POLICY BEATS BIAS? AN EVALUATION OF THE IMPACT OF OPERATIONAL DEFINITIONS ON DISPROPORTIONATE DISCIPLINARY OUTCOMES FOR BLACK STUDENTS

by

KARA NYSTROM BOULAHANIS

A DISSERTATION

Presented to the Department of Special Education and Clinical Sciences and the Graduate
School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2020

DISSERTATION APPROVAL PAGE

Student: Kara Nystrom Boulahanis

Title: Policy Beats Bias? An Evaluation of the Impact of Operational Definitions on Disproportionate Disciplinary Outcomes for Black Students

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Special Education and Clinical Sciences by:

Roland Good, PhD	Co-Chairperson
Kent McIntosh, PhD	Co-Chairperson
Robert Horner, PhD	Core Member
Claudia Vincent, PhD	Core Member
Erik Girvan, PhD	Institutional Representative

and

Kate Mondloch	Interim Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2020

DISSERTATION ABSTRACT

Kara Nystrom Boulahanis

Doctor of Philosophy

Department of Special Education and Clinical Sciences

June 2020

Title: Policy Beats Bias? An Evaluation of the Impact of Operational Definitions on Disproportionate Disciplinary Outcomes for Black Students

Office discipline referrals (ODRs), suspensions, and expulsions are exclusionary disciplinary practices commonly used in U.S. schools that are associated with decreased student achievement and a host of negative school and life outcomes. This study examined the impact of operational definitions, race, behavior class, level of behavior concern, and cultural context on disproportionate disciplinary outcomes through an evaluation of educators' consistency with experts in rating problem behavior using a randomized control, pre-test/post-test intervention study. It was hypothesized that more consistency with experts may reduce disproportionality in disciplinary outcomes due to more accurate identification of problem behaviors requiring out of classroom disciplinary practices. Participants' consistency with expert ratings of students' misbehaviors was examined by measuring participant responses pre- and post-test on four questions regarding their reaction to a series of video vignettes depicting student misbehavior selected by the researcher. No discernable impact of operational definition condition, race or level of behavior concern on participants' accuracy in rating student problem behavior was found. Behavior class, that is objective vs subjective behaviors, explained 33% of the variance in participants' consistency with expert ratings of student misbehavior. Participants were more consistent with experts in rating video vignettes depicting behaviors classified as subjective such as defiance or disrespect than they were in rating videos depicting behaviors classified as objective such as physical aggression or smoking. Implications and future directions are discussed.

CURRICULUM VITAE

NAME OF AUTHOR: Kara Nystrom Boulahanis

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, Oregon
National Louis University, Chicago, Illinois
University of Chicago, Chicago, Illinois

DEGREES AWARDED:

Doctor of Philosophy, school psychology, 2020 (expected), University of Oregon
Master of Arts in Teaching, special education, 2012, National Louis University
Bachelor of Arts, law, letters & society, 2008, University of Chicago

AREAS OF SPECIAL INTEREST:

My research and clinic interest are in the follow areas: (1) systemic oppression in educational practice and policy, (2) culturally relevant practices & pedagogy, and (3) systems change in educational settings.

PROFESSIONAL EXPERIENCE:

Pre-doctoral Intern, School Psychology, Eugene 4J School District, Eugene, OR. 2019-Present

Graduate Teaching Fellow, Critical and Socio-Cultural Studies in Education, University of Oregon, Eugene, OR. 2018-2019

Graduate Teaching Fellow, Family and Human Services, University of Oregon, Eugene, OR. 2017-2018

Graduate Teaching Fellow, Service Learning Program, University of Oregon, Eugene, OR. 2015-2017

Skills Coach, Oregon Social Learning Center (OSLC), Eugene, OR. 2017 – 2018

Lead Network Case Manager, LEARN Charter Network, Chicago, IL. 2012-2014

Special Education Continuum Instructor, Gallistel Language Academy, Chicago Public Schools, Chicago, IL. 2010-2012

GRANTS, AWARDS, AND HONORS:

Dynamic Measurement Group Dissertation Award, University of Oregon, 2019
Dynamic Measurement Group Award, University of Oregon, 2015, 2016, 2017, 2018, 2019
Chicago Teaching Fellow, Chicago Public Schools, 2010

ACKNOWLEDGMENTS

I would like to express my most sincere gratitude to each member of my committee. Claudia, Erik and Rob – each of you contributed immeasurably to this study. I will forever be grateful for your guidance and support. Kent, your encouragement and assistance in each phase of my study from design, to execution, to analysis, was critical not only to this study but to my development as a researcher. I cannot thank you enough for your guidance. Roland, you are the best advisor, support system and mentor I could possibly ask for. I would not be the scholar or person I am today without your influence and I could not be more grateful. I will carry your lessons and kindness with me on all of my adventures; hoping to live up to the example you have set!

I would also like to thank IRIS Educational Media, the Systemic Supervision project and Project ReACT for allowing the use of their media in the study. Without their support this study wouldn't have been possible. This investigation was also supported in part by the Dynamic Measurement Group Dissertation Award from the University of Oregon College of Education. Thank you to Roland, Ruth and the entire initial DIBELS early literacy team for your support of my work.

Lastly, I would like to acknowledge that this work was completed on unceded Kalapuya territory and the land should be returned to its rightful stewards.

To my mother, Karen, father, George, brother, Brian and sister, Bridgit – Thank you for your unconditional love and support from my first days of schools to every day since. Brian, thank you for inspiring me to fight to ensure all children have access to the education they deserve. Bridgit, your support, commiseration and unwavering belief in me are the only things that got me through some days. I am so lucky to be your big sister. Mom & Dad, without your guidance and enduring belief in my ability to change the world, I would have never had the nerve to even attempt a doctorate. Mom, you set the bar for fighting for what you believe in so incredibly high I had to complete a dissertation to meet it. You will forever be an inspiration for my fire. Dad, every time I forgot whose daughter I am, you were there to remind me. Your steadfast support is the foundation to my success. All of you are the reason this was possible for me.

