary to isolate all studies related to the research question (Berman & Parker, 2002). Studies were identified through an electronic search using the Psycinfo database, Education Full Text and Wilson database. The following Boolean string searches were conducted: bias and behavior rating* and teacher. This search resulted in 173 articles from the years 1975 to 2010. The search was updated until February 3, 2012 with one additional article meeting selection criteria prior to search termination (Kelter & Pope, 2011) and a final total of 174 articles.

A hand search was then conducted by the author. The tables of contents and abstracts were searched from the year 2002 to the issue available on February 3, 2012 in the following education and assessment journals with published work in rater bias: *Journal of Abnormal Child Psychology*, *Journal of Emotional and Behavioral Disorders*, *Journal of Psychoeducational Assessment*, *Psychology in the Schools*, *School Psychology Quarterly*, and *School Psychology Review*. The hand search resulted in no additional articles that had not been identified in the electronic search. A reference search of the 174 included articles was then conducted, with 15 articles identified as

potentially meeting inclusion criteria added to the search list for a final total of 189 articles.

Inclusion criteria. Following the search results, each article was evaluated to determine if the criteria for inclusion in the literature review had been met. The studies had to meet the following inclusion criteria: (1) the study was published in an English-language, peer–reviewed journal, (2) behavior ratings were collected from classroom teachers rather than preservice teachers or college students, (3) ratings of behavior were used as the dependent measure (likelihood of referral, perceptions of achievement were excluded), (4) a criterion measure of behavior was used or collected in the study that allowed for comparison against teacher ratings, and (6) the criterion was not simply a score on another rating scale unless it was the same scale used by the same teacher and measured post-experimental condition. (7) Results allowed for comparison of mean differences and subsequent assessment of bias direction.

Inter-rater agreement for inclusion/exclusion. Twenty percent of studies (38/189) were reviewed by two evaluators to evaluate the reliability of coding articles for criterion and determine if the criteria for inclusion had been met. A third evaluator reviewed any studies for which the first two evaluators disagreed and/or one evaluator was undecided in how to code based on the publication text. All evaluators were doctoral candidates in educational psychology. The decision made by two of the three evaluators was the final decision using a majority agreement method. Agreement for inclusion or exclusion was 97%. Only one article required a third evaluator to determine exclusion/inclusion decision, and by majority agreement (2 of 3 reviewers agreed on

criteria) was included in the review. Of the original 189 identified articles, 25 met the criteria for inclusion in the study. The majority of excluded articles did not collect behavior ratings from teacher (n=69), or did not include a criterion measure of behavior (n=60).

Coding

Six descriptive characteristics of each study were coded in the database to be assessed in answering the research question. The six characteristics were as follows: (a) the author (s) and year of publication, (b) the research sample (teachers filling out the rating forms), (c) the behavior rating scale used and construct measured, (d) the criterion measure used in the study, (e) the bias type investigated, and (f) study results (finding or non-finding of bias). If multiple ratings occurred (e.g. behavior, academic, likelihood of referral), then only the behavior rating results were coded for review inclusion. Operational definitions for coding of each bias type are presented in Table 4.

Table 4.

Operational definitions for coding of bias types

	Bias type	Operational definition
1.	**Age bias**	Ratings of student behavior attributed to the age of the rater, the rated student, or an interaction between the two
2.	**Cultural bias**	Ratings of student behavior attributed to differences in the rater's cultural expectations and beliefs (e.g. country of origin, region of origin, socioeconomic status of the rater)
3.	**Ethnic bias**	Ratings of student behavior attributed to the ethnicity of the rater, the rated student, or an interaction between the two.
4.	**Examiner bias**	Ratings of student behavior attributed to the rater's level of knowledge, vocational or educational experience (e.g. knowledge of learning disability, coursework in special education, non-example: specific rating scale experience)
5.	**Expectancy bias**	Ratings of student behavior attributed to unobserved behaviors or characteristics assigned to the rated students (e.g. labels of categorical disability but not ethnicity, age, or gender)
6.	**Gender bias**	Ratings of student behavior attributed to differences in student gender
7.	**Halo effects**	Ratings of behavior attributed to the influence of unrelated student behavior (e.g. defiant behaviors increasing ratings of inattention)
8.	**Language bias**	Ratings of behavior attributed to differences in the language of the rater, the student, or the linguistic complexity of the rating scale
9.	**Order effects**	Ratings of behavior attributed to differences in the order of presentation (e.g. if students with more severe behaviors are rated early or later)
10.	**Practice effects**	Ratings of behavior attributed to differences in experience with the rating scale
11.	**Projection bias**	Ratings of behavior attributed to differences the rater's level of the latent construct being measured (e.g. raters with higher levels of depression rating students as being more depressed)
12.	**Rating-style bias**	Ratings of behavior attributed to differences in rating styles (e.g. harsh raters, lenient raters, midpoint raters)
13.	**Regression effects**	Ratings of behavior attributed to the effects of regression on ratings (e.g. more severe ratings are less harsh at follow-up rating)
14.	**Situational bias**	Ratings of behavior attributed to environmental factors (e.g. classroom activity, classroom layout, time of day)

Coder reliability. Twenty percent of research reports were coded by two graduate student coders. As recommended by Cooper (1998), coding disagreements were resolved by discussion. If the disagreement could not be resolved by discussion, a third coder was consulted. The decision made by two of the three coders was the final one. Coder agreement was determined in two ways. Overall agreement of all possible codes (25 studies, 6 codes per study) was 98% (147/150). The only disagreement occurred on the "type of bias" and this disagreement occurred in 3 of the 25 articles. Reliability of bias type was 88% (22/25). Disagreement occurred with identifying categories including the following: cultural bias/ethnic bias (2/3) and expectancy bias/halo effects (5/7).

As one example of overlap between categories, Neal et al. (2003) collected teacher ratings of video vignettes of students engaged in a "standard" walk or a "stroll" representing the movement style of inner city, urban youth. The videos rated were of either a similarly dressed Caucasian or African American student. While coder 1 correctly identified two bias types—the ethnic bias associated with the target student as well as the cultural bias representing the movement style itself, Coder 2 coded only ethnic bias. Coder 3 correctly identified both bias types, resolving the discordant codes.

Results

The research question: *what evidence of teacher bias in behavior ratings exists when a criterion measure of behavior is included* required a comprehensive and broad search that resulted in a small number of studies with a strong external or conclusion validity. Results are first presented descriptively in table format, followed by brief

summaries of teacher characteristics, dependent measures, and criterion measures. Then, a more thorough discussion of individual studies is completed under the categories of bias types. Quantitative data in the form of p values for statistically significant differences in teacher ratings are presented for all studies that found differences.

Descriptive Summary

Descriptive summary for each of the 25 studies can be found in Appendix A.

Rater characteristics. Twenty one of the twenty five included studies reported teacher characteristics in the method section. For the twenty one reporting studies, 1,844 total teachers were included in the study review. Of the articles that included sufficient information to determine teacher placement, all but three used elementary or primary teachers. Bahr et al. (1991) and Neal et al. (2003) used middle school teachers, while Saunders and TiLullo (1972) used a K-8 teaching sample. Secondary teachers were not used in any study that met inclusion criteria and provided teacher information. Trained observers were used to independently rate student behavior in ten studies, and all observer data were collected in classroom settings.

Dependent measure. The behaviors most commonly assessed in the literature review were ADHD and related hyperactivity/inattentive behaviors, representing 13 total studies. Total problem scores or similar ratings of aggregate behavioral symptomology were also common, representing eight studies. Similarly, composite scores of externalizing or internalizing problems were used in five studies. Two studies investigated teacher ratings of student aggression.

Criterion measure. The criterion measures used to validate behavior ratings fell into three broad categories--scripted vignette, video vignette, or direct observation with two exceptions. Vignettes were used in 14 of 25 studies, and at least one statistically significant result was found for 12 of the 14. However, differences were found according to category. Scripted vignettes were used in five studies, with three of the five showing statistically significant results (and consequently harsher ratings) indicating biasing effects of at least one examined variable. Video vignettes were used in ten studies, with some measure of behavior showing significantly harsher ratings in nine of the ten studies. The only study in which video vignettes were used that resulted in nonsignificant results was also the only study to use both scripted and video vignettes with mixed results (Dukes & Saudargas, 1989). Direct observation served as the criterion measure for 10 of 25 studies, with significant results found in seven. Two studies used a repeated measures format in order to assess changes in ratings across the experimental condition, with one showing significant results (Brandon et al., 1990) and one resulting in none (Saunders and Di Tullio, 1972).

Bias Type

Expectancy. Expectancy bias was the most common bias type and was investigated in nine studies, with six demonstrating statistically significant differences in behavior rating scores (Brandon et al., 1990; Chang & Sue, 2003; Dukes & Saudargas, 1989; Foster & Ysseldyke, 1976; Phillips & Lonigan, 2010; Saunders & Di Tullio, 1972; Stevens, 1980; Sonuga-Barke et al., 1993; Walker, Bettes, & Ceci, 1984). Chang and Sue (2003) used scripted vignettes to assign normal behavior, externalizing

(undercontrolled) problems, and internalizing (overcontrolled) problems to hypothetical students. In teacher ratings of behavioral severity, 197 teachers rated externalizing problems as being more severe than internalizing problems ($p<.001$). This finding suggested that the authors' attempts to match severity of internalizing and externalizing problems in vignettes had little impact on teacher ratings. Similar teacher beliefs about severity of behavior problems were investigated by Walker, Bettes, and Ceci (1984). In their study, 100 preschool teachers rated behavior severity for scripted vignettes, with aggression rated as significantly more serious than hyperactivity, which in turn was rated as more serious than withdrawal ($p<.01$). A significant interaction effect was also found for withdrawal, as teachers rated vignettes for a five year old's withdrawal symptoms as more serious than that of a three year olds ($p<.01$) despite behaviors that were identically described.

