Running head: CONCURRENT VALIDITY OF OFFICE DISCIPLINE REFERRALS

Concurrent Validity of Office Discipline Referrals and Cut Points Used in School-Wide Positive Behavior Support

Kent McIntosh

University of British Columbia

Amy L. Campbell

Grand Valley State University

Deborah Russell Carter

Boise State University

Bruno D. Zumbo

University of British Columbia

Authors' Note: This research was supported by U.S. Department of Education Grants No. H324D020031, H326S980003, and Q215E050001. Opinions expressed herein do not necessarily reflect the policy of the Department of Education, and no official endorsement by the Department should be inferred.

Abstract

Office discipline referrals (ODRs) are commonly used by school teams implementing school-wide positive behavior support to indicate individual student need for additional behavior support. However, little is known about the technical adequacy of ODRs when used in this manner. In this study, the authors assessed a) the concurrent validity of number of ODRs received with a contemporary standardized behavior rating scale (the BASC-2 Teacher Report Form), and b) the validity of common cut points to determine level of support needed (i.e., 0 to 1, 2 to 5, and 6 or more ODRs). Results indicated strong correlations between ODRs and rating of externalizing behavior and statistically and clinically significant differences in behavior ratings based on existing ODR cut points, but no significant relation between ODRs and ratings of internalizing problems. Results are discussed in terms of recommended use of ODRs as a screening measure to indicate level of behavior support required.

Concurrent Validity of Office Discipline Referrals and Cut Points Used in School-Wide Positive Behavior Support

Youth violence has risen dramatically in the past few years (Federal Bureau of Investigation, 2006; Statistics Canada, 2007), and problem behavior remains a top concern of school personnel (Markow, Moessner, & Horowitz, 2006). Clearly, valid and reliable measurement of behavior is a necessity for school teams seeking to assess the adequacy of school-wide support, identify students for additional intervention, and monitor the effectiveness of those interventions (McIntosh, Reinke, & Herman, in press). There are a variety of approaches available to measure levels of problem behavior in schools, and determining what data to collect is an important prerequisite for effective data-based decision making (Newton, Horner, Algozzine, Todd, & Algozzine, 2009).

Measurement of behavior typically falls into two categories, direct and indirect. Direct observation (given acceptable inter-observer agreement) is often viewed as the gold standard of assessment by behavioral researchers, due to its low level of inference (Cone, 1997; Johnston & Pennypacker, 1993; Shriver, Anderson, & Proctor, 2001; Volpe & McConaughy, 2005). However, direct observation is often viewed by school personnel as too time-consuming for regular use, particularly if the purpose is to identify the level of support needed for an entire student body (Briesch & Volpe, 2007). Observations are a necessary component of a comprehensive evaluation process, but usually only once students have been referred for additional support beyond the school-wide level.

In place of direct observation, school personnel often rely on a variety of indirect measures of behavior. A popular indirect measure of problem behavior is the standardized behavior rating scale. School personnel prefer these measures in general because of their

efficiency and documented technical adequacy (Merrell, 2007). However, these rating scales also have several weaknesses, including the teacher time needed to complete individual rating scales for multiple students (H. M. Walker & Severson, 1994), and lack of treatment utility to inform intervention (McIntosh et al., in press).

In an effort to reduce the resource demands of assessment, school personnel often identify existing information already collected by schools for use. Information pertaining to behavior includes rates of special education referral and eligibility for behavior disorders, rates of suspension, and reports of behavioral incidents. The most commonly used type of extant data to assess behavior is the office discipline referral, or ODR. ODRs are forms that document serious behavioral incidents in a systematic manner (Sugai, Sprague, Horner, & Walker, 2000). School personnel issue ODRs when they observe events of serious problem behavior (such as defiance, fighting, harassment, or possession of weapons or banned substances) that necessitate administrative action. Students are often sent to the office when they receive an ODR, though this may not always be the case (e.g., a teacher handles the behavior in the classroom but is required to report serious events to the office for documentation and review). Most schools have some form of incident reporting tool, but ODRs represent a systematic process with the following features: a) a common form that details important information about the incident (e.g., location, time of day, others involved), b) clear definitions of what behaviors warrant a referral, c) clear definitions of what behaviors are expected to be handled without a referral, d) regular training on use and discrimination between reportable and non-reportable behaviors, and e) a system for compiling and analyzing ODR data. Nearly 5000 schools use ODRs in this manner through the web-based data entry and analysis application *School-Wide Information System* (SWIS; May et al., 2008).

Use of Office Discipline Referrals for Decision Making

Informal measures such as ODRs and suspensions are being increasingly used as indicators of problem behavior. This increase in use is based on three primary factors: (a) ready availability, (b) difficulty of direct observation for many cases of problem behavior, and (c) utility for making a wide range of decisions at the school and individual level.

One of the primary advantages to using ODR data for decision-making at individual student and school-wide levels is that many schools already collect these data (Kern & Manz, 2004; Rusby, Taylor, & Foster, 2007; Tobin, Sugai, & Colvin, 2000; Wright & Dusek, 1998). Many types of discipline data reports are required by districts, provinces and states, and federal governments (U.S. Department of Education, 2002). As these data are routinely collected by most schools, they potentially represent an efficient and time-saving method for assessing, evaluating, and planning behavior support, particularly in times of scarce resources (Irvin et al., 2006; Irvin, Tobin, Sprague, Sugai, & Vincent, 2004; Sugai et al., 2000; U.S. Department of Education, 2002). In a survey of ODR data users in 32 schools, Irvin and colleagues (2006) found that school staff using *SWIS* to enter and analyze ODR data reported an increase in efficiency of decision making regarding student behavior in schools.

Another advantage of using ODRs is the capability to sample behavior that is difficult to observe directly. Some problem behaviours, such as low-frequency, high-intensity problem behavior (Sprague & Horner, 1999) and relational aggression (Merrell, Buchanan, & Tran, 2006), are challenging to measure using direct observation procedures, primarily because of the voluminous time required to generate accurate rates of behavior. Traditional observation systems suffer decreased accuracy when problem behavior occurs at low rates or is difficult to observe (Hintze & Matthews, 2004), making it difficult to document both baseline rates of behavior and

response to intervention. Using ODRs and suspension data can allow for a method of gathering information on low-frequency, high-intensity behaviors that is more realistic in school settings than conducting direct observation and waiting for these behaviors to occur.

