A New Method of Measuring Discipline Disproportionality Applied to Title 1 Montessori and Non-Montessori Schools

Lee P. LeBoeuf

Dayton, OH

B.A., Ohio Wesleyan University, 2017

A Thesis presented to the Graduate Faculty

of the University of Virginia in Candidacy for the Degree of

Master of Arts

Department of Psychology

University of Virginia

December, 2021 Degree **will be Conferred**

Abstract

Research on discipline disproportionality has been hindered by the lack of a reliable measure. In this study, we propose a novel way to measure discipline disproportionality using multilevel modeling, and we demonstrate how it resolves many of the issues inherent in more common methods—namely, relative rate ratios and risk differences. One previous study suggests there is less discipline disproportionality in Montessori schools, so we used our new method, along with more common methods, to compare discipline disproportionality in a sample of Title 1 Montessori and non-Montessori schools that was identified using propensity score matching. Results showed that overall suspension rates were significantly lower in Montessori schools but that discipline disproportionality was similar across school settings.

A New Method of Measuring Discipline Disproportionality Applied to Montessori and Non-Montessori Title 1 Schools

In the United States, Black students are two to three times more likely than White students to receive an out-of-school suspension (U.S. Department of Education, 2018). This phenomenon is termed *discipline disproportionality*. Discipline disproportionality 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). Even for very similar behaviors, students of color tend to receive harsher punishments than 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). Teachers are most likely to act on racial biases and stereotypes when evaluating student behavior when they are stressed or have competing demands on their time (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). Moreover, exclusionary disciplinary practices—like suspension—are risk factors for negative long-term outcomes like school drop-out and involvement in the criminal justice system (Skiba et al., 2014). Eliminating discipline disparities is critical for an equitable education system (Resh & Sabbagh, 2016).

Researchers' understanding of discipline disparities, and 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 method of measuring racial disparities in school discipline using multilevel modeling. To demonstrate the utility of this new method, we compared disciplinary data in Montessori and non-Montessori Title 1 schools. We used propensity score matching to identify our sample, and we compared overall suspension rates for White, Black, and Hispanic students at those schools, 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, 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 frequency of a disciplinary infraction across two student subgroups. The RD is simply the difference in the proportion of students in two subgroups who received a disciplinary infraction. For example, to calculate the RRR for Black and White students receiving out-of-school suspensions, one divides the proportion of Black students who received an out-of-school suspension by the proportion of White students who received an out-of-school suspension. 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. Following the same example, to calculate the RD, one would subtract the proportion of White students who received an out-of-school suspension from the proportion of Black students who received an out-of-school suspension.

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 complicates understanding of the overall disciplinary climate in a school, and it means that neither measure provides information about the number of students impacted by a disparity (Curran, 2020). 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 in the previous example). RRRs are 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 reveals little, as zero divided by 0.01 is equivalent to zero divided by 0.90. 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 multiple racial groups at once would need to run separate models estimating the RRR or RD for each pair of racial groups, which is cumbersome when examining diverse schools. Finally, a major limitation of both measures is that they are unreliable over time, muddling longitudinal analyses (Curran, 2020; Girvan et al., 2019).

Here, we show how multilevel modeling can be used to measure discipline disproportionality, and we demonstrate how it addresses many of these issues. In this method, one calculates the proportion of students suspended within each racial group and treats those values as repeated measures within a school. One then creates a dummy variable for racial group (White, Black, and so on.) and constructs the following model (here using White students as the reference category to compare with Hispanic and Black students):

$$ProportionSuspended_{ij} = \beta_0 + \beta_1 Hispanic_{ij} + \beta_2 Black_{ij} + \mu_{0j} + \varepsilon_{ij} \quad (1)$$

Where i = measurements within schools;

j = individual schools;

μ_{0j} = the random intercept for the jth school

In this model, the coefficients for Hispanic and Black denote the average difference between suspension rates for each of those racial groups and White students. Therefore, one can calculate the predicted suspension rates for Black and Hispanic students by adding the corresponding

coefficient to the intercept (which represents the estimated suspension rate for White students). This approach offers an advantage over RRRs and RDs by making the proportion of students affected by a given disparity more apparent. One could also add any covariates of interest, like the percent of free/reduced price lunch–qualifying students in the school, to control for school characteristics related to discipline disproportionality.