Foster and Ysseldyke (1976) measured the impact of assigned labels (normal, learning disabled, emotionally disturbed, educable mentally retarded) on behavior ratings of 100 experienced elementary teachers (mean years of experience= 9.7 years). Significantly more severe ratings were found for disability labels in the scripted format ($p<.01$). Additionally, these rating differences were maintained even when rating a video vignette of identical (and typical) child behavior. Notably, statistically significant differences were found not just for the categorical labels, but between them, as raters given the student label of "educable mentally retarded" rated the student more harshly than those receiving the learning disabled or emotionally disturbed labels ($p<.05$) which in turn were harsher than the ratings of those receiving the normal label ($p<.05$).

In a like design, Dukes and Saudargas (1989) investigated the impact on behavior ratings by teachers first for assigned labels (learning disabled or normal) in scripted vignettes and then in videotaped behavior of a normal 8 year old child in individual seatwork or a group work setting. Using a seasoned group of 80 teachers with an average experience of 13 years, significant differences were found in scripted ($p<.01$) but not video vignette ratings, suggesting expectancy bias may be modified by the transition from hypothetical to actual student behavior for experienced teachers. In a study using 20 primary teachers, Fogel and Nelson (1981) assigned labels (normal, learning disabled, educable mentally retarded, and emotionally disturbed) to a typically-behaving student in a video vignette and found significantly harsher behavior ratings for every disability label compared to both the normal label and the subject receiving no label at all ($p<.01$). Teacher-coded behavior observation data did not differ across label conditions as did checklist-ratings.

In a study in which teachers rated ADHD behaviors for video vignettes of children exhibiting mild, moderate, and severe ADHD, Brandon et al. (1990) found ratings from 60 teachers and school personnel did not differ after being told the child had just received medication for ADHD prior to the video recording. This use of expectancy as a measure of expected change in behavior rather than expected behavior was the only study of its kind that met inclusion criteria.

Saunders and Di Tilullo (1972) collected ratings of student behavior for nine teachers (grades K-8), suggested that three of the six students receiving the harshest behavior ratings in each class were of high "achievement potential," and returned in

three months to collect a second round of teacher ratings. The study authors' hypothesis that behavior ratings would be improved at three month follow-up due to Pygmalion Effects was not confirmed.

In a study using Head Start teachers, Philips and Lonigan (2010) investigated expectancy effects based on the socioeconomic status of children for 98 teachers, but found SES did not impact teacher ratings of preschoolers' behavior compared to those obtained by trained observers. It was noted that while SES did not appear to contribute to rating differences, teachers rated all students more severely than did observers for five of six ratings constructs ($p<.001$). Conversely, in the first of the remaining studies that found significant differences in behavior ratings based on expectancy effects, Stevens (1980) varied child characteristics in biographical packets describing students prior to teacher viewing of child behavior and found expectancy bias due to assigned socioeconomic status. In the study, 27 teachers rated child behavior in silent videos more harshly if the child had received the low socioeconomic status background rather than the middle-class backgrounl in the pre-rating packet ($p<.001$).

Ethnic bias. Ethnic bias in teacher ratings of behavior was measured in eight studies (Bahr, Fuchs, Strecker, & Fuchs, 1991; Chang & Sue, 2003; Epstein et al., 2005; Hosterman, DuPaul, & Jitendra, 2008; Milfort & Greenfield, 2002; Neal, McCray, Webb-Johnson, & Bridges, 2003; Sonuga-Barke, et al., 2003; Stevens, 1980). Significant differences in behavior ratings due to student ethnicity were found in three of the seven. Stevens (1980) found harsher ratings ($p<.01$). of student behavior by 27 elementary teachers when the student was African American rather than Hispanic despite videotaped

behavior that was pre-rated as identical by 10 graduate students. Milfort & Greenfield (2002) found 22 Head Start teachers rated African American children's play interaction more positively than Hispanic children's play interaction while observers found the reverse. Teachers and observers agreed on higher ratings of disruption for African American children compared to Hispanic children. Additionally, teachers gave generally higher ratings of disconnection (internalizing behaviors) for all subjects—a finding that did not reach significance.

Sonuga-Barke et al. (2003) investigated ethnic bias in teacher ratings of hyperactivity by comparing British teacher ratings of 99 British children (47 "hyperactive, 52 controls) and 30 Asian (10 "hyperactive," 20 controls) children with direct observation data. While the teacher ratings did not differ significantly for the two ethnic groups, significant differences were found in comparisons with direct observation data. The group of English "hyperactives" were rated more harshly by direct observers than the Asian group ($p<.05$). Moreover, the differences in direct observation data were stark enough that the Asian "hyperactive" group's ratings of ADHD did not differ significantly differ from the ratings of the British control group.

Hosterman and colleagues (2008) compared teacher ratings with direct observation for 124 elementary students meeting ADHD criteria and 48 students not meeting ADHD criteria. Direct observation data from trained graduate students suggested teachers were actually more accurate in rating behavior of African American and Hispanic students than Caucasian children, and thus ethnic bias was not found. The authors posited that disproportionately harsh behavior ratings for ethnic students may at

least partially due to underrating of Caucasian student behavior. Neal et al., (2003) assessed ethnic bias in ratings of aggression by 136 middle school teachers using video vignettes. In a novel design, the authors used video vignettes of Caucasian and African American students engaged in a standard walk or a "stroll" representing the movement style of urban black youth to determine if ethnicity impacted teacher ratings. While significant effects were found for other variables in the study, teacher-rated aggression did not differ due to ethnicity or ethnicity-movement interactions.

Epstein et al., (2005) investigated ethnic bias in teacher ratings of behavior of 528 African American and Caucasian students with diagnosed ADHD by collecting direct observation data for all included students. While ratings of ADHD behaviors for African American subjects were more severe than Caucasian subjects, direct observation data supported real differences in classroom behavior that suggested a lack of teacher bias. It was noted that direct observation data collected for a comparison peer nominated as having typical behavior suggested higher levels of average misbehavior in the classroom setting for African American students as well, a setting difference that may contribute to exacerbation of ADHD symptoms.

Ethnic Bias was investigated by Bahr et al., (1991) in a study collecting teacher ratings and direct observation data for 40 students nominated as extremely difficult to teach (DTT). While ratings were collected from the 40 middle school teachers on a variety of variables, only the behavioral severity rating met inclusion criteria and was included in the review. No significant differences were found for behavior ratings between teachers and trained observers, suggesting highly problematic behavior may

sufficiently override any potential confound related to group membership. Using a similar variable of teacher-ratings of behavioral severity, Chang and Sue (2003) varied student ethnicity (Asian, Caucasian, and African American) and problem type (no problem, externalizing problems or internalizing problems) in scripted vignettes that were rated by 197 teachers. No bias was found in teacher ratings due to student ethnicity.

Cultural bias. Cultural bias in teacher ratings of behavior was investigated in five studies (Alban-Metcalfe, Cheng-Lai, & Ma, 2002; Mueller at al., 1995; Neal et al., 2003; Puig et al., 1999; Weisz et al., 1995), with all five resulting in significant differences in ratings of behavior across cultural categories. Trained observer ratings of Jamaican (n=27) and African American (n=24) student behavior were compared with ratings collected from their respective teachers in a study by Puig et al., (1999). Significant main effects were found for rater type, with teachers rating more harshly than observers in both settings ($p<.0001$). Additionally, main effects were found for rater nationality, with U.S. teachers rating students more harshly than Jamaican teachers ($p<.0001$). This finding was in direct conflict with trained observer ratings denoting more classroom misbehavior from Jamaican students ($p<.0001$). A higher student-teacher ratio in the Jamaican sample (45:1) was offered as a possible explanation for the greater teacher tolerance for misbehavior by the Jamaican teachers.

A similarly-designed cross-cultural study of Thai and U.S. students rated by teacher and trained observer was completed by Weisz et al. (1995). Using a sample of teachers from both countries along with trained observers, the researchers found Thai teachers rated Thai student behavior more harshly than their U.S. counterparts rated U.S.

student behavior ($p<.0001$). Similar to Puig et al. (1999), observer ratings resulted in contrary findings, with U.S. student behavior ratings denoting much more sever classroom misbehavior than those obtained from Thai students ($p<.0001$). Interestingly, nearly all of the score variance lay in externalizing problems, with teacher ratings of internalizing disorders showing little difference across Thai and U.S. teacher ratings.

Mueller et al. (1995) investigated cultural bias by collecting behavior ratings of four videotapes of child behavior by 130 teachers from five countries--China, Indonesia, Japan, Thailand, and the United States. The videotapes presented children in both group and individual settings displaying varying degrees of hyperactive, inattentive, and oppositional behaviors. Significant effects were found for ratings by country ($p<.01$) as well as by videotape ($p<.01$), suggesting cultural factors impacted ADHD ratings in general but also differentially affected subsets of behavior subsumed under the ADHD domain (e.g. inattention, hyperactivity). Similarly, Alban-Metcalf and colleagues (2002) used video vignettes of a nine-year old child diagnosed with ADHD to compare behavior ratings from 130 teachers in Mainland China, Hong Kong, and the United Kingdom. As expected, main effects for teacher origin were significant ($p<.001$), and intra-group effects supported the hypothesis that Hong Kong's more "Westernized" culture would lead to more lenient behavior ratings. Despite sharing national heritage, significant differences between mainland China and Hong Kong were found for all three subscales--inattention, impulsivity, and hyperactivity ($p<.007$).

