Multilevel modeling resolves ambiguities in analyses of discipline disproportionality: A demonstration comparing Title 1 Montessori and non-Montessori schools

Lee LeBoeuf[1], Jacob Goldstein-Greenwood[2], Angeline S. Lillard[1]

Department of Psychology[1]

Library Research Data Services[2]

University of Virginia

Corresponding Author: Lee LeBoeuf, Department of Psychology, University of Virginia, P.O. Box 400400, Charlottesville, VA 22904. (434) 982-5232. (ll3jt@virginia.edu)

This article was published online in the Journal of Research on Educational Effectiveness on March 21, 2023

Acknowledgements: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B200005 to the University of Virginia. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Many thanks to Sofia Mohammed and Clay Ford for their help with this project.

Abstract

Common methods of measuring discipline disproportionality can produce contradictory results and obscure base-rate information. In this paper, we show how using multilevel modeling to analyze discipline disparities resolves ambiguities inherent in traditional measures of disparities: relative rate ratios and risk differences. One previous study suggests there is less racial discipline disproportionality in Montessori schools, so we used our new approach, along with relative rate ratios and risk differences, to compare discipline disproportionality in a sample of Title 1 Montessori and non-Montessori schools identified using propensity score matching. Using the multilevel model clarified results from other measures: discipline disproportionality was similar across school settings, even though overall rates were significantly lower in the Montessori schools.

Keywords: Discipline disproportionality, equity, Montessori

Multilevel modeling resolves ambiguities in analyses of discipline disproportionality: A demonstration comparing Title 1 Montessori and non-Montessori schools

In the United States, Black students are two to three times more likely than White students to be suspended (U.S. Department of Education, 2018). This phenomenon is termed *discipline disproportionality*, and it results in both Black and Hispanic students missing significantly more school days due to suspensions than White students (Vincent et al., 2012). Critically, previous research does not suggest that these disparities are caused by Black or Hispanic students misbehaving more than White students; rather, it suggests that they are caused by teachers and school personnel administering more suspensions to children of color (relative to White students) for subjectively defined misbehavior, such as being disrespectful or "too loud" (Fabelo et al., 2011; Girvan et al., 2017; Skiba et al., 2002; Smolkowski et al., 2016). Moreover, even for similar behavioral offenses, educators tend to give harsher punishments to students of color than to White students (Lewis et al., 2010). Disparities in disciplinary outcomes are therefore thought to be driven by racial bias and stereotypes that lead teachers to perceive behaviors of children of color as more problematic than when those same behaviors are performed by White students (Okonofua et al., 2016). Racial bias is more likely to influence teachers' decisions regarding student behavior when teachers are stressed or have competing demands on their time, so disciplinary outcomes in environments where teachers are less stressed or have more time to make thoughtful disciplinary decisions might be more equitable (McIntosh et al., 2014).

In addition to being unfair, discipline disparities are associated with lower academic performance for many children of color (Lewis et al., 2010), and exclusionary disciplinary practices—like suspension—are risk factors for negative long-term outcomes like school dropout and involvement in the criminal justice system (Skiba et al., 2014). Given these relations, reducing the use of exclusionary disciplinary practices for all children is an appropriate end goal, but at the very least, eliminating racial disparities in disciplinary outcomes is critical for an equitable education system (Resh & Sabbagh, 2016).

Researchers' understanding of discipline disparities, and of how to reduce them, has been hindered by a lack of a consistent and reliable measure of discipline disproportionality (Curran, 2020; Girvan et al., 2019; Larson et al., 2019; Nishioka et al., 2017; Petrosino et al., 2017). Some of the most common methods—relative rate ratios and risk differences—can lead to misleading conclusions, which can cloud accurate evaluation of interventions meant to reduce disparities. In this paper, we address this problem by proposing a new approach to analyzing racial disparities in school discipline using multilevel modeling. To demonstrate the utility of this new approach, we compared disciplinary data from Montessori and non-Montessori Title 1 schools. We used propensity score matching to identify our sample, and we compared average overall suspension rates for White, Black, and Hispanic students taken together, as well as discipline disproportionality between those same racial groups, across school types. We measured discipline disproportionality using multilevel modeling as well as relative rate ratios and risk differences, as traditionally calculated, and we discuss the advantages of multilevel modeling over those more common methods.

Measuring Discipline Disproportionality

There are a variety of ways to quantify discipline disproportionality, and no single measure paints a complete picture, but relative rate ratios (RRRs) and risk differences (RDs) are used most often (Petrosino et al., 2017). The RRR is the relative suspension frequency across two student subgroups. The RD is simply the difference in the proportion of students in two subgroups who received a suspension. For example, to calculate the RRR of Black and White students, one divides the proportion of Black students suspended by the proportion of White students suspended. A ratio of one indicates that Black and White students are suspended at equal rates; a ratio of two would suggest Black students are suspended twice as often as White students. To calculate the RD between those groups, one would subtract the proportion of suspended White students from the proportion of suspended Black students.

Both measures can be used to identify a disparity, but neither captures complete information about the disciplinary climate of a school. The following example from Larson et al. (2019) makes clear why neither measure is sufficiently informative: Suppose school A has suspended 30% of its Black students and 10% of its White students, and school B has suspended 3% of its Black students and 1% of its White students. Both schools would have a relative rate ratio of 3, meaning that Black students are three times as likely to be suspended as White students, but the disciplinary climates in the two schools are clearly very different. School A would have a RD of 20%, whereas school B would have a RD of 2%. School B's RD is relatively small, but one could also imagine a third school, school C, that would have an equally low RD as school B if it suspended 10% of its Black students and 8% of its White students. Comparing schools A and B on their RRR's alone, or comparing schools B and C on their RDs alone, would give the false impression of similar disciplinary climates within each school because calculating the RRR or RD inherently obscures base-rate information. This obfuscation means that neither measure provides information about the number of students impacted by a disparity (Curran, 2020), which complicates understanding of the overall disciplinary climate in a school. As a result, using one measure without the other, or failing to interpret one measure within the context of overall disciplinary rates, shrouds the true magnitude of the problem.

Both measures also have other limitations. RRR's are negatively correlated with overall suspension rates, so they can appear inflated in schools with low suspension rates (Curran, 2020; Girvan et al., 2019). Risk differences have the opposite problem—they lack the sensitivity to show meaningful disparities in relatively low probability events (Petrosino et al., 2017) (as shown by School B). These sensitivity issues could make evaluating the effect of an intervention aimed at reducing the use of exclusionary discipline and discipline disproportionality particularly challenging: If RRRs or RDs are the outcome of interest, observed changes (or a lack of changes) in either could be misleading if overall suspension rates have also changed. This limitation also means that RRRs and RDs can lead to different conclusions about changes in discipline disproportionality when analyzed longitudinally, muddling understanding of trends over time (Curran, 2020; Girvan et al., 2019). RRRs are also difficult to calculate and interpret when schools have suspended zero students from a given racial group: It's impossible to divide by zero, and switching the numerator and denominator in such a case reveals little, as zero divided by 0.01 is equivalent to zero divided by 0.99. Another limitation of both measures is that they only allow for pairwise comparisons (e.g., Black and White students or Hispanic and White students). A researcher interested in evaluating disparities between more than two racial groups at once would need to calculate separate RRRs or RDs for each pair of racial groups, which is burdensome when examining diverse schools. All of the issues above present challenges when evaluating discipline disproportionality in a single school, and they magnify when trying to understand disproportionality in a sample of schools—how, for example, ought one best account for schools of different sizes?

To compensate for the shortcomings of RRRs and RDs as outcome measures when evaluating average discipline disproportionality across a sample of schools, current guidance is to consider both the RRRs and RDs alongside the overall suspension rates (Curran, 2020; Girvan et al., 2019; Larson et al., 2019). This process can be cumbersome, and average RRRs and RDs across multiple schools can still be misleading or ambiguous for the reasons discussed above.

Here, we show how multilevel modeling can be used to assess discipline disproportionality in a set of schools, and we demonstrate how it addresses many of the issues with RRRs and RDs described above. In this method, we treat suspension counts for different racial groups as repeated measures within schools, and we then construct a multilevel negative binomial regression model predicting suspension counts, including the log of each racial group's size (i.e., the total number of students in a racial group in a schools) as an offset to account for differences in racial group sizes. Unlike Poisson models, negative binomial models include a dispersion parameter which allows for violation of the assumption that the outcome follows a Poisson distribution (with a mean equal to its variance). We include dummy variables for each racial group (Black, Hispanic, and so on.), which estimate the difference in the log of suspension counts for each racial group compared to the reference category. For example, using White students as the reference group, one could construct the following model:

$$\log(suspension\ count)_{ij} = \eta_{ij} = \gamma_{00} + \beta_1 Montessori_j + \beta_2 Hispanic_{ij} + \beta_3 Black_{ij} + \log\left(racial\ group\ size_{ij}\right) + \mu_{0j}, \quad (1)$$

where i indexes measurements (subgroups) within schools, j indexes individual schools, and μ_{0j} is the random intercept for the j^{th} school. The coefficients for Hispanic and Black denote the model-estimated differences in average log suspension counts between each of those racial groups and White students. By exponentiating the coefficients, one can derive the multiplicative difference between suspension counts for each racial group relative to White students—in other words, one can derive the relative ratio of each racial group to White students. The method offers the same ease of interpretation of RRRs as described above, but we do not sacrifice base-rate information, as estimated average rates for each racial group are easily derived from the model, and unlike when using RRRs as the outcome measure, schools that gave zero suspensions to one or more racial groups can be included in the model. In this example, we use observations within schools (number of students suspended from each racial group), but one could also use observations within classrooms or districts depending on the data available and the research question.