To test whether disparities in the proportion of students suspended from each racial group differ across levels of an intervention (like whether a school is a Montessori school or a conventional school), one adds an interaction term between each racial group and a dummy variable indicating intervention status (in the example, Montessori classification):

$$ProportionSuspended_{ij} = \beta_0 + \beta_1 Montessori_j + \beta_2 Hispanic_{ij} + \beta_3 Black_{ij} +$$

$$\beta_4(Montessori_j * Hispanic_{ij}) + \beta_5(Montessori_j * Black_{ij}) + \mu_{0j} + \varepsilon_{ij} \quad (2)$$

In this model, the coefficients for Hispanic and Black are again the estimated differences between suspension rates between each of those racial groups and White students, but in conventional schools only. The coefficients for the interaction terms between each racial groups and Montessori classification are the estimated differences in disparities in suspension rates across school types. For example, if the data used in the model are percentages, a value of 2.5 for the Black coefficient and a value of –1.0 for the Black*Montessori interaction would indicate that Black students are suspended at 2.5% higher rates than White students in conventional schools and at 1.5% higher rates than White students in Montessori schools.

The primary benefit of using this multilevel model is that it resolves many of the issues presented by RRRs and RDs. First, unlike 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; further, the model makes it easy to calculate suspension rates for each

racial group, allowing for estimation of the magnitude of the issue (i.e., how many students are affected by a disparity). 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 rates for any racial group. Additionally, the user is no longer limited to evaluating pairwise comparisons in each model (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); therefore, if the disciplinary climate of Montessori schools is similar to national trends, 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 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 to date 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 support and instruction. During these observations, educators "must not start, for example, from any dogmatic ideas which we 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.

If a child's behavior is disruptive to another student's learning, Montessori teachers are trained to help the child find personally interesting work to do. Should that fail, Montessori wrote that the instructor should give the child space in the classroom where they can see other students working, which she believed would help the child realize the benefits of contributing to the classroom community and motivate them to participate productively (Montessori, 1912).

This approach differs substantially from typical exclusionary disciplinary practices, where students are removed from the classroom entirely, and it might lead to lower suspension rates.

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 warrant suspension.

Discipline Disproportionality and Montessori

As already mentioned, students of color receive referrals 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 much more likely to make racially biased discipline-related 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 in itself might create stress and result in biased discipline 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 lead to fewer Vulnerable Decision Points throughout a Montessori teacher's day and thus reduce the likelihood that racial bias would influence disciplinary decisions. If this is true, 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 correspond to 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 failed to measure 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). 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 has measured discipline disproportionality in Montessori schools. Brown and Steele (2015) used RRRs to compare discipline disproportionality in three public Montessori schools to that of 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 significantly 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 like school size and average socioeconomic status. The authors acknowledged that these characteristics could be important confounding variables, but they did not account these characteristics in their estimation of discipline disproportionality differences. Propensity score matching Montessori schools to non-Montessori schools based on school-level characteristics, and checking for balance on variables related to disciplinary outcomes, would likely result in a less biased sample for reasons described later.

The Current Study

This study used propensity score matching to estimate the effect of the Montessori method on overall in-school-suspension (ISS) and out-of-school-suspension (OSS) rates, as well as discipline disproportionality in Title 1 schools. Discipline disproportionality was measured in three ways: RRRs, RDs, and a multilevel model as in equation (2). These measures were used to compare disparities between Black and White students and between Hispanic and White students. Based on the theoretical reasons discussed above, we expected 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 both of 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) the raw counts of students who received an ISS, students who received one OSS, and students who received more than one OSS. For the purposes of this study, the two categories for OSS were summed to calculate the overall OSS rates. 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 qualify for free/reduced price lunch (FRPL), and binary variables indicating magnet or charter status.