Finally, ODR data can be used to answer a broad range of important questions for school behavior support teams (Irvin et al., 2006; Irvin et al., 2004; Rusby et al., 2007; Sugai et al., 2000; B. Walker, Cheney, Stage, & Blum, 2005). At the school-wide level, ODR data are used to indicate the behavioral climate of schools, identify and track school-wide patterns of problem behavior, help target and evaluate reform efforts, and monitor compliance with school mission and safety goals (Holcomb, 1998; Irvin et al., 2004; Stephens, 2000; Sugai et al., 2000; Tobin et al., 2000). These uses allow schools to prevent problem behavior in specific areas (Kartub, Taylor-Greene, March, & Horner, 2000; Putnam, Handler, Ramirez-Platt, & Luiselli, 2003) and identify students who require targeted and individualized support as soon as problems emerge (McIntosh et al., in press). At the individual student level, discipline data are used to monitor and analyze student problem behavior (Tobin & Sugai, 1999), evaluate intervention effectiveness (March & Horner, 2002), and improve the efficiency and effectiveness with which schools provide and assess individualized support (Lewis-Palmer, Bounds, & Sugai, 2004).

Evidence Supporting the Use of Office Discipline Referrals

Beyond the ready availability and numerous uses of ODR data, there is a growing body of evidence supporting their validity and utility for decision making regarding problem behavior (Irvin et al., 2004; Irvin et al., 2006; Rusby et al., 2007; Tobin & Sugai, 1999). In an archival review of 526 randomly selected students over a 6-year period, Tobin and Sugai found that ODRs received in sixth grade served as a significant predictor of chronic discipline problems, violent behavior, and school failure. Rusby and colleagues found that ODRs in kindergarten

were more effective than family income in predicting problem behavior in first grade and ODRs in first grade predicted parent- and teacher-reported problem behavior at the end of the school year. H. M. Walker and colleagues (1990) found that students labeled as antisocial had statistically significantly higher levels of ODRs than students at risk for such classifications. Further, higher levels of ODRs at the school-wide level have been associated with more problematic social climates in schools (Irvin et al., 2004). Finally, the number of ODRs has been shown to be significantly related to academic underachievement, particularly in reading (McIntosh, Flannery, Sugai, Braun, & Cochrane, 2008; McIntosh, Horner, Chard, Boland, & Good, 2006; McIntosh, Sadler, & Brown, 2009; Nelson, Gonzalez, Epstein, & Benner, 2003).

Irvin and colleagues (2004) reviewed the existing research base and presented evidence regarding the psychometric properties of ODRs. When the process and problem behaviors for referral are clearly identified, ODRs have been shown to possess sufficient construct validity as a behavioral measure (Irvin et al., 2004). Irvin and colleagues also presented preliminary evidence for concurrent validity of ODRs through moderate to strong correlations with more established measures of problem behavior, such as teacher ratings of student problem behavior and self-report of delinquent behavior. In addition, the researchers noted strong stability of scores over time, equivalent to or stronger than some standardized behavior rating scales. More recent research also demonstrates the stability of ODRs across elementary, middle, and high schools (McIntosh et al., 2008; McIntosh, Horner et al., 2006)

Concerns regarding the Validity and Reliability of Office Discipline Referrals

Despite the preliminary evidence supporting the technical adequacy of ODRs, there are significant validity concerns that must be considered when they are used. First, the evidence regarding psychometric properties is limited to a small number of studies. Second, there are

threats to validity that are inherent in the use of ODRs in general. Though these are not specific to ODRs alone (i.e., these threats exist to some extent with all indirect measurement of behavior), they are particularly applicable to ODRs (Kern & Manz, 2004; Morrison, Peterson, O'Farrell, & Redding, 2004; Rusby et al., 2007; Sugai et al., 2000).

ODRs as a behavioral chain. ODRs do not simply occur as a result of student problem behavior, but rather at the end of a chain of behaviors (Sugai et al., 2000). When a student engages in a behavior that warrants an ODR, an adult must observe the behavior, then determine if an ODR should be issued, and file the ODR form into the system for analysis. At the point of observation, the level of supervision may contribute to the rate of referral. For example, a school with either more adults supervising in key areas or more adults using active supervision techniques (Colvin, Sugai, Good, & Lee, 1997) may report a higher rate of referrals, even though the actual rate of problem behavior may be lower than in schools with less effective supervision. The point of determination is also a critical time for accuracy of ODRs. Adults in the school may use varying criteria for determining whether the behavior observed deserves an ODR or a simple correction (Kern & Manz, 2004; Nelson et al., 2003). Finally, the point of ODR entry (usually into an electronic database) can be a source of error. Schools without consistent ODR submission or entry policies may not enter all ODRs, resulting in an underestimated rate of ODRs for individuals or the school as a whole (Rusby et al., 2007).

Bias. The provision of ODRs may also be influenced by intentional or unintentional bias on the part of adults in the school. Teachers may provide higher rates of ODRs if a large number of referrals is a path to additional classroom or individual student support. On the other hand, teachers may be reluctant to issue ODRs if a school administrator perceives the use of ODRs as an indicator of poor teaching. This phenomenon of underreporting may occur at a systems level

if there is pressure from administrators to rely on level of ODRs as the sole determination of success of school-wide efforts. If lower ODRs are emphasized over accurate data, school staff may not issue ODRs in efforts to give the appearance of improvement (Kern & Manz, 2004).

Another clear and concerning source of bias is the disproportionate use of ODRs with students from culturally diverse backgrounds. Though issues of cultural bias have traditionally focused on disproportional special education eligibility (Oswald, Coutinho, Best, & Singh, 1999; Skiba et al., 2008), there is evidence of differential rates of ODRs by ethnicity as well (see Nelson et al., 2003; Skiba et al., 2008). Students from particular cultural minority backgrounds (e.g., students of African-American or Native American heritage) often receive higher rates of ODRs, particularly for more subjective and culturally defined types of behavior (e.g., defiance, disrespect), even when SES is included in analyses (Krezmien, Leone, & Achilles, 2006; Shaw & Braden, 1990; Skiba, Michael, Nardo, & Peterson, 2002). School teams can assess the level of potential cultural bias by calculating risk ratios for ODRs by ethnicity (Skiba et al., 2008), and some ODR computer applications, such as *SWIS*, can automatically generate reports to assess disproportionality in referral rates.

Many of the abovementioned concerns can potentially be mitigated through professional development focusing on strategies to improve the accuracy of ODRs. These strategies focus on standardizing the use of ODRs through clear and consistent definitions of incidents, regularly scheduled data summaries, accuracy checks, and training in cultural responsiveness (Irvin et al., 2004; U.S. Department of Education, 2002; Weinstein, Tomlinson-Clarke, & Curran, 2004). In addition, as staff become more skilled at implementing school-wide interventions, ODR data can become more accurate and useful (Irvin et al., 2006). Yet though training and experience may

increase the validity and reliability of ODRs for assessing behavior, it should not be assumed that these sources of variance can be eliminated completely.