Finally, theNeal et al. (2003) study investigated cultural bias by examining the impact on behavior ratings of aggression when a Caucasian and African American

student engaged in a "standard" walk or a "stroll" depicting a gait considered to be associated with inner-city, African American culture. The study found 136 middle school teachers rated the student as more aggressive if he engaged in the "stroll" rather than the "standard" walk ($p=.001$) regardless of ethnicity.

Examiner bias. Teacher experience and/or knowledge as an independent source of bias were variables investigated in three studies (Abikoff et al., 1993; Mueller et al., 1995; Stevens et al., 1998). In a video vignettes study examining ratings of child actors engaging in oppositional, ADHD, or typical behavior, Abikoff et al. (1993) found equivalent ratings of oppositional behavior by 72 regular education and 67 special education teachers but more severe ratings of ADHD behavior by regular educators ($p<.001$).

Special education experience was an independent bias variable assessed by Stevens, Quittner, and Abikoff (1998) in a study that collected ADHD behavior ratings of video vignettes by 105 elementary teachers. Videos depicted students engaging in oppositional, ADHD or typical behaviors, and teachers with more special education experience were reported as ratings students less harshly on the inattentive/passive construct (p value not supplied). Additionally, professional experience with ADHD, knowledge of ADHD, and educational experiences related to ADHD were assessed for raters, with only educational experiences showing significant differences ($p<.05$). In the only study to assess for experiential variables related to parenting, parental experience was investigated as an independent variable in a study of 130 teachers from five countries (Mueller et al., 1995). In the study, teachers with children of their own rated

students exhibiting ODD and ADHD behaviors less severely than did teachers without children ($p<.01$). While the authors were unable to determine the precise mechanism of parenting's effects on ratings, it was robust across national samples. Additionally, the authors investigated experience with ADHD as a biasing variable, with nonsignificant results.

Situational bias. A second bias variable that was added post hoc due to review of search results included situational bias in which the context of the rated behavior occurred. Situational bias was investigated in three studies (Dukes & Saudargas, 1989; Jacob, O'Leary, & Rosenblad, 1978; Mueller et al., 1995. The relationship between behavior setting and teacher ratings for a sample of 80 experienced teachers was investigated by Dukes and Saudargas (1989). After an expectancy condition was applied, teachers rated student behavior in a large group or individual seatwork condition. As the behavior sampled was typical for grade, ratings were both harsher and more accurate in the large group setting ($p<.01$).

In the third, behavior ratings of students across different classroom organization were compared (Jacob, et al., 1978). Eight "hyperactive" and 16 control students were rated on hyperactive behavior by a study-assigned teacher and two trained observers. Significant differences in teacher accuracy were found in ratings of hyperactivity for formal and informal (i.e. "open" classroom) settings, with teacher ratings being more accurate in the formal setting than the informal setting. Direct observation frequency data suggested less overall ADHD behavior in the informal setting.

Halo effects. Halo effects were investigated in five studies (Abikoff et al., 1993; Epstein et al., 2005; Jackson & King, 2004; Schachar, Sandberg, & Rutter, 1986; Stevens, Quittner, & Abikoff, 1998). All five studies investigated unidirectional or bidirectional halo effects on ratings of aggression or defiance and ratings of ADHD, with significant effects found for all but one study. Epstein et al, (2005) measured teacher and observer ratings of ADHD as well as observer ratings of aggression in a large, multisite study including 528 students with and without ADHD. While observers found elevated ADHD ratings for students that exhibited classroom aggression, no such relationship was found for teachers, indicating a lack of halo bias in the only study of its kind and also suggesting teachers were *underreporting* child aggression in the study for African American students.

Abikoff et al. (1993) showed videos of child actors portraying ODD, ADHD, and typical behavior to 139 elementary regular and special education teachers and collected behavior ratings immediately post-viewing. Confirming the research hypothesis, inflated ratings of ADHD were found when teachers rated the ODD exemplar as evidenced by a lack of significant difference—teachers rated ADHD similarly for both the ADHD and the ODD exemplar. Consequently, 40% of the ratings of the ODD video met clinical criteria for ADHD. Interestingly, these effects were unidirectional in nature, with no significant differences in ODD ratings of ADHD models. Using the same video vignettes, Stevens et al. (1998) collected ADHD ratings from 105 elementary teachers and found similarly elevated ADHD ratings of the ODD exemplar. However, the effects were less pronounced as the ADHD exemplar was rated

as having more severe ADHD symptomology than the ODD model ($p<.05$). The authors posited that this finding of partially attenuated bias may have been due to the greater frequency of inclusive school practices in the eight years between studies that led to more comfort and familiarity with ODD behaviors.

Schachar, Sandberg, & Rutter, 1986 found multiple statistically significant halo effects in a study using both behavior ratings and direct observation data from observers blind to the experimental condition. Using a sample of 33 boys culled from a screening of all first grade students in six London schools, the researchers selected three groups based on low (1 *sd* below the mean), high (1 *sd* above the mean) or midpoint (any point between the high or low groups) on the CTRS hyperactivity index. First, students with a greater number of negative peer interactions received undeserved scores in the hyperactive range ($p<.03$). Second, students that self-vocalized were rated lower on hyperactivity than was merited by direct observation data ($p<.05$). Third, students with more positive peer interactions were rated higher in behavior problems than similarly behaving peers without those positive interactions ($p<.05$). Last, the traditional finding of higher scores of teacher rated inattention resulting from defiant or aggressive student behaviors was present ($p<.05$).

Jackson and King (2004) found complex interactions behavior ratings by 80 middle school teachers of male and female video vignettes depicting typical, ODD, and ADHD behaviors. Using a balanced sample of 40 regular education and 40 special education teachers, the authors found the exemplar displaying oppositional behaviors

was rated as having more hyperactivity than the normal exemplar despite no more ADHD behaviors ($p<.05$).

Gender bias. Gender bias was empirically investigated in four studies, with significant differences in ratings by gender found in two of the four. Jackson and King (2004) investigated the interaction between gender and problem behaviors in the aforementioned video vignette study of expectancy bias. In the study, males displaying oppositional behaviors were rated as having more hyperactivity than did females ($p<.05$). The inverse was also found, with students displaying ADHD symptoms receiving harsher ratings as having more oppositional behavior than the typically-acting exemplar ($p<.000$). Contrary to the oppositional findings, females displaying ADHD behaviors received harsher ratings of oppositional behavior than did males exhibiting similar behaviors ($p<.05$). Thus, the study found complex interactions between problem type, problem severity, and child gender.

Walker, Bettes, and Ceci (1984) examined ratings of behavioral severity by 100 preschool teachers for vignettes describing boys and girls with varied age and problem type (aggression, hyperactivity, and withdrawal). In the study, teacher ratings differed for other varied child characteristics, but not for gender. Similarly, Kelter and Pope (2011) investigated gender influences on teacher ratings in a study using scripted vignettes to vary ethnicity for hypothetical students displaying oppositional behaviors. In a study examining 145 experienced elementary teacher ratings of child behavior, the authors found no gender effects on behavior ratings.

In the previously discussed study of observer and teacher ratings of children in Head Start (Milfort & Greenfield, 2002), teachers rated boys' disruptive behavior more harshly than did observers ($p<.001$). While both teachers and observers rated boys as displaying more disruptive classroom behaviors than did girls, the difference in harshness across raters was not present for girls. In sum, gender influences on teacher ratings of behavior were uneven across studies, and difficult to assess at this time. It is noted that significant effects were found in the two studies that included teacher ratings of actual students or videos of actual behavior but not in the two that used scripted vignettes, suggesting a potential stereotype activation effect that requires real students and real behaviors.

Discussion

The research question called for a best-evidence synthesis of the of teacher bias when rating student behavior. A comprehensive literature review was conducted and articles were included if bias was investigated with two central characteristics—first, that teacher ratings of student behavior were included and second, that those ratings were accompanied by a criterion measure that allowed for a reasonable interpretation of the accuracy of those ratings. Given these two parameters, 25 articles that met inclusion criteria resulted in the following findings. Evidence of bias due to student ethnicity was mixed, while evidence of effects due to culture, expectancy, and halo effects were strong. These findings are outlined below in greater detail..

Evidence for ethnic bias in teacher ratings of behavior was mixed. In the two studies that employed the strongest sampling and data collection methods, no ethnic bias

was found for ratings of ADHD (Epstein et al., 2005; Hosterman et al., 2008). In fact, Hosterman and colleagues found ratings of African American students were *more* accurate than those of Caucasian students. Interestingly, the strongest evidence of ethnic bias was shown by Sonuga-Barke et al. (2007) as a potential handicapping effect of positive ethnic stereotypes. In their study of Asian and British students rated as varying in levels of ADHD, Asian students rated as high in ADHD did not differ from British controls. These findings suggest teacher rating bias may be impacted by a perceived gap between positive ethnic stereotypes and perceived misbehavior by a member of that ethnic group. However, concluding that ethnic bias in behavior ratings based on negative stereotypes is not supported may be premature. First, teachers often knew which students were being assessed for which behaviors, and that may have influenced rating authenticity. Second, studies that used scripted vignettes with assigned ethnicities may have alerted teachers to study intent. Third, ethnicity and culture are difficult to disentangle, as shown in the Neal et al. (2003) study of movement styles. The Neal finding suggests cultural biases that may have originally functioned as ethnic biases may exert effects *independent from* the associated ethnic group. Culture may be seen as the culture of the teacher (as seen in the studies of cross cultural ratings of video vignettes), culture of the student, or the interaction between the two. To date, surprisingly little work has been done in this area.