Propensity Score Matching

To reduce the potential for self-selection to bias results here, we used propensity score matching to identify our comparison sample. Propensity score matching (PSM) can be an effective way to reduce bias in treatment-effect estimates when treatment is self-selected or non-randomly assigned (Gu & Rosenbaum, 1993; Harris & Horst, 2016; Steiner et al., 2015; Stuart, 2010). PSM involves estimating the probability of each unit of observation receiving the treatment (i.e., each unit's propensity score) based on selected covariates. From there, treated units are "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 in the results, but it can help ensure that the control group is similar or equivalent to the treatment group on known and 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 [blinded for review] laboratory to confirm that the schools were indeed Montessori schools. To allow for adequate comparison across racial groups, we limited the sample to schools that had 5% or more White students and 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). 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 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, 647 schools remained in the pool of potential matches. Two additional schools were removed from the pool of potential matches because they were deemed 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 linear propensity scores predicting treatment condition (i.e., whether a school is a Montessori school) using the following equation:

$$D_i = \beta_0 + \theta' X_i + e_i \quad (3)$$

where:

D_i = log-odds of a school being a Montessori school

θ' = a vector of coefficients

X_i = 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 provided school services through the Individuals with Disabilities Education Act (IDEA); the proportion of students with disabilities provided 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.

With each iteration, we tried identifying a match for each Montessori school using both optimal pair and nearest neighbor matching with the *MatchIt* R package (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 final 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 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).

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 which offered sixth grade.

Once the sample was finalized, we retrieved disciplinary data from the CRDC. The CRDC tabulates school-level disciplinary data as count data. We converted the raw count data into rates for each school by dividing the number of disciplinary events in each racial group 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 there were very few disciplinary events for students with other racial identities. For simplicity, the overall disciplinary rates reported here refer to the proportion of the total number of Black, White, and Hispanic students suspended at each school.

Analysis Plan

This study sought to compare the 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 test 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 (vs. the alternative binary outcomes of giving at least one ISS or one OSS):

$$For\ ISS: \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 Montessori_i + \theta' X_i + \varepsilon_i \quad (4a)$$

$$For\ OSS: \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 Montessori_i + \theta' X_i + \varepsilon_i \quad (4b)$$

where:

θ' = a vector of coefficients

X_i = a vector of the following school characteristics: number of students in the school, charter and magnet classification, the proportion of Black, Hispanic, and White students, and the proportion of students with disabilities (IDEA) and who qualify for FRPL

π_i = probability of giving zero ISS or the probability of giving zero OSS

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 other school characteristics included as covariates.

To estimate differences in overall ISS and OSS rates between Montessori and non-Montessori schools, we ran the following linear models:

$$ISSrate_i = \beta_0 + \beta_1 Montessori_i + \theta' X_i + \varepsilon_i \quad (5a)$$

$$OSSrate_i = \beta_0 + \beta_1 Montessori_i + \theta' X_i + \varepsilon_i \quad (5b)$$

θ' = a vector of coefficients

X_i = a vector of the following school characteristics: charter and magnet classification, the proportion of Black, Hispanic, and White students, and the proportion of students with disabilities (IDEA) and who qualify for FRPL.

For models 5a and 5b, Shapiro-Wilk tests and Bartlett tests revealed that both the error-normality assumption and the homogeneity of variance assumption between the Montessori and non-Montessori schools were not met. Therefore, we bootstrapped both models 1,000 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 rates of ISS and OSS between racial groups and school types using multilevel models as exemplified in equation (2).

To examine discipline disproportionality using RRRs and RDs, we first subset the sample to include just schools for which meaningful RRRs and RDs 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. Risk differences for ISS and OSS were calculated for schools that gave at least one of the relevant disciplinary sanctions to White students. For both metrics, 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 RDs and RRRs for these schools 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. However, forming the subsets was a necessary given the constraints of using RRRs and RDs.

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

$$For\ ISS\text{: } (RRR\ or\ RD)_i = \beta_0 + \beta_1 Montessori_i + \ \theta' X_i + \varepsilon_i \quad (6a)$$

$$For\ OSS\text{: } (RRR\ or\ RD)_i = \beta_0 + \beta_1 Montessori_i + \ \theta' X_i + \varepsilon_i \quad (6b)$$

θ' = a vector of coefficients

X_i = a vector of the following school characteristics: charter and magnet classification, the proportion of Black, Hispanic, and White students, and the proportion of students with disabilities (IDEA) and who qualify for FRPL.