Given that ODR and suspension measures are used for a variety of data collection and decision-making purposes for individual students and school-wide interventions, it is critical to examine the validity of these measures. Establishing concurrent validity with a contemporary standardized behavior rating scale with strong psychometric properties would provide evidence for the validity of inferences made from ODRs and suspensions (Hintze, 2005; Kern & Manz, 2004; Rusby et al., 2007). Nelson and colleagues (2002) completed a recent study assessing the relation between non-standardized referrals and the teacher report form of the Child Behavior Checklist (Achenbach & Rescorla, 2001). They concluded that there was limited evidence supporting their use. However, the type of referral Nelson and colleagues tested was a less systematic referral that does not meet the definition of an ODR (as described above) for the following reasons: a) the form was open-ended (no specific categories of behavior were included), b) there were no formal criteria for what behaviors did and did not warrant a referral, and c) no process of regular training to monitor the fidelity of referral use was in place. Results indicated that an unsystematic referral process resulted in a poor measure of behavior. Thus, the concurrent validity of ODRs used in a systematic manner is still uncertain.

Use of ODR Cut Points to Determine Level of Support Needed

One area of ODR use that has been understudied is the method of using ODRs to determine the level of support required by students. Gordon (1983) proposed a three-tiered model of disease prevention that provides support and care for all members of the population, matching their level of need with the level of support they receive. This model has been adopted to support school-wide discipline (H. M. Walker et al., 1996), and researchers and school teams

alike are currently using the number of ODRs received to define the level of support required based on the following cut points (from Sugai et al., 2000): students receiving 0 to 1 ODRs per year are deemed to be supported adequately by Tier I support (universal, school-wide interventions focused on developing a consistent environment and teaching appropriate behavior to all students); students receiving 2 to 5 ODRs are indicated to receive Tier II support (selected interventions that provide some additional behavior support but are efficient to implement); and students receiving 6 or more ODRs are indicated to receive Tier III support (intensive interventions that are individualized to meet student needs).

These cut scores were initially determined by logic and theoretical estimates, and the proportions of ODR distributions in a large sample of schools closely replicate the theoretical public health percentages of 80% in Tier I, 15 to 20% in Tier II, and 1 to 5% in Tier III (Horner, Sugai, Todd, & Lewis-Palmer, 2005). However, very little research has examined whether students have different levels of problem behavior based on these cut scores. In one research study to date, B. Walker and colleagues (2005) compared students with 1 to 2 ODRs to students with 2 or more ODRs, using the *Social Skills Rating System* (Gresham & Elliott, 1990). Results showed no statistically significant differences in ratings of social skills, but significant differences in ratings of problem behavior. Though there were some drawbacks (the study did not compare students at all levels of the three-tier model, and the rating scale used had norms over ten years old), the results provide initial evidence of statistically significant differences based on the existing cut points.

The Present Study

As informal measures of problem behavior are being used more often in schools, it is critical to know relations among these measures to be able to use them with more confidence.

School personnel are using ODRs in schools due to their ready accessibility and ease of use, yet little is known to validate these uses. Clearly, more information is needed regarding the use of ODRs as a measure of individual problem behavior. In particular, evidence is needed to document the validity of ODRs to measure problem behavior and identify the level of support required by students.

The authors explored these questions by assessing the concurrent validity of ODRs with a recently normed standardized behavior rating scale to help identify whether these measures should be used as they are being used in schools today. The study examined overall relations between informal and standardized, norm referenced measures, then addressed the validity of ODR cut scores to divide students into groups with significantly different levels of problem behavior. The specific research questions addressed in this study were as follows:

1. What are the correlations among ODRs, out-of-school suspensions, and scores from the BASC-2 teacher report form?
2. Do students have different BASC-2 ratings and numbers of out-of-school suspensions based on existing ODR cut points used in school-wide positive behavior support?

Method

Setting and Participants

The study was conducted in all five elementary and one of two K-8 schools in a public school district in the Pacific Northwest. Total district K-12 enrollment was 5,410 students. The district's ethnic composition was 3% African American, 3% Asian American/Pacific Islander, 78% European American, 14% Latino/a, and 3% Native American/Native Alaskan. Five of the six schools qualified for Title I services, and 53% of students were eligible for free or reduced

lunch. In addition, each school in the district had a durable, sustained system of School-wide Positive Behavior Support (SWPBS; Horner et al., 2005) for over a decade, which has been merged with a school-wide reading improvement model (Simmons et al., 2002) to provide combined support in academics and behavior (McIntosh, Chard, Boland, & Horner, 2006). To assess implementation of SWPBS, external evaluators administered the *School-wide Evaluation Tool* (Sugai, Lewis-Palmer, Todd, & Horner, 2001), a research-validated measure with evidence of sufficient validity and reliability to assess fidelity of SWPBS implementation (Horner et al., 2004). Each school scored over the 80% mean implementation criterion, indicating adequate implementation of SWPBS.

Participants in the study (n = 40) were identified through the district's typical pre-referral process. All students were referred by their elementary classroom teachers to the pre-referral team for additional behavior support. When students were identified as needing additional support, school personnel contacted the students' caregivers and obtained consent for them to participate in the study. All caregivers consented to having their children participate in the study. These students were enrolled in grades 1 to 5 (mean grade = 2.78) and ranged in age from 6 to 11. There were 34 male students (85%). The percentage (and number) of students receiving special education services was 35% (14 of 40 students), in the areas of autism spectrum (1), communication (7), other health impairment (1), and specific learning disability (5). The ethnic backgrounds of the participants were in the following proportions: 3% African American, 88% European American, 8% Latino/a, and 3% Native American/Native Alaskan/First Nations.

Measures

Behavior Assessment Scale for Children – Second Edition Teacher Report Scale – Child Form (BASC-2). The BASC-2 (Reynolds & Kamphaus, 2004) is a standardized, norm referenced

behavior rating scale used to assess levels of behavior in relation to a broad normative sample. This form produces a broad number of scores, and for this study, three composite scales were used: the Externalizing Composite (used to measure disruptive, aggressive, and rule-violating behavior), the Internalizing Composite (used to measure anxiety, depression, and withdrawal), and the Adaptive Composite (used to measure social skills, leadership, and communication). Each scale is reported as a T-score, with a mean of 50 and a standard deviation of 10. Higher scores on the Externalizing and Internalizing composites indicate higher levels of problem or maladaptive behavior, and higher scores on the adaptive scale indicate higher levels of prosocial, desired behavior. The BASC-2 was selected for use in this study because of its strong psychometric properties and large, representative normative group. The *BASC-2* test manual reports high correlations with other well-regarded behavior rating scales and strong technical adequacy figures for the composite scales used in this study, indicating that these scores are suitable for important individual decision-making (Salvia & Ysseldyke, 2001).