The evidence for expectancy bias suggested categorical labels assigned to students exerted strong effects on ratings independent of actual student behavior. Reinforcing concerns that categorical labels influence teacher's beliefs about students is

the finding that for both scripted and video vignettes, teachers rated student behavior differently based on assigned labels. In the case of scripted vignettes, this rating difference could be considered an exercise in stereotyping of a "typical" child with a categorical disability that might not influence ratings of actual children. However, similar effects have been shown in video vignette studies, suggesting teacher perceptions are modified to accommodate those same stereotypes with real children and real behaviors. Evidence of halo effects were also quite strong, and suggestive that multiple overlaps may occur for bidirectional inflation of ratings of defiance /opposition and ADHD as well as harsher or more lenient ADHD ratings due to unrelated neutral or positive social behaviors. These effects were robust across teachers and studies and suggestive of classic, systematic bias in teacher ratings.

Implications for Practice

Screening. While Jackson and King (2004) were the only authors to address simultaneously more than one set of rating-related halo effects (e.g. defiance and hyperactivity), they were also the only researchers to study the impact positive behaviors may contribute to harsher behavior ratings if the positive behaviors are poorly timed. The implications for screening are compelling. If a behavioral rating form used as in school-wide screening captures a broad set of externalizing behaviors and those behaviors are subject to halo effects, then one negative behavioral category may be sufficient to tip a student into an at-risk category. Additionally, if behaviors that are only negative due to timing or context (e.g. peer interaction during classroom instruction) result in greater ratings on an ADHD measure, then these problems will only be

exacerbated. Thus, broadband measures that capture multiple behaviors for screening purposes may be at risk of multiple levels of criterion contamination. One possible remedy may be to use mixed formats in screening. As an example, defiance and ADHD are particularly susceptible to halo effects when typical behavior rating formats are used, but quantitatively based data collection appears more resistant to these effects when the label is known in advance. Therefore, the use of a Likert-type format for one category (e.g. defiance) and a more quantitative format for the other category that is unintentionally linked (e.g. ADHD) may help to reduce reciprocal influences on ratings, although retrospective frequency counts have shown to be influenced by labels in prior research (Yates, Klein, & Haven, 1978).

Assessment. Expectancy effects in which teachers rated similar behaviors differently due to assigned disability labels were robust across studies and settings, and may impact reevaluations by assessment personnel (e.g. diagnosticians, psychologists) that may be influenced by *a priori* knowledge of child referral question and diagnosis. By extension, teachers are almost certainly aware of categorical labels assigned directly through the school assessment process and indirectly through informal channels of school communication. Consequently, behavior ratings using information from teachers that are already aware of the prior classification may contribute systematic variance in the direction of continued qualification for students *absent of behavioral evidence*. While there is no perfect solution to this problem, collection of behavioral data from the student of interest as well as a peer nominated by the teacher as "typically behaving"

may assist evaluators in sorting out which behaviors are actually present when a limited sample of behavioral data is all that may be collected.

Limitations

The most salient study limitations are related to sample size and teacher characteristics. The sample size is a result of a combination of strict inclusion criteria chosen to improve external/conclusion validity but unintentionally reduces conclusion validity through a reduction in studies. This number was reduced in two ways—first by using only studies with a meaningful and defensible criterion, and second by excluding preservice teachers. While the first was necessary for results interpretation, the second was predicated upon the idea that the use of practicing teachers with classroom experience offers indispensable information that may not be obtainable from preservice teachers that have not experienced classroom instruction experiences. It is possible that classroom experience has little effect on bias in behavior ratings and can only be determined empirically. A final sample of 25 articles is not small for a typical synthetic review, nor was it small in participating subjects (total n= over 1880 teachers). Rather, the sample size was most limiting due to the broad number of bias categories investigated, with some bias types receiving one study and some receiving seven or more. As an example, Pygmalion effects were only empirically studied in one article that met criteria and rating-style biases in only two, while expectancy or ethnic bias was investigated in over half of included studies. Second, the vast majority of teachers were

female, Caucasian, and teaching elementary grade levels. While this is a problem inherent to study of teacher behaviors in the field as a whole, the lack of any secondary teachers and the limited number of males or teachers of other ethnic backgrounds reduces external validity to the extent the findings are applied to those groups.

Future Research

It is striking that studies of bias in teacher ratings of behavior are almost certainly influenced by implicit and explicit beliefs, yet these beliefs are rarely, if ever, measured. It would be rare for a social psychologist to attempt to capture evidence of a latent construct (e.g. just-world beliefs) impacting a behavioral variable without attempting to measure the latent construct prior to the execution of the study. In order to understand fully the source of biases present in teacher ratings of behavior, the next step will likely demand a better understanding of the engine that drives the harsher ratings of certain groups under certain conditions—the beliefs themselves. The tendency for education to avoid this method with teachers was perhaps most striking in the area of projection bias. While multiple research articles in clinical and counseling psychology have investigated the impact maternal depression has on maternal parent rating of child depression as well as related child psychopathology, yet no articles that met inclusion criteria had investigated this with teachers. It is perhaps uncomfortable for researchers, but critical to the understanding of the phenomenon.

While narrow scope was necessary for this review, the restriction of category to behavior ratings has inevitably masked evidence of bias *subsequent to* the behavior rating**.** As an example, Bahr et al. (1991) and Chang and Sue (2003) found no evidence

of bias in behavior ratings and thus were coded as not evidencing ethnic bias for the purposes of this review, yet both found bias in referral across ethnic groups. Referral practices, like behavior rating processes, are summative in nature and likely include multiple implicit beliefs about student characteristics and outcomes, as well as personal and organizational values. The conceptual gaps in the bias literature are most evident when unbiased scores can lead to biased outcomes based on those scores. Going forward, investigation of these processes using real teachers in real settings is most likely to yield meaningful answers to disparities in intervention and placement.

CHAPTER IV CONCLUSION: IMPLICATIONS FOR PRACTICE AND FUTURE RESEARCH

Universal screening for behavior offers the opportunity for local educational authorities to meet social and emotional needs for students that might otherwise go unnoticed. While screening for behavior is not new to communities and schools, there has been a recent expansion in research and practice that demands greater scrutiny of screening processes and measures. Universal screening has been implemented successfully for both literacy and mathematics, with screening instruments serving to screen for intervention as well as to monitor progress. However, there is limited evidence that behavioral screeners can be used to monitor progress with sufficient reliability to inform decision making. The most salient barriers to the use of behavioral screeners for progress-monitoring are: first, the lack of an evidence base supporting identifiable effects on screener scores from interventions known to produce behavioral effects, and second, insufficient evidence regarding teacher biases that may impact screener ratings even when behavioral differences are present. These two questions are, at their core, questions of validity.

This dissertation addresses this gap in the literature in two ways. First, universal screener scores were collected for preschool attenders and matched non-attenders in a school district with a public preschool option for students meeting admission criteria. Preschool participation has a long history of resulting in positive behavioral benefits for participating students (Barnett, 1990; Coleman, 1966), and if screeners are to be used to measure intervention outcomes, preschool participation is within the class of

interventions where such effects may be found due to its duration, intensity, and targeted nature (students of lower socioeconomic status). Second, a best-evidence synthesis of teacher ratings of bias in teacher ratings of behavior was conducted to provide an overview of the current evidence base supporting (or refuting) bias in teacher ratings of behavior.

The results of study one suggested no differences in student behavioral risk as judged by teachers for 69 preschool attenders and 69 non-attenders matched for ethnicity, gender, and a gross measure of socioeconomic status (free or reduced lunch). While this result does not suggest screeners cannot be used as measures of interventional benefits, it does not support it either. Of interventions available to learning communities, a yearlong introduction to the rules, procedures, and cultural mores of schools is one that comes at a higher fiscal cost but one that has a high degree of face validity. As academic effects fade by third grade, behavioral benefits should explain at least a portion of the variance and in the study conducted, no significant variance for teacher ratings of behavioral risk was explained by the preschool participation. More positively, the lack of difference found for Hispanic and Caucasian samples suggested a structured screening process may be more suitable for Hispanic students, a group that otherwise may not receive needed services.

The results of study two suggested the presence of multiple bias types in teacher ratings of behavior. While evidence of ethnic and gender bias was mixed, strong effects of teacher culture, expectancies due to disability label, halo effects from unrelated behaviors, and teacher training and experience with special education populations were

found. Additionally, multiple bias types had not been investigated with teachers. These included language bias (related to rating questions), and projection bias in which ratings of a construct are affected by the rater's own level of that construct. These systematic effects on behavior ratings are undetectable unless explicitly investigated and may actually raise reliability of measures at the psychometric level, rendering detection by publishers an unnecessary (and possibly counterproductive) endeavor as detection would negatively impact ratings validity for consumers.

Implications for Practice

Both sets of findings have implications for practitioners initiating universal screening for behavior. First, the same measures used to predict risk have not yet been proven to possess sufficient sensitivity to change for use as an outcome measure post-intervention. Accordingly, the use of screening measures as progress monitors is also unsupported, as the sensitivity to change required for progress monitoring is greater than the sensitivity required for measurement of intervention effects post-implementation. Thus, local educational authorities should be aware that the stability of risk status over time suggests that changing risk status is a difficult task and it is likely that more narrow and focused measures should be used for intervention measurement.

