

Establishing Research-Based Trajectories of Office Discipline Referrals for Individual Students

Kent McIntosh
University of British Columbia

Jennifer L. Frank
University of Oregon

Scott A. Spaulding
University of Washington

Abstract. The purpose of this study was to examine the technical (psychometric) adequacy of office discipline referrals (ODRs) as a behavioral assessment tool for individual student, data-based decision making within a problem-solving model. Participants were 990,908 students in 2,509 elementary schools. Students were grouped into ODR cut points of 0–1, 2–5, and 6 or more total ODRs. Descriptive analyses indicated consistent mean ODR growth trajectories for each group. Logistic regression analyses showed that intermediate cut points (1 ODR, 2 or more ODRs) were moderately accurate in predicting total ODRs received, and accuracy increased throughout the fall. Specific problem behaviors were less predictive of total ODRs, although adding specific behaviors to intermediate cut points enhanced prediction to some extent. Results are discussed in terms of using ODRs to identify students in need of intervention and monitor progress.

The adoption of systematic, evidence-based practices for identifying students at risk or in need of support has allowed schools to implement effective academic and behavioral strategies (e.g., National Center on Response to Intervention, www.rti4success.org; School-wide Positive Behavior Support, www.pbis.org). Such practices have helped schools address increases in student violence and disruption

that preoccupy educators with management of discipline rather than academic curricula and prevent meaningful student engagement. However, the success of these initiatives depends on the ability of schools to screen and track student behavior performance efficiently. This data-based problem-solving approach—collecting, evaluating, and using data to promote appropriate behavior—is practiced by school-

This research was supported by a discretionary grant from the BC Ministry of Education and U.S. Department of Education IES Grant No. R305B060014. Opinions expressed herein do not necessarily reflect the policy of either organization, and no official endorsements should be inferred.

Correspondence regarding this article should be addressed to Kent McIntosh, The University of British Columbia, 2125 Main Mall, Vancouver, BC V6T 1Z4, Canada; E-mail: kent.mcintosh@ubc.ca

Copyright 2010 by the National Association of School Psychologists, ISSN 0279-6015, which has nonexclusive ownership in accordance with Division G, Title II, Section 518 of P.L. Law 110-161 and NIH Public Access Policy

based teams that assess student behavior across several levels: whole school, groups of students who exhibit similar behaviors or behaviors in similar contexts and locations, and individual students. These school teams are most effective at decision making when the core academic and social competence outcomes targeted by their schools are clearly defined and measured (Newton, Horner, Algozzine, Todd, & Algozzine, 2009).

Measurement of student achievement and social behavior is essential, and reduction in problem behavior (or improvement in discipline) requires a system that documents and tracks specific behaviors. Some schools and districts may record and track their events of problem behavior using narratives that describe specific incidents (e.g., Kaufman et al., 2010), but such data may lack consistency across teachers and students within a school and can be cumbersome for generating indices and reports for decision making. In contrast, office discipline referrals (ODRs), standardized records of events of problem behavior that occur in schools (Sugai, Sprague, Horner, & Walker, 2000), have the potential to provide school personnel with a systematic and observable index of student problem behavior that can be measured, compiled, and analyzed reliably across different contexts, students, and behaviors.

For the purpose of this article, ODRs are contrasted with unstandardized incident reports, which are not coded or used systematically and have questionable validity (Nelson, Benner, Reid, Epstein, & Currin, 2002). Unlike the unstructured narrative of an incident report, well-designed ODR forms contain categories with predefined choices, such as location (e.g., playground, cafeteria, hallway) and problem behavior (e.g., physical aggression, disrespect) so that a teacher can quickly and consistently complete a referral using a series of checkboxes. With these critical features, ODRs provide increased consistency and efficiency for summaries and interpretation (Wright & Dusek, 1998).

At their lowest level of inference, ODRs measure the rate of specific problem behaviors (e.g., fighting, disrespect) for both individual

students and schools. ODRs are also considered indicators of student behavior problems, particularly externalizing problem behavior (McIntosh, Campbell, Carter, & Zumbo, 2009). In addition, they have been shown to be associated with broader social constructs (e.g., student and teacher perceptions of school climate, school engagement, classroom orderliness, effectiveness of school-wide intervention) and predictive of negative student outcomes (e.g., behavior disorders, delinquency, dropout, use of illegal substances, academic failure, family conflict; see Irvin, Tobin, Sprague, Sugai, & Vincent, 2004).

Although ODRs should not be the sole index of the social culture of a school, the ease with which they can be collected has contributed to their prominence as a behavioral outcome measure in schools (Wright & Dusek, 1998). Whereas a complete behavioral assessment might include the multiple methods of direct observation, teacher interview, and rating scales or checklists of student behavior, the effort involved in such a thorough assessment makes regular use for tracking behavior at the school level unwieldy. In contrast, ODRs are an efficient naturally occurring data source (i.e., already collected, easily available, cost-effective), making their use much more likely than data that require unique or additional collection procedures (McIntosh, Reinke, & Herman, 2009).

Uses of ODRs in a Problem-Solving Model

When used as a school-wide measure, ODRs have many potential uses within a problem-solving model (McIntosh, Reinke et al., 2009). ODRs can be used in problem identification to identify whether base rates of problem behavior indicate the need to change practices (Wright & Dusek, 1998). During problem analysis, ODRs can serve as a tool to identify patterns of concern in school locations (e.g., fighting on the playground, running in the hallways), times of day (e.g., before school, lunchtime), or with subgroups of students (e.g., Grade 4 boys in music class; Tobin, Sugai, & Colvin, 2000). ODRs may also

be used during the plan evaluation phase to determine the effectiveness of Tier 1 (universal) interventions (i.e., supporting a large proportion of students; Irvin et al., 2004).