We set White students as the reference category to be consistent with previous research and recommendations (Larson et al., 2019; Morris et al., 2017; Skiba et al., 2011). The designation of White students as the reference category is not meant to suggest that the suspension rate of White students is an appropriate standard for all students—as already mentioned, eliminating all suspensions would be an appropriate end goal of school discipline reform—but given the current use of suspension throughout schooling, designating White students as the reference category makes the determination of racial disparities simple, as White students are reliably suspended at lower average rates than Black students and often Hispanic students (Losen et al., 2015; U.S. Department of Education, 2018).

To evaluate whether disparities in the average proportion of students suspended from each racial group differ across levels of an intervention (like whether a school is a Montessori school or a non-Montessori school), we add an interaction term between a dummy variable indicating intervention status (in the example, Montessori classification) and the dummy variables for each racial group:

$$\log(suspension\ count)_{ij} = \eta_{ij} = \gamma_{00} + \beta_1 Montessori_j + \beta_2 Hispanic_{ij} \\ + \beta_3 Black_{ij} + \beta_4 \left(Montessori_j * Hispanic_{ij}\right)$$

$$+ \beta_5(Montessori_j * Black_{ij}) + \log(racial\ group\ size_{ij}) + \mu_{0j}, \qquad (2)$$

In this model, the coefficients for Hispanic and Black are again the estimated average differences in the log suspension counts between each of those racial groups and White students, but in non-Montessori schools only. The coefficients for the interaction terms between each racial group and Montessori classification are the estimated differences in disparities in the average log suspension counts across school types.

The primary benefit of using this multilevel modeling approach is that it resolves many of the issues presented by using RRRs and RDs as outcome measures. First, unlike with RRRs and RDs, the user does not lose base-rate information in the calculation and thus derives a more complete view of schools' overall disciplinary climates, as suspension rates for each group are easily recoverable from the model. The resulting estimates are also not prone to the sensitivity issues of RRRs and RDs in schools with low suspension rates, making the model more accommodating to schools with low or zero suspension counts for any racial group (though depending on the specific data, one may opt for a Poisson model instead of the negative binomial model we use here, and either model may potentially be zero-inflated). Additionally, the user is no longer required to run separate models for each pairwise racial group comparison as one is when using RRRs or RDs); rather, one can estimate each disparity in one model. Finally, the model offers the flexibility to analyze longitudinal data (we elaborate on this point in the discussion section).

Next, we explain why we applied this model to examine discipline disproportionality in Montessori schools.

Montessori

Physician-educator Maria Montessori and her collaborators developed the Montessori system based on their experiences working with children with disabilities and, later, children living in poverty (Lillard, 2019; Montessori, 1912). Today, Montessori is the most prevalent and most enduring alternative pedagogy in the world (Lillard, 2019). In the United States alone, there are over 500 public Montessori schools, and the majority of public Montessori students are children of color (Debs, 2016). If the disciplinary climate of Montessori schools is similar to national trends, then the majority of Montessori students are at risk for unfair punishment. To date, very few studies have investigated the disciplinary climate of Montessori schools, but those that have suggest that Montessori schools tend to have more racially equitable and less punitive disciplinary climates than non-Montessori schools (Brown & Steele, 2015; Culclasure et al., 2018). More research in Montessori schools could be valuable because if discipline disparities are consistently smaller in Montessori schools, further study of the Montessori approach to discipline could reveal more-equitable disciplinary practices. We first discuss theoretical reasons why disciplinary climates in Montessori schools might be different from those in non-Montessori schools, and we then review the two studies we know of that have investigated Montessori student disciplinary outcomes.

Montessori and Discipline

Among other things, Montessori's attitudes toward discipline differed from those of her predecessors. She wrote, "The task of the educator lies in seeing that the child does not confound *good* with *immobility* and *evil* with *activity*, as often happens in the case of the old-time discipline. And all this because our aim is to discipline *for activity, for work, for good;* not for *immobility*, not for *passivity*, not for *obedience*" (Montessori, 1912, p. 74; italics in original). Montessori trained teachers to diligently observe their students because she believed that students' behaviors reflect their developmental needs and that student behavior should therefore inform instruction. During these observations, educators "must not start, for example, from any dogmatic ideas which [they] may happen to have held upon the subject of child psychology" (Montessori, 1912, p. 38). A large portion of Montessori teacher training focuses on learning to become an unbiased observer (Montessori, 1912; Whitescarver & Cossentino, 2007). The focus on objectivity in evaluating student behavior could cause Montessori teachers to react less punitively to students' behavior, in which case one would predict lower overall rates of exclusionary discipline, like suspension, in Montessori schools.

Montessori classrooms are also characterized by high degrees of student self-determination and free choice. Montessori students, relative to conventional school students, report feeling a stronger sense of classroom community at school (Lillard et al., 2006; Rathunde & Csikszentmihalyi, 2005) and enjoying schoolwork more (Lillard et al., 2017). Increased student self-determination, higher student engagement, and stronger classroom community could correspond to fewer disruptive behaviors. If so, these differences would likely also lead to lower overall suspension rates simply by reducing student behaviors that school personnel believe warrant suspension.

Discipline Disproportionality and Montessori

As already mentioned, educators refer students of color for suspension at disproportionality high rates relative to White children for subjectively defined misbehavior (Fabelo et al., 2011; Girvan et al., 2017; Skiba et al., 2002; Smolkowski et al., 2016). Even for very similar behaviors, students of color tend to receive harsher punishments than White students (Lewis et al., 2010). Discipline disparities are therefore thought to be a product of racially biased disciplinary decisions (Skiba et al., 2002). Teachers are most likely to make racially biased disciplinary decisions when they do not have the ability or motivation to do otherwise, whether it be due to stress or limited time (McIntosh et al., 2014; Okonofua et al., 2016). McIntosh and colleagues (2014) call these moments "vulnerable decision points" and argue that one way to reduce discipline disproportionality is to reduce the number of times teachers are forced to make snap decisions regarding student behavior.

In a conventional classroom, where a large portion of the work is guided by the teacher, intervening with one disruptive student likely means putting the rest of the class's learning on hold. This fact in itself might create stress and result in biased disciplinary decisions, as it often leads teachers to view student disruptions as a threat to keeping the rest of the class's learning on track (Fenning & Rose, 2007). By contrast, in a Montessori classroom, students are taught to work primarily independently or in small groups (Lillard, 2019; Montessori, 1912), so one disruptive student is less likely to interfere with the entire class. Assuming levels of racial bias are similar, on average, across all teachers, Montessori teachers may be less likely to be influenced by racial bias while making disciplinary decisions simply because Montessori teachers may feel less rushed. Whereas a conventional teacher might feel pressure to redirect a disruptive student quickly so that they can resume whole-class instruction, a Montessori teacher might have more time to work with the student one-on-one because the rest of the class would likely already be working independently. This would likely lead to fewer vulnerable decision points throughout a Montessori teacher's day and thus reduce the likelihood that racial bias would influence disciplinary decisions. Finally, most Montessori classrooms have an assistant teacher (Montessori, 1912), which might further alleviate pressure to swiftly curtail a disruptive student and has been associated with smaller racial disparities in conventional classrooms

(Gregory et al., 2019). If so, one would predict lower discipline disproportionality in Montessori schools (McIntosh et al., 2014; Okonofua et al., 2016).

Research on Discipline in Montessori Schools

Some studies suggest Montessori schools have both lower overall exclusionary discipline rates and lower discipline disproportionality. However, estimating the true effect of the Montessori method on discipline outcomes is difficult due to the infeasibility of randomly assigning children to Montessori schools. The most obvious source of potential bias in comparing Montessori schools to non-Montessori schools is self-selection. Nearly all public Montessori schools are school-choice programs, meaning that most are either magnet or charter schools (Debs, 2016). Most Montessori parents have therefore elected to enroll their child in a public Montessori school, and that decision might be associated with an average difference in parenting practices related to discipline. Two previous studies that investigated potential differences between American Montessori and non-Montessori parents observed no meaningful average differences (Dreyer & Rigler, 1969; Fleege et al., 1967), but these studies alone cannot rule out the potential for self-selection bias; the researchers might have simply not measured certain characteristics that are associated with selecting a Montessori school. Montessori research requires careful consideration of self-selection bias during sampling procedures and the selection of control variables.

In the one previous attempt to estimate the effect of Montessori education on overall discipline outcomes, Culclasure and colleagues (2018) compared the suspension rates of all (over 7000) South Carolina public Montessori students to the suspension rates of demographically matched conventional school students. They found that, after controlling for family income, race, gender, English as a second language status, special education status, and grade, suspension rates among Montessori students were 1–2% lower than among non-Montessori students. The authors benefited from extensive state data and were able to exact-match each Montessori student with a non-Montessori student in the same district on demographic variables and their previous year's test scores. Exact matching can be a powerful tool for estimating treatment effects when random assignment is impossible (Stuart, 2010). However, this study was limited to the state of South Carolina; disciplinary practices vary heavily by region (Losen et al., 2015), so results might not generalize to the rest of the United States. This study also did not investigate discipline disproportionality.