To my partner, Tyler – you are my best friend and my rock. I cannot begin to express my gratitude for your selflessness while I have pursued my dream. Thank you for constantly building me up when I stumble and sharing each moment of failure or triumph with me. I would not have made it through this process without you by my side.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

CHAPTER I

Disproportionate Use of Exclusionary Disciplinary Practices

Office discipline referrals (ODRs), suspensions, and expulsions are exclusionary disciplinary practices commonly used in U.S. schools as a consequence for student problem behavior. These disciplinary practices are associated with decreased student achievement as well as increased risk for drop out and increased likelihood of future arrest (Wishman & Hammer, 2014; Rausch & Skiba, 2005; Archambault, Janosz, Fallu, & Pagani, 2009; Cranerm, Gonzalez, & Pellegrini-Lafont, 2014; Fenning & Rose, 2007; Rocque & Paternoster, 2011; Monahan, VanDerhei, Bechtold, & Cauffman, 2014).

Black students are at particularly high risk for being suspended. Suspension rates for Black students remained unchanged, at two to three times that of their White peers, from the first examination of differential rates of disciplinary outcomes, the 1975 Children's Defense Fund Report, to the 1990 and 2012 U.S. Department of Education's Office for Civil Rights' (OCR) surveys. Black children make up 18% of our nation's preschool population but comprise more than 48% of preschool children who receive more than one out of school suspension. (USDOE, 2017). In the 2011 – 2012 school year, nationally, 23% of Black secondary students and 7.6% of Black elementary students were suspended while only 6.7% of White secondary students and 1.6% of White elementary students were suspended (Losen, Hodson, Keith, Morrison & Belway; 2015).

Black students are not only being excluded from the school building at an increased rate, but they also experience increased rates of office discipline referrals. Roque (2010) found that Black students have 2.27 greater odds of being referred to the office than other racial groups, even within the same schools. These effects persisted even when grades, socioeconomic status, special education status, age and gender were controlled for. These findings drive the urgent question: why do these disproportionate outcomes exist?

Research on Causes and Contributing Factors

Although attempts to explain disproportionate discipline outcomes through student-teacher ratios or student attendance have failed to produce consistent significant results, research on the impact of poverty, differential rates of misbehavior, and implicit bias on disproportionate disciplinary outcomes has provided significant results that help illuminate the contributing factors to this persistent problem. For example, poverty accounts for some of the differences in discipline by race, but it does not explain most of the variance (Wu et al., 1982; Skiba, Michael, Nardo, & Peterson, 2002; Wallace et al., 2008).

A substantial body of evidence also suggests that Black students do not misbehave at a higher rate than their White peers. Across a wide range of methodologies and data sets, this theory has failed to be validated. Evaluations of differences in severity of misbehavior have consistently failed to show racial difference in the severity of discipline referrals, with White students being similarly likely to commit serious violations (McCarthy & Hodge, 1987; Gregory & Weinstein, 2008). Statistical analysis where the type of infraction is held constant and racial differences are examined have similarly failed to show the expected racial differences if race is a significant predictor of student misbehavior (Skiba et al., 2011; Eitle & Eitle, 2004; Peguero & Shekarkhar; 2011). Further, controlling for teacher's own ratings of students misbehavior has failed to show a significant effect of race, indicating that even when teachers rate a student as less disruptive they are still more likely to refer them to the office if they are Black (Bradshaw, Mitchell, O'Brennan, & Leaf, 2010). Black students are more likely to be referred to the office for the same behavior, and they are more likely to receive exclusionary disciplinary consequences when they are referred (Bradshaw, Mitchell, O'Brennan, & Leaf, 2010; Skiba, Horner, Chung, Rausch, May, & Tobin, 2011; Wallace, Goodkind, Wallace, & Bachman, 2008; Skiba, Michael, & Peterson, 2000; Elliott, Ageton, & Huizinga, 1980).

Given the evidence that Black students do not misbehave at higher rates, researchers have turned to social psychology's theory of implicit bias for a possible explanation of this phenomenon. Although explicit bias is conscious prejudicial behaviors, implicit bias refers to unconscious attitudes and beliefs that impact perceptions, judgments, decision-making, and behavior (Mendoza, Gollwitzer, & Amodio, 2010). Most residents of the US express little explicit bias, but many may simultaneously harbor implicit biases against non-dominant groups. (Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Pearson, Dovido & Gaertner, 2009). While explicit bias is often considered more harmful, there may be a significant impact of implicit bias on real world outcomes, such as medical, housing, and hiring decisions (Blair, Steiner & Havranek, 2011; Ollinger, Capatoso & McKay, 2017; Bertrand & Mullainathan, 2004).

The risk of implicit bias affecting discipline decisions increases (a) when the demands of the situation outpace the available information, (b) when cognitive resources are limited, and (c) when there is greater individual discretion in decision making (McIntosh, Girvan, Horner, & Smolkowski, 2014). However, the work on implicit bias was primarily conducted with adults in laboratory settings and cannot be easily generalized to education settings. Research connecting implicit bias to educational decisions is in its infancy, but promising. For example, an initial evaluation found that teacher implicit bias predicted the magnitude of the achievement gap between a teacher's dominant and non-dominant students (Van den Bergh, Denessen, Hornstra, Voeten, & Holland, 2010).