Many schools use online database systems to collect and report ODR data, such as the School-wide Information System (SWIS; May et al., 2008). SWIS is a web-based application that allows school staff to enter, analyze, and monitor ODR data. Prior to adopting SWIS, schools must establish specific readiness criteria that ensure a clear procedure for identifying student problem behavior, managing ODR entry, and providing training and support for staff (Todd, Horner, & Dickey, 2009). In addition, schools adopt an office referral form with specific aspects of operationally defined problem behaviors (e.g., location, time of day, perceived function, administrative decision) that are compatible with SWIS data entry. Certified facilitators provide all school staff with ongoing training and district support around the collection, input, and use of ODR data within SWIS.

As part of SWIS implementation, facilitators provide school staff with training to clarify specific referral form definitions that are recorded as part of every ODR, including definitions for 24 distinct major problem behaviors (Todd, Horner, & Tobin, 2010). Examples of these behaviors include the following: “Physical Aggression/Fighting” (student engages in actions involving serious physical contact where injury may occur [e.g., hitting, punching, hitting with an object, kicking, hair pulling, scratching]); “Gang Affiliation Display” (student uses gesture, dress, and/or speech to display affiliation with a gang); and “Disruption” (student engages in behavior causing an interruption in a class or activity. Disruption includes sustained loud talk, yelling, or screaming; noise with materials; horseplay or roughhousing; and/or sustained out-of-seat behavior.).

With the growth and popularity of these web-based systems, a number of recent studies have evaluated the validity of ODRs for making decisions about the behavior of students, establishing tentative support for the use of ODRs as a broad indicator of school-wide

levels of problem behavior (Irvin et al., 2004; Spaulding, Irvin et al., 2010). However, there is less evidence of the properties of ODRs for measuring individual behavior (Nelson, Gonzalez, Epstein, & Benner, 2003).

Utility of ODRs for Measuring Individual Student Behavior

Beyond their use as indicators of problem behavior at the school level, there are three common uses for ODRs in measurement of individual student behavior. First, ODRs have been used as part of a multisource approach to identify the operant function of problem behavior, a critical component for intervention selection (McIntosh, Brown, & Borgmeier, 2008). Research has identified how patterns of ODRs can be used to indicate a hypothesized behavioral function (March & Horner, 2002; McIntosh, Horner, Chard, Dickey, & Braun, 2008; Stage et al., 2006). Second, ODRs have been used as progress monitoring measures to determine response to intervention. There are several examples of the use of ODRs as a secondary measure of intervention effectiveness, particularly for targeted Tier 2 interventions (e.g., McIntosh, Campbell, Carter, & Dickey, 2009; Todd, Campbell, Meyer, & Horner, 2008). However, this use of ODRs is somewhat hampered by the lack of normative ODR growth rates for students requiring different levels of support. Third, ODRs have been used as screening measures to identify students who require additional behavior support beyond universal interventions (Tobin, Sugai, & Colvin, 1996; Tobin & Sugai, 1999).

ODR Cut Points

Recent research has examined and provided tentative support for common cut points to identify response to universal behavior support and the level of behavior support required in a three-tier response to intervention model (McIntosh, Campbell, Carter, & Zumbo, 2009; B. Walker, Cheney, Stage, & Blum, 2005). In particular, these studies examined the adequacy of the typical cut points used in School-wide Positive Behavior Support (0–1, 2–5,

and 6 or more total ODRs per year; Horner, Sugai, Todd, & Lewis-Palmer, 2005) to determine whether 2 or more total ODRs indicate the need for additional behavior support beyond universal support (Tier 2) and whether 6 or more total ODRs indicates the need for intensive individual behavior support (Tier 3). B. Walker et al. (2005) found that the 2 or more total ODRs cut point identified students with significantly higher scores on the Problem Behavior Scale of the Social Skills Rating System (Gresham & Elliott, 1990). McIntosh, Campbell, Carter & Zumbo (2009) found significant differences on the Externalizing Composite of the Behavior Assessment Scale for Children 2 (Reynolds & Kamphaus, 2004) for the 0–1, 2–5, and 6 or more total ODR cut points. Means for students with 0–1 ODRs were at the cutoff between the “average” and “at-risk” classifications, students with 2–5 ODRs were at the cutoff between “at risk” and “clinically significant,” and students with 6 or more ODRs were far above the “clinically significant” classification.

Although these ODR cut points show promise as an efficient measure of externalizing problem behavior, these categories are calculated from a year’s worth of data, meaning that teams would need to wait until the criterion was met (potentially most of the year) before screening would indicate a problem. As this length of time is unacceptable for proactive behavior support, it would be helpful to identify whether an intermediate number of ODRs at the start of school might predict the number of total ODRs received by the end of the school year.

ODRs as Screening Measures

Two studies by Tobin and colleagues have examined the use of ODRs as individual student screening measures and predictors of long-term student outcomes in middle school. Tobin et al. (1996) found that receiving 2 or more ODRs in the fall of the first year of middle school was a statistically significant predictor of chronic ODRs throughout middle school. In addition, the type of ODR received (in this study, for harassment) was an addi-

tional, although less powerful, predictor. Based on their results, they recommended that students receive additional behavior support if they are given either (a) 2 ODRs in the fall or (b) an ODR for harassment in the fall. Tobin and Sugai (1999) followed up with a larger sample to assess the predictive validity of ODRs in middle and high school and found strong evidence for ODRs at the start of middle school predicting chronic behavior challenges throughout middle school. For example, the frequency of ODRs in Grade 6 predicted both ODRs in Grades 7 and 8 and a trajectory of academic failure leading to dropout. ODRs for fighting and harassment in particular raised these risks.