Only one previous study to our knowledge has measured discipline disproportionality in Montessori schools. Brown and Steele (2015) used RRRs to compare discipline disproportionality in three public Montessori schools to that in 14 conventional schools in a single district in the southeastern United States. They found that discipline disproportionality was present in both school systems, but the disparity between Black and White students' suspension rates was smaller in the Montessori schools than the conventional schools. However, as already discussed, relying on a single measure of discipline disproportionality does not offer a clear picture of the full disciplinary climate. The study was also limited in that the two samples differed dramatically on some important characteristics such as school size and average socioeconomic status, and these variables were not controlled for in the estimation of discipline disproportionality differences. Because the Montessori and non-Montessori schools in this sample differed on multiple characteristics, it's possible that it wasn't Montessori pedagogy alone that explained differences in the outcome: It would be preferable to compare samples of schools that are more similar, which can be done via propensity score matching.

The Current Study

This study used propensity score matching to estimate the effect of the Montessori method on average overall in-school-suspension (ISS) and out-of-school-suspension (OSS) frequency, as well as on discipline disproportionality, in Title 1 schools. Using school-level data disaggregated by racial group, we analyzed discipline disproportionality in three ways: Using RRRs, RDs, and multilevel models as in model (2). These approaches were used to compare disparities between Black and White students and between Hispanic and White students. Based on the reasons discussed above, we expected average overall rates of exclusionary discipline, as well as mean RRRs and RDs for both sets of race comparisons, to be lower for the Montessori schools. We expected these results to be supported and clarified by the multilevel model.

Method

Data

All data were collected as part of the Civil Rights Data Collection's (CRDC) 2017 survey, which only includes school-level data. The CRDC disaggregates (by race, sex, and disability status) the raw counts of students who received one or more ISS, students who received one OSS, and students who received more than one OSS. For the purposes of this study, we aggregated all students who received one OSS or more than one OSS for each racial group, and we considered ISS and OSS counts separately. So, the number of Black students who received an ISS or OSS represents the total number, regardless of sex, disability status, and whether they received one or multiple ISS/OSS. The data also include the total number of students in the school, the proportion of students with different racial identities and disability statuses, the proportion of students who qualified for free/reduced price lunch (FRPL), and binary variables indicating magnet or charter status.

Racial groups reported by the CRDC are broad—Black, White, Hispanic, etc.—and they do not include more granular classifications based on skin complexion. These categories prevented us from considering the relation between colorism and suspension rates, which is a limitation because previous research suggests that even among students of color, students with darker complexions tend to be suspended at higher rates than students with lighter complexions (Blake et al., 2017; Hannon et al., 2013). Additionally, for simplicity in this demonstration, we collapsed male and female suspensions together in each racial group. Doing so makes us unable to comment on the intersectional role of gender and race in contributing to suspension rates, which is another limitation because previous work suggests that the discipline gap is larger between White girls and Black girls specifically than it is between White boys and Black boys (Morris & Perry, 2017). Despite these limitations, the CRDC data provided ample opportunity to investigate racial disparities in suspension rates across both school systems.

Propensity Score Matching

Propensity score matching (PSM) can be an effective way to reduce bias in treatment-effect estimates when treatment is self-selected or otherwise non-randomly assigned (Gu & Rosenbaum, 1993; Harris & Horst, 2016; Steiner et al., 2015; Stuart, 2010). Since schools are not randomly assigned to be a Montessori school (or not), we used propensity score matching to ensure that the distributions of student characteristics that are likely related to students' disciplinary outcomes were similar in the Montessori and non-Montessori schools in our sample. PSM involves estimating the probability of each unit of observation receiving the treatment (i.e., each unit's propensity score) based on selected covariates. In this study, "treatment" was whether or not a school is a Montessori school. From there, treated units were "matched" to a sample of control units based on their propensity score. The end goal is to minimize differences between the treated and control samples on whichever available covariates might be related to both the outcome and the treatment selection. PSM alone cannot eliminate the potential for selection bias to confound results, but it can help ensure that the control group is similar or equivalent to the treatment group on available covariates.

For this study, we identified an initial sample of Montessori schools by searching the Civil Rights Data Collection (CRDC) database for Title 1–classified schools with “Montessori” in the name. We cross-referenced this initial list with a list of public Montessori schools maintained by the University of Virginia’s Early Development Laboratory to confirm that the schools were indeed Montessori schools. To reduce the chance of comparing suspension rates between schools with only a small handful of students of a given race (rendering estimates of the proportion suspended noisy), we limited the sample to schools that had (a) 5% or more White students and (b) 5% or more Black students and/or 5% or more Hispanic students (e.g., a school with 90% White students, 3% Hispanic students, and 7% Black students would have been included, as would a school with 80% White students, 10% Hispanic students, and 10% Black students). The specific cutoff of 5% was used to be consistent with Brown and Steele (2015). Using a cutoff at all was primarily useful for calculating RDs and RRRs, but future research using count data with our multilevel approach would likely not need to use a similar cutoff, as modeling raw counts (with a group size offset) instead of converting the raw counts to rates in advance of modeling helps avoid introducing noise in the case of small group sizes.

We identified 151 Title 1 Montessori schools in the CRDC’s database. From this group, 20 were removed due to a lack of racial diversity in the school; two were removed due to naming inconsistencies between the CRDC’s datafile and the schools’ websites; and four were removed because they were missing percent FRPL data (which we used in propensity score matching) on the CRDC's website. This process left 125 Title 1 Montessori schools in the sample.

The initial pool of potential matches for the Montessori sample included all schools in the same ZIP Codes as the Montessori schools, comprising a total of 1,268 non-Montessori schools. We selected potential matches from the same geographic areas as the Montessori schools because students living in the same area are likely to be similar on unobserved variables that could influence school discipline (for example, district/state leadership). There were non-Montessori schools within each ZIP Code where there was a Montessori school, but ZIP Code itself was not prioritized in the PSM model (described later), so the final sample does not include an equal number of Montessori and non-Montessori schools from each ZIP Code.

We applied the same criteria regarding racial diversity and Title 1 status to the non-Montessori schools as we applied to the Montessori schools. After removing non–Title 1 schools and schools without sufficient racial diversity, 648 schools remained in the pool of potential matches. Two additional schools were removed from the pool of potential matches because we deemed them inappropriate matches (one was entirely virtual, and the other was exclusively for deaf and blind children), and 17 were removed because they were missing FRPL data. The final list of potential matches included 629 schools.

Following Stuart's (2010) recommendation, we calculated propensity scores predicting treatment condition (i.e., whether a school is a Montessori school) using the following logistic regression model:

$$D_i = B_0 + \theta' X_i, \tag{3}$$

where D_i is the log odds of a school being a Montessori school, θ' is a vector of coefficients, and X_i is a vector of school characteristics. Ideally, estimation of propensity scores would include covariates related to self-selection into a Montessori school (Harris & Horst, 2016). In the absence of such information, we used the “kitchen sink” approach (Steiner et al., 2015) and calculated propensity scores using an array of available predictors. To dial in a well-matched sample, we estimated propensity scores multiple times using different combinations of the following covariates: the number of students in each school; the proportion of Black, White, and Hispanic students; the proportion of students with disabilities who received school services through the Individuals with Disabilities Education Act (IDEA); the proportion of students with disabilities who received school services through Section 504 of the Rehabilitation Act (504); the proportion of students who qualified for FRPL; binary variables indicating whether each school offered each grade from preschool to twelfth; and binary variables indicating magnet and charter school classification.

On each iteration, we tried identifying a match for each Montessori school using both optimal pair and nearest neighbor matching with the *MatchIt* R package (v4.2.0; Ho et al., 2011) to see which method yielded a list of matches most-similar to the Montessori schools. One match for each Montessori school was identified without replacement such that there was an equal number of non-Montessori and Montessori schools in the sample.

After each iteration of the matching routine, we checked the balance between the Montessori and non-Montessori schools by computing standardized mean differences (as recommended by Stuart, 2010) on each of the variables listed above (also see Table 1). In the final iteration, using the nearest neighbor method, the Montessori schools and the matched non-Montessori schools had no significant standardized mean differences on any of the variables mentioned above except for the proportion of schools of each type that offered sixth grade. There is not empirical evidence (to our knowledge) to suggest that sixth grade disciplinary outcomes would be meaningfully different from other nearby grades, so this is unlikely to bias the results.

The final sample included 250 schools (125 Montessori) in 25 states, representing 100,204 students (Table 1 includes additional sample information). Once the sample was finalized, we retrieved disciplinary data from the CRDC. We converted the raw counts of ISS and OSS available in the CRDC data into rates for each racial group within each school (for use in the RRR and RD analyses). For both ISS and OSS, we divided the number of students in each racial group who were suspended by the total number of students in that racial group. Roughly 90% of students in every school in this sample were either Black, White, or Hispanic, and very few students from other racial groups were suspended. For simplicity, the overall disciplinary rates reported here refer to the proportion of the total number of Black, White, and Hispanic students suspended (one or more times) at each school.