However, any research into implicit bias must be interpreted with caution as the assessment used to evaluate individuals' implicit bias, the Implicit Association Test or IAT, has a test-re-test reliability of .44, indicating that it is insufficiently reliable for even group decision making (Gawronski, Morrison, Phills, Curtis & Galdi 2017). Further, meta-analyses have indicated that the IAT is a weak predictor of behavior (Cameron, Brown-Iannuzzi & Payne, 2012; Greenwald, Poelhman, Ulhmann & Banaji; 2009; Carlsson & Agerstrom; 2015). One evaluation, Oswald, Mitchell, Blanton, Jaccard & Tetlock (2012), indicated that the IAT was not better at predicting behavior better than even simple explicit measures of biased behavior. Forscher et al. (2017) conducted a meta-analysis of 484 studies with more than 80,000 total participants on studies of interventions designed to decrease participant implicit bias. They found no evidence that interventions to address implicit bias had a significant impact on behavior. Further studies have shown that participants are able to predict their score on the IAT with accuracy, indicating that individuals are not nearly as unaware of their biases as "unconscious" implies (Hahn, Judd, Hirsh, & Blair; 2014). However, the consistent impact of race on disproportionate disciplinary outcomes indicates some level of bias influencing decisions, regardless of whether that bias is conscious, unconscious or simply the pervasive, culturally permissible bias against Black individuals known as systemic oppression (Feagin, 2006). It is clear that addressing the issue of disproportionality will require new solutions that address bias.

Despite the concerns with the methods used to assess implicit bias and the impact of interventions to ameliorate implicit bias, a promising theoretical model, based on the work of social psychologist and legal scholar Erik Girvan, provides a framework for evaluating the setting events and antecedents of teacher disproportionate disciplinary decisions to develop interventions that directly address environmental and personal factors that impact educational decision making.. The "Vulnerable Decision Points" (VDP) model applies implicit bias theory to real world decision making through the identification of VDPs, or contextual events and elements of an immediate situation that increase the likelihood of bias in educational decision making (McIntosh, Girvan, Horner, & Smolkowski, 2014). A preliminary investigation of the model identified situations where disciplinary disproportionality was more likely to occur in order to assess their fit with the theoretical model. Smolkowski, Girvan, McIntosh, Nese, and Horner (2016) found that students were more likely to be referred for (a) subjectively-defined behaviors, (b) that occurred in classrooms where a teacher is often the sole decision maker, and (c) when the level of the behavior is considered more severe. This finding provided initial support for the model, though more research is needed to validate and inform specific future directions for interventions.

Interventions to Address Disproportionality in Discipline Outcomes

Despite the lack of malleable factors identified by research, both practitioners as well as state and federal policy makers are eager to address racial disparities in exclusionary discipline practices. The U.S. Department of Education issued a comprehensive Guiding Principles document that recommends ongoing data collection and analysis coupled with two primary interventions to address disproportionate discipline outcomes. The recommended interventions include a focus on positive climate and prevention as well as the development of clear, appropriate, and consistent expectations and consequences to address student behavior (U.S. DOE, 2014).

School-wide Positive Behavior Support (SWPBS) is a framework for delivering a whole-school socio-culture intervention that includes systematic data collection and analysis, a focus on positive climate and prevention, as well as clear and consistent expectations with planned consequences. SWPBS fulfills all of the guidance document recommendations and has already been adopted by thousands of schools in every state (Horner & Sugai, 2015; Johnson, Foxx, Jacobson, Green, & Mulick, 2006). As a program, SWPBS strives to achieve "universal expectations" for student behavior through operational definitions (ODs) of behavior. An OD is an observable description of a target behavior that is designed to provide a universal understanding of the topography of a behavior that is sufficiently clear to allow persons other than the definer to independently measure or test for the behavior at their desire. This also functions to limit teacher discretionary judgement in decision making (Todd, Horner, & Tobin, 2010). Operational definitions in SWPBS were designed to provide a "universal language" about student behavior expectations, ensuring that all students and teachers theoretically have a shared understanding of what is appropriate and inappropriate behavior. Operational definitions should therefore decrease disproportionate disciplinary outcomes as teachers decrease inappropriate referrals for non-problem behaviors and students decrease misbehavior through decreased inadvertent rule breaking due to lack of knowledge.

Unfortunately, no component analysis research has been conducted to validate the use of operational definitions in communicating expectations to students or teacher or on their impact on disciplinary outcomes.

SWPBS has shown remarkably positive impacts on wide ranging student outcomes including feelings of school safety, academic outcomes, and rates of exclusionary discipline (Horner, Sugai, Smolkowski, Todd, Nakasato, & Esperanza, 2009; Bradshaw, Koth, Thornton, & Leaf, 2009; Bradshaw, Mitchell, & Leaf, 2010). SWPBS has even demonstrated a statistically significant impact on the extent of disproportionality within a school, but it has not been associated with the elimination of disproportionate discipline outcomes (Vincent, Swain-Broadway, Tobin, & May, 2011). Therefore, typical implementation of SWPBS is a first step to reducing disproportionality, but further strategies may be necessary to fully address this pressing concern.

Vulnerable Decision Points and Schoolwide Positive Behavior Support

The VDP model offers an opportunity to examine the typical implementation of SWPBS and identify areas where implicit bias may impact disciplinary outcomes. SWPBS emphasizes universal expectations and consequences through an objective, observable, and measurable description of a problem behavior, known as an operational definition (OD). These definitions are *intended* to be universal and help limit teacher discretionary judgement, but research on SWPBS indicates that ODs may not always achieve these goals. Smolkowski, Girvan, McIntosh, Nese, & Horner (2016) found that in schools implementing SWPBS to criterion, disproportionate discipline outcomes for Black students were more frequently related to behaviors considered more subjective or discretionary, such as disrespect and defiance, and were less related to behaviors considered more objective or mandatory, such as vandalism, physical aggression, and swearing. This provides compelling evidence that the application of subjective operational definitions may represent a vulnerable decision point (See *Figure 1*). Further analysis is necessary to fully explain the factors or conditions related to operational definitions that may influence the activation of implicit bias and therefore, the extent to which they contribute to differences in student outcomes.