Office Discipline Referrals (ODRs). The district in the present study uses ODRs, as described in the previous section, as standardized documentation of events of serious problem behavior. The district has identified a common ODR form and list of problem behaviors that warrant ODRs as separate from minor problem behaviors, which can be handled by the staff member witnessing the behavior. To mitigate threats to inter-rater reliability, the district conducts regular trainings on discriminating between behaviors that do and do not warrant a referral, based on existing definitions from *SWIS* documentation. All schools in the participating district use the *School-Wide Information System* (SWIS, May et al., 2002), a web-based ODR data entry and reporting system. Results of the *School-wide Evaluation Tool* indicated that each school possessed a standardized ODR form, there was agreement among randomly selected staff

and administrators on what behaviors warranted ODRs, and ODR data were regularly used for decision-making.

The authors used the total number of ODRs issued to each student during the school year as a variable in the analyses. The number of ODRs in this study ranged from 0 to 13, with a mean of 2.28. In addition, the location of the ODR provides whether the incidents occurred in the classroom or another area of the school. In this sample, the majority of ODRs (63%) were administered for incidents in non-classroom locations, such as playgrounds, hallways, and other common areas, indicating that the majority of ODRs were administered by staff other than the students' classroom teachers.

Out of school suspensions. Suspension is a common procedure used as an attempt to punish students by excluding them from school, and it can be a marker of intense, dangerous behavior in schools. Though their use is often detrimental and can occasion future problem behavior (Hemphill, Toumbourou, Herrenkohl, McMorris, & Catalano, 2006), suspensions remain a frequently used procedure to address serious problem behavior. The authors used the total number of out of school suspensions issued during the school year. The number of suspensions in this study ranged from 0 to 7, with a mean of .80.

Procedure

Once caregivers consented for their children to participate in the study, their regular classroom teachers were asked to participate as well; all teachers elected to participate. The teachers completed the BASC-2 teacher report scale before any additional support was provided. At the conclusion of the school year, district administrators provided the authors with the students' total number of ODRs and suspensions for the entire year. The authors then merged and analyzed the data using *SPSS 13.0 for Windows*.

Design

Relations among measures. To examine overall relations among the variables, the authors ran bivariate correlations for all pairs of the following variables: total ODRs, total out of school suspensions, and BASC-2 Externalizing, Internalizing, and Adaptive Composite Scales. Given the strong skews of the ODR and suspension score distributions and mild skews of the BASC-2 score distributions, rank order correlations were also computed. The analysis logic used was to run both sets of analyses and compare results. If all variables were related in the same direction and at the same level of significance, the bivariate correlations would be reported, as Pearson's *r* is more commonly interpreted. These analyses had an *n* of 40 and an α level of .05.

ODR cut points. The authors examined the utility of commonly used ODR cut points by analyzing differences among groups of students divided by 0 to 1, 2 to 5, and 6 or more ODRs received per year. The authors used multiple univariate analyses of variance, with ODR level as the independent variable and suspensions and BASC-2 Externalizing, Internalizing, and Adaptive Composite Scales as dependent variables. Due to the skewness of the score distributions, the authors used rank transformations of the dependent variables to allow for accurate analyses. To control for family-wise error, the authors used a Bonferroni correction. These analyses had an *n* of 40 and a Bonferroni-corrected α level of .0125. To identify the magnitude of difference between the groups, effect sizes (in terms of Cohen's *d*; Cohen, 1988) were computed, comparing students with 2 to 5 and 6 or more ODRs to students with 0 to 1 ODRs.

Results

For the correlation analyses, both bivariate and rank order correlations were computed and compared. All individual correlations were in the same direction and of the same

significance level. The bivariate correlations between the pairs of variables are presented in Table 1. As seen, the strongest correlation was between ODRs and suspensions (r = .76). Correlations between the BASC-2 Externalizing Composite and both ODRs and suspensions were identically strong and statistically significant (r = .51). These variables shared a considerable amount of variance, indicating a strong relation among these three variables. Correlations with the BASC-2 Adaptive Composite were in the anticipated direction (negative correlations), but neither reached the level of statistical significance. All correlations with the BASC-2 Internalizing Composite were weak, even with the Externalizing Composite, suggesting that this composite measures a different, unrelated set of behaviors than the Externalizing Composite, ODRs, and suspensions.

Descriptive statistics for ODR cut point groups are provided in Table 2, and analysis results are presented in Table 3. There was a statistically significant relation between ODR cut point and suspensions, $F(2, 37) = 12.23$, $p < .001$, as well as the BASC-2 Externalizing Composite, $F(2, 37) = 6.55$, $p < .01$. There were not statistically significant relations between cut points and the BASC-2 Internalizing Composite, $F(2, 37) = .01$, $p = .99$, or the BASC-2 Adaptive Composite, $F(2, 37) = 2.15$, $p = .13$. These results showed significant differences in suspensions and BASC-2 Externalizing scores based on the existing three-tier ODR cut points. Effect sizes showed large magnitudes of difference in comparison to students with 0 to 1 ODRs. For suspensions, the effect sizes for students with 2 to 5 ODRs (d = 1.14) and 6 or more ODRs (d = 3.87) were well above the standard “large effect” criterion (Cohen, 1988). Likewise, effect sizes for the Externalizing scores were also above the same criterion (d = .87 and 2.08, respectively).

These differences can clearly be seen in graphs of the differences in mean suspensions, and BASC-2 Externalizing and Adaptive Composites (see Figures 1 and 2). In terms of BASC-2 Externalizing score classifications, the mean score for students with 0 to 1 ODRs was closest to the cutoff between "average" and "at-risk," the mean for students with 2 to 5 ODRs was just below the cutoff between "at-risk" and "clinically significant," and the mean for students with 6 or more ODRs was one standard deviation above the "clinically significant" cutoff, or three standard deviations above the mean. In terms of Adaptive classifications, the mean score for students with 0 to 1 ODRs was in the "average" range, and the means for students with 2 to 5 and 6 or more ODRs were in the "at-risk" range.

Discussion

This study assessed the concurrent validity of office discipline referrals, commonly-used informal behavior measures in schools, by comparing results to those of a standardized, norm-referenced rating scale (the BASC-2) and out-of-school suspensions. Bivariate correlations were used to examine shared variance and ANOVA analyses were used to determine whether existing ODR cut points divided students into significantly different groups. Results showed strong, statistically significant correlations between the BASC-2 Externalizing Composite and both ODRs and suspensions, providing evidence of adequate concurrent validity for ODRs and suspensions for this domain of behavior, but not for internalizing problems. ODR cut point analyses showed that students had significantly different BASC-2 Externalizing Composite scores and levels of suspensions based on their indicated tier in the three-tier model, with substantial effect sizes. These results provide further evidence that the existing ODR cut points divide students into groups with significantly different levels of externalizing problem behavior.