Analysis Plan

This study sought to compare the average overall rates of ISS and OSS between Montessori and non-Montessori schools, as well as racial discipline disproportionality in both of those outcomes between school types.

Overall Suspension Rates

Upon initial inspection of the data, we realized that more of the schools in the sample had given zero suspensions than we were expecting: 66 of the Montessori schools and 48 of the non-Montessori schools had zero ISS, and 44 of the Montessori schools and 26 of the non-Montessori schools had zero OSS. In light of this, we decided to additionally assess whether Montessori schools or non-Montessori schools were more likely to give zero suspensions during an entire school year. We ran the following binary logistic regressions predicting the log odds of giving zero ISS and the log odds of giving zero OSS (versus the alternative of giving at least one ISS and giving at least one OSS):

$$For\ ISS\ or\ OSS\text{: } \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 Montessori_i + \theta' X_i, \tag{4}$$

where π_i is the probability of giving zero ISS or the probability of giving zero OSS, θ' is a vector of coefficients, and X_i is a vector of the following school characteristics: number of students in the school, charter and magnet classification, the proportions of Black, Hispanic, and White students, and the proportions of students with disabilities (IDEA) and who qualify for FRPL. In these models, the Montessori coefficient represents the average difference in the log odds of giving zero ISS/OSS between the Montessori and non-Montessori schools, controlling for the school characteristics indicated above.

To estimate differences in average overall ISS and OSS rates between Montessori and non-Montessori schools, we ran the following negative binomial model, once for ISS rates and once for OSS rates:

$$\log(suspension\ count)_i = \eta_i = \beta_0 + \beta_1 Montessori_i + \theta' X_i + \log(number\ of\ students_i), \tag{5}$$

where η_i is the log count of ISS or OSS, θ' is a vector of coefficients, and X_i is a vector of the following school characteristics: charter and magnet classification, the proportions of Black, Hispanic, and White students, and the proportions of students with disabilities (IDEA) and who qualify for FRPL. We again included the number of students in each school as an offset to account for different school sizes. Model diagnostics using the DHARMa package (v0.4.6; Hartig, 2022) in R indicated imperfect alignment of the expected and observed residual distributions, so we bootstrapped both models 5,000 times and calculated bias-corrected 95% confidence intervals around the coefficients.

Discipline Disproportionality

To compare racial discipline disproportionality across school types, we first analyzed differences between Montessori and non-Montessori schools on conventional metrics—RRRs and RDs for ISS and OSS. Then, to demonstrate the utility of our proposed method, we compared average frequencies of ISS and OSS between racial groups and school types using multilevel models as exemplified in model (2).

To examine discipline disproportionality using RRRs, we first subset the sample to include just schools for which meaningful RRRs could be calculated. Calculating RRRs only made sense for schools that gave at least 1 ISS or OSS to Black and White or Hispanic and White students to avoid having zero in the numerator or denominator. For both RRRs and RDs, comparisons between racial groups were only done in schools that had at least 5% either Black or Hispanic students (depending on which racial groups were being compared; all schools in the sample had greater than or equal to 5% White students). Forming subsamples to calculate the RRRs meant a significant amount of data loss because of how many schools gave zero suspensions (see Table 4), which again highlights a limitation of RRRs. This limitation also means that the assumptions of propensity score matching hold less well for these analyses, as the covariate balance for each subset was not prioritized in the original propensity score estimation.

Once RRRs and RDs were calculated, we compared them across school types using the following linear models:

$$(RRR\ or\ RD\ for\ ISS\ or\ OSS)_i = \beta_0 + \beta_1 Montessori_i + \theta' X_i + \varepsilon_i, \tag{6}$$

where θ' is a vector of coefficients and X_i is a vector of the following school characteristics: charter and magnet classification, the proportions of Black, Hispanic, and White students, and the proportions of students with disabilities (IDEA) and who qualify for FRPL. In all, model (6) was used to predict (a) Black and White student ISS and OSS RRRs and RDs and (b) Hispanic and White student ISS and OSS RRRs and RDs. Residual Q-Q plots, Shapiro-Wilk tests, and plots of fitted values against residuals suggested that the error normality and homogeneity of variance assumptions were violated in all models. In response, for each model, we bootstrapped the sample 5,000 times, re-estimated the model on each resample, and calculated bias-corrected 95% bootstrap confidence intervals around the coefficients.

Finally, after calculating models for RRRs and RDs, we then used model (2) to estimate differences in discipline disproportionality across school types. As before, we bootstrapped these models 5,000 times each to generate bias-corrected 95% confidence intervals around the coefficients and predicted values. In addition to the advantages already discussed, the multilevel models offered the benefit of being able to include all schools in our sample in the estimations of average ISS and OSS log counts for each racial group. We also ran versions of models with the same control variables used in previous models (charter and magnet classification, the proportions of Black, Hispanic, and White students, and the proportions of students with disabilities [IDEA] and who qualify for FRPL). The maximum absolute log-count change in predictor coefficients across both models once those controls were added was 0.077 (for the Montessori coefficient in the OSS model), and there were no changes in the significance patterns of the predictors. Given the small difference, we report the results from model (2), which did not include those control variables.

Results

Overall Disciplinary Rates

Descriptive statistics—average overall ISS/OSS rates and average RRRs/RDs for each disciplinary outcome—for the Montessori and non-Montessori schools are shown in Table 2.

Overall, the Montessori schools (on average) gave 66% as many of their students an ISS and half as many students an OSS as the non-Montessori schools. Results from the logit models predicting the log odds of whether a school gave zero ISS or OSS are shown in Table 3. Exponentiating the coefficients for Montessori classification indicated that, holding all other variables in the model constant, Montessori schools had 1.74 times higher odds of giving zero ISS and 1.87 times higher odds of giving zero OSS than non-Montessori schools ($p = 0.04$ for ISS and $p = 0.05$ for OSS). Results from the negative binomial models estimating overall ISS and OSS counts are shown in Table 2 along with bias-corrected 95% bootstrap confidence intervals. In the ISS model, exponentiating the coefficient for the dummy variable indicating Montessori classification suggested that, holding constant all the indicated controls, Montessori schools gave, on average, 0.62 times as many ISS compared to the non-Montessori schools (although the 95% confidence interval included zero). The Montessori schools also gave 0.59 times as many OSS as the non-Montessori schools, again holding all the covariates constant, and the 95% confidence interval excluded zero. In the ISS model, the percentage of Black students and the percentage of IDEA students were both positively associated with ISS rates. In the OSS model, the percentage of IDEA students and the percentage of FRPL-qualifying students were positively associated with OSS counts, and the percentage of Hispanic students was negatively associated with OSS counts.

Relative Rate Ratios

Results from the models predicting ISS and OSS RRRs for Black and White students and for Hispanic and White students are shown in Table 4 along with bias-corrected 95% bootstrap confidence intervals. Unlike in previous research, RRRs between Black and White students were higher on average in the Montessori schools than in the non-Montessori schools, and the difference was statistically significant for both ISS and OSS. These results are the opposite of what was expected. For Hispanic and White student RRRs, there was not a statistically significant difference between the Montessori and non-Montessori schools for either ISS or OSS. Charter school classification was negatively associated with Black/White RRRs for OSS and with Hispanic/White RRRs for both OSS and ISS. Percent FRPL-qualifying students was negatively associated with Black/White RRRs for OSS.

Risk Differences

Results from the models predicting ISS and OSS RDs for Black and White students and for Hispanic and White students are shown in Table 5 along with bias-corrected 95% bootstrap confidence intervals. There were no significant differences between Montessori and non-Montessori RDs for either ISS or OSS across both the Black/White and Hispanic/White comparisons. In all RD models, the estimated coefficients for all racial demographic variables, percent of IDEA students, and percent FRPL-qualifying students were equal to or near zero.

Multilevel Models

Results from the multilevel negative binomial model predicting the ISS counts for each racial group, with random intercepts for schools and including racial group sizes as an offset, are presented in Table 6 along with bias-corrected 95% bootstrap confidence intervals for each coefficient. The estimated counts per 100 students in each racial group (interpretable as the rates) from this model are shown in Figure 1, accompanied by bias-corrected 95% bootstrap confidence intervals. The same results for OSS are presented in Table 6 and Figure 1.

Exponentiating the coefficients from the ISS model indicated that the estimated average ISS rate for Black students in non-Montessori schools was 2.45 times higher than for White students, which was a statistically significant difference. There was not a significant difference between Hispanic and White student average ISS rates in non-Montessori schools (the estimated average ISS rate for Hispanic students was 0.87 times that of White students). The Montessori coefficient indicated that the average ISS rate for White Montessori students was 0.43 times the average rate for their White non-Montessori school peers, and this was a significant difference. The disparity between Black and White student ISS rates was slightly, but non-significantly, larger at Montessori schools than non-Montessori schools (in Montessori schools, the estimated Black ISS rate was 3.07 times larger than the White ISS rate; it was 2.45 times larger in non-Montessori schools). The disparity in estimated ISS rates between Hispanic and White students did not significantly differ in Montessori schools compared to non-Montessori schools, although the direction of the disparity switched: The estimated ISS rate for Hispanic students in non-Montessori schools was 0.87 times that of the estimated rate for White students, but in Montessori schools, it was 1.30 times that of White students. These differences in disparities appear to be driven by lower average ISS rates for White students in Montessori schools relative to White students in non-Montessori schools: As shown in Figure 1, estimated average ISS rates for Hispanic students were nearly equivalent across school types, whereas the estimated average Black and White student ISS rates were lower in the Montessori schools than in the non-Montessori schools.