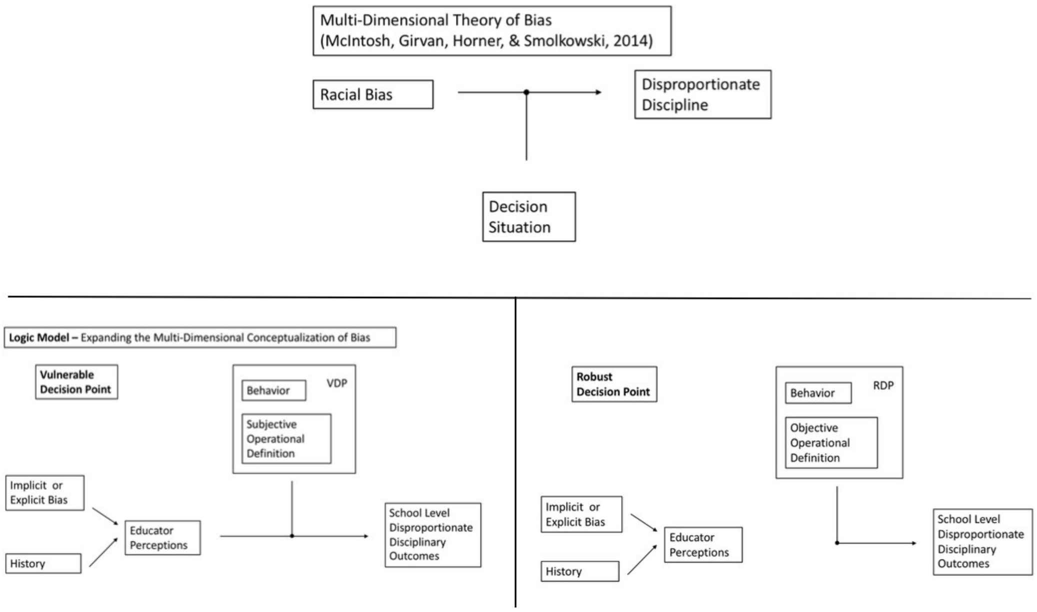

Multi-Dimensional Theory of Bias
(McIntosh, Girvan, Horner, & Smolkowski, 2014)
Racial Bias
Disproportionate Discipline
Decision Situation
Logic Model – Expanding the Multi-Dimensional Conceptualization of Bias
Vulnerable Decision Point
Behavior
VDP
Subjective Operational Definition
Implicit or Explicit Bias
History
Educator Perceptions
School Level Disproportionate Disciplinary Outcomes
Robust Decision Point
Behavior
RDP
Objective Operational Definition
Implicit or Explicit Bias
History
Educator Perceptions
School Level Disproportionate Disciplinary Outcomes

Figure 1. The Multi-Dimensional Conceptualization of Bias provides an initial explanation for the role of bias in discipline decisions. This study seeks to expand the Multi-Dimensional Conceptualization of Bias to include additional relevant factors. In this model, on the left, a Vulnerable Decision Point is depicted as being comprised of educator perceptions, which themselves are made up of bias and educator history, both with the student as well as the teacher's history with similar individuals or individuals who were perceived to be similar, as well as the student's behavior and a subjective operational definition of that behavior class, leading to school level disproportionate disciplinary outcomes. Conversely, in a Robust Decision Point, the relationship between educator perceptions and school level disproportionate disciplinary outcomes is interrupted by an objective operational definition of the behavior class.

Summary

The over referral of Black students for disciplinary consequences is a pervasive problem with far-reaching consequences including decreased academic achievement, increased likelihood of dropping out as well as an increased likelihood of future involvement with the criminal justice system. This disproportionality in disciplinary outcomes has been found in studies from as early as 1975 to 2019. There has been a significant effort to explain these outcomes through research. However, efforts to explain disproportionate disciplinary outcomes through analyses of poverty, student attendance, student teacher-ratio, and most importantly, through disproportionate rates of misbehavior between White and Black students, have failed to explain most, or in most cases, any of the variance. More recently, a promising theory based on the social psychology concept of implicit, or unconscious, bias, has emerged to possibly explain the different rates of disciplinary referral for Black and White students. The Vulnerable Decision Point (VDP) model argues that implicit bias may impact teacher disciplinary decision making under specific contextual and intrapersonal conditions, known as vulnerable decision points. Preliminary analyses of this model have been promising, but more research is warranted.

The failure to uncover a specific causal mechanism has not slowed the push for solutions to this concerning problem. Federal policy makers have published guidance for addressing disproportionate disciplinary outcomes that recommends robust data collection, a focus on prevention, positive climate and the development of clear, specific behavioral expectations and consequences. These recommendations are frequently met through the implementation of School-wide Positive Behavioral Interventions and

Supports (SWPBS), a whole school social and cultural support systemic intervention. SWPBS supports the develop of clear and specific behavioral expectations for both teachers and students through the implementation of operational definitions (ODs) or clear, observable definitions of expected and non-expected behaviors that are taught to both students and staff. Operational definitions may increase teachers' accuracy in referring students for disciplinary consequences by reducing the likelihood of implicit bias entering their decision making. Within the VDP model, operational definitions are a policy factor that reduces the contextual factors that increase the likelihood of implicit bias impacting decision making, thereby decreasing the likelihood of disproportionate disciplinary outcomes.

This study evaluated the impact of operational definitions on teacher disciplinary decision making in the context of misbehaviors of different types, levels of concern about the behavior and performed by students of different races. Further, this study evaluated differences in the impact of operational definitions based on the cultural context of the educator making the disciplinary decision.

Research Questions

1) To what extent does providing operational definitions and decision-making support improve educators' accuracy in rating behavior, as compared to an expert panel?
2) Does the impact of operational definition condition depend on the level of concern about the behavior, the race of the target student, and the behavior class?
3) Does this impact depend on cultural context?