Externalizing behavior. The present study contributes to the current literature base supporting the validity of ODRs as a meaningful measure of externalizing problem behavior in school settings. Results were similar to those found in the Irvin et al. (2004) study, which examined the concurrent validity of ODRs with other measures of problem behavior in school settings. An important finding was the large magnitude of the relation between ODRs and the BASC-2 Externalizing Composite. According to Cohen (1988), this relation may be as strong as can be expected for a single event-based (rather than multiple item-based) measure, especially when considering the "real world" nature of ODRs. This finding is particularly important for schools, given current demands for accountability in the area of school safety and increasing responsibility for a wide range of initiatives.

It is important to note that the relation between ODRs and the BASC-2 was observed to be much stronger than results reported by Nelson and colleagues (2002). However, there are some important differences between the type of referral studied. Nelson and colleagues tested the validity of a quite different type of referral, one that more approximates an unstandardized "incident report" in schools. It is logical to conclude that the results from the Nelson and colleagues study (2002) may document the validity of an unsystematic use of referrals, and results from the present study may show the validity of standardized, systematic use of ODRs. In addition, the school studied by Nelson was only in its first year of implementing a systematic approach to school-wide behavior support. It is predictable that a non-systematic referral process would result in a measure with unacceptably low validity, and use of non-systematic referrals as data sources is not recommended.

Internalizing behavior. No significant relation was found between ODRs and the BASC-2 Internalizing Composite. Correlations were similar between the BASC-2 Externalizing and

Internalizing Composites. These results are consistent with the findings from Nelson and colleagues (2002), who found lower relations between referrals and the internalizing scale than externalizing scale, suggesting that, systematic or unsystematic, ODRs are not a valid measure of internalizing problems. Results indicate that the BASC-2 Internalizing Composite may measure distinctly different behaviors than both the BASC-2 Externalizing Composite and ODRs.

These results and logic indicate that students rarely receive referrals or suspensions for internalizing problems. One conclusion is that though ODRs may be used as an efficient indicator of externalizing behavior, using ODRs alone is likely to underestimate the levels of anxiety and depressive symptoms in schools. This finding is not surprising given that internalizing problems are difficult to observe and often go unnoticed by school personnel (Davis, 2005).

Adaptive behavior. No significant relation was found between ODRs and the BASC-2 Adaptive Composite. However, because this study utilized a referred sample, the range may have been artificially truncated, particularly for students with 0 to 1 ODRs. It is possible that teachers referred these students for additional support not because of problem behavior, but because of low adaptive skills. Results may have been different if the sample included all students in the district with 0 or 1 ODRs, not only students who were referred for additional behavior support. Even so, there were non-significant mean differences in the anticipated patterns (see Figure 2).

ODR cut points. With regard to the ODR cut points commonly used in SWPBS to divide students by need for support, results support the use of the existing ODR cut points (0 to 1, 2 to 5, and 6 or more) to indicate the level of support needed. When interpreting BASC-2 Externalizing Composite classification ranges, these scores show significantly different clinical levels of behavioral symptoms and different levels of behavioral functioning in general. As such,

these results provide evidence that—when using a systematic ODR process—the use of existing ODR cut points is a valid screening method for classifying students by need in the area of externalizing behavior.

The finding that no students who received 0 to 1 ODRs received any suspensions is not surprising given the nature of referrals. ODRs are typically used to record a behavior that resulted in some type of consequence. If no behavior was recorded, logically there would be no consequence (i.e., suspension). Therefore, one would expect that students with 0 to 1 ODRs would not have been suspended.

Limitations. The present study has several significant limitations; therefore results must be interpreted with some caution. The first major limitation regards measurement of problem behavior. The study solely examined relations among indirect measurements of problem behavior. Although indirect measures provide some insight into overall levels and topography of behavior, none of the measures used are as valid as direct measures (i.e., direct observations). As such, future research should assess the concurrent validity of ODRs and behavior rating scales with direct observations of student behavior.

The second major limitation regards the sample studied. The sample was extremely small (particularly for students with 6 or more ODRs) and may not adequately reflect the general school population across North America. In particular, the proportion of students from European American backgrounds (though reflective of the school district's demographics) was higher than is found in large, urban school districts. Further, the participants in the study were students referred for additional behavior support, and therefore did not include students who were judged by teachers to be adequately supported by school-wide program, as well as students already identified for support, including students identified with emotional and/or behavior disorders. It

is unclear whether similar results would be found in a sample that included all students, rather than only students who were referred for problem behavior. As such, further research is needed to replicate these results in a larger, more diverse student population.

The third major limitation regards the ODR system in the district studied. The school district had been implementing School-wide Positive Behavior Support and the same ODR system for over ten years, with regular instruction and feedback on the appropriate use of referrals. The school district had clear definitions of what does and does not constitute an ODR, a system for recording and summarizing data, and ongoing procedures to increase fidelity of use. School personnel experience using ODRs may have also contributed to the results obtained. Other school districts may not have the same level of expertise in using ODRs, and thus these results may only pertain to schools with clear and comprehensive ODR systems in place.

Conclusion and Recommendations

In summary, this study provides further evidence that when ODRs are defined and used systematically, they can be valid measures of the level of support needed for elementary students in the area of externalizing behavior. The existing ODR cut scores divide students into significantly different groups. However, there is evidence in the literature that when ODRs are not defined clearly or used systematically, ODRs are not valid measures and are prone to bias, particularly on the basis of ethnicity (Skiba et al., 2002). Furthermore, it appears that ODRs do not provide a valid measure of internalizing problems.

Given these results within the context of the research base, the following recommendations for the use of ODRs can be supported. First, school teams using ODRs to make decisions regarding individual student support will make more accurate decisions when they use a systematic process, including clear descriptions of behaviors that should result in a

referral (and those that should be handled by the observing adult) and ongoing training of school personnel to improve accuracy of use. Administrators and researchers should emphasize the accurate use of data, rather than reductions in ODRs, as goals for school teams. In addition, it is critical to assess whether students are disproportionally receiving ODRs on the basis of ethnicity. Second, other measures assessing internalizing behavior are needed to supplement the use of ODRs. Finally, best practice indicates the use of multi-method, multi-source assessment procedures. It is advised to use ODRs within a broader system of behavior assessment, including teacher referral, multiple-gating screening systems, and direct observation.