For OSS, the average estimated OSS rate for Black non-Montessori students was significantly higher—2.68 times higher—than the average rate for their White peers. Average estimated rates for Hispanic students in non-Montessori schools were also 1.10 times higher than for White non-Montessori students. The estimated average OSS rate for White students in Montessori schools was 0.43 times the rate for White students in non-Montessori schools, and this was a significant difference. The interaction term for Black students and Montessori schools (which, exponentiated, was 1.37) suggested that the relative rate of Black student OSS to White student OSS was slightly larger at Montessori schools than at non-Montessori schools, but we interpret this with caution because the bias-corrected 95% bootstrap confidence interval barely excluded zero. The confidence interval on the interaction term for Hispanic students and Montessori schools comfortably captured zero. As shown in Figure 1, estimated average OSS rates for each racial group were lower in Montessori than non-Montessori schools.

Discussion

This study introduces a new approach to analyzing discipline disproportionality, and it compares average overall disciplinary rates and discipline disproportionality in a propensity score matched sample of Title 1 Montessori and non-Montessori schools. For overall ISS and OSS rates, our results are consistent with results reported by Culclasure and colleagues (2018), who found lower suspension rates among Montessori students in their sample relative to the non-Montessori students. Montessori schools in this sample were associated with lower average ISS and OSS rates, resulting in an average of 66% and 50% as many students receiving ISS and OSS, respectively, in the Montessori schools in our sample. Montessori schools in this sample were also significantly more likely to give zero OSS than non-Montessori schools. Consistency across those results helps build confidence that Montessori schools yield lower suspension rates for Montessori students than their non-Montessori counterparts serving similar populations.

When comparing discipline disproportionality for Hispanic and White students, there were no significant differences in RRRs or RDs between the Montessori and non-Montessori schools. The mean RDs suggest that Hispanic students received ISS and OSS at roughly equivalent rates as White students in both Montessori and non-Montessori schools. These results are consistent with the multilevel model results: For both ISS and OSS, the interaction terms between Hispanic students and Montessori status suggests that the disparities between White and Hispanic ISS and OSS rates were not significantly different between school types. As can be seen in Figure 1, in both school types, the estimated ISS and OSS rates for White and Hispanic students were similar (though lower for both racial groups in the Montessori schools). These results are consistent with previous reports on suspension rates for Hispanic or Latino students (U.S. Department of Education, 2018).

The results comparing Black and White students are slightly more complex. Only one previous study (to our knowledge) has attempted to measure discipline disproportionality in Montessori schools (Brown & Steele, 2015); the average RRR between Black and White students for the 3 Montessori schools in that study was 2.61. Here, the average RRR was much higher: 4.54 and 5.45 for ISS and OSS respectively. Nationally, the average RRR for Black and White students is around 2 or 3 (U.S. Department of Education, 2018), which is nearly identical to the RRR calculated for the non-Montessori schools in this sample. Based on the RRRs alone, these results suggest that discipline disproportionality between Black and White students was *worse* in Montessori schools than non-Montessori schools. However, as previously noted, RRRs tend to increase as overall suspension rates decrease (Curran, 2020), and the Montessori schools in this sample had significantly lower suspension rates than the non-Montessori schools, making them more likely to have inflated and misleading RRRs. Moreover, the average RDs calculated here for the Montessori and non-Montessori schools are not significantly different.

The results from the multilevel models help clarify these seemingly contradictory findings: As shown in Figure 1, average OSS rates for all three racial groups were lower in the Montessori schools than the non-Montessori schools. The results of the multilevel model suggest that the difference between Black and White student OSS rates was slightly larger in the

Montessori sample, but the confidence interval nearly includes zero, so we interpret this result with caution. These results suggest that Black students were suspended at higher average rates than White students in Montessori schools, and the disparity between White and Black student suspension rates was slightly larger in the Montessori schools. Critically, the difference in disparity across school types appears to be driven primarily by significantly *lower* OSS rates for White Montessori students as opposed to higher rates for Black Montessori students (see Figure 1). In other words, Montessori schools seem to be associated with dramatically lower OSS rates for White students, and less dramatically lower rates for Black students, so the disparity between White and Black students appears slightly larger than in the non-Montessori schools. However, it's important to keep in mind that Black students were still suspended at lower rates in Montessori schools relative to non-Montessori schools, which would likely be important to a Black parent deciding between Montessori and non-Montessori schooling; the Montessori schools might still be preferrable given the lower overall OSS rates even though Black students were still given OSS at higher rates than their White peers in Montessori schools. Interview data suggests that Black and Latino parents of Montessori students report feeling as though consequences for misbehavior at their children's school are appropriate (Golann et al., 2019). Presumably, the Black and Latino parents interviewed in that study would not have reported satisfaction with their children's schools' disciplinary practices if they observed racial bias in those practices.

It is also important to note that RRRs were only calculated in schools that had suspended at least one Black student and at least one White student, and the Montessori schools in this sample were significantly more likely to give zero suspensions, so they are underrepresented in the RRR comparisons. Moreover, the assumptions of the propensity score matching process do not hold as strongly for the subsamples for which RRRs were calculated, so results based on those measures might not be an appropriate comparison. A major benefit of using the multilevel model is that it allows inclusion of all schools, so the estimates are more representative of all the schools in the sample.

Ultimately, the seeming contradiction in the RRR and RD results supports recommendations to use overall disciplinary rates to contextualize RRRs and RDs, and to not use either measure in isolation (Curran, 2020; Larson et al., 2019; Petrosino et al., 2017). These results also showcase how multilevel modeling offers a more elegant and straightforward way to analyze discipline disparities without obscuring important base-rate information.

Strengths and Limitations

Measuring discipline disproportionality. This study presents a novel way to analyze discipline disproportionality. The limitations of the commonly used RRRs and RDs have been discussed in detail (Curran, 2020; Girvan et al., 2019; Larson et al., 2019; Nishioka et al., 2017; Petrosino et al., 2017), but a straightforward solution has not been forthcoming. The approach used here offers a reasonable solution to many of the issues raised by other researchers. This approach allows for the same intuitive interpretation of disparities as RRRs, but this approach is not prone to the same insensitivity or inflation issues as RRRs and RDs in schools with low suspension rates. This approach also allows for the inclusion of schools with zero suspensions, and it allows for comparisons across multiple racial groups in one model. Ultimately, the results from this study showcase how RRRs and RDs can lead to contradictory findings, and how multilevel modeling can be used to resolve these contradictions.

Although not applicable to this study, another significant benefit of measuring discipline disproportionality (or another form of racial disparity in an outcome of count incidences) using multilevel modeling is that one could also assess how racial disparities have changed over time by using longitudinal data and introducing time as a predictor, optionally with random slopes. With a time variable representing the time period during which data were collected (e.g., school year) and dummy variables for racial groups, one could run the following model predicting the proportion of students suspended in each racial group:

$$\log(suspension\ count)_{ij} = \eta_{ij} = \gamma_{00} + \beta_1 Time_{ij} + \beta_2 Hispanic_{ij} + \beta_3 Black_{ij} + \beta_4\left(Time_{ij} * Hispanic_{ij}\right) + \beta_5\left(Time_{ij} * Black_{ij}\right) + \log(racial\ group\ size)_{ij} + u_{0j}. \quad (7)$$

Assuming that White students are again made the reference racial group and that their suspension rates are lower than the other two racial groups, the interactions between time and the other racial groups would denote the average change in the size of disparities between that group and White students from one measurement to the next. A positive coefficient would signify a growing disparity; a negative coefficient would signify an equalizing trend. One could also add a three-way interaction between time, a racial group of interest, and an intervention to evaluate whether the disparities are changing at different rates based on intervention status. This offers an advantage over RRRs and RDs, which are relatively unreliable measures in longitudinal analysis; as described previously, they can often result in contradictory conclusions because of their relationship to base suspension rates (Curran, 2020; Girvan et al., 2019).

Self-selection bias. Separate from the overall multilevel modeling approach, interpreting the results of our specific demonstration requires consideration of potential self-selection on the part of the parents who did (or did not) chose to enroll their child in a Montessori school. These limitations may not apply to future research that uses this multilevel modeling approach to evaluate disciplinary interventions when random assignment is possible—for instance, in studies using student-level data from students who were admitted (or not admitted) to a Montessori school via a lottery admissions process.

To our knowledge, this study is the first to use propensity score matching to estimate the effect of Montessori schools on suspension rates for White, Black, and Hispanic students, and on discipline disproportionality between those same racial groups. Montessori schools are typically choice schools, so selection bias is an ongoing concern in studying differences between Montessori and non-Montessori student outcomes. To date, only two studies we know of have attempted to measure differences between Montessori and non-Montessori American parents that could explain subsequent differences in their children, and those studies revealed no significant differences (Dreyer & Rigler, 1969; Fleege et al., 1967). However, the possibility of such differences cannot be ruled out based on existing research, so it remains possible that differences in Montessori and non-Montessori students are related to differences in parents who self-select into those programs for their children. This study, and Montessori research more generally, would benefit from more research on Montessori parents' beliefs, parenting behaviors, and motivations for enrolling their children in Montessori schools. Propensity score matching was our choice for reducing potential self-selection bias in the estimates because other quasi-experimental techniques for causal inference and random assignment were not feasible or applicable.