Second, teacher bias is present in behavior ratings scales at multiple levels, and if referral systems for special education intervention and placement use them at multiple levels (screening, referral, progress monitoring), the impact of teacher bias is multiplied. Consequently, any procedure that relies *solely* on behavior rating scales for identification will be susceptible to that bias. While scales that detect harshness in teacher ratings are

helpful, they are not necessarily indicative of biased ratings. However, if schools add multiple layers to schoolwide screening mechanisms, those mechanisms may become cost prohibitive. At this point, it is most important that practitioners are aware of potential teacher biases until more research is conducted.

Limitations

The study was limited in the number of students as well as background information provided for nonattending students that would allow for comparison of the public preschool option to the setting for nonattending students (e.g. home, center-based child care, private school). Consequently, is unknown if the lack of differences in behavioral risk stems from a lack of difference in outcomes, a lack of difference in essential characteristics of public preschool and the alternative setting for nonparticipants, or a measurement that was insensitive to the changes associated with public preschool.

Future Research

Future research that explores the use of screening procedures that utilize teacher ratings of behavior should focus on the root source of differences in teacher ratings. Discovering the factors behind differences in teacher ratings will assist in correcting for bias in screening procedures. Additionally, as Hispanic growth in student populations continue, more work on how teacher culture and student culture interact to influence ratings of behavior is needed.

More broadly, these two studies indicate screening procedures that use teacher ratings of behavior alone to screen for intervention is open to potential rating bias and

more research is needed in this area. This research should focus not just on identification of at-risk groups, but on the predictive utility of those ratings. Additionally, the use of screeners as measures of interventional efficacy was not supported in the current study but may perhaps be supported using a different screening measure, a more powerful intervention, or a different sample. A screening tool sensitive enough to be used for screening and progress monitoring that is not time-intensive will assist greatly in maximizing screening utility in schools.

REFERENCES

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

Achenbach, T. M. (1966). The classification of children's psychiatric symptoms: A factor-analytic study. *Psychological Monographs, 80, (No. 615).*

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

Achenbach, T. M. & Rescorla, L.A. (2001). *Manual for the ASEBA School-Age Forms & Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Alban-Metcalfe, J., Cheng-Lai, A., & Ma, T. (2002). Teacher and student teacher ratings of attention-deficit/hyperactivity disorder in three cultural settings. *International Journal of Disability, Development & Education*, *49*(3), 281 - 299.

Alliger, G. M., & Williams, K. J. (1992). Relating the internal consistency of scales to rater response tendencies. *Educational & Psychological Measurement*, *52*(2), 337.

Aloise-Young, P. A., & Chavez, E. L. (2002). Not all school dropouts are the same: Ethnic differences in the relation between reason for leaving school and adolescent substance use. *Psychology in the Schools*, *39*(5), 539-547.

American Academy of Pediatrics, Committee on Children with Disabilities (1999). The pediatrician's role in the development and implementation of an Individual Education Plan (IEP) and/or an Individual Family Service Plan (IFSP). *Pediatrics,* 104:124–127.

Angold, A., Costello, E. J., Farmer, E. M. Z., Burns, B. J., & Erkanli, A. (1999). Impaired but undiagnosed. *Journal of the American Academy of Child & Adolescent Psychiatry, 38*(2), 129-137. doi:10.1097/00004583-199902000-00011

Atkins, M. S., Pelham, W. E., & Licht, M. H. (1985). A comparison of objective classroom measures and teacher ratings of attention deficit disorder. *Journal of Abnormal Child Psychology: An Official Publication of the International Society for Research in Child and Adolescent Psychopathology*, *13*(1), 155-167. doi:10.1007/BF00918379

Bahr, M. W., Fuchs, D., Stecker, P. M., & Fuchs, L. S. (1991). Are teachers' perceptions of difficult-to-teach students racially biased? *School Psychology Review*,*20*, 599–608.

Barnett, W.S., & Hustedt, J. T. (2005). Head start's lasting benefits. *Infants and Young Children*, *18*, 16–24.

Beaver, B. R., & Busse, R. T. (2000). Informant reports: Conceptual and research bases of interviews with parents and teachers. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools: Theory, research, and clinical foundations* (2nd ed., pp. 257–287). New York: Guilford Press.

Belfield, C.R. (2006). *Does It Pay to Invest in Preschool for All? Analyzing Return-on-Investment in Three States*. New Brunswick, NJ: National Institute for Early Education Research.

Belsky, J., & MacKinnon, C. (1994). Transition to school: Developmental trajectories and school experiences. *Early Education and Development, 5*(2), 106–119.

Berman, N.G., & Parker, R.A. (2002) Meta-analysis: neither quick nor easy. *BMC Medical Research Methodology, 2* (10), 1–9.

Bland, J.M, & Altman, D.G. (1996).Transforming data. *British Medical Journal*, 307-313.

Brandon, K. A, Kehle, T.J., Jenson, W. R, & Clark, E. (1990). Regression, practice, and expectation effects on the Revised Conners Teacher Rating Scale. *Journal of Psychoeducational Assessment, 8,* 456-466.

Burns BJ, Costello EJ, Angold A, Tweed, D., Stangl, D., Farmer, E.M., & Erklani, A. (1995). Children's mental health service use across service sectors. *Health Affairs, 14*(3):147–159.

Caldarella, P., Young, E. L., Richardson, M. J., Young, B. J., & Young, K. (2008). Validation of the systematic screening for behavior disorders in middle and junior high school. *Journal of Emotional and Behavioral Disorders*, *16*(2), 105-117. doi:10.1177/1063426607313121

Campbell, D. T., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Chafouleas, S.M., Riley-Tillman, T.C., & McDougal, J.L. (2002). Good, bad, or in-between: How does the daily behavior report card rate? *Psychology in the Schools, 39,* 157–169.

Chang, D. F., & Sue, S. (2003). The effects of race and problem type on teachers' assessments of student behavior. *Journal of Consulting and Clinical Psychology, 7,* 235–242.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement. 5,* 115-124.

Cole, M., & Bruner, J. S (1971). Cultural differences and inferences about psychological processes. *American Psychologist,* 26, 867-876.

Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Wein-feld, F., & York, R. (1966). Equality of educational opportunity. Washington, DC: U.S. Department of Health, Education, and Welfare.

Cook, C. R., Rasetshwane, K. B., Truelson, E., Grant, S., Dart, E. H., Collins, T. A., & Sprague, J. (2011). Development and validation of the "student internalizing behavior screener": Examination of reliability, validity, and classification accuracy. *Assessment for Effective Intervention, 36*(2), 71-79. Retrieved from http://dx.doi.org.lib-ezproxy.tamu.edu:2048/10.1177/1534508410390486

Cook, C. R., Volpe, R. J., & Livanis, A. (2010). Constructing a roadmap for future universal screening research beyond academics. *Assessment for Effective Intervention, 35*(4), 197-205. Retrieved from http://dx.doi.org.lib-ezproxy.tamu.edu:2048/10.1177/1534508410379842

Costello, E.J., Mustillo, S., Erkanli, E., Keeler, G., & Angold, A. (2003) Prevalence and development of psychiatric disorders in childhood and adolescence. *Archives of General Psychiatry*. *60*(8), 837-844.

Crocker, L, & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace Jovanovich College Publishers.

Currie, J., & Neidell, M. (2003). *Getting inside the "black box" of Head Start quality: What matters and what doesn't* (Working Paper 10091). Cambridge, MA: National Bureau of Economic Research.

DeVellis R.F. (2006). Classical test theory. *Medical Care*. *44*(3),50–59.

Diamond, A., Barnett, W.S., Thomas, J., & Munro, S. (2007). Preschool program improves cognitive control. *Science*, 318, 1387–1388.

Diamond, K., & Squires, J. (1993). The role of parental report in the screening and assessment of young children. *Journal of Early Intervention, 17.* 107-115.

Distefano, C. A., & Kamphaus, R. W. (2007). Development and validation of a behavioral screener for preschool-age children. *Journal of Emotional and Behavioral Disorders, 15*(2), 93-102. doi:10.1177/10634266070150020401

Duhig, A.M., Renk, K., Epstein, M.E., & Phares, V. (2000). Interparental agreement on internalizing, externalizing, and total behavior problems: A meta-analysis. *Clinical Psychology: Science and Practice, 7,* 435-453.

Dukes, M. & Saudargas, R.A. (1989). Teacher evaluation bias toward LD children: Attenuating effects of the classroom ecology. *Learning Disability Quarterly*, (*12*)2, 126-132.

Duncan, G.J., Dowsett, C.J., Claessens, A., Magnuson, K., Huston, A.C., Klebanov, P., et al. (2007). School readiness and later achievement. *Developmental Psychology 43*, 1428–1446.

Eklund, K., Renshaw, T. L., Dowdy, E., Jimerson, S. R., Hart, S. R., Jones, C. N., & Earhart, J. (2009). Early identification of behavioral and emotional problems in youth: Universal screening versus teacher-referral identification. *California School Psychologist, 14*, 89-95.

Elliott, S. N., Huai, N., & Roach, A. T. (2007). Universal and early screening for educational difficulties: Current and future approaches. *Journal of School Psychology, 45*(2), 137-161. doi:10.1016/j.jsp.2006.11.002

Epkins, C. C., & Meyers, A. W. (1994). Assessment of childhood depression, anxiety, and aggression: Convergent and discriminant validity of self-, parent, teacher, and peer-report measures. *Journal of Personality Assessment, 62,* 364-381.