In all, models 6a and 6b were used to predict each of the following outcomes: Black and White student ISS RD, Hispanic and White student ISS RD, Black and White student OSS RD, and Hispanic and White student OSS RD. Once again, the error normality and homogeneity of variance assumptions were violated in all models. In response, for each model, we bootstrapped the sample 1,000 times, re-estimated the model on each resample, and calculated bias-corrected 95% bootstrapped confidence intervals around the coefficients.

Finally, after calculating models for RRRs and RDs, we then calculated multilevel models as proposed to estimate differences in discipline disproportionality across school types. We ran the models below:

$$ISSrate_{ij} = \beta_0 + \beta_1 Montessori_j + \beta_2 Hispanic_{ij} + \beta_3 Black_{ij} + \beta_4(Montessori_j * Hispanic_{ij}) + \beta_5(Montessori_j * Black_{ij}) + \mu_{0j} + \varepsilon_{ij} \quad (7a)$$

$$OSSrate_{ij} = \beta_0 + \beta_1 Montessori_j + \beta_2 Hispanic_{ij} + \beta_3 Black_{ij} + \beta_4(Montessori_j * Hispanic_{ij}) + \beta_5(Montessori_j * Black_{ij}) + \mu_{0j} + \varepsilon_{ij} \quad (7b)$$

Once again, Shapiro-Wilk tests indicated that errors were non-normal for each of the models, so for each model, we bootstrapped the sample 1000 times, recalculated the given model on each resample, and generated bias-corrected 95% confidence intervals around each coefficient. In addition to the benefits already discussed, the multilevel models offered the benefit of being able to account for the entire sample in the estimations for ISS and OSS rates for each racial group. For this reason, the results from this model should offer the most accurate comparison of Montessori and non-Montessori disciplinary outcomes given the propensity score matching

process. We also ran versions of models 7a and 7b with the same control variables used in previous models (charter and magnet classification, the proportion of Black, Hispanic, and White students, and the proportions of students with disabilities (IDEA) and who qualify for FRPL but did not observe meaningful differences in any coefficients. For simplicity, we report the results from models 7a and 7b, which did not include those control variables.

Results

Overall Disciplinary Rates

Descriptive statistics—overall ISS/OSS rates and RRRs/RDs for each disciplinary outcome—for the Montessori and non-Montessori schools are shown in Table 2. Overall, the Montessori schools had 66% as many ISS and half as many OSS as the conventional schools. Results from the logit model predicting the log odds of whether a school gave zero ISS or OSS are shown in Table 3.

Exponentiating the coefficients for Montessori classification indicates that, holding all other variables in the model constant, Montessori schools were 1.74 times more likely to give zero ISS and 1.87 times more likely to give zero OSS ($p = 0.04$ for ISS and $p = 0.05$ for OSS). Results from the models estimating overall ISS and OSS rates are shown in Table 2 along with bias-corrected 95% confidence intervals. The results suggest that Montessori schools gave 1.0% fewer of their students an ISS (although the 95% confidence interval includes zero). Montessori schools also gave 2.3% fewer of their students an OSS, and the 95% confidence interval suggests that this is a significant difference. The only other statistically significant covariates were charter school classification and percent of IDEA students, both of which were positively associated with ISS and OSS rates.

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	Risk Dif.	RRR	Risk Dif.		RRR	Risk Dif.	RRR	Risk Dif.
Montessori	0.02	4.54	0.06	2.05	0.00	0.03	5.45	0.06	2.19	0.01
	(0.05)	(3.80)	(0.09)	(1.98)	(0.04)	(0.05)	(4.79)	(0.08)	(1.90)	(0.05)
Non-Montessori	0.03	2.68	0.04	1.99	-0.02	0.06	3.21	0.07	1.85	0.00
	(0.09)	(2.38)	(0.09)	(3.01)	(0.13)	(0.08)	(2.67)	(0.09)	(2.13)	(0.06)
# of M. (non-M.) schools	125 (125)	27 (37)	33 (41)	18 (30)	33 (46)	125 (125)	43 (53)	49 (58)	38 (43)	52 (59)

Note. Mean (standard deviation). RRR = relative rate ratio, which can 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 gave at least one of the relevant disciplinary sanctions to a White student, and which had greater than or equal to 5% of whichever other racial group was involved in the comparison.