These studies by Tobin and colleagues provide evidence that the number and type of ODRs can be used as screening measures for additional behavior support, but less is known about prediction in elementary school. Currently, SWIS is used in over 6,400 schools, and 59% of these are elementary schools (May et al., 2008). Further, SWIS is used to collect and analyze ODR data by many districts that implement school-wide behavior support systems. The National Technical Assistance Center on Positive Behavioral Interventions and Supports reported that of 7,953 schools implementing school-wide positive behavior support, 62% are elementary schools. Moreover, much of the data from the two studies by Tobin and colleagues are over 20 years old. Since that time, the use of ODRs for measuring student behavior has increased exponentially (Irvin et al., 2006; Spaulding, Irvin et al., 2010), and the technology for efficient entry and visual analysis of ODR data has been enhanced by applications such as SWIS.

The purpose of this article is to examine the properties of ODRs as they relate to measuring individual student problem behavior in elementary schools. To justify the technical adequacy of ODRs as a tool for individual student data-based decision making, additional research is needed. Although ODRs easily meet the feasibility standard of the third generation of behavioral assessment measures (Chafouleas, Volpe, Gresham, & Cook, 2010), their psychometric adequacy for producing

valid conclusions about individual students is less clear. To answer this question, the authors examined a large sample of students from the SWIS (May et al., 2008) database to (a) identify normative ODR growth trajectories for progress monitoring purposes and (b) examine the stability and accuracy of trajectories toward different levels of total ODRs. Such evidence would better inform the use of ODRs as a behavioral assessment tool within a problem-solving model. The specific research questions were as follows:

1. What are the ODR trajectories, by month, for students in each total ODR category (0–1, 2–5, and 6 or more total ODRs)?
2. How accurately can the total ODR category be predicted by using intermediary criteria (one ODR for predicting 2 or more total ODRs, and 2 or more ODRs for predicting 6 or more total ODRs)?
3. How accurately can the total ODR category be predicted using the type of problem behaviors (e.g., fighting, arson) in ODRs received by the end of September?
4. How accurately can the total ODR category be predicted using a combination of number and type of ODRs?

Method

Participants and Settings

Participants in this study were 990,908 students in Grades K–6 from 2,509 public elementary schools in the United States using the SWIS to gather and analyze student ODRs during the 2007–2008 academic year. The schools represented 890 school districts across 42 states, and all agreed to share their data for research purposes. Private schools, alternative/juvenile justice schools, and year-round schools¹ were excluded from analyses. Thirty-seven percent of schools were considered urban ($n = 924$), 29% were suburban ($n = 724$), and 34% were rural ($n = 861$). Socioeconomic indicators were available for 96% of schools in the sample. Eight percent ($n = 182$) had 10% or fewer students eligible for free or reduced-price lunch, 13% of

schools ($n = 310$) had between 11% and 25% of students eligible, 31% of schools ($n = 758$) had between 26% and 50% of students eligible, 30% ($n = 732$) had 51% to 75% of students eligible, and 18% ($n = 432$) had more than 75% of students eligible for free or reduced-price lunch. Average student enrollment across schools was 394.88 ($SD = 201.76$). The average number of full-time classroom teachers was 31.32 ($SD = 14.51$). Regarding grade level, 8% of the sample was in kindergarten, and between 13% and 19% was in Grades 1–6.

To ensure data collection integrity, all schools were required to meet each of the requirements listed in the SWIS Readiness Checklist (Todd et al., 2009) before beginning data collection. This checklist includes the following five criteria: (a) adopting standardized ODR forms as previously described, (b) establishing a coherent office discipline referral procedure, (c) engaging in timely data entry, (d) identifying a SWIS Facilitator who will coach school personnel on data collection and decision-making procedures, and (e) participating in ongoing training related to the use of SWIS. These criteria were judged by certified SWIS Facilitators. Of the schools included in this sample, 70% ($n = 1749$) had been using SWIS for 1 year, 27% ($n = 670$) had been using SWIS for 2–4 years, and 4% ($n = 89$) had been using SWIS for 5 or more years.

The total number of ODRs for the entire year was 581,775, received by 272,455 students. The number of students without ODRs was calculated using each school's enrollment data from the National Center for Education Statistics. Using the common ODR cut points described by Horner et al. (2005), the proportion of participants in each category was as follows: 90% ($n = 892,981$) received 0–1 total ODRs, 7% ($n = 72,641$) received 2–5 total ODRs, and 3% ($n = 25,286$) received 6 or more total ODRs.

Measures

The number of ODRs received by each student for major offenses according to SWIS classifications (e.g., disrespect) was used in

analyses. Given their wide variation in use in practice, minor ODRs were not used in analyses. ODRs were calculated by number and type of problem behavior per month, and total ODRs for the school year. The average number of total ODRs received was 0.59, with a standard deviation of 2.24 and a range of 0–154.

Analyses

The first set of analyses involved generating descriptive statistics for students based on the total ODR cut point categories (0–1, 2–5, and 6 or more). Mean increases in ODRs per month and percent of students in their final category by month (e.g., the proportion of students with 2 or more total ODRs who had already accumulated 2 or more cumulative ODRs by that month) were calculated. For the prediction analyses, logistic regression was used to generate beta weights for testing statistical significance. The dichotomous outcome variable was based on the number of ODRs received by students: either 2 or more total ODRs, signifying need for additional support, or 6 or more total ODRs, signifying need for intensive support. Predictor variables included intermediate cut points—receipt of an ODR (predicting 2 or more total ODRs) or receipt of 2 or more ODRs (predicting 6 or more total ODRs)—and type of ODR received (i.e., specific problem behaviors). The final analysis combined intermediate cut points and the most predictive types of ODRs to assess the extent to which adding ODR type enhanced the prediction of 6 or more total ODRs.