References

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

Briesch, A. M., & Volpe, R. J. (2007). Important considerations in the selection of progress-monitoring measures for classroom behaviors. *School Psychology Forum, 1*, 59-74.

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

Colvin, G., Sugai, G., Good, R. H., & Lee, Y. (1997). Effect of active supervision and precorrection on transition behaviors of elementary students. *School Psychology Quarterly, 12*, 344-363.

Cone, J. D. (1997). Issues in functional analysis in behavioral assessment. *Behavior Research and Therapy, 35*, 259-275.

Davis, C. A. (2005). *Effects of in-service training on teachers' knowledge and practices regarding identifying and making a focus of concern students exhibiting internalizing problems*. Unpublished doctoral dissertation, University of Oregon

Federal Bureau of Investigation (2006). *Preliminary annual uniform crime report, 2005*. Washington, DC: Author. Retrieved June 12, 2006, from http://www.fbi.gov/ucr/2005preliminary/index.htm.

Gordon, R. S. (1983). An operational classification of disease prevention. *Public Health Reports, 98*, 107-109.

Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system*. Circle Pines, MN: American Guidance Service.

Hemphill, S. A., Toumbourou, J. W., Herrenkohl, T. I., McMorris, B. J., & Catalano, R. F. (2006). The effect of school suspensions and arrests on subsequent adolescent antisocial behavior in Australia and the United States. *Journal of Adolescent Health, 39*, 736-744.

Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review, 34*, 507–519.

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.

Holcomb, E. L. (1998). *Getting excited about data: How to combine people, passion, and proof*. Thousand Oaks, CA: Corwin.

Horner, R. H., Sugai, G., Todd, A. W., & Lewis-Palmer, T. (2005). School-wide positive behavior support. In L. Bambara & L. Kern (Eds.), *Individualized supports for students with problem behaviors: Designing positive behavior plans* (pp. 359-390). New York: Guilford Press.

Horner, R. H., Todd, A. W., Lewis-Palmer, T., Irvin, L. K., Sugai, G., & Boland, J. B. (2004). The School-wide Evaluation Tool (SET): A research instrument for assessing school-wide positive behavior support. *Journal of Positive Behavior Interventions, 6*, 3-12.

Irvin, L. K., Horner, R. H., Ingram, K., Todd, A. W., Sugai, G., Sampson, N. K., et al. (2006). Using office discipline referral data for decision making about student behavior in elementary and middle schools: An empirical evaluation of validity. *Journal of Positive Behavior Interventions, 8*, 10-23.

Irvin, L. K., Tobin, T. J., Sprague, J. R., Sugai, G., & Vincent, C. G. (2004). Validity of office discipline referral measures as indices of school-wide behavioral status and effects of school-wide behavioral interventions. *Journal of Positive Behavior Interventions, 6*, 131-147.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of human behavioral research* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Kartub, D. T., Taylor-Greene, S., March, R. E., & Horner, R. H. (2000). Reducing hallway noise: A systems approach. *Journal of Positive Behavior Interventions, 2*, 179-182.

Kern, L., & Manz, P. (2004). A look at current validity issues of school-wide behavior support. *Behavioral Disorders, 30*, 47-59.

Krezmien, M. P., Leone, P. E., & Achilles, G. M. (2006). Suspension, race, and disability: Analysis of statewide practices and reporting. *Journal of Emotional and Behavioral Disorders, 14*, 217-226.

Lewis-Palmer, T., Bounds, M., & Sugai, G. (2004). Districtwide system for providing individual student support. *Assessment for Effective Intervention, 30*, 53-65.

March, R. E., & Horner, R. H. (2002). Feasibility and contributions of functional behavioral assessment in schools. *Journal of Emotional and Behavioral Disorders, 10*, 158-170.

Markow, D., Moessner, C., & Horowitz, H. (2006). *The MetLife survey of the American teacher, 2005-2006: Expectations and experiences*. New York: MetLife Insurance Company. Available at http://www.metlife.org.

May, S., Ard, W. I., Todd, A. W., Horner, R. H., Glasgow, A., Sugai, G., et al. (2002). School-Wide Information System. Educational and Community Supports, University of Oregon, Eugene, OR.

May, S., Ard, W. I., Todd, A. W., Horner, R. H., Glasgow, A., Sugai, G., et al. (2008). School-Wide Information System homepage Retrieved 2 July 2008, from http:/www.swis.org

McIntosh, K., Chard, D. J., Boland, J. B., & Horner, R. H. (2006). Demonstration of combined efforts in school-wide academic and behavioral systems and incidence of reading and behavior challenges in early elementary grades. *Journal of Positive Behavior Interventions, 8*, 146-154.

McIntosh, K., Flannery, K. B., Sugai, G., Braun, D., & Cochrane, K. L. (2008). Relationships between academics and problem behavior in the transition from middle school to high school. *Journal of Positive Behavior Interventions, 10*, 243-255.

McIntosh, K., Horner, R. H., Chard, D. J., Boland, J. B., & Good, R. H. (2006). The use of reading and behavior screening measures to predict non-response to School-Wide Positive Behavior Support: A longitudinal analysis. *School Psychology Review, 35*, 275-291.

McIntosh, K., Reinke, W. M., & Herman, K. E. (in press). School-wide analysis of data for social behavior problems: Assessing outcomes, selecting targets for intervention, and identifying need for support. In G. G. Peacock, R. A. Ervin, E. J. Daly & K. W. Merrell (Eds.), *The practical handbook of school psychology*. New York: Guilford.

McIntosh, K., Sadler, C., & Brown, J. A. (2009). Kindergarten reading skill and response to instruction as risk factors for problem behavior. *Manuscript submitted for publication*.

Merrell, K. W. (2007). *Behavioral, social, and emotional assessment of children and adolescents* (3rd ed.). Mahwah, NJ: Erlbaum.

Merrell, K. W., Buchanan, R. S., & Tran, O. K. (2006). Relational aggression in children and adolescents: A review with implications for school settings. *Psychology in the Schools, 43*, 345-360.

Morrison, G. M., Peterson, R., O'Farrell, S., & Redding, M. (2004). Using office referral records in school violence research: Possibilities and limitations. *Journal of School Violence, 3*, 39-61.

Nelson, J. R., Benner, G. J., Reid, R. C., Epstein, M. H., & Currin, D. (2002). The convergent validity of office discipline referrals with the CBCL-TRF. *Journal of Emotional and Behavioral Disorders, 10*, 181-188.

Nelson, J. R., Gonzalez, J. E., Epstein, M. H., & Benner, G. J. (2003). Administrative discipline contacts: A review of the literature. *Behavioral Disorders, 28*, 249-281.