CHAPTER II

Researchers first identified disproportionate disciplinary outcomes for students of color as a concern in the 1975 Children's Fund Report. Research into the cause of these outcomes began in earnest shortly thereafter. Some of the earliest research into disproportionality implicated bias as a primary cause. However, limitations of methodology, sample sizes and generalizability plague much of the research in this area, even into the modern era, and intervention studies remain extremely rare. The majority of the research on the causes of disproportionality has examined permanent product data without providing connection to concrete next steps to address the concerns of bias in educational decision making.

Early Evaluations of the Causes of Disproportionate Disciplinary Outcomes

As early as 1981 there were case presentations from researcher-practitioners at the regional meetings of the American Educational Research Association (AERA) in Boston and Los Angeles, on the causes of disproportionality in the suspensions and expulsions of "male and Black" students (Bickel & Qualls, 1981; Bennet, 1981). One of the earliest published studies followed shortly thereafter in 1982, when Drs. Bennett and Harris published, "Suspensions and expulsions of male and Black students: A study of the causes of disproportionality," in Urban Education. The study included a wide range of measures and methods to assess the school factors impacting disciplinary outcomes in two school corporations, comprised of 5-6 school sites each. The methods included taped interviews of students, parents, teachers and administrators; student cumulative file reviews, school disciplinary file reviews, paper-and-pencil questionnaires administered to students, teachers and administrators, as well as third party collected enrollment, withdrawal, suspension and expulsion data broken down by school, sex and race. The study examined both student factors as well as school factors that may have impacted the disciplinary outcomes. Interestingly, the authors found that there were not significant differences between students who had been suspended and rated as "serious disruptors" by school staff and non-disruptors on ratings of the positivity of the school climate, indicating that as a whole the students shared a common perception of the school climate They also found that schools with higher rates of disproportionality in school discipline outcomes had lower scores on the positive school climate index, lower scores on the interracial environment index, and higher scores on the white predominance index. Lower scores on the positive school climate index combined with higher scores on the interracial environment index indicates an environment that is not welcoming to students of color. Add the higher scores on the White predominance index and this study provides evidence that schools with higher levels of disproportionality have higher levels of bias, supporting the bias explanation of disproportionate disciplinary outcomes. The authors note that their findings show that an "overall orientation of White predominance which includes institutional and individual racism," is a cause of the disparities. They go on to elaborate that, "sources of racism are difficult to pinpoint because they originate in a social context beyond the school, but this does not relieve the school from taking action to mediate racism."

The authors suggested their findings indicate school programs should focus on building feelings of school efficacy and a "stake in the school," which are wonderful concepts that lack clear action steps for administrators and teachers seeking to implement an intervention in their school. While their findings are stark, there were limitations to their results. All of the measures were surveys. There was no direct collection of data on student behavior or teacher decision making. Only two districts were included in the study, which significantly limits the generalizability of the findings. Finally, the measures were primarily researcher-designed and lacked psychometric evaluation, making it impossible to evaluate their reliability and validity at assessing they constructs they claim.

Addressing some of these concerns, Wu, Pink, Crain and Moles (1982) analyzed representative national data from more than 4,500 elementary and secondary schools from a congressionally mandated “safe school study” (National Institute of Education, 1978). Using a regression model Wu, Pink, Crain and Moles (1982) evaluated the student and school factors that contributed to disproportionality in their sample. Unlike other studies, they did not find that race significantly predicted disproportionate disciplinary outcomes. They did find however that school factors accounted for 15.2 to 32.5% of the variance in in student suspension rate, whereas student factors accounted for only 1.2 to 13.3% of the variance. The school factors they identified as increasing a student’s chances of being suspended were teachers’ perceptions and beliefs, the school’s administrative structure for handling disciplinary matters, and the presence of institutional bias, or racism, in the school. This evidence supports the current study, indicating again that bias is a factor in disproportionate disciplinary decision making and that administrative structures for disciplinary matters may impact educators’ decision making.

Wu, Pink, Crain and Moles (1982) found that when disciplinary matters are largely handled by administrative rules, disproportionality in discipline outcomes is likely to be greater. Wu, Pink, Crain and Moles (1982) do not offer any recommendations other than the suggestion that we must ask not what students did wrong, but “what kind of school did that student go to and how was it run?”

In 1987 McCarthy and Hodge published “the social construction of school punishment: Racial disadvantage out of universalistic process.” This study examined student perceptions of delinquency and racial differences in their school using 3 years of longitudinal data from 6 public schools in a Mid-Atlantic city. The schools were chosen to produce a sample with maximal representational variation in socioeconomic status that included both boys and girls as well as Black and White students. That is to say, the study did not seek to replicate the real world with their sample but instead to sought to achieve the maximum possible diversity in their sample with regards to gender, race and socioeconomic status. Using a multiple regression analysis and controlling for the amount of misbehavior by a student, they found that knowledge of a student’s past punishment was the strongest predictor of future punishment. Additional factors included teachers’ perceptions of the student’s general level of good behavior and the student’s past grades, indicating that prior knowledge about a student may be a type of bias that impacts educational decision making specifically. The authors concluded that “the social construction process, based on the central understandings of school authorities as to the meaning of proper student school behavior, offer the most plausible account of our findings” (McCarthy & Hodge, 1987). This specifically supports the need for operational definitions that are contextually valid, as we are unable to remove information about students from teachers’ perceptions, but we could potentially intervene on their actions based on that information.

McCarthy and Hodge (1987) included 1,200 7th, 9th and 11th, graders in 6 public schools in the Mid-Atlantic region, which was about 25% of the area's population at the time. Their evaluations of the differences between populations were not statistical analyses and they only provided an "estimation" that racial and social bias impacted the rate of consent obtained. However, they concluded that those biases did not impact the results of their analyses without providing further information. These flaws may significantly impact the validity of the reported results, but without further information it is not possible to evaluate the potential impact.