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

Ervin, R.A., Schaughency, E., Goodman, S.D., McGlinchey, M.T., & Matthews, A. (2006). Merging research and practice agendas to address reading and behavior schoolwide. *School Psychology Review, 35,* 198–223.

Essex, M. J., Kraemer, H. C., Slattery, M. J., Burk, L. R., Boyce, W. T., Woodward, H. R., & Kupfer, D. J. (2009). Screening for childhood mental health problems:

Outcomes and early identification. *Journal of Child Psychology and Psychiatry, 50*(5), 562-570. Retrieved from http://www.blackwell-synergy.com/doi/abs/10.1111/j.1469-7610.2008.02015.x

Farmer, E. M. Z., Burns, B. J., Phillips, S. D., Angold, A., & Costello, E. J. (2003). Pathways into and through mental health services for children and adolescents. *Psychiatric Services, 54*(1), 60-66. doi:10.1176/appi.ps.54.1.60

Fennerty, D., Lambert, C., & Majsterek, D. (2000). Behavior rating scales: An analysis. U.S. Department of Education; Educational Resources Information Center (ERIC). Retrieved from www.eric.ed.gov

Fogel, L.S., & Nelson, R.O. (1983). The effects of special education labels on teachers behavioral observations, checklist scores, and grading of academic work, *Journal of School Psychology, 3*(21), 241-251.

Foster, G., & Ysseldyke, J. (1976). Expectancy and halo effects as a result of artificially induced teacher bias. *Contemporary Educational Psychology, 1*, 37–45.

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling, 13,* 378−402.

Glover, T., & Albers, A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45*, 7−25

Gregorich, S.E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*. *44*(3),78–94.

Gresham, F.M., MacMillan, D. L., & Bocian, K. (1996). "Behavioral earthquakes": Low frequency, salient behavioral events that differentiate students at-risk for behavioral disorders. *Behavioral Disorders, 21*, 277−292.

Gresham, F. M., & Elliott, S. N. (2008). *Social Skills Improvement System—Rating Scales.* Minneapolis: Pearson Assessments.

Guilford, J. P. (1954). *Psychometric Methods*, (2nd Ed). New York, NY: McGraw-Hill Book Company, Inc.

Hanish, L. D. & Guerra, N. G. (2000). The Roles of Ethnicity and School Context in Predicting Children's Victimization by Peers. *American Journal of Community Psychology*, *28*, 201–223.

Harris, F.C. & Lahey, B.B. (1982). Subject reactivity in direct observational assessment: A review and critical analysis, *Clinical Psychology Review*, Volume 2, (4), 523-538. Retrieved from http://dx.doi.org/10.1016/0272-7358(82)90028-9

Harris, J., Tyre, C., & Wilkinson, C. (1993). Using the child behavior checklist in ordinary primary schools. *British Journal of Educational Psychology, 63*, 245-60.

Haynes, S. N., Smith, G., & Hunsley, J. (2011). *Scientific Foundations of Clinical Assessment*. New York, NY: Taylor & Francis.

Hinshaw, S. P. (1992). Externalizing behavior problems and academic underachievement in childhood and adolescence: Causal relationships and underlying mechanisms. *Psychological Bulletin, 111*(1), 127-155.

Hintze, J.M., & Matthews, W.J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review*, *33*, 258–270.

Hoge, R. D. & McKay, V. (1986). Criterion-related validity data for the Child Behavior Checklist-Teacher's Report Form. *Journal of School Psychology*, *24*, 387–393.

Hooman, H.A. (1982). Study of validity of ratings. *Psychological Reports*, *51*(3, 1263-1270.

Hosp, J. L., Howell, K. W., & Hosp, M. K. (2003). Characteristics of behavior rating scales: Implications for practice in assessment and behavioral support. *Journal of Positive Behavior Interventions, 5*, 201-208.

Hosterman, S. J., DuPaul, G. J., & Jitendra, A. K. (2008). Teacher ratings of ADHD symptoms in ethnic minority students: Bias or behavioral difference? *School Psychology Quarterly, 23*, 418-435.

Howell, A. J., & Watson, D. C. (2009). Impairment and distress judgments of symptoms composing childhood externalizing and internalizing syndromes. *Journal of Child and Family Studies, 18*(2), 172-182. doi:10.1007/s10826-008-9217-y

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

Hughes, C., & Ensor, R. (2011). Does executive function matter for preschoolers' problem behavior? *Journal of Abnormal Child Psychology*, 36, 1–14.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings.* Newbury Park, CA: Sage.

Individuals with Disabilities Education Improvement Act of 2004 (IDEA), Pub. L. No. 108– 446, 118 Stat. 2647 (2004). [Amending 20 U.S.C. §§1400 et seq.].

Intille, S. S., Tapia, E. M., Rondoni, J., Beaudin, J., Kukla, C., Agarwal, S.,& Bao, L. (2003). Tools for studying behavior and technology in natural settings (pp. 157–174). In A. K. Dey, A. Schmidt, & J. F. McCarthy(Eds.), *Proceedings of UBICOMP 2003: Ubiquitous Computing, vol.LNCS 2864.* Berlin: Springer.

Jacob, R. G., O'Leary, K. D., & Rosenblad, C. (1978). Formal and informal classroom settings: Effects on hyperactivity. *Journal of Abnormal Child Psychology*, *6*, 47-60.

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

Kahn, B. E., & Baron, J. (1995). An exploratory study of choice rules favored for high-stakes decisions. *Journal of Consumer Psychology, 4*,305–328.

Kamphaus, R. W. (1999). Intelligence test interpretation: Acting in the absence of evidence. In A. Prifitera & D. Saklofske (Eds.). *WISC-III Clinical Use and Interpretation: Scientist — Practitioner Perspectives* (pp. 39-57). San Diego, CA: Academic Press.

Kamphaus, R.W., & Reynolds, C.R. (2007). *Behavior assessment system for children—second edition (BASC-2): Behavioral and Emotional Screening System (BESS)* Bloomington, MN: Pearson.

Kamphaus, R.W., Thorpe, J., Winsor, A.P., Kroncke, A.P., Dowdy, E.T., & VanDeventer, M, (2007). Development and Predictive Validity of a Teacher Screener for Child Behavioral and Emotional Problems at School. *Educational and Psychological Measurement*, *67*(2), 342-356.

Kauffman, J.M. (2001). *Characteristics of Emotional and Behavioral Disorders of Children and Youth* (7th ed.). Upper Saddle River, NJ: Merrill Prentice-Hall.

Kelter, J.D., & Pope, A.W. (2011). The effect of child gender on teachers' responses to oppositional defiant disorder. *Child & Family Behavior Therapy*, *33*(1), 49-57.

Kessler, R.C., McGonagle, K.A., Zhao, S., Nelson, C.B., Hughes, M., Eshleman, S., Wittchen, H.U., & Kendler, K.S. (1994). Lifetime and 12-month prevalence of *DSM-III-R* psychiatric disorders in the United States: results from the National Comorbidity Survey. *Arch Gen Psychiatry, 51,* 8-19.

Kindler, A. L. (2002). Survey of the states' limited English proficient students and available programs and services 2000–2001 summary report. Washington, DC: National Clearinghouse for English Language Acquisition.

Kingstrom, P. O., & Bass, A. R. (1981). A critical analysis of studies comparing behaviorally anchored rating scales (BARS) and other rating formats. *Personnel Psychology, 34,* 263–289.

Kroes, G., Veerman, J. W., & De Bruyn, E. J. (2005). The impact of the big five personality traits on reports of child behavior problems by different informants. *Journal of Abnormal Child Psychology*, *33*(2), 231-240.

Law, J., Boyle, J., Harris F, Harkness, A., & Nye, C. (1998). Screening for primary speech and language delay: a systematic review of the literature. *International Journal of Language and Communication Disorders*. *33*, 21-23.

Levitt, J.M., Saka, N., Romanelli, L.H., & Hoagwood, K. (2007). Early identification of mental health problems in schools: The status of instrumentation. *Journal of School Psychology, 45*, 163–191.

Loeber, R., Green. S., Lahey, B., & Stouthamer-Loeber, M. (1990). Optimal informants on childhood disruptive behaviors. *Development and Psychopathology*, *1*, 317-337.

Love, J. M., Chazan-Cohen, R., Raikes, H., Aber, J. L., Bishop-Josef, S. J., Jones, S. M., et al. (2007). Forty years of research knowledge and use: from Head Start to Early Head Start and beyond. In J. L. Aber, S. J. Bishop-Josef, S. M. Jones, K. T. McLearn, & D. A. Phillips (Eds.), *Child Development and Social Policy: Knowledge for Action* (pp. 79-95). Washington, DC: American Psychological Association.

Maag, J. W., & Katsiyannis, A. (2008). The medical model to block eligibility for students with EBD: A response-to-intervention alternative. *Behavioral Disorders, 33*(3), 184-194. Retrieved from

http://www.ccbd.net/behavioraldisorders/Journal/Journal.cfm?error=BD&BDID=542E656C-3048-7B00-419CB9E2B7B7AD54

Magnuson, K., Ruhm, C., & Waldfogel, J. (2004). *Does prekindergarten improve school preparation and performance?* Cambridge, MA: National Bureau of Economic Research.