For each of the predictor analyses, a number of diagnostic accuracy statistics were generated. These statistics included sensitivity, specificity, area under the curve, and conditional probabilities (for the individual types of ODRs). A brief description of each term follows, and readers are directed to articles by Swets and colleagues (e.g., Swets, Dawes, & Monahan, 2000) for more detailed discussions.

Sensitivity. Sensitivity is the rate of true-positives, or in this case, the proportion of

students who were at or above each total ODR criterion (2 or more, or 6 or more) at the end of the year and were identified by the intermediate predictor as needing support. High sensitivity indicates a high ratio of true-positives to false-negatives.

Specificity. Specificity is the rate of true-negatives, or in this case, the proportion of students who did not need additional support and were not identified as needing additional support. High specificity indicates a high ratio of true-negatives to false-positives.

Area under the curve. Because a useful prediction cut point represents a balance between sensitivity and specificity, and each decision depends on the factors involved in the decision (such as the relative risks of false-positives and false-negatives), there are no agreed upon criteria for adequate sensitivity or specificity. Generated through receiver operating characteristic curve analyses, the area under the curve (AUC) statistic takes both sensitivity and specificity into account. AUC models account for both terms for an overall assessment of accuracy. Rice and Harris (1995) provide the following criteria for diagnostic accuracy using AUC: 0.50–0.59 represents low diagnostic accuracy, 0.60–0.65 represents medium diagnostic accuracy, and 0.66–1.00 represents high diagnostic accuracy.

Conditional probabilities. Conditional probabilities describe the percent probability that an individual identified by the predictor will also meet the outcome. In this study, it is the probability that a student receiving an ODR for a specific problem behavior will be at or above the total ODR criterion (2 or more, or 6 or more) at the end of the year.

Results

Analyses were conducted to identify referral trajectories for students with 0 to 1, 2 to 5, and 6 or more total ODRs and determine the month by which students met these criteria. Table 1 and Figure 1 present increases in cumulative ODRs per month by ODR group. As seen, the data show surprisingly

Table 1
Growth Trajectories by ODR Cut Point Group

Group	n	ODR Growth Trajectories per Month										Mean (SD)
		Sept.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	Jun.	
0–1 ODRs	892,981	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.00	.01 (.005)
2–5 ODRs	72,641	0.21	0.33	0.29	0.24	0.27	0.35	0.32	0.39	0.38	0.06	.28 (.01)
6 or more ODRs	25,286	0.87	1.35	1.21	0.95	1.18	1.29	1.21	1.39	1.25	0.19	1.09 (.36)

Note. ODR = office discipline referral.

steady mean slopes, with a generally consistent mean rate for each group. However, the results in Table 1 indicate more overlap and less differentiation among the students by total ODR category for the first half of the school year. For example, only 20% of students were in their final total ODR category by the end of November, 50% by the end of February, and 80% by the end of April.

Predicting 2 or More Total ODRs With 1 or More ODRs by Month

The upper portion of Table 2 presents results from the accuracy of 1 or more ODRs by the end of each month as predictors of 2 or more total ODRs. Although this criterion was a statistically significant predictor at each month (likely because of the large sample

size) and specificity was extremely high, sensitivity remained moderate until the end of December, indicating a number of false-negatives early in the year (e.g., students receiving 2 or more total ODRs who did not have an ODR in September). According to criteria from Rice and Harris (1995), AUC results indicate that 1 or more ODRs was a moderately accurate predictor by the end of September and a highly accurate predictor in later months (see Table 2).

Predicting 6 or More Total ODRs With 2 or More ODRs by Month

The lower portion of Table 2 presents results from the accuracy of 2 or more ODRs at each month as predictors of 6 or more total ODRs. This prediction was more accurate than in the previous analyses. Specificity remained strong for each month, and sensitivity remained moderate until the end of November. The AUC results show that receiving 2 or more ODRs was a moderately accurate predictor at the end of September and a highly accurate predictor in October and beyond. Moreover, using the intermediate cut point of 2 or more ODRs by month resulted in a larger proportion of students identified early in the year: 50% of students with 6 or more total ODRs had 2 or more ODRs by the end of October, and 79% had 2 or more in December. This pattern indicates that the intermediate cut point of 2 or more ODRs results in earlier identification of students with 6 or more total ODRs.

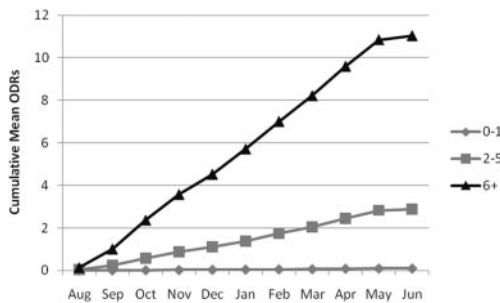


Figure 1. Growth in ODRs for students in total ODR category by month (N = 990,908, n [0 to 1 ODRs] = 892,981, n [2 to 5 ODRs] = 72,641, n [6 or more ODRs] = 25,286).

Table 2
Accuracy of Prediction of Total ODR Category

Outcome and predictor	β	OR	Sensitivity	Specificity	AUC
2 or more total ODRs ($n = 97,927$)					
1 or more ODRs by end of Sept.	3.63***	37.73	.26	.99	.63
1 or more ODRs by end of Oct.	3.76***	42.82	.48	.98	.73
1 or more ODRs by end of Nov.	3.90***	49.16	.61	.97	.79
1 or more ODRs by end of Dec.	4.05***	57.35	.70	.96	.83
6 or more total ODRs ($n = 25,286$)					
2 or more ODRs by end of Sept.	4.47***	87.21	.24	1.0	.62
2 or more ODRs by end of Oct.	4.57***	96.64	.51	.99	.75
2 or more ODRs by end of Nov.	4.78***	118.46	.69	.98	.84
2 or more ODRs by end of Dec.	5.00***	148.65	.79	.98	.88

Note. ODR = office discipline referral; OR = odds ratio; AUC = area under the curve.