Table 3

Logit Model Predicting Zero Suspensions and Bootstrapping Results Predicting Overall Disciplinary Rates

	Logit Model		Overall Disciplinary Rates	
	Zero ISS	Zero OSS	ISS	OSS
Montessori	0.554*	0.624*	-0.009	-0.023
	(0.270)	(0.318)	[-0.027, 0.008]	[-0.039, -0.008]
# of students	-0.001*	-0.002*	--	--
	(0.001)	(0.001)	--	--
Charter	0.260	0.591	0.024	0.022
	(0.369)	(0.453)	[0.001, 0.052]	[0.004, 0.043]
Magnet	-0.163	-0.039	0.011	-0.005
	(0.376)	(0.483)	[-0.015, 0.034]	[-0.025, 0.014]
% Black	-0.0003	0.002	0.000	-0.001
	(0.013)	(0.017)	[-0.001, 0.001]	[-0.002, 0.000]
% White	-0.003	-0.005	0.000	-0.002
	(0.014)	(0.018)	[-0.002, 0.001]	[-0.003, -0.001]
% Hispanic	0.007	0.028	0.000	-0.002
	(0.013)	(0.017)	[-0.002, 0.000]	[-0.002, -0.001]
% IDEA	-0.063**	-0.035	0.003	0.003
	(0.023)	(0.027)	[0.001, 0.006]	[0.001, 0.004]
% FRPL	0.005	-0.017*	0.000	0.003
	(0.007)	(0.008)	[-0.001, 0.001]	[0.000, 0.000]
Constant	0.426	-0.229	0.004	0.117
	(1.297)	(1.677)		
Observations	250	250	250	250
Log Likelihood	-158.913	-124.589	--	--
Akaike Inf. Crit.	337.826	269.178	--	--

Note. Coefficients and SEs given for logit model; *p**p***p<0.001. Betas and bias-corrected 95% confidence intervals shown for the overall disciplinary rates results.

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% confidence intervals. Unlike in previous research, RRRs for Black and White students were higher on average in the Montessori schools than in the conventional 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. None of the other covariates were statistically significant predictors of RRRs for Black/White or Hispanic/White comparisons, except for charter school classification, which was negatively associated with Black/White OSS RRRs, and Hispanic/White OSS and ISS RRRS.

Risk Differences

Bootstrapping results 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% confidence intervals. Surprisingly, although the RRRs for ISS and OSS were significantly higher in the Montessori schools when comparing Black and White students, the RDs were not significantly different between the two school systems for either disciplinary event. For the Hispanic and White student RDs, there was no significant difference between Montessori and non-Montessori schools for ISS or OSS. Finally, the estimated coefficients for all racial demographic variables, percent of IDEA students, and percent of FRPL-qualifying students were equal to or near zero.

Table 4
Relative Rate Ratios

	Black and White Students				Hispanic and White students			
	ISS		OSS		ISS		OSS	
	β	95% CI	*β*	95% CI	*β*	95% CI	*β*	95% CI
Montessori	1.531	[0.354, 2.972]	2.213	[0.733, 3.764]	-0.443	[-2.687, 0.993]	0.361	[-0.502, 1.105]
Charter	0.762	[-1.690, 3.250]	-1.888	[-3.790, -0.516]	-1.688	[-3.513, -0.072]	-1.819	[-3.114, -0.750]
Magnet	0.604	[-1.628, 2.633]	0.669	[-1.606, 2.868]	-0.508	[-2.553, 1.444]	-0.742	[-2.147, 0.505]
% Black	-0.030	[-0.092, 0.009]	-0.003	[-0.041, 0.038]	--	--	--	--
% Hispanic	--	--	--	--	-0.026	[-0.059, 0.010]	-0.014	[-0.036, 0.009]
% White	0.029	[-0.034, 0.079]	0.040	[-0.005, 0.098]	0.025	[-0.013, 0.078]	0.012	[-0.011, 0.034]
% IDEA	0.127	[-0.007, 0.287]	0.048	[-0.066, 0.197]	0.059	[-0.053, 0.385]	0.012	[-0.051, 0.072]
% FRPL	-0.006	[-0.047, 0.034]	-0.029	[-0.061, 0.004]	0.004	[-0.028, 0.036]	-0.020	[-0.052, 0.008]
# of M. (non-M.) schools	27 (37)	--	43 (53)	--	18 (30)	--	38 (43)	--

Note. 95% confidence intervals are bias-corrected.