For the purposes of this study, it is unclear how potential differences between Montessori and non-Montessori parents might have biased the results. Race and SES are both related to disciplinary outcomes, and public Montessori schools are more likely than conventional schools to be racially diverse and enroll economically advantaged students (Debs, 2016). However, the two groups of schools in this study were very similar in terms of racial demographics and the proportion of students qualifying for free/reduced price lunch (a proxy for SES), which lowers the likelihood that race or SES could be a meaningful source of bias. Empirical evidence on how characteristics of Montessori parents might relate to school disciplinary outcomes for their children is scarce, so it is unclear how parenting differences might have biased these results (if at all). It is therefore difficult to know whether the estimates here are more likely to be an overestimate or underestimate of the true effect of the Montessori curriculum. It is impossible with these data to verify that the two samples were equivalent on *all* variables related to disciplinary outcomes and selection into Montessori schools, but the propensity score matching process helped achieve balance on a number of important covariates, resulting in a less biased sample. At the very least, this study provides a rich descriptive analysis of disciplinary outcomes in Title 1 Montessori schools. Our results suggest that Montessori schools have a less punitive disciplinary climate than non-Montessori schools, although racial disparities remain.

Conclusions and Future Directions

For practitioners and school leaders, RRRs and RDs still offer measurements of discipline disproportionality that are easy to calculate and interpretable, so long as they are interpreted alongside overall disciplinary rates (Curran, 2020). However, for researchers, multilevel models like the ones used here offer a more elegant, reliable, and straightforward solution for comparing disparities between multiple racial groups. A consistent and standardized method for measuring discipline disproportionality will be critical as researchers continue to search for ways to reduce exclusionary disciplinary practices and discipline disproportionality, which is an urgent goal given the negative outcomes associated with exclusionary discipline (Lewis et al., 2010; Skiba et al., 2014). We believe that future research on discipline disproportionality should consider using multilevel models à la those presented here to avoid the issues inherent in more common measures. Our approach could be easily applied to compare discipline disproportionality between other alternative school systems or interventions and business-as-usual approaches.

For researchers hoping to understand which school characteristics are associated with fewer exclusionary discipline practices in general, research in Montessori schools may still be an informative place to start. Any of the characteristics of Montessori classrooms discussed in this paper—high student engagement, student self-determination, individualized learning, and strong classroom community—are potential causes of the lower average suspension rates we observed. Future research using student-level data to predict individual disciplinary sanctions (referrals, suspensions, detentions, etc.) from student reports of engagement, student-teacher relationships, and classroom community would help clarify the extent to which any of those variables explain Montessori disciplinary rates. Researchers could also make classroom observations of Montessori teachers to better understand how Montessori teachers interact with students exhibiting disruptive behavior and to determine what constitutes a vulnerable decision point in a Montessori context. Classroom observations of Montessori teachers' disciplinary practices might also reveal why lower overall disciplinary rates in Montessori schools do not always correspond to lower discipline disproportionality. In any case, disciplinary practices in Montessori classrooms warrant future research given the negative association between exclusionary discipline and a wide range of outcomes, which are all disproportionately experienced by students of color (Lewis et al., 2010; Skiba et al., 2014; Vincent et al., 2012).

Data Availability Statement

The data and code that support the findings of this study are available in a GitHub repository, Title1-Montessori, at https://github.com/leboeuf77/Title1-Montessori

References

Blake, J. J., Keith, V. M., Luo, W., Le, H., & Salter, P. (2017). The role of colorism in explaining African American females' suspension risk. *School Psychology Quarterly*, *32*(1), 118–130. https://doi.org/10.1037/spq0000173

Brooks M. E., Kristensen K., van Benthem K. J., Magnusson A., Berg C. W., Nielsen A., Skaug H. J., Maechler M., Bolker B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.32614/RJ-2017-066

Brown, K. E., & Steele, A. S. (2015). Racial discipline disproportionality in Montessori and traditional public schools: A comparative study using the relative rate index. *Journal of Montessori Research, 1*(1), 14–27. https://doi.org/10.17161/jomr.v1i1.4941

Culclasure, B., Fleming, D. J., Riga, G., & Sprogis, A. (2018). *An evaluation of Montessori education in South Carolina's public schools*. The Riley Institute at Furman University.

Curran, F. C. (2020). A matter of measurement: How different ways of measuring racial gaps in school discipline can yield drastically different conclusions about racial disparities in discipline. *Educational Researcher, 49*(5), 382–387. https://doi.org/10.3102/0013189X20923348

Debs, M. C. (2016). Racial and economic diversity in US public Montessori schools. *Journal of Montessori Research*, *2*(2), 15–34. https://doi.org/10.17161/jomr.v2i2.5848

Dreyer, A. S., & Rigler, D. (1969). Cognitive performance in Montessori and nursery school children. *The Journal of Educational Research*, *62*(9), 411–416. https://doi.org/10.1080/00220671.1969.10883885

Fleege, U. H., Black M., & Rackauskus, J. (1967). *Montessori pre-school education*

(ED017320). ERIC. https://files.eric.ed.gov/fulltext/ED017320.pdf

Fabelo, T., Thompson, M. D., Plotkin, M., Carmichael, D., Marchbanks, M. P., & Booth, E. A. (2011). *Breaking schools' rules: A statewide study of how school discipline relates to students' success and juvenile justice involvement*. Council of State Governments Justice Center.

Fenning, P., & Rose, J. (2007). Overrepresentation of African American students in exclusionary discipline: The role of school policy. *Urban Education, 42*(6), 536–559. https://doi.org/10.1177/0042085907305039

Girvan, E. J., Gion, C., McIntosh, K., & Smolkowski, K. (2017). The relative contribution of subjective office referrals to racial disproportionality in school discipline. *School Psychology Quarterly*, *32*(3), 392–404. https://doi.org/10.1037/spq0000178

Girvan, E. J., McIntosh, K., & Smolkowski, K. (2019). Tail, tusk, and trunk: What different metrics reveal about racial disproportionality in school discipline. *Educational Psychologist*, *54*(1), 40–59. https://doi.org/10.1080/00461520.2018.1537125

Golann, J. W., Debs, M., & Weiss, A. L. (2019). "To be strict on your own": Black and Latinx parents evaluate discipline in urban choice schools. *American Educational Research Journal*, *56*(5), 1896–1929. https://doi.org/10.3102/0002831219831972

Gomila, R. (2021). Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General, 150*(4), 700–709. https://doi.org/10.1037/xge0000920

Gregory, A., Ruzek, E. A., DeCoster, J., Mikami, A. Y., & Allen, J. P. (2019). Focused classroom coaching and widespread racial equity in school discipline. *AERA Open, 5*(4). https://doi.org/10.1177/2332858419897274

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, *2*(4), 405–420. https://doi.org/10.2307/1390693

Hannon, L., DeFina, R., & Bruch, S. (2013). The relationship between skin tone and school suspension for African Americans. *Race and Social Problems*, *5*(4), 281–295. https://doi.org/10.1007/s12552-013-9104-z

Harris, H., & Horst, S. J. (2016). A brief guide to decisions at each step of the propensity score matching process. *Practical Assessment, Research, and Evaluation*, *21*, Article 4. https://doi.org/10.7275/yq7r-4820

Hartig, F. (2022). DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models. R package version 0.4.6. https://CRAN.R-project.org/package=DHARMa

Ho D. E., Imani K., King G., Stuart E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, 42*(8), 1–28. http://doi.org/10.18637/jss.v042.i08

Larson, K. E., Bottiani, J. H., Pas, E. T., Kush, J. M., & Bradshaw, C. P. (2019). A multilevel analysis of racial discipline disproportionality: A focus on student perceptions of academic engagement and disciplinary environment. *Journal of School Psychology*, *77*, 152–167. https://doi.org/10.1016/j.jsp.2019.09.003

Lewis, C. W., Butler, B. R., Bonner, F. A., III, & Joubert, M. (2010). African American male discipline patterns and school district responses resulting impact on academic achievement: Implications for urban educators and policy makers. *Journal of African American Males in Education, 1*(1), 7–25.

Lillard, A. S. (2019). Shunned and admired: Montessori, self-determination, and a case for radical school reform. *Educational Psychology Review, 31*(4), 939–965. https://doi.org/10.1007/s10648-019-09483-3

Lillard, A. S., & Else-Quest, N. (2006). Evaluating Montessori education. *Science, 313*(5795), 1893–1894. https://doi.org/10.1126/science.1132362

Lillard, A. S., Heise, M. J., Richey, E. M., Tong, X., Hart, A., & Bray, P. M. (2017). Montessori preschool elevates and equalizes child outcomes: A longitudinal study. *Frontiers in Psychology, 8*(OCT), Article 1783. https://doi.org/10.3389/fpsyg.2017.01783

Losen, D. J., Hodson, C. L., Keith II, M. A., Morrison, K., & Belway, S. (2015). *Are we closing the school discipline gap?* The Civil Rights Project.