As the decade drew to a close, two additional studies were conducted evaluating racial bias in the use of corporal punishment in Florida school districts through record reviews. Shaw and Braden (1990) utilized 6,244 discipline files from 16, K-12 schools in a Central Florida school district to conduct a multiple regression analysis. They found that neither frequency nor severity of misbehavior significantly predicted the use of corporal punishment. When race was added to the regression, they found the combination of race with frequency and severity was a significantly better predictor than either frequency or severity alone ($R^2 = .131, p < .003$), indicating the effect of frequency and severity depends on race. Externally identifiable bias accounted for 22% of the variation between White and Black students' likelihood of receiving corporal punishment. The authors note that a limitation of their study was the evaluation only of bias at the administrative level of disciplinary decision-making and not if there was disproportionality in the teacher's referral process. Therefore, it is likely that teacher bias additionally compounds the administrator bias identified in Shaw and Braden's model.

McFadden and Marsh (1992) used a chi-square analysis to evaluate if disproportionate discipline rates were representative of bias in the referral process using a sample of 4,391 discipline files from 9, K-12 schools in South Florida. They found that Black students accounted for only 36.7% of the disciplinary referrals and just 23.0% of the internal suspensions. However, Black students received 54.1% of the corporal punishment and 43.9% of the school suspensions, replicating the bias in administrative decision making found in Shaw and Braden (1990). The authors examined whether Black students were more likely to commit a serious offense warranting more severe disciplinary consequences. However, data indicated that corporal punishment was most often administered for defiance or disrespectful behaviors, both behaviors that are considered "subjective," and for which White students were referred at a much higher rate. The authors again concluded "that some form of bias does appear to have existed." They went on to state that "to the extent that firm, fair, and uniform practices of discipline and respect for both teachers and students may be implemented in policy and action, the more ably will schools deal with the issue of behavioral control of students." This directly supports the argument in support of culturally responsive operational definitions that provide respectful, contextual and fair expectations.

Recent Studies

By the end of the 1980's a picture of the causes of disproportionality was beginning to emerge. Although student factors continued to require exploration, school level factors and issues of institutionalized racism were consistently implicated throughout the findings. The findings were largely centralized in a few small districts, with only one nationally representative sample to serve as a literature base for future researchers. Moving forward to the 21st century, interest in the causes of disproportionate disciplinary practices rose as studies continued to document the stagnant or climbing rates of disproportionality despite more than 10 years of research.

In 2002, the next major study to evaluate the causes of disproportionality, "The Color of Discipline: Sources of Racial and Gender Disproportionality in School Punishment," examined the disciplinary records of 11,001 students. The students were attending 19 middle schools in a large, urban midwestern district during the 1994-95 school year (Skiba, Michael, Nardo & Peterson; 2002). Confirming prior results, Skiba et al, found that students of color and those from low-income backgrounds were more likely to experience a variety of school punishments, but race remained a significant predictor over and above socioeconomic status.

More relevantly, their results also suggested that contrary to both Shaw and Braden (1990) and McCarthy and Hodge (1987), gender and racial disparities in school suspension were not due to administrative decision making but instead to differences in rates of initial referrals to the office by classroom teachers. Their analyses suggested that, when controlling for the rate of office referrals, significant racial differences in the rate of suspension disappeared. Skiba, Michael, Nardo, and Peterson (2002) determined that "absent support for any plausible alternative explanation, these data lend support to the conclusion that racial disproportionality in school discipline, originating at the classroom level, is an indicator of systematic racial discrimination." Although these findings support the theory that bias is a significant factor in disproportionate disciplinary outcomes, the study is an evaluation of extant data and so must be interpreted with caution. It does not include information about the administrator or teacher who made the disciplinary decisions, which limits the utility of this information to address the impact of systemic racism on disciplinary outcomes in schools.

Gregory and Weinstein (2007) provided a novel addition to the body of evidence on bias: the concept that students behave differently for different instructors. This new research dimension allowed for a correlational evaluation of the interaction between student behavior and perceived teacher characteristics. In their first study they evaluated factors that were related to increased disciplinary disproportionality. Their collection of ODRs from a district in a mid-sized urban Midwestern city revealed that 67% of all referrals could be grouped under “defiance of adult authority.” Black students comprised 30% of the student body, but 55% of all ODRs for defiance. This study provided early evidence of the concept of “subjective” behaviors. Defiance is now commonly referred to as a subjective behavior but at this point disciplinary outcomes were not being considered as distinct by behavior class, but instead as an overall group in most cases. This may have artificially deflated the effects found in prior studies, but Gregory and Weinstein (2007) opened the door for future research into the impact of the type of behavior on disciplinary outcomes.

Study 2 sought to further explore the patterns identified in Study 1 by examining individual student’s defiant and cooperative behavior in two classrooms. This study was quite small, with only 33 student participants and 43 teacher participants, limiting the conclusions that can be drawn from it. Student participants were matched to a “referring” and a “nominating” teacher, specifically the instructor who had most recently referred them for in-school suspension as well as the instructor that the student reported getting along with the best. Teachers were blind to their status to prevent affecting their ratings.