Table 5

Risk Differences

	Black and White Students				Hispanic and White students			
	ISS		OSS		ISS		OSS	
	β	95% CI	β	95% CI	β	95% CI	β	95% CI
Montessori	0.020	[-0.015, 0.056]	-0.008	[-0.044, 0.022]	0.009	[-0.012, 0.046]	0.014	[-0.005, 0.034]
Charter	-0.032	[-0.094, 0.030]	-0.010	[-0.052, 0.030]	-0.047	[-0.120, -0.011]	-0.022	[-0.055, 0.008]
Magnet	-0.024	[-0.100, 0.035]	-0.012	[-0.061, 0.031]	0.006	[-0.041, 0.074]	0.013	[-0.019, 0.047]
% Black	0.000	[-0.002, 0.001]	0.000	[-0.001, 0.001]	--	--	--	--
% Hispanic	--	--	--	--	0.001	[-0.001, 0.002]	0.000	[-0.000, 0.001]
% White	0.000	[-0.001, 0.008]	0.00	[-0.001, 0.001]	0.002	[-0.001, 0.005]	0.001	[-0.001, 0.003]
% IDEA	0.003	[-0.001, 0.008]	0.00	[-0.002, 0.004]	-0.001	[-0.008, 0.002]	0.001	[-0.001, 0.004]
% FRPL	0.000	[-0.000, 0.002]	0.00	[-0.000, 0.001]	0.001	[-0.000, 0.004]	0.001	[-0.000, 0.002]
# of M. (non-M.) schools	33 (41)	--	49 (58)	--	33 (46)	--	52 (59)	--

Note. 95% confidence intervals are bias-corrected.

Multilevel Models

Bootstrapping results of the multilevel model predicting the proportion of students who received an ISS from each racial group, with random intercepts for schools, are presented in Table 6 along with bias-corrected 95% confidence intervals for each coefficient. The estimated means from this model are also shown in Figure 1. The same results and means for OSS are presented in Table 6 and Figure 2.

For ISS, results showed that Black students in conventional schools received ISS at 2.4% higher rates than White students, which was a statistically significant difference. There was not a significant difference between White and Hispanic student ISS rates in conventional schools. The coefficient for Montessori suggests that White Montessori students received an ISS at 1.1% lower rates than their White conventional school peers, although this was not a significant difference. The disparity between both racial groups and White student ISS rates was slightly, but non-significantly, larger at Montessori schools than non-Montessori schools. As shown in Figure 1, Black and White students ISS rates were lower in the Montessori schools than the conventional schools; ISS rates for Hispanic students were nearly equivalent across school types.

For OSS, Black conventional school students received OSS at significantly higher rates than their White peers (estimated 5% higher rate). There was not a significant difference in OSS rates in conventional schools between White and Hispanic students. White Montessori students received OSS at 1.8% lower rates than White conventional school students, and this difference was statistically significant. There was a 1.5% smaller disparity in the Black and White suspension rates at Montessori schools than at the conventional schools, but this difference was not statistically significant. The difference in OSS rates between White and Hispanic students

was virtually equivalent at both school types. As shown in Figure 2, OSS rates for each racial group were lower in Montessori than non-Montessori schools.