*** $p < .001$.

Predicting 2 or More Total ODRs With Referred Behavior

The upper portion of Table 3 presents the ten ODR behaviors that most strongly predicted receiving 2 or more total ODRs. ODRs for gang display (i.e., gang-affiliated signs or clothing) was by far the strongest predictor of receiving 2 or more total ODRs, followed by a series of less severe behaviors, such as inappropriate displays of affection, dress code violations, and inappropriate use of technology (e.g., prohibited use of cell phones or portable music players). All listed predictors were statistically significant, but the AUC results indicated that screening with any of these specific ODRs had low diagnostic accuracy. Probabilities of receiving 2 or more ODRs are presented in the far-right column. For example, students who received an ODR for gang display had a 94% probability of receiving 2 or more total ODRs.

Predicting 6 or More Total ODRs With Referred Behavior

The lower portion of Table 3 presents the ten ODR behaviors that most strongly predicted receiving 6 or more total ODRs. Again, gang display was the strongest predictor, followed by possession of alcohol, disrespect, skipping class, and bomb threat. Less severe

behaviors, such as inappropriate displays of affection and use of technology, were not as strong predictors of 6 or more total ODRs. Like in the previous analysis, all of the listed behaviors were statistically significant predictors but had poor diagnostic accuracy. Only disrespect and physical aggression/fighting met the minimum criterion for moderate predictive accuracy.

Predicting 6 or More Total ODRs With 2 or More ODRs and Type of Referred Behavior

Table 4 presents the results of the combined prediction of number and type of ODRs. At the end of October, the point with highly accurate prediction was used. The 2 or more ODR intermediate cut point was used as a base predictor model, and receipt of any referrals for the most predictive type of referred behaviors (either disrespect or physical aggression/fighting) was added to the model to test whether adding type of behavior led to incremental increases in accuracy.

Results show that including the type of problem behavior increased the sensitivity and AUC of prediction at the end of October, with minimal loss of specificity. Adding either disrespect or physical aggression/fighting resulted in a small increase in sensitivity and

Table 3
Prediction of Total ODRs by Referred Behavior (by End of September)

Outcome and behavior	β	OR	AUC	Probability
2 or more total ODRs ($n = 97,927$)				
Gang affiliation display	5.17***	175.81	.50	.94
Inappropriate display of affection	3.78***	43.66	.50	.78
Dress code violation	3.71***	41.07	.50	.77
Disrespect	3.71***	41.01	.54	.77
Tardy	3.69***	39.84	.50	.77
Disruption	3.66***	38.85	.52	.76
Skipping class	3.58***	35.82	.50	.75
Inappropriate use of technology	3.45***	31.42	.50	.72
Physical aggression/fighting	3.32***	27.62	.55	.70
Bomb threat	3.31***	27.34	.50	.69
6 or more total ODRs ($n = 25,286$)				
Gang affiliation display	4.00***	54.68	.50	.47
Possession of alcohol	3.45***	31.37	.50	.33
Disrespect	3.12***	22.61	.60	.27
Skipping class	3.07***	21.49	.51	.26
Bomb threat	3.05**	21.19	.50	.25
Dress code violation	2.91***	18.36	.50	.23
Tardy	2.84***	17.14	.50	.21
Disruption	2.78***	16.15	.55	.20
Physical aggression/fighting	2.75***	15.71	.60	.20
Inappropriate location/out of bounds area	2.72***	15.15	.50	.19

Note. Only the strongest 10 predictors are presented. Behavior definitions can be obtained at <http://www.swis.org/index.php?page=resources;rid=10121>. ODR = office discipline referral; OR = odds ratio; AUC = area under the curve.

** $p < .01$.

*** $p < .001$.

AUC, and adding both led to the largest increase, an increase in sensitivity of 0.13 and AUC of 0.05.

Discussion

The purpose of this study was to examine properties of ODRs when used to measure student behavior—specifically ODR growth trajectories for students by total ODR category and prediction of ODR category by the number and type of ODRs received during the fall. Results from a sample of 990,908 students indicated a steady, linear mean increase in ODRs for students in each category. Intermediate cut points were highly accurate predictors of total ODR category by the end of October. Prediction using specific behaviors

had poor predictive power overall, but identified important behaviors for teams to track and showed differences in the types of problem behaviors exhibited by students receiving 2–5 and 6 or more ODRs. The most accurate prediction of 6 or more total ODRs came from using a combination of either (a) 2 or more ODRs for any behavior or (b) an ODR for either disrespect or physical aggression/fighting.

ODR Growth Trajectories

This study provides the first attempt at documenting growth trajectories of ODRs with a large sample of students. The mean increases show differences in slopes by total ODR group, indicating that these data could potentially be used to identify trajectories of

Table 4
Combined Prediction of Total ODR Category

Outcome and predictor	β	OR	Sensitivity	Specificity	AUC
6 or more total ODRs ($n = 25,286$)					
2 or more ODRs by end of Oct.	4.57***	96.64	.51	.99	.75
2+ ODRs and/or ODR for Disrespect	4.26***	70.52	.57	.98	.78
2+ ODRs and/or ODR for Physical aggression	3.99***	54.54	.58	.98	.78
2+ ODRs and/or ODR for Physical aggression or Disrespect	3.96***	52.32	.64	.97	.80

Note. ODR = office discipline referral; OR = odds ratio; AUC = area under the curve.
 *** $p < .001$.