Unfortunately, the authors chose not to collect data in classrooms directly, but instead opted for survey measures from both the teachers and students as well as a review of their school records. Students completed the Defiance Scale, the Teacher Caring Scale, the Academic Expectations Scale and an adapted scale on teacher authority while teachers provided demographic information and ratings of the behavior of the defiance-referred student. Gregory and Weinstein (2007) found that teachers and students were largely congruent in their ratings of their behavior across classrooms and that teacher characteristics were significantly associated with student behavior. Specifically, both students and teachers rated their behavior as significantly more rule breaking and defiant in their referring teachers' classrooms (M = 2.79, SD = 0.75) than in their nominated teachers' classrooms (M = 1.75, SD = 0.59; $t(25) = 6.22$, $p < .001$). Additionally, students' perceptions of teachers' care ($\gamma_{03} = 0.94$, HLM $t(30) = 6.97$, $p < .001$) and student's perceptions of academic expectations were significant predictors of trust ($\gamma_{04} = 0.48$, HLM $t(30) = 5.45$, $p < .001$). Last, an interaction effect between academic expectations and caring ($\gamma_{05} = 0.15$, HLM $t(28) = 2.81$, $p < .005$) was found. The authors concluded that defiance among Black students in classrooms may be an interaction effect between students and teachers. This once again highlights the teacher as the potentially vulnerable point for effective intervention on disproportionate disciplinary practices.

The Gregory and Weinstein (2007) studies are limited by the single year data analysis in Study 1 as well as the use of HLM in Study 2 with only 33 students nested in classrooms. The authors described using HLM in an "innovative manner" to utilize the 33 students as 94 reports but provide no further explanation, limiting future researchers' ability to evaluate their findings. Despite these concerns, this study provided further evidence for the complex relationship between teacher behavior and disproportionate disciplinary outcomes. The authors specifically suggest policy change and teacher professional development as future avenues of research for potential interventions to address these outcomes.

A significant resurgence of research into the causes of disproportionality started in 2011. At least four studies were published addressing the topic that year alone. Two of the studies directly evaluated disproportionality within a SWPBS framework and so will be discussed in a future section focused on that framework. Two more general studies on disproportionality, Shirley and Cornell (2011) and Gregory, Cornell, and Fan (2011) evaluated disproportionality in traditional school settings. Shirley and Cornell (2011) built on Gregory and Weinstein (2007)'s work, whereas Gregory, Cornell, and Fan (2011) departed from the prior research base to primarily examine factors at schools with low rates of disproportionality.

Shirley and Cornell (2011) evaluated school climate factors related to disproportionate disciplinary outcomes for students of color using the School Climate Bullying Survey. They note that there is no single definition of school climate, but that it "refers to the quality and character of school life, typically as reflected in the nature of interactions among adults and students" (Shirley & Cornell, 2011, p. XX). The form assessed incidences of bullying at students' schools as well as school climate factors known to reduce bullying. They administered the survey to 400 middle school students in a suburban area of Virginia. In addition to the survey data, the researchers collected the total number of discipline referrals and out-of-school suspensions for each student via school records.

Black students were significantly less willing to seek help from their teachers and peers than their White counterparts, significantly more likely to endorse receiving aggressive attitudes from peers and significantly more likely to report being teased over clothing or physical appearance. However, results from the hierarchical regression were less conclusive. The Aggressive Attitudes Scale and the Willingness to Help Scale were significantly related to disciplinary outcomes but the relationship was small with only 8% of the variance explained ($R^2 = .08$, p < .001). When race was entered as the second step in the regression, it accounted for 11% of the variance in discipline outcomes alone. However, when entered into the model with the Aggressive Attitudes Scale and the Willingness to Help Scale, the variance explained by race was once again reduced to 8%, $F(1, 391)$, $p < .001$). This may be an indication of the inadequacy of the School Climate Bullying Survey, which was designed to address school bullying concerns and developed by a dominant population research team, for evaluating the concerns of students of color about potentially biased behavior in their schools.

Gregory, Cornell, and Fan (2011) used the authoritative parenting theory to examine the relationship between school structure and support and disproportionate disciplinary outcomes. The primary focus of the study was an investigation of schools with low rates of disproportionality in disciplinary outcomes. The study collected data from 25 randomly selected students from each of 289 out of the 314 high schools across Virginia. However only 199 schools were used in the final analysis due to outliers and other data issues.

The authors posited that teachers whose behavior management aligned with authoritative parenting styles, which they described as “highly demanding and high responsive," would lead to better disciplinary outcomes for students of color. They utilized the Student Perceptions of the School Survey to measure student perceptions of supportiveness while two measures, the Academic Press Scale and the Experience of School Rules sub scale were used to assess structure. All of the surveys used a Likert-type scale to rate agreement with statements. The Student Perceptions of the School Survey asks students how much they agree that adults "really care about students," and "treat all students fairly" (Austin & Duerr, 2005). The Academic Press Scale contains 6 items assessing how much teachers press students to use their "full effort" and attempt challenging work (Middleton & Midgley, 2002). Finally, the Experience of School Rules sub scale is a 7-item scale that is designed to measure perceptions of school rules as fair and "universally applied" (NCES, 2005). School data on suspensions and other disciplinary records was collected from the district.

First, they found that the sociodemographic characteristics of schools accounted for only 5% of the variance in the disparity in suspension rates. The authors then divided schools into high/low structure and high/low support, using a mean split approach, to evaluate the significant interaction effect of structure and support on disproportionate disciplinary outcomes. They found that schools with lower levels of both structure and support had a greater degree of disproportionality in their disciplinary outcomes. However, the patterns were not consistent and no statistical significance information is reported. Although this study attempted to provide a specific set of teacher behaviors that were consistently associated with disproportionate disciplinary outcomes, the results of this study are inconclusive and require further research.

Research into the causes of disproportionality in the 2000 – 2010s largely confirmed the initial findings of the studies of the 80s and 90s. They contributed further evidence of the role of bias in disparities in disciplinary outcomes for White and Black students. However, they also branched out and began to examine behavior types, further school factors and teacher specific factors that may impact disciplinary outcomes. These researchers provided a strong basis for future research directions but with many remaining questions.