McIntosh, K., Girvan, E. J., Horner, R., & Smolkowski, K. (2014). Education not incarceration: A conceptual model for reducing racial and ethnic disproportionality in school discipline. *Journal of Applied Research on Children: Informing Policy for Children at Risk*, *5*(2), Article 4.

Montessori, M. (1912). *The Montessori method.* Frederick A. Stokes Company.

Morris, E. W., & Perry, B. L. (2017). Girls behaving badly? Race, gender, and subjective evaluation in the discipline of African American girls. *Sociology of Education*, *90*(2), 127–148. https://doi.org/10.1177%2F0038040717694876

Nishioka, V., Shigeoka, S., & Lolich, E. (2017). *School discipline data indicators: A guide for districts and schools* (REL 2017–240). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northwest.

Okonofua, J. A., Walton, G. M., & Eberhardt, J. L. (2016). A vicious cycle: A social-psychological account of extreme racial disparities in school discipline. *Perspectives on Psychological Science*, *11*(3), 381–398. https://doi.org/10.1177/1745691616635592

Petrosino, A., Fronius, T., Goold, C. C., Losen, D. J., & Turner, H. M. (2017). *Analyzing student-level disciplinary data: A guide for districts* (REL 2017–263). Regional Educational Laboratory Northeast & Islands.

Rathunde, K., & Csikszentmihalyi, M. (2005). Middle school students' motivation and quality of experience: A comparison of Montessori and traditional school environments. *American Journal of Education*, *111*(3), 341–371. https://doi.org/10.1086/428885

Resh, N. & Sabbagh, C. (2016). Justice and education. In C. Sabbagh and M. Schmitt (Eds.), *Handbook of Social Justice Theory and Research* (pp. 349-367). Springer.

Skiba, R. J., Arredondo, M. I., & Williams, N. T. (2014). More than a metaphor: The contribution of exclusionary discipline to a school-to-prison pipeline. *Equity & Excellence in Education*, *47*(4), 546–564. https://doi.org/10.1080/10665684.2014.958965

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*(4), 317–342. https://doi.org/10.1023/A:1021320817372

Smolkowski, K., Girvan, E. J., McIntosh, K., Nese, R. N., & Horner, R. H. (2016). Vulnerable decision points for disproportionate office discipline referrals: Comparisons of discipline for African American and White elementary school students. *Behavioral Disorders*, *41*(4), 178–195. https://doi.org/10.17988/bedi-41-04-178-195.1

Steiner, P. M., Cook, T. D., Li, W., & Clark, M. H. (2015). Bias reduction in quasi-experiments with little selection theory but many covariates. *Journal of Research on Educational Effectiveness*, *8*(4), 552–576. https://doi.org/10.1080/19345747.2014.978058

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1–21. https://doi.org/10.1214/09-STS313

U.S. Department of Education. (2018). 2015-2016 *Civil rights data collection: School climate and safety*. Office for Civil Rights. https:// www2.ed.gov/about/offices/list/ocr/docs/school-climate-and-safety.pdf.

Vincent, C. G., Sprague, J. R., & Tobin, T. J. (2012). Exclusionary discipline practices across students' racial/ethnic backgrounds and disability status: Findings from the Pacific Northwest. *Education and Treatment of Children*, *35*(4), 585–601. http://doi.org/10.1353/etc.2012.0025

Whitescarver, K., & Cossentino, J. (2007). Lessons from the periphery: The role of dispositions in Montessori teacher training. *Journal of Educational Controversy*, *2*(2), Article 11.

Table 1

Sample Demographics

	Pre-matching			Post-matching				
	Non-Montessori (*N* = 629)			Montessori (*N* = 125)		Non-Montessori (*N* = 125)		
	M	*SD*	*d*	*M*	*SD*	*M*	*SD*	*d*
Total students	561.64	408.88	-0.45***	388.29	235.44	413.34	210.54	-0.11
% Black	25.00	23.20	-0.04	24.14	23.30	26.93	25.01	-0.12
% Hispanic	34.31	27.77	-0.41***	23.44	21.40	23.95	20.88	-0.02
% White	30.68	24.44	0.47***	42.18	24.64	38.13	26.82	0.16
% IDEA	16.93	16.87	-0.37***	11.17	7.14	11.77	6.38	-0.09
% 504	2.12	2.56	-0.31**	2.96	3.33	2.40	2.75	0.19
% FRPL	66.44	22.13	-0.74***	49.90	23.45	54.60	25.86	-0.19
Preschool	38.79	-	0.57***	66.40	-	58.40	-	0.17
Kindergarten	61.05	-	0.69***	92.80	-	88.80	-	0.14
First	62.16	-	0.67***	92.80	-	88.80	-	0.14
Second	61.84	-	0.66***	92.00	-	88.00	-	0.13
Third	61.84	-	0.64***	91.20	-	86.40	-	0.15
Fourth	60.89	-	0.56***	87.20	-	84.00	-	0.09
Fifth	59.94	-	0.53***	84.80	-	83.20	-	0.04
Sixth	37.20	-	0.66***	68.80	-	50.40	-	0.38**
Seventh	29.57	-	0.38***	47.20	-	40.80	-	0.13
Eighth	29.73	-	0.29**	43.20	-	38.40	-	0.09
Ninth	21.78	-	-0.37***	7.20	-	11.20	-	-0.14
Tenth	21.78	-	-0.42***	5.60	-	9.60	-	-0.15
Eleventh	21.78	-	-0.42***	5.60	-	8.00	-	-0.09
Twelfth	21.30	-	0.39***	6.40	-	8.00	-	-0.06
Magnet	17.97	-	0.37***	32.80	-	39.20	-	-0.13
Charter	13.51	-	0.90***	47.20	-	37.60	-	0.19

Note. IDEA = students with disabilities covered by the Individuals with Disabilities Education Act. 504 = students with disabilities covered by Section 504 of the Rehabilitation Act. FRLP = students who qualify for free/reduced price lunch. *D* = standardized mean difference. After propensity score matching, the only statistically significant standardized mean difference between the two groups was in the proportion of schools that offered sixth grade.

Table 2

Descriptive Statistics

	ISS					OSS				
	Overall rates	Black and White students		Hispanic and White students		Overall rates	Black and White students		Hispanic and White students	
		RRR	RD	RRR	RD		RRR	RD	RRR	RD
Montessori	0.02	4.54	0.02	2.05	0.00	0.03	5.45	0.04	2.19	0.01
	(0.05)	(3.80)	(0.06)	(1.98)	(0.02)	(0.05)	(4.79)	(0.07)	(1.90)	(0.04)
Non-Montessori	0.03	2.68	0.02	1.99	-0.01	0.06	3.21	0.05	1.85	0.00
	(0.09)	(2.38)	(0.06)	(3.01)	(0.08)	(0.08)	(2.67)	(0.08)	(2.13)	(0.05)
# of M. (non-M.) schools	125 (125)	27 (37)	102 (93)	18 (30)	104 (101)	125 (125)	43 (53)	102 (93)	38 (43)	104 (101)

Note. Mean (standard deviation). RRR = relative rate ratio, which could only be calculated for schools which gave at least one of the relevant disciplinary sanctions to a student in both racial groups. Risk differences only calculated for schools which had greater than or equal to 5% of whichever racial groups were involved in a given comparison.

Table 3

Logistic Regression Results Predicting Zero Suspensions and Negative Binomial Regression Results Predicting Overall Suspension Counts

	Logistic regression models		Negative binomial models	
	Zero ISS	Zero OSS	ISS	OSS
Intercept	0.426	-0.229	-7.117	-3.986
	(1.297)	(1.677)	[-9.707, -4.545]	[-5.428, -2.411]
Montessori	0.554*	0.624*	-0.472	-0.520
	(0.270)	(0.318)	[-1.049, 0.057]	[-0.865, -0.190]
# of students	-0.001*	-0.002*	--	--
	(0.001)	(0.001)	--	--
Charter	0.260	0.591	0.266	0.107
	(0.369)	(0.453)	[-0.333, 1.007]	[-0.300, 0.564]
Magnet	-0.163	-0.039	0.282	-0.263
	(0.376)	(0.483)	[-0.427, 0.900]	[-0.710, 0.218]
% Black	-0.000	0.002	0.026	0.010
	(0.013)	(0.017)	[0.004, 0.060]	[-0.005, 0.024]
% White	-0.003	-0.005	0.015	-0.006
	(0.014)	(0.018)	[-0.012, 0.043]	[-0.020, 0.010]
% Hispanic	0.007	0.028	0.006	-0.030
	(0.013)	(0.017)	[-0.021, 0.038]	[-0.045, -0.015]
% IDEA	-0.063**	-0.035	0.092	0.053
	(0.023)	(0.027)	[0.042, 0.149]	[0.026, 0.077]
% FRPL	0.005	-0.017*	0.011	0.012
	(0.007)	(0.008)	[-0.004, 0.029]	[0.004, 0.021]
Observations	250	250	250	250
Log Likelihood	-158.913	-124.589	--	--

Note. Log-odds coefficients and standard errors shown for logistic regression models; $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Log-counts coefficients and bias-corrected 95% bootstrap confidence intervals shown for the negative binomial models.