Marchant, M., Anderson, D. H., Caldarella, P., Fisher, A., Young, B. J., & Young, K. R. (2009). Schoolwide screening and programs of positive behavior support: Informing universal interventions. *Preventing School Failure, 53*(3), 131-144.

Marchant, M., Brown, M., Caldarella, P., & Young, E. (2010). Effects of strong kids curriculum on students with internalizing behaviors: A pilot study. *Journal of Evidence-Based Practices for Schools*, *11*, 123-143.

Marks, K., Glascoe, F. P., Aylward, G. P., Shevell, M. I., Lipkin, P. H., & Squires, J. K. (2008). The thorny nature of predictive validity studies on screening tests for developmental-behavioral problems. *Pediatrics, 122*(4), 866-868. doi:10.1542/peds.2007-3142

Massa, I. (2011) *Underrepresentation of Hispanic/Latino students identified with emotional disturbance in IDEIA: What's the teacher's role*? Unpublished doctoral dissertation. Texas A&M University; College Station, TX: 2011.

Meisels, S.J. & Provence, S. (1989) *Screening and Assessment. Guidelines for Identifying Young Disabled and Developmentally Vulnerable Children and Their Families*. Washington, DC: National Center for Clinical Infant Programs.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*, 525−543.

Merrell, K.W. (1994). Assessment of behavioral, social, & emotional problems. New York: Longman.Merrell, K. W. (2003). *Behavioral, Social, and Emotional Assessment of Children and Adolescents.* Mahwah, NJ: Lawrence Erlbaum.

Merrell, K. W. (2000). Informant reports: Theory and research in using child behavior rating scales in school settings. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral Assessment in Schools: Theory, Research, and Clinical Foundations* (2nd ed., pp. 233–256). New York, NY: The Guilford Press.

Merz, E.C., & McCall, R.B. (2010). Behavior problems in children adopted from psychosocially depriving institutions. *Journal of Abnormal Child Psychology*. *38*, 459–470.

Milfort, R., & Greenfield, D. B. (2002). Teacher and observer ratings of Head Start children's social skills. *Early Childhood Research Quarterly, 17,* 581–595.

Milich, R., & Landau, S. (1988). Teacher ratings of inattention/overactivity and aggression: Cross-validation with classroom observations. *Journal of Clinical Child Psychology, 17*(1), 92-97.

Milich, R., Roberts, M. A., Loney, J., & Caputo, J. (1980). Differentiating practice effects and statistical regression on the Conners' Hyperkinesis Index. *Journal of Abnormal Psychology, 8,* 549-552.

Mueller, C.W., Mann, E.M., Thanapum, S., Humris, E., Ikeda, Y., Takahashi, A., Tao, K.T., & Li, B.L. (1995). Teachers' ratings of disruptive hehaviour in five countries. *Journal of Clinical Child Psycholo*gy, *24*(4) 434-42,

Neal, L. V., McCray, A. D., Webb-Johnson, G., & Bridges, S. T. (2003). The effects of African American movement styles on teachers' perceptions and reactions. *Journal of Special Education, 37,* 49–57..

Nores, M., & Barnett, S. (2009). Benefits of early childhood interventions across the world: Investing in the very young *Economics of Education Review*, 1-12.

Nunnally, J. C. (1978). *Psychometric Theory*. New York: McGraw-Hill.

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric Theory* (3rd ed.). New York, NY: McGraw-Hill.

Phares, V., Compas, B. E., & Howell, D. C. (1989). Perspectives on child behavior problems: Comparisons of children's self-reports with parent and teacher reports. *Psychological Assessment, 1,* 68–71.

Phillips, B. M., & Lonigan, C. J. (2010). Child and informant influences on behavioral ratings of preschool children. *Psychology in the Schools, 7,* 311.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science: Research and recommendations on how to control it. *Annual Review of Psychology, 63*(1), 539-569.

Polo, A.J., & Lopez, S.R. (2009). Culture, context, and the internalizing distress of Mexican American youth. *Journal of Clinical Child & Adolescent Psychology, 38*, 273-285.

Prieto, A.G., & Zucker, S.H. (1981). Teacher perception of race as a factor in the placement of behaviorally disordered children. *Behavioral Disorders, 7*, 34-38.

Puig, M., Lambert, M. C., Rowan, G. T., Winfrey, T., Lyubansky, M., Hannah, S. D., & Hill, M. F. (1999). Behavioral and emotional problems among Jamaican and African American children, ages 6 to 11: Teacher reports versus direct observations. *Journal of Behavioral and Emotional Disorders, 7*, 240-250.

Ramey, C. T., & Ramey, S. L. (1998). Early intervention and early experience. *American Psychologist, 53*, 109–120.

Reid, R. (1995). Assessment of ADHD with culturally different groups: The use of behavioral rating scales. *School Psychology Review, 24*, 537.

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

Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavior Assessment System for Children—Second Edition*. Circle Pines, MN: AGS.

Reynolds, C. R., Lowe, P. A., & Saenz, A. L. (1999). The problem of bias in psychological assessment. In C. R.Reynolds & T. B. Gutkin (Eds.). *The Handbook of School Psychology* (pp. 549–595). New York, NY: Wiley.

Riggs, N.R, Blair, C.B., & Greenberg, M.T. (2003). Concurrent and two-year longitudinal relations between executive function and the behavior of first- and second-grade children. *Child Neuropsychology, 9,* 267–276.

Sandoval J (1977). The measurement of the hyperkinetic syndrome in children. *Review of Educational Research,* 47:293–318.

Sattler, J. M. (1992). *Assessment of Children* (3rd ed.). San Diego, CA: Author.

Saudino, K.J. (2005). Behavioral genetics and child temperament. *Journal of Developmental and Behavioral, Pediatrics. 26*(3), 214-223.

Saunders, & Di Tullio, (1972). The failure of biased information to affect teacher behavior ratings and peer sociometric status of disturbing children in the classroom. *Psychology in the Schools*, *9*(4), 440-445.

Schachar, R., & Logan, G.D. (1990). Impulsivity and inhibitory control in normal development and childhood psychopathology. *Developmental Psychology, 26,* 710-720.

Schachar, R.J., Sandberg, S., & Rutter, M. (1986). Agreement between teachers' ratings and observations of hyperactivity, inattentiveness and *defiance. Journal of Abnormal Child Psychology, 14,* 331-345.

Schweinhart, L. J., & Weikart, D. P. (1997). The High/Scope preschool curriculum comparison study through age 23. *Early Childhood Research Quarterly, 12,* 117–143.

Seifer, R., Sameroff, A.J., Barrett, L.C., & Krafchuk, E. (1994). Infant temperament measured by multiple observations and mother report. *Child Development, 65,* 1478 –1490.

Shaffer, D., Fisher, P.W., Dulcan ,M., Davies, M., Piacentini, J., Schwab-Stone, M., Lahey, B.B., Bourdon, K., Jensen, P., Bird, H., Canino, G., & Regier, D. (1996) The NIMH diagnostic interview schedule for children (DISC 2.3): Description, acceptability, prevalences, and performance in the MECA study. *Journal of the American Academy of Child and Adolescent Psychiatry. 35,* 865-877.

Shin, Y. M., Chung, Y. K., Lim, K. Y., Lee, Y. M., Oh, E. Y., & Cho, S. M. (2009). Childhood predictors of deliberate self-harm behavior and suicide ideation in korean adolescents: A prospective population-based follow-up study. *Journal of Korean Medical Science, 24*(2), 215-222.

Slavin, R.E. (1986). Best-evidence synthesis: an alternative to meta-analytic and traditional reviews. *Educational Researcher*. *15*, 5-11.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12*(1), 102-111. doi:10.1037/1040-3590.12.1.102

Sonuga-Barke, E., Minocha, K., Tayor, E., & Sandberg, S. (1993). Inter-ethnic bias in teachers' ratings of childhood hyperactivity. *British Journal of Developmental Psychology*, *11*, 187-200.

Sourander, A., Klomek, A. B., Niemelä, S., Haavisto, A., Gyllenberg, D., Helenius, H., Gould, M. S. (2009). Childhood predictors of completed and severe suicide

attempts: Findings from the Finnish 1981 birth cohort study. *Archives of General Psychiatry, 66*(4), 398-406. doi:10.1001/archgenpsychiatry.2009.21

Stanger, C., & Lewis, M. (1993). Agreement among parents, teachers, and children on internalizing and externalizing behavior problems. *Journal of Clinical Child Psychology, 22*, 107–115.

Stevens, J., Quittner, A.L., & Abikoff, H. (1998). Factors influencing elementary school teachers' ratings of ADHD and ODD behaviors. *Journal of Clinical Child Psychology, 27*, 406–414.

Stevens, G. (1980). Bias in attributions of positive and negative behavior in children by school psychologists, parents, and teachers. *Perceptual and Motor Skills, 50*, 1283-1290.

Stevens, G. (1981). Bias in the attribution of hyperkinetic behavior as a function of ethnic identification and socioeconomic status. *Psychology in the Schools, 18*, 99-106.

Thorndike, Edward (1920). The constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25-29.

U. S. Department of Education, National Center for Education Statistics. *Status and trends in the education of Hispanics*. Washington, DC: 2003. (NCES 2003-008)

United States Public Health Service. (2000). *Report of the Surgeon General's Conference on Children's Mental Health: A national action agenda*. Washington, DC: Department of Health and Human Services.

Vélez, W., & Saenz, R. (2001). Toward a comprehensive model of the school leaving process among latinos. *Social Psychology Quarterly* 16:445–67.