ODRs. Yet the averages hide individual variation in slope and imply more differentiation than may exist, particularly early in the year. Results were consistent with previous research showing that the early rate of ODRs can be used to differentiate between patterns of persistent but moderate challenges, as opposed to chronic severe challenges (Tobin et al., 1996). There is also tentative support for ODRs as a long-term progress monitoring measure for students with high rates of ODRs (i.e., at the Tier 3 level). Because of the low base rates of ODRs in general, there appears to be little sensitivity as a measure for individual students with fewer total ODRs, but the growth rate of over 1 per month for students with 6 or more total ODRs indicate that it could be used as a secondary measure of long-term intervention outcomes. Nevertheless, there is a powerful degree of individual variation within the data, so ODRs would be a far less adequate primary measure of short-term intervention effectiveness than other options, such as direct observation, direct behavior rating, or brief behavior rating scales. For students with high rates of ODRs, it may be more accurate to compare an individual's rate of ODRs to her or his previous rates than rely on these rates.

Intermediate Cut Points as Predictors of Total ODRs

The steady growth in ODRs seen in Table 1 and Figure 1 show that there is no sharp rise in mean ODRs in any particular month

and no visually obvious point for screening with the existing total ODR criteria. These results were consistent with previous research showing low rates of ODRs early in the year for students with both high and low total ODRs (Taylor-Greene et al., 1997), and the trend in these data may be from differences in context, such as a focus on teaching school-wide expectations at the start of the school year. As a result, relying on total ODR cut points to identify students provides late identification of students for additional support, but intermediate ODR cut points yield a sharper increase at the start of the school year and hence more promise for screening.

Prediction of 2 or more total ODRs from 1 ODR only becomes adequate as a screening measure by the end of December. At this point, the school year is nearly half over, and providing additional support so late in the year can be less effective (H. M. Walker & Sprague, 1999). The inaccuracy of this criterion likely comes from the challenge of predicting future behavior from one recorded event. A single ODR represents one instance of problem behavior. Little can be interpreted accurately from a single event (Cohen, 1988), so it is more accurate to use ODR data when patterns of behavior emerge. As a student receives more ODRs, the ODR data become more valuable to interpret. For example, March and Horner (2002) and McIntosh and colleagues (2008) used patterns of ODRs as one source to verify teacher-identified func-

tion of problem behavior, but only for students who had a sufficient number of ODRs to show a pattern.

The intermediate cut point was much more accurate when predicting 6 or more total ODRs. Prediction of total ODRs from intermediate cut points had high specificity throughout the fall but suffered from lack of sensitivity early in the fall (e.g., the end of September). After each month, prediction became more accurate, and vastly more students were correctly identified. Analyses show that by the end of October, prediction is statistically significant and highly accurate, with high specificity and moderate sensitivity. Over half of the students with 6 or more total ODRs had 2 or more ODRs by then, so 2 or more ODRs seems a useful intermediate predictor, and screening at the end of October seems to represent the best time to screen in terms of the balance between early and more accurate identification (Swets et al., 2000).

Screening at the end of October produced few false-positives but a moderate number of false-negatives. A low rate of false-positives is beneficial, in that resources are not allocated unnecessarily and students are not mislabeled as having chronic behavior problems. Knowing this information is particularly helpful for decision making (Swets et al., 2000). The 2 or more ODRs intermediate criterion signals significant risk and a low likelihood of accidentally misidentifying a student as requiring support, but school personnel need to be aware that ODRs will not identify all students. Because an adult must observe or receive a student report of problem behavior, it is unclear how well ODRs capture covert behavior, such as relational aggression. In addition, ODRs under-report challenges with internalizing behavior (McIntosh, Campbell, Carter, & Zumbo, 2009). Given this information, use of intermediate ODRs makes for an efficient but not perfect screening tool for externalizing behavior problems in elementary school. These results were consistent with those obtained at the middle school level by Tobin and colleagues (Tobin et al., 1996; Tobin & Sugai, 1999), whereby 2 or more ODRs

in the fall signaled high rates of ODRs later that year and in subsequent years.

Combined Prediction

Also consistent with the research by Tobin and colleagues (1996, 1999), adding the most predictive type of ODRs enhanced the accuracy of prediction of 6 or more ODRs, although not substantially. Nevertheless, the use of ODR databases makes screening by the type of ODR easy and straightforward, so the added accuracy seems to make the extra step worthwhile.

It is interesting to note that the most predictive behaviors in elementary school were different than those in middle school. In middle school, physical aggression/fighting and harassment were powerful additional predictors (Tobin et al., 1996). In elementary school, physical aggression/fighting and disrespect were moderate additional predictors. It is likely that these behaviors signal different levels of severity between elementary and middle school. For example, shoving another student in Grade 1 might cause concern, but perhaps more concern when it occurs in Grade 8. And although disrespect may be an obvious warning sign to an elementary classroom teacher, that same classroom teacher may be unaware of physical aggression in nonclassroom areas, such as the playground, without a systematic approach to documenting ODRs. Using these data may indicate to teams and teachers that problems that may seem specific to one setting are more pervasive than one teacher might observe.

Specific Problem Behaviors

Other than as part of the combined prediction, the AUC results show that specific ODR behaviors were not accurate screeners in and of themselves, owing to the high rate of false-negatives for each behavior. This result was not surprising, given the fact that a student who receives multiple ODRs in September for fighting but none for other behaviors would count as a false-negative for every other behavior. However, some behaviors are clearly red flags for chronic problem behavior

and the need for additional support. As noted, physical aggression/fighting and disrespect were moderately accurately indicators. Gang display was a particularly strong indicator of challenges for each outcome, but it occurred too rarely to be a useful broad screener. For each behavior, school personnel can use the conditional probabilities in Table 4 to identify the possibility of particular ODR outcomes and work to decrease these odds through additional intervention.