Table 4

Relative Rate Ratios

	Black and White students				Hispanic and White students			
	ISS		OSS		ISS		OSS	
	β	CI	*β*	CI	*β*	CI	*β*	CI
Intercept	1.227	[-2.421, 6.050]	3.154	[-0.570, 6.530]	1.802	[-5.357, 5.596]	3.522	[0.879, 6.767]
Montessori	1.531	[0.249, 2.973]	2.213	[0.734, 3.743]	-0.443	[-2.545, 1.022]	0.361	[-0.530, 1.170]
Charter	0.762	[-1.746, 3.339]	-1.888	[-3.651, -0.290]	-1.688	[-3.553, -0.119]	-1.819	[-2.998, -0.699]
Magnet	0.604	[-1.471, 2.621]	0.669	[-1.597, 2.856]	-0.508	[-2.521, 1.405]	-0.743	[-2.016, 0.409]
% Black	-0.030	[-0.089, 0.010]	-0.003	[-0.046, 0.036]	--	--	--	--
% Hispanic	--	--	--	--	-0.026	[-0.062, 0.011]	-0.014	[-0.037, 0.007]
% White	0.029	[-0.032, 0.080]	0.040	[-0.012, 0.095]	0.025	[-0.012, 0.077]	0.012	[-0.013, 0.034]
% IDEA	0.127	[-0.008, 0.299]	0.048	[-0.069, 0.198]	0.059	[-0.051, 0.391]	0.012	[-0.042, 0.081]
% FRPL	-0.006	[-0.046, 0.033]	-0.029	[-0.059, -0.001]	0.004	[-0.027, 0.037]	-0.020	[-0.051, 0.007]
# of M. (non-M.) schools	27 (37)	--	43 (53)	--	18 (30)	--	38 (43)	--

Note. CI = bias-corrected 95% bootstrap confidence intervals.

Table 5

Risk Differences

	Black and White students				Hispanic and White students			
	ISS		OSS		ISS		OSS	
	β	CI	*β*	CI	*β*	CI	*β*	CI
Intercept	0.005	[-0.051, 0.068]	0.006	[-0.053, 0.061]	-0.061	[-0.208, 0.022]	-0.034	[-0.116, 0.017]
Montessori	0.001	[-0.015, 0.017]	-0.016	[-0.038, 0.005]	0.008	[-0.004, 0.029]	0.005	[-0.006, 0.018]
Charter	-0.016	[-0.049, 0.009]	-0.009	[-0.039, 0.023]	-0.022	[-0.058, -0.002]	-0.012	[-0.033, 0.009]
Magnet	-0.008	[-0.048, 0.019]	-0.012	[-0.041, 0.018]	0.000	[-0.017, 0.025]	-0.001	[-0.019, 0.019]
% Black	0.000	[-0.001, 0.001]	0.000	[0.000, 0.001]	--	--	--	--
% Hispanic	--	--	--	--	0.000	[0.000, 0.002]	0.000	[0.000, 0.001]
% White	0.000	[-0.001, 0.001]	0.000	[0.000, 0.001]	0.001	[0.000, 0.003]	0.000	[0.000, 0.001]
% IDEA	0.001	[-0.001, 0.002]	0.001	[-0.001, 0.003]	-0.001	[-0.004, 0.001]	0.001	[-0.001, 0.002]
% FRPL	0.000	[0.000, 0.001]	0.000	[0.000, 0.001]	0.001	[0.000, 0.002]	0.000	[0.000, 0.001]
# of M. (non-M.) schools	102 (93)	--	102 (93)	--	104 (101)	--	104 (101)	--

Note. CI = bias-corrected 95% bootstrap confidence intervals.

Table 6

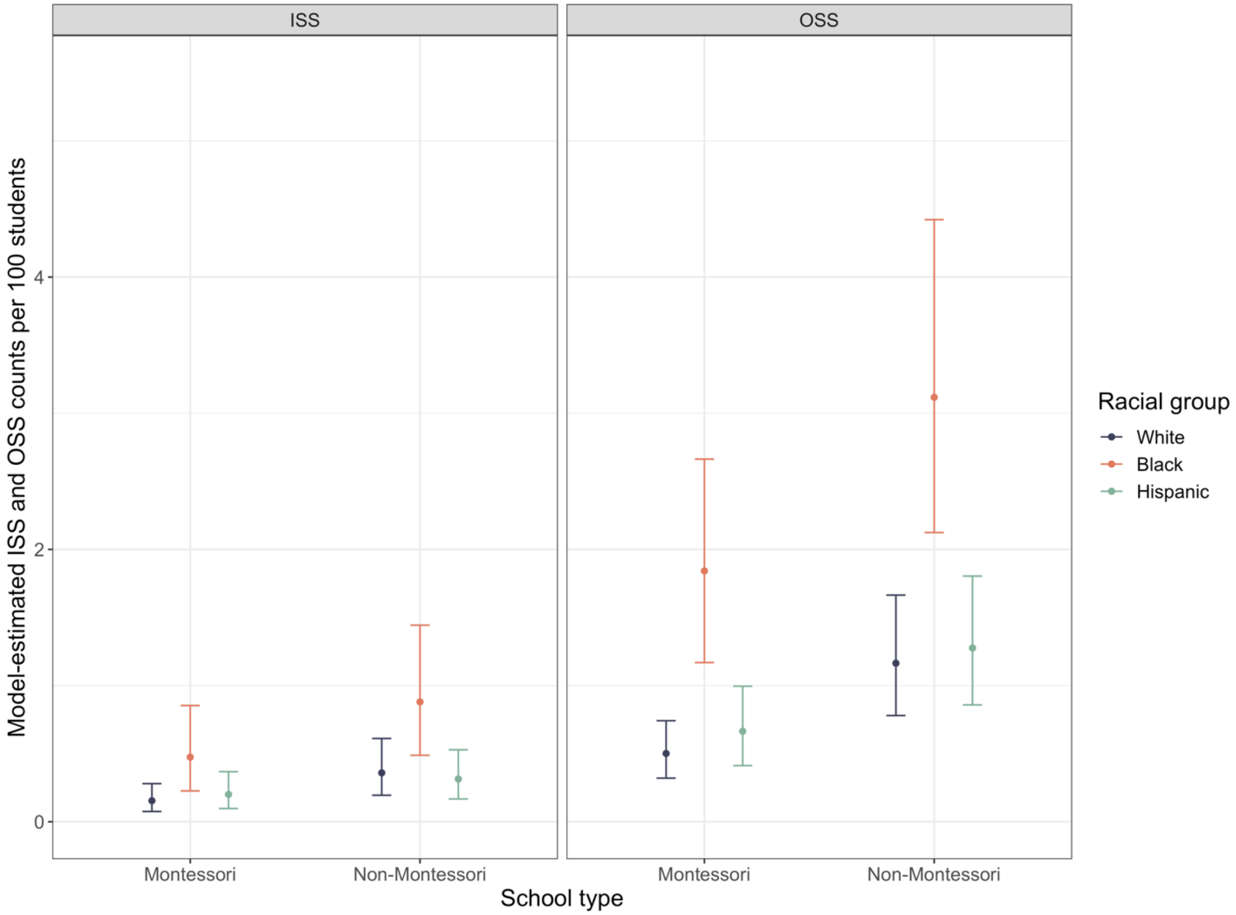
Coefficients and Bootstrap (5000×) Bias-Corrected 95% Confidence Intervals for Multilevel Negative Binomial Models Predicting ISS and OSS Counts

	ISS		OSS	
	β	CI	*β*	CI
Intercept	-5.629	[-6.241, -5.097]	-4.453	[-4.854, -4.096]
Black	0.896	[0.682, 1.118]	0.985	[0.809, 1.185]
Hispanic	-0.134	[-0.518, 0.156]	0.092	[-0.115, 0.282]
Montessori	-0.844	[-1.620, -0.113]	-0.842	[-1.371, -0.343]
Black × Montessori	0.226	[-0.152, 0.560]	0.316	[0.004, 0.622]
Hispanic × Montessori	0.396	[-0.023, 0.863]	0.189	[-0.149, 0.525]
N_{schools}	250		250	
$N_{\text{observations}}$	738		738	
Var(μ_{0j})	5.43		2.90	

Note. Coefficients are in terms of log counts. CI = bias-corrected 95% bootstrap confidence intervals. We used the glmmTMB package (v1.1.4; Brooks et al., 2017) in R to estimate the multilevel negative binomial models. We received model convergence warnings on 10 out of 5,000 bootstrap iterations (0.2%) for the ISS model and on one out of 5,000 bootstrap iterations (0.02%) for the OSS model. The data used to estimate the multilevel models were clustered within schools, so we bootstrapped clusters to preserve the nested data structure in the bootstrap replicates. Further, although most schools had ISS and OSS rates for all three racial groups, twelve had rates for only two (if there were no Black students or no Hispanic students at the school). As such, the exact size of the bootstrap replicates varied slightly. For the ISS model, the N ranged from 722 to 747 (median = 738, IQR = 4); for the OSS model, the N ranged from 723 to 748 (median = 738, interquartile range = 4).

Figure 1

Estimated ISS and OSS Counts per 100 Students for Each Racial Group Across Montessori and non-Montessori Schools



Note. Error bars represent bias-corrected 95% bootstrap confidence intervals.