Newton, J. S., Horner, R. H., Algozzine, R. F., Todd, A. W., & Algozzine, K. M. (2009). Using a problem-solving model to enhance data-based decision making in schools. In W. Sailor, G. Dunlap, G. Sugai & R. H. Horner (Eds.), *Handbook of positive behavior support* (pp. 551-580). New York: Springer.

Oswald, D. P., Coutinho, M. J., Best, A. M., & Singh, N. N. (1999). Ethnic representation in special education: The influence of school related economic and demographic variables. *Journal of Special Education, 32*, 194-206.

Putnam, R. F., Handler, M. W., Ramirez-Platt, C. M., & Luiselli, J. K. (2003). Improving student bus-riding behavior through a whole-school intervention. *Journal of Applied Behavior Analysis, 36*, 583–590.

Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavior Assessment Scale for Children* (2nd ed.). Circle Pines, MN: AGS Publishing.

Rusby, J. C., Taylor, T. K., & Foster, E. M. (2007). A descriptive study of school discipline referrals in first grade. *Psychology in the Schools, 44*, 333-350.

Salvia, J., & Ysseldyke, J. (2001). *Assessment in special and inclusive education*. Boston: Houghton Mifflin.

Shaw, S. R., & Braden, J. P. (1990). Race and gender bias in the administration of corporal punishment. *School Psychology Review, 19*, 378-383.

Shriver, M. D., Anderson, C. M., & Proctor, B. (2001). Evaluating the validity of functional behavior assessment. *School Psychology Review, 30*, 180-192.

Simmons, D. C., Kame'enui, E. J., Good, R. H., Harn, B. A., Cole, C., & Braun, D. (2002). Building, implementing, and sustaining a beginning reading model: Lessons learned school by school. In M. R. Shinn, H. M. Walker & G. Stoner (Eds.), *Interventions for academic and behavioral problems II: Preventive and remedial approaches* (pp. 403-432). Bethesda, MD: National Association of School Psychologists.

Skiba, R. J., Michael, R. S., Nardo, A. C., & Peterson, R. L. (2002). The color of discipline: Sources of racial and gender disproportionality in school punishment. *The Urban Review, 34*, 317-342.

Skiba, R. J., Simmons, A. B., Ritter, S., Gibb, A. C., Rausch, M. K., Cuadrado, J., et al. (2008). Achieving equity in special education: History, status, and current challenges. *Exceptional Children, 74*, 264-288.

Sprague, J. R., & Horner, R. H. (1999). Low-frequency high-intensity problem behavior: Toward an applied technology of functional assessment and intervention. In A. C. Repp & R. H. Horner (Eds.), *Functional analysis of problem behavior: From effective assessment to effective support* (pp. 98-116). Belmont, CA: Wadsworth Publishing.

Statistics Canada (2007). *Crime statistics in Canada, 2006*. Ottawa, ON: Author.

Stephens, R. D. (2000). Safe school planning. In D. S. Elliot, B. A. Hamburg & K. R. Williams (Eds.), *Violence in American schools* (pp. 253-291). New York: Cambridge University Press.

Sugai, G., Lewis-Palmer, T. L., Todd, A. W., & Horner, R. H. (2001). *School-wide Evaluation Tool (SET)*. Eugene, OR: Educational and Community Supports. Available at http://www.pbis.org.

Sugai, G., Sprague, J. R., Horner, R. H., & Walker, H. M. (2000). Preventing school violence: The use of office discipline referrals to assess and monitor school-wide discipline interventions. *Journal of Emotional and Behavioral Disorders, 8*, 94-101.

Tobin, T. J., Sugai, G., & Colvin, G. (2000). Using disciplinary referrals to make decisions. *NASSP Bulletin, 84*, 106-117.

Tobin, T. J., & Sugai, G. M. (1999). Using sixth-grade school records to predict school violence, chronic discipline problems, and high school outcomes. *Journal of Emotional and Behavioral Disorders, 7*, 40-53.

U.S. Department of Education (2002). *Safety in numbers: Collecting and using incident data to make a difference in schools*. Washington, DC: U.S. Government Printing Office.

Volpe, R. J., & McConaughy, S. H. (2005). Systematic direct observational assessment of student behavior: Its use and interpretation in multiple settings: An introduction to the miniseries. *School Psychology Review, 34*, 451-453.

Walker, B., Cheney, D., Stage, S. A., & Blum, C. (2005). Schoolwide screening and Positive Behavior Supports: Identifying and supporting students at risk for school failure. *Journal of Positive Behavior Interventions, 7*, 194-204.

Walker, H. M., Horner, R. H., Sugai, G., Bullis, M., Sprague, J. R., Bricker, D., et al. (1996). Integrated approaches to preventing antisocial behavior patterns among school-age children and youth. *Journal of Emotional and Behavioral Disorders, 4*, 194-209.

Walker, H. M., & Severson, H. (1994). Replication of the systematic screening for behavior disorders (SSBD) procedure for the identification of at-risk children. *Journal of Emotional and Behavioral Disorders, 2*, 66-78.

Walker, H. M., Steiber, S., & O'Neill, R. E. (1990). Middle school behavioral profiles of antisocial and at-risk control boys: Descriptive and predictive outcomes. *Exceptionality, 1*, 61-77.

Weinstein, C. S., Tomlinson-Clarke, S., & Curran, M. (2004). Toward a conception of culturally responsive classroom management. *Journal of Teacher Education, 55*, 25-38.

Wright, J. A., & Dusek, J. B. (1998). Compiling school base-rates for disruptive behavior from student disciplinary referral data. *School Psychology Review, 27*, 138-147.

Table 1

Intercorrelations among Behavior Variables and BASC-2 Composite Scales

	ODRs	SUSPENSIONS	BASC-EXT.	BASC-INT.	BASC-ADAPT.
ODRs	--	.76**	.51**	-.05	-.22
SUSPENSIONS		--	.51**	.17	-.10
BASC-EXT.			--	.10	-.31
BASC-INT.				--	-.10
BASC-ADAPT.					--

Note. EXT. = Externalizing Composite, INT. = Internalizing Composite, ADAPT. = Adaptive Composite.

*$p < .05$. **$p < .01$.

Table 2

Means (and Standard Deviations) of Variables by ODR Cut Point Group

GROUP	N	ODRs	SUSPENSIONS	BASC-EXT.	BASC-INT.	BASC-ADAPT.
0 to 1 ODRs	20	.10 (.31)	0 (0)	62.05 (9.20)	56.20 (11.01)	41.45 (6.13)
2 to 5 ODRs	17	3.47 (1.33)	1.12 (1.50)	69.94 (9.42)	55.76 (10.79)	38.71 (4.71)
6 or more ODRs	3	10.00 (2.65)	4.33 (3.79)	80.00 (7.00)	55.33 (9.87)	35.67 (2.89)

Note. EXT. = Externalizing Composite, INT. = Internalizing Composite, ADAPT. = Adaptive Composite.