The results also highlight some differences between the groups of students who received either 2 or more or 6 or more total ODRs. Although 8 of the 10 behaviors were the same for each analysis, the 2 behaviors that separated the subset of 6 or more total ODRs from the 2 or more group were possession of alcohol and inappropriate location/out of bounds, both significant causes of concern in elementary school. In contrast, the predictors of 2 or more total ODRs that were not on the 6 or more list—inappropriate displays of affection and use of technology—are less severe, pointing to a pattern of less serious problem behaviors for the students with between 2 and 5 total ODRs. These results complement those obtained by Tobin et al. (1996), who found that students with high levels of ODRs had significantly more ODRs for violent behavior and that the number of violent and nonviolent ODRs were not correlated for middle school students. Taken together, these studies provide further evidence of the discriminant validity of these cut points, that they divide students based on both the frequency and severity of problem behavior.

A final note regarding prediction and the use of ODRs is that the majority of schools in this study accessed their ODR data at least monthly for decision-making purposes (Spaulding, McIntosh, Horner, & Frank, 2010). Given this information, it is logical to assume that some students received additional support based on early receipt of ODRs or ODRs for specific behaviors. Hence, the false-positives observed in the data set may represent either poor prediction or additional intervention that broke the prediction. However, the steady trajectory for students with 6 or more total ODRs

indicates that for too many students, the prediction was not broken.

Limitations

Although results of this analysis provide some indication of the utility of ODRs in a very large sample, some limitations are noteworthy. First, all of the schools included in this sample were using the SWIS system, which mandates training, adoption of clear definitions for ODRs, and standardized ODR forms. Data from schools using less standardized procedures may not have the same results (Nelson et al., 2002). However, because this study used extant data, little is known about the consistency of referral and data entry practices for the schools in this sample. At the individual level, teacher decisions regarding whether to issue an ODR can be subjective and influenced by ethnic bias (Kern & Manz, 2004; Skiba, Michael, Nardo, & Peterson, 2002), and schools in this sample may have been at different stages of standardizing their referral practices. In addition, although all schools were required to meet SWIS Readiness Criteria before using SWIS, schools may have had varying levels of implementation of overall School-wide Positive Behavior Support practices. Future research may explore how the level of School-wide Positive Behavior Support implementation influences the validity of ODR data. Finally, although the hypotheses were based on logic and practice, the results should be viewed with some caution, as the same variable was used as both a predictor and an outcome. This aspect of the design limited variability and may have inflated the test statistics (Swets et al., 2000).

Future Research Directions

Results of this study reveal further information regarding the strengths and weaknesses of ODRs as a screening and progress monitoring measure. Because of their widespread and increasing use, it is clear that continued technical adequacy research is needed. Because this study focused only on elementary schools, it will be important to examine results of similar analyses in middle and high schools,

where ODRs are even more frequently used. Moreover, additional validity research on ODRs and the common cut points is needed. Although there has been some promising research regarding concurrent validity with traditional behavior rating scales (McIntosh, Campbell, Carter, & Zumbo, 2009; B. Walker et al., 2005), it would be important to assess concurrent validity with measures in this special issue, such as direct behavior rating and brief behavior rating scales, for progress monitoring purposes. In addition, more research is needed regarding relating ODRs with general student outcomes, such as academic achievement. Linking online ODR data to other large-scale individual student databases is an important avenue for this research. For example, by linking individual reading skill scores with ODR data, questions about the relation between ODRs and academic performance can be explored on a larger scale. Finally, there is a need to examine how to improve ODR reliability and reduce subjectivity or bias in use. There is already information regarding what types of forms and systems lead to more accurate data (e.g., use of checkboxes as opposed to blank fields, clear definitions of behaviors; McIntosh, Campbell, Carter, & Zumbo, 2009; Nelson et al., 2003), but the type and amount of training is needed to ensure reliability has yet to be tested. Given the value of reliable ODR data to school teams, answering these questions will greatly inform their use within a problem-solving model.

Implications for Practice

These results support the use of standardized ODRs as an efficient, but not early, screening measure and a secondary measure that can be analyzed for student response to intervention within a multimeasure approach to assessing individual student behavior. Thus we recommend their use in addition to (although certainly not in place of) other measures, such as those covered in this special issue.

Common sense should also be used when analyzing ODRs. Regardless of numbers or types, a rapid accumulation of ODRs

should trigger a school team to examine a student's degree of success and needed level of support. Screening with ODRs can be a helpful warning sign, but students might be identified more quickly with an add-on, multiple-gate screening system, such as Systematic Screening for Behavior Disorders (H. M. Walker & Severson, 1992), which can screen for internalizing challenges and uses direct observation for students with the greatest risk, or brief behavior rating screeners, such as the Behavioral and Emotional Screening System (Kamphaus & Reynolds, 2007). However, this increase in early identification comes at the cost of reduced efficiency (McIntosh, Reinke et al., 2009). There may be promise in integrating ODR data into existing multiple-gate screening systems, which could enhance screening accuracy with little increase in effort, or as an adjunct to teacher referrals for additional support. Students receiving multiple ODRs could be automatically referred for support, which could help identify students who display low-level challenges across settings and teachers. Given that ODRs are already collected in many schools, it seems prudent to use them as an additional source of data, especially if they are collected through a standardized procedure with ongoing training.