Verhulst, F. C, & Akkerhuis, G. W. (1989). Agreement between parents' and teachers' ratings of behavioral/emotional problems of children aged 4-12. *Journal of Child Psychology and Psychiatry, 30,* 123-136.

Verhulst, F. C, & van der Ende, J. (1991). Assessment of child psychopathology: Relationships between different methods, different informants and clinical judgment of severity. *Acta Psychiatrica Scandinavica, 84,* 155-159.

Volpe, R. J., Briesch, A. M., & Chafouleas, S. M. (2010). Linking screening for emotional and behavioral problems to problem-solving efforts: An adaptive model of behavioral assessment. *Assessment for Effective Intervention, 35*(4), 240-244. Retrieved from http://dx.doi.org.lib-ezproxy.tamu.edu:2048/10.1177/1534508410377194

Volpe, R. J., Briesch, A. M., & Gadow, K. D. (2011). The efficiency of behavior rating scales to assess inattentive-overactive and oppositional-defiant behaviors: Applying generalizability theory to streamline assessment. *Journal of School Psychology, 49*(1), 131-155. doi:10.1016/j.jsp.2010.09.005

Walker, E., Bettes, B., & Ceci, S. (1984). Teachers' assumptions regarding the severity, causes, and outcomes of behavioral problems in preschoolers: Implications for referral. *Journal of Consulting and Clinical Psychology, 52,* 899-902.

Walker, H.M., & Severson, H.H. (1992). Systematic screening for behavior disorders.

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science, 36*(3), 409–422. doi:10.1007/s11747-007-0077-6

Weisz, J.R., Chaiyasit, W., Weiss, B., Eastman, K.L., & Jackson, E.W. (1995). A multimethod study of problem behavior among Thai and American children in school: teacher reports versus direct observation. *Child Development,66*, 402–15.

Wilson, J.M.G., & Jungner, G. (1968). *Principles and Practice of Screening for disease*. Geneva: WHO.

Wu, P., Hoven, C. W., Bird, H. R., Moore, R. E., Cohen, P., Alegria, M., et al. (1999). Depressive and disruptive disorders and mental health service utilization in children and adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 38*, 1081–1090. doi:10.1097/00004583-199909000-00010

Yates, B.T., Klein, S.B. & Haven, W.G. (1978). Psychological nosology and mnemonic reconstruction: Effects of diagnostic labels on observers' recall of positive and negative behavior frequencies. *Cognitive Therapy and Research, 2*, 377–387.

APPENDIX A

DESCRIPTIVE SUMMARY OF STUDIES INCLUDED IN BIAS REVIEW

Author/ Year (alpha)	*Raters*	*Rating forms/ Dependent measure*	*CM*[a]	*Bias variable*	*Results*
Abikoff, Courtney, Pelham, & Koplewicz, 1993	139 TCHRs (72 RegEd, 67 SPED)	CTRS, IOWA Conners/ ADHD and ODD ratings	VV	Examiner, halo effects	Examiner bias and halo effects
Alban-Metcalfe, Cheng-Lai, & Ma, 2002	175 TCHRs from China, Hong Kong, and the UK;	SNAP/Ratings of HYP, Inattention, and Impulsivity,	VV	Cultural	Cultural bias found
Bahr, Fuchs, Stecker, & Fuchs, 1991	40 middle school TCHRs and 10 school-based OBSs	RBPC/Bx ratings	DO	Ethnic	Ethnic bias not found
Brandon, Kehle, Jenson, & Clark, 1990	60 elementary school SPED and RegEd TCHRs	CTRS-R/ ADHD ratings	VV, RM	Expectancy, order effects, practice, regression,	Expectancy bias not found, order effects found, practice effects not found, regression effects not found
Chang & Sue, 2003	197 TCHRs (75% elementary)	Likert-scale/Bx problem severity	SV	Ethnic, expectancy	Ethnic bias not found; expectancy bias found
Dukes & Saudargas, 1989	80 elementary TCHRs	CBI-modified/ Ratings of Distractibility, Orientation to Task, and Introversion	SV, VV	Expectancy, situational bias	Expectancy bias found, situational bias found

Epstein, et al, 2005	Elementary TCHRs and trained OBSs	CTRS-R, CBCL-TRF, SNAP-IV/ADHD Bx	DO	Ethnic, halo effects	Ethnic bias not found halo effects not found
Fogel & Nelson, 1981	30 primary-grade teachers	Bx Checklist/ Ratings of LD, MR and ED Bx	VV	Expectancy	Expectancy bias found
Foster & Ysseldyke, 1976	100 elementary TCHRs	Bx checklist/Bx ratings	SV, VV	Expectancy[b]	Expectancy bias found
Hosterman, DuPaul, & Jitendra, 2008	120 elementary TCHR and OBSs	CTRS, ADHD-IV/ ADHD ratings	DO	Ethnic	No ethnic bias found
Jacob, O' Leary,& Rosenblad, 1978	1 Study-assigned TCHR and two trained OBSs	CTRS/Ratings of HYP Bx	DO	Situational bias	Situational bias found
Jackson & King, 2004	80 TCHRS (40 SPED, 40 RegEd)	CTRS-R; DBD/ Ratings of ODD, ADHD	VV	Halo effects	Halo effects found
Kelter & Pope, 2011	145 elementary TCHRs	Likert scale/ Bx severity	SV	Gender	Gender bias not found
Milfort & Greenfield, 2002	21 Head Start TCHRs and 11 OBS	PIPPS/ Ratings of: Disruption, Play Interaction, Disconnection	DO	Ethnic, gender	Ethnic bias found Gender bias found
Mueller et al., 1995	130 elementary TCHRs from five countries	17 item disruptive Bx checklist/ Ratings of disruptive Bx and HYP	VV	Cultural, examiner, situational bias,	Cultural bias found examiner bias found situational bias found
Neal, McCray, Webb-Johnson, & Bridges, 2003	136 middle school TCHRs	AGG scales of the ACL/Ratings of AGG	VV	Ethnic, cultural	Ethnic bias not found Cultural bias found

Phillips & Lonigan, 2010	98 TCHRs and 36 OBSs for: of preschool children	CTRS, EASI/behavior ratings	DO	Expectancy	Expectancy bias not found
Puig, et al., 1999	51 U.S. and Jamaican TCHRs and 4 OBSs	CBCL, JTRF/TPS, EP,IP	DO	Cultural, gender	Cultural bias found gender bias not found
Saunders & Di Tullio, 1972	9 TCHRs (grades 3-6)	BRS/Ratings of Bx post-treatment	RM	Expectancy	Expectancy bias not found
Schachar, Sandberg, & Rutter, 1986	Elementary TCHRs and OBSs in London	CTRS/Ratings of HYP, Bx problems	DO	Halo effects	Halo effects found
Sonuga-Barke, et al., 1993	Elementary TCHRs and OBSs in London	Rutter questionnaire/ ratings of HYP	DO	Ethnic	Ethnic bias found
Stevens, 1980; Stevens, 1981[c]	27 elementary school TCHRs	Likert-scale/Ratings of positive and negative Bx	VV	Expectancy, ethnic	Expectancy bias found ethnic bias found
Stevens, Quittner, & Abikoff, 1998	105 elementary school TCHRs	CTRS, SNAP-IV/Ratings of: HYP, ODD, Inattention	VV	Examiner, halo effects, order effects, rating-style	Examiner bias found halo effects found, order effects not found rating-style effects not found
Walker, Bettes, & Ceci, 1984	100 preschool TCHRs	Likert-scale/ Ratings of Bx severity of AGG, HYP, Withdrawal	SV	Expectancy, gender	Expectancy bias found gender bias not found
Weisz,Chaiyasit, Weiss, Eastman, & Jackson, 1995	Thai and U.S. elementary TCHRs; OBSs (n=5)	CBCL-modified/ Ratings of TPS,EP, IP	DO	Cultural, gender	Cultural bias found gender bias not found

Note. ACL= Adjective Checklist; ADHD= Attention Deficit Hyperactive Disorder; AGG= Aggression; BPC=Behavior problem checklist; BRS= Behavior Rating of Pupils; Bx= behavior; CBCL= Child Behavior Checklist; CBI= Classroom Behavior inventory; CDI= Children's Depression Inventory; CRS= Conners Rating Scale; CTRS= Conners Teacher Rating Scale; DBD= Disruptive Behaviors Disorder Rating Scale; DO= direct observation; EASI= The Emotionality, Activity, Sociability & Impulsivity Temperament Survey; EP= Externalizing problems; HYP= Hyperactive; IP= Internalizing problems; JTRF= Jamaican Teacher's Report Form; ODD= Oppositional/Defiant Disorder; PIPPS= Penn Interactive Peer Play Scale; RBPC= Revised Behavior Problem Checklist; RCMAS = Revised Children's Manifest Anxiety Scale; RegEd= regular education; RM= Repeated measure; SDQ= Strengths and Difficulties Questionnaire; SNAP= Swanson, Nolan, and Pelham Rating Scale; SPED= Special education; SV= scripted vignette; TCHR= Teacher; TPS= Total problem score; TRS= Teacher's Rating Scale of Child's Actual Behavior; VV= video vignette

[a] CM= the criterion measure (if any) used to determine accuracy of behavior ratings
[b] While Foster & Ysseldyke (1976) referred to separate study conditions as expectancy bias and halo effects, both were coded as expectancy due to operational definitions chosen for this review
[c] Stevens, 1980 and Stevens, 1981 were combined due to identical sample and method