Table 3

Univariate Analysis of Variance Summary Table for ODR Cut Point and Rank Ordered Dependent Variables

Dependent Variable	*df*	*F*	*p*
Suspensions	2	12.23***	< .001
Error	37	(53.58)	
BASC-2 Externalizing	2	6.55**	.004
Error	37	(106.2)	
BASC-2 Internalizing	2	.01	.99
Error	37	(143.53)	
BASC-2 Adaptive	2	2.15	.13
Error	37	(128.42)	

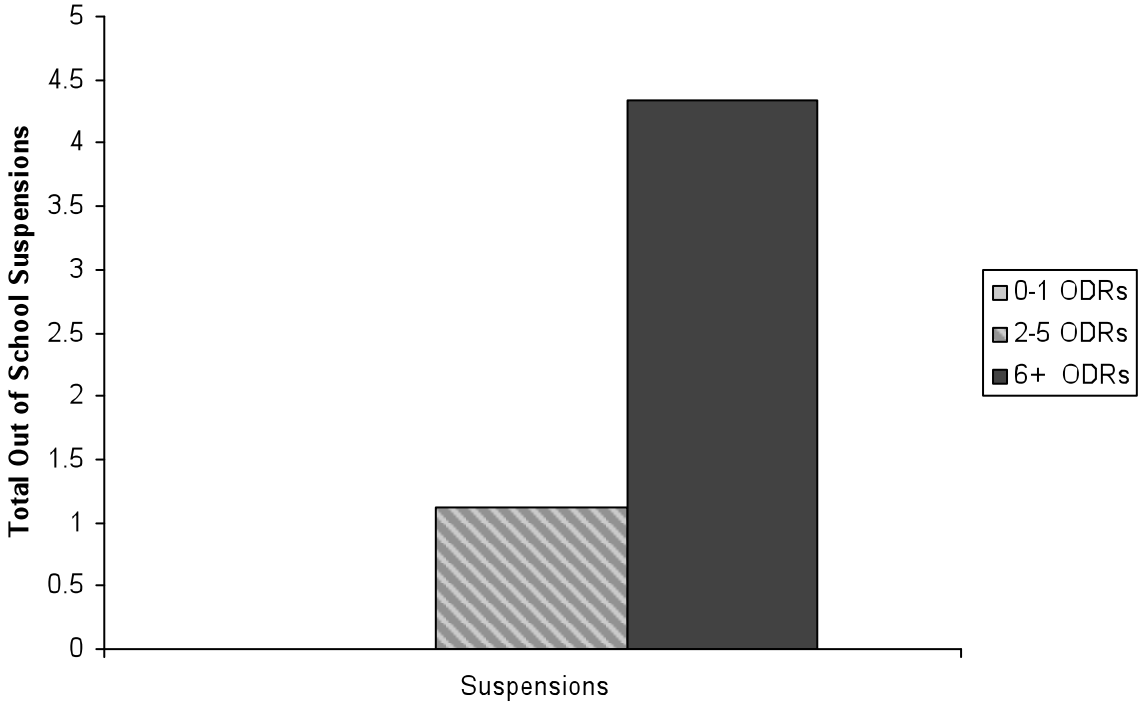
Note. Values enclosed in parentheses represent mean square errors.

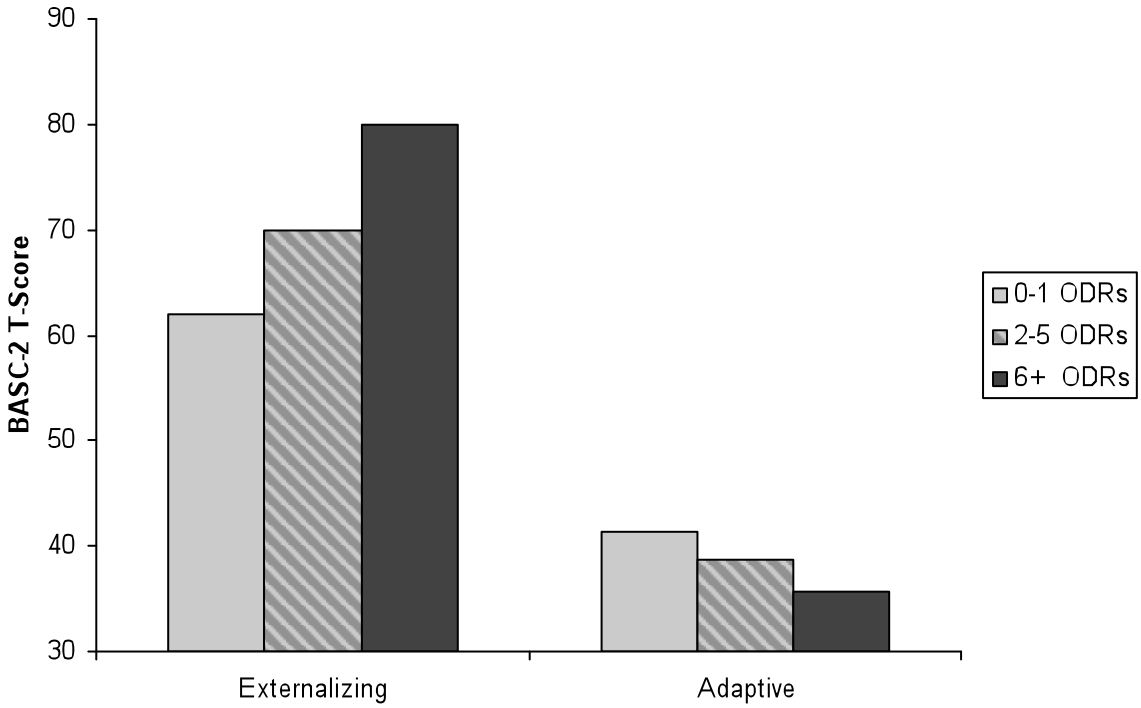
*$p < .0125$. **$p < .01$. ***$p < .001$.

Figure Captions

Figure 1: *Differences in mean out of school suspensions by ODR cut scores.*

Figure 2: *Differences in mean BASC-2 T-Scores by ODR cut scores.*


Total Out of School Suspensions
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
Suspensions
0-1 ODRs
2-5 ODRs
6+ ODRs


90
80
70
60
50
40
30
BASC-2 T-Score
Externalizing
Adaptive
0-1 ODRs
2-5 ODRs
6+ ODRs