Footnotes

¹Although year-round schools were not readily identifiable, potential year-round schools using SWIS were estimated as any schools reporting at least 1 ODR in June, July, and August. This calculation resulted in the exclusion of 15 schools. In addition, June was considered the last month for counting total ODRs.

References

- Chafouleas, S. M., Volpe, R. J., Gresham, F. M., & Cook, C. R. (2010). School-based behavioral assessment within problem-solving models: Current status and future directions. *School Psychology Review, 39*, 343–349.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system*. Circle Pines, MN: American Guidance Service.
- 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.
- 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.
- Kamphaus, R. W., & Reynolds, C. R. (2007). *Behavioral and Emotional Screening System*. Bloomington, MN: Pearson.
- Kaufman, J. S., Jaser, S. S., Vaughan, E. L., Reynolds, J. S., DiDonato, J., Bernard, S. N., et al. (2010). Patterns in office referral data by grade, race/ethnicity, and gender. *Journal of Positive Behavior Interventions, 12*, 44–54.
- Kern, L., & Manz, P. (2004). A look at current validity issues of school-wide behavior support. *Behavioral Disorders, 30*, 47–59.
- 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.
- May, S., Ard, W. I., Todd, A. W., Horner, R. H., Glasgow, A., Sugai, G., et al. (2008). *School-Wide Information System*. University of Oregon, Eugene: Educational and Community Supports.
- McIntosh, K., Brown, J. A., & Borgmeier, C. J. (2008). Validity of functional behavior assessment within an RTI framework: Evidence and future directions. *Assessment for Effective Intervention, 34*, 6–14.
- McIntosh, K., Campbell, A. L., Carter, D. R., & Dickey, C. R. (2009). Differential effects of a tier two behavior intervention based on function of problem behavior. *Journal of Positive Behavior Interventions, 11*, 82–93.
- McIntosh, K., Campbell, A. L., Carter, D. R., & Zumbo, B. D. (2009). Concurrent validity of office discipline referrals and cut points used in schoolwide positive behavior support. *Behavioral Disorders, 34*, 100–113.
- McIntosh, K., Horner, R. H., Chard, D. J., Dickey, C. R., & Braun, D. H. (2008). Reading skills and function of problem behavior in typical school settings. *Journal of Special Education, 42*, 131–147.
- McIntosh, K., Reinke, W. M., & Herman, K. E. (2009). 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* (pp. 135–156). New York: Guilford Press.
- 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.
- Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavior Assessment Scale for Children* (2nd ed.). Circle Pines, MN: AGS Publishing.
- Rice, M. E., & Harris, G. T. (1995). Methodological development: Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology, 63*, 737–748.
- 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.
- Spaulding, S. A., Irvin, L. K., Horner, R. H., May, S. L., Emeldi, M., Tobin, T. J., et al. (2010). Schoolwide social-behavioral climate, student problem behavior, and related administrative decisions: Empirical patterns from 1,510 schools nationwide. *Journal of Positive Behavior Interventions, 12*, 69–85.
- Spaulding, S. A., McIntosh, K., Horner, R. H., & Frank, J. L. (2010). *A measure of implementation for office discipline referrals: Input, access, and use of school-wide data*. Manuscript in preparation.
- Stage, S. A., Jackson, H. G., Moscovitz, K., Erickson, M. J., Thurman, S. O., Jessee, W., et al. (2006). Using multimethod-multisource functional behavioral assessment with students with behavioral disabilities. *School Psychology Review, 35*, 451–471.
- 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.
- Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*, 1–26.
- Taylor-Greene, S., Brown, D., Nelson, L., Longton, J., Gassman, T., Cohen, J., et al. (1997). School-wide behavioral support: Starting the year off right. *Journal of Behavioral Education, 7*, 99–112.
- Tobin, T. J., Sugai, G., & Colvin, G. (1996). Patterns in middle school discipline records. *Journal of Emotional and Behavioral Disorders, 4*, 82–94.
- 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.
- Todd, A. W., Campbell, A. L., Meyer, G. G., & Horner, R. H. (2008). The effects of a targeted intervention to reduce problem behaviors: Elementary school implementation of check-in-check-out. *Journal of Positive Behavior Interventions, 10*, 46–55.
- Todd, A. W., Horner, R. H., & Dickey, C. R. (2009). *SWIS 4.3 Readiness Checklist*. Eugene: University of Oregon. Available at <http://www.swis.org/index.php?page=resources;rid=10014>
- Todd, A. W., Horner, R. H., & Tobin, T. J. (2010). *SWIS documentation project: Referral form definitions (version 4.3)*. Eugene: University of Oregon. Available at

- <http://www.swis.org/index.php?page=facresources;rid=10013>
- 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., & Severson, H. (1992). *Systematic screening for behavior disorders* (2nd ed.). Longmont, CO: Sopris West.
- Walker, H. M., & Sprague, J. R. (1999). The path to school failure, delinquency and violence: Causal factors and some potential solutions. *Intervention in School and Clinic, 35*, 67–73.
- 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.

Date Received: October 29, 2009

Date Accepted: May 18, 2010

Action Editor: Sandra Chafouleas ■

Kent McIntosh is an assistant professor of school psychology at the University of British Columbia. His current research interests include the sustainability of evidence-based practices in schools, integrated academic and behavior response to intervention models, and culturally responsive positive behavior support for indigenous Canadian students.

Jennifer L. Frank is a postdoctoral research fellow at the University of Oregon. Her research interests include issues related to the identification and implementation of evidence-based prevention and intervention practices in school settings.

Scott Spaulding is a research scientist at the Haring Center for Applied Research and Training in Education at the University of Washington. His current research interests include applied behavior analysis and function-based supports in educational settings, and evidence-based interventions and practices for children with severe problem behaviors.