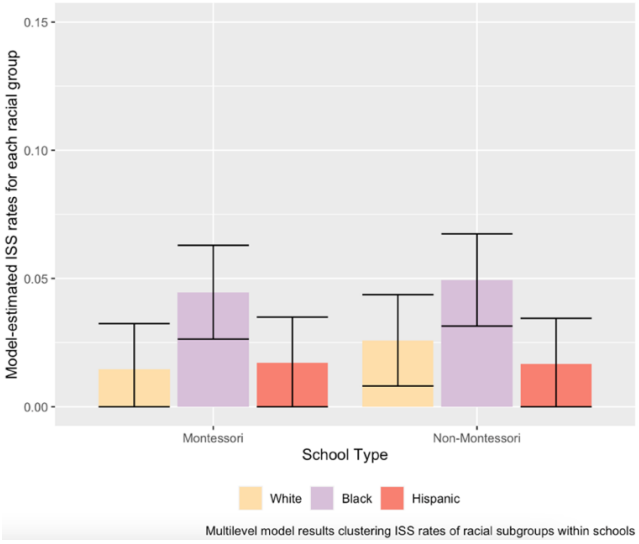
Table 6

Bootstrapped (1000×) bias-corrected 95% confidence intervals and coefficients for multilevel models predicting ISS and OSS rates

	ISS		OSS	
	β	95% CI	*β*	95% CI
Intercept	-0.026	[0.015, 0.035]	0.035	[0.026, 0.045]
Black	0.024	[0.008, 0.042]	0.050	[0.031, 0.074]
Hispanic	-0.009	[-0.029, 0.003]	-0.003	[-0.017, 0.010]
Montessori	-0.011	[-0.026, 0.007]	-0.018	[-0.030, -0.007]
Black × Montessori	0.006	[-0.021, 0.051]	-0.015	[-0.045, 0.010]
Hispanic × Montessori	0.012	[-0.007, 0.035]	0.007	[-0.011, 0.023]

Figure 1

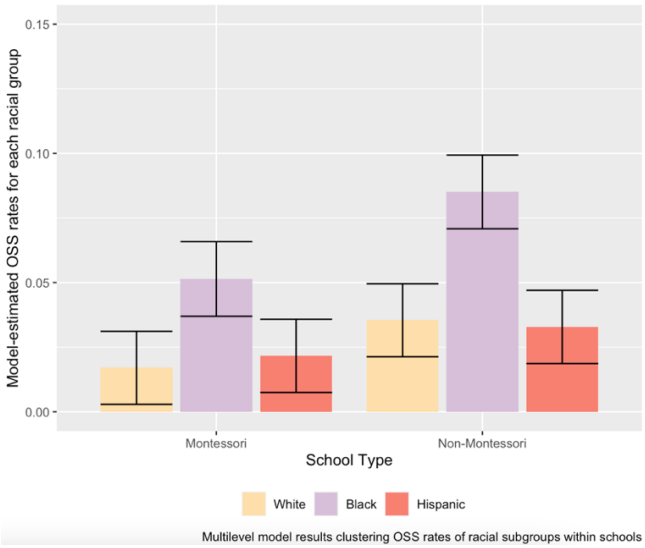
Predicted ISS rates for each racial groups across Montessori and non-Montessori schools



Note. Bars represent confidence intervals. Confidence intervals with lower bounds below zero have been truncated to zero.

Figure 2

Predicted OSS rates for each racial groups across Montessori and non-Montessori schools



Note. Bars represent confidence intervals. Confidence intervals with lower bounds below zero have been truncated to zero.

Discussion

This study introduces a new method to measure discipline disproportionality and compares 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, the estimated effect of Montessori on lowering ISS and OSS rates here is identical to results reported by Culclasure and colleagues (2018). Montessori schools in this sample were associated with 1% and 2% lower ISS and OSS rates respectively; Culclasure and colleagues (2018) reported 1–2% fewer suspensions for the Montessori students in their sample relative to the non-Montessori students. Because suspension is a rare event, these differences are meaningful (66% and 50% as many ISS and OSS, respectively). 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 the true effect of Montessori schools

on discipline yields lower suspension rates for Black, White, and Hispanic Montessori students than their non-Montessori counterparts.

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 receive ISS and OSS at roughly equivalent rates as White students in both Montessori and non-Montessori schools. These results are also corroborated by the lack of significant interaction between Montessori status and Hispanic students in the multilevel model for both ISS and OSS, and they are consistent with previous reports on suspension rates for Hispanic or Latino students (U.S. Department of Education, 2018).

The discipline disproportionality results comparing Black and White students are slightly more complex. Only one previous study 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 2-3 (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 is *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 Figures 1 and 2, suspension rates for all three racial groups were lower in the Montessori schools than the non-Montessori schools. Although just shy of traditional statistical significance, the bootstrapping results of the multilevel model suggest that the disparity between Black and White student OSS rates is slightly smaller in the Montessori sample. These results suggest that although Black students are suspended more often than White students in Montessori schools, the disparity between White and Black student suspension rates is slightly lower in the Montessori schools. Discipline disproportionality certainly exists in the Montessori schools, but it is unclear from this study whether the disparity is meaningfully different than it is in conventional schools. Because the suspension rates in Montessori schools were lower, on average, than non-Montessori schools for all three racial groups, Black Montessori students were less likely to be suspended than Black non-Montessori students.

It is also important to note that RRRs and RDs were only calculated in 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 and RDs comparisons. Moreover, the assumptions of the propensity score matching process do not hold as strongly for the subsamples for which RRRs and RDs 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, not just those who had suspended at least one student, 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

Self-selection bias. 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. Because Montessori schools are typically choice schools, selection bias is an ongoing concern in studying differences between Montessori and non-Montessori student outcomes. To date, only two studies have attempted to measure average 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. This study, and Montessori research more generally, would benefit from more research on Montessori parents' beliefs, parenting behaviors, and motivations for enrolling their child(ren) in a Montessori school. Propensity score matching was the best choice for reducing potential self-selection bias in the estimates simply 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 samples in this study were very similar in terms of racial demographics and proportion of

students qualifying for free/reduced price lunch (a proxy for SES), so neither race nor SES is likely to be a meaningful source of bias. Qualitative data suggest that low-income Montessori parents of color value the Montessori method for its focus on self-discipline (Golann et al., 2019). It is possible that Montessori parents, on average, value self-discipline in their children more, and seek out Montessori schools as a result. Accordingly, it is possible that Montessori parents discipline their children differently from non-Montessori parents, on average, in ways that elicit less disruptive classroom behavior from their children. If so, one would predict fewer disciplinary infractions in Montessori schools based on parenting differences alone. However, parents from the same qualitative study who chose to send their children to "no-excuse" charter schools reported similar beliefs related to self-discipline as the Montessori parents, despite enrolling their children in schools with drastically different disciplinary practices. Empirical evidence to 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 over- or under-estimate 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 the richest descriptive analysis to date of disciplinary outcomes in Title 1 Montessori schools.

Measuring discipline disproportionality. In addition to the analysis of Montessori disciplinary outcomes, this study also presents a novel way to measure 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 no straightforward solution has been offered. The approach used here offers a reasonable solution to many of the issues raised by other researchers. This approach allows for a more intuitive interpretation of the magnitude of a disparity by making overall suspension rates for each racial subgroup easily calculable from the resulting coefficients. Further, this approach is not prone to the same insensitivity or inflation issues as RRRs and RDs in schools with low suspension rates; it allows for inclusion of schools with zero suspensions; and it allow for comparisons across multiple racial groups in one model. Finally, 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 any other form of racial disparity) using multilevel modeling is that one could also test for how racial disparities have changed over time by using longitudinal data and introducing a random slope for the time variable. 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:

$$ProportionSuspended_{ij} = \beta_{0j} + \beta_1 Time_{ij} + \beta_2 Hispanic_{ij} + \beta_3 Black_{ij} + \quad (8)$$

$$\beta_4(Time_j * Hispanic_{ij}) + \beta_5(Time_j * Black_{ij}) + \mu_{0j} + \mu_{1j} + \varepsilon_{ij}$$

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-level interaction between time, a racial group of interest, and an intervention to easily test 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 because they can often result in contradictory conclusions (Curran, 2020; Girvan et al., 2019).

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 one 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 a way 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). Future research on discipline disproportionality should consider using multilevel models à la those presented here to resolve the issues with more common measures.

Discipline disproportionality findings from this study did not replicate those reported by Brown and Steele (2015). However, 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 already listed—high student engagement, student self-determination, individualized learning, and strong classroom community—are potential causes of the lower

disciplinary rates observed here. Future research using student-level data to predict individual disciplinary infractions 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-teacher 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).

References

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

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).

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

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

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.

Nishioka, V., Shigeoka, S., & Lolich, E. (2017). *School discipline data indicators: A guide for districts and schools* (REL 2017–240). Washington, DC: 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*. Washington, DC: 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.