Studies Evaluating Disproportionality in SWPBS

In the 2010s a specific strand of disproportionality research emerged; studies evaluating disproportionality within SWPBS and its impact on disproportionality. The system offered a major move from poorly defined recommendations for addressing disproportionality to the provision of evidence-based systemic practices that highlight the importance of positive behavior supports and consistent expectations. Four studies have been published specifically evaluating the relationship between SWPBS and disproportionate disciplinary outcomes: one in 2010, two in 2011 and one in 2012. These studies also represent some of the first forays into schoolwide interventions to address disproportionate disciplinary practices.

In 2010, Bradshaw, Mitchell, O'Brennan, and Leaf published the first evaluation of disparate disciplinary outcomes within 21, K-5 schools participating in a randomized trial of School-wide Positive Behavior Supports (SWPBS). Using a sample of 381 teachers and 6,988 children who were either Black or White and enrolled in the trial, they evaluated rates of office discipline referrals, teacher ratings of student behavior, classroom average ratings of disruptive behavior as well as the percentage of students in the classroom with one or more of the ODR categories. ODR data was collected using the School Wide Information System (SWIS), utilizing the same behavior categories and operational definitions as the current study. Again, they found that controlling for the teacher's rating of the severity of the behavior as well as the average level of classroom disruption, race remained a significant predictor of an office discipline referral.

Interestingly, the authors found that Black students with a Black teacher were more likely to be referred for a major ODR and less likely to receive a minor ODR when compared to a White counterpart, which provides preliminary evidence against racial/ethnic match as a sufficient solution to address these concerns. The authors argue that despite the implementation of SWPBS in the study schools, the, "discrepancies in disciplinary practices observed in the current study are likely perceived by the students as biased and may lead to negative student-teacher interactions as well as a diminished sense of school climate," (Bradshaw, Mitchell, O'Brennan, & Leaf; 2010).

Bradshaw, Mitchell, O'Brennan, and Leaf noted that their use of teacher and student ethnicity as a proxy for culture was inadequate, as further "cultural, contextual or economic" factors not addressed in the study may be more salient. They also note the limitation in their teacher report data, as the teacher report data may be providing noise in the data from unmeasured teacher characteristics. Additionally, the study only included elementary schools, where behavior concerns are consistently less severe than secondary settings. They conclude with a recommendation that additional empirical work is needed "to develop evidence-based skills-focused programs to promote cultural competency among teachers in order to reduce disproportionality in ODRs and provide increased opportunities for student learning," (Bradshaw, Mitchell, O'Brennan, & Leaf, 2010, p. 518).

Turning to the two studies published in 2011, Skiba, Horner, Chung, Rausch, May, and Tobin (2011) built on Gregory and Warner's (2007) work on behavior types with their work, "Race is Not Neutral: A National Investigation of African American and Latino Disproportionality in Schools," while Vincent, Swain-Broadway, Tobin, and May (2011) conducted one of the first extant evaluation studies of disproportionality through an examination of the impact of the implementation of SWPBS on the discipline gap.

Skiba, Horner, Chung, Rausch, May, and Tobin (2011) again found that Black students are significantly overrepresented in ODRs across all infraction types. However, a comparison of rates of referral for behavior types found that the highest rates of disproportionality between White and Black students were for the infraction types: Tardy/Truancy, Disruption, and Noncompliance. Both disruption and non-compliance continue to post some of the highest rates of disparities in discipline outcomes to this day. Tardiness and Truancy are considered objective behaviors, however they are significantly impacted by family economic and cultural circumstances, which may explain the increased rates of disproportionality.

Importantly, this study was conducted in 436 schools nationwide that were implementing SWPBS for at least the full academic year prior to the study and collecting behavioral data using the School Wide Information System (SWIS). The authors additionally examined the impact of race on administrative decision through an analysis of the probability of a given consequence for each behavior, and examination of the actual rates of consequences for each behavior by race. The results revealed a significant relationship between race and administrative decision. This indicates that administrators may also need additional specific interventions to address their biased behavior, however if most inappropriate referrals are prevented at the teacher level, the administrative impact becomes moot.

The authors suggest a “graduated model of discipline,” where the severity of the consequences is appropriately matched to the severity of the infraction. This may be a more effective method of organizing school disciplinary policy and practice. Operational definitions would clearly be a necessity in any discipline model, but especially one where behaviors and their consequences are clearly laid out. However, the caution across the studies that have evaluated the impact of administration so far have indicated that rigid policies are equally as ineffective at battling disproportionate disciplinary outcomes.

In the first quasi-experimental evaluation study, Vincent, Swain-Broadway, Tobin, and May (2011) examined the impact of the implementation of SWPBS on the discipline gap. Using a sample of 72 schools implementing SWPBS with fidelity and 81 who were not, the authors examined rates of ODRs for White and non-White students, both with and without an identified disability. Across the 3-year implementation of SWPBS and data collection, researchers found no significant impact on the magnitude of the gap between Black students enrolled and Black students receiving an ODR, indicating no significant impact of the implementation of SWPBS as designed nor the operational definitions contained within SWIS on disparate disciplinary outcomes for students of color. However, Vincent, Swain-Bradway, Tobin and May (2011) did discover an interesting pattern in their results; when comparing schools implementing SWPBS to those who were not, the discipline gap differs. For Black and White students, the discipline gap significantly differed by implementation status at each point of implementation. This indicates the potential for SWPBS to impact disproportionate disciplinary rates but suggests that SWPBS as currently designed is likely insufficient on its own. In an article published the same year, Vincent, Randall, Cartledge, Tobin, and Swain-Broadway (2011) argued that an integration of cultural responsiveness and SWPBS will likely provide a better foundation for addressing the divergent needs of non-dominant students than a framework that lacks that responsiveness. Taken together these articles make a strong argument for the incorporation of best practice operational definitions embedded within the cultural context.

Looking Forward

The body of evidence built over the last 30 years, across the nation, and both within and outside of SWPBS frameworks, consistently indicates that bias plays a role in the persistence of disparate rates of discipline for White and Black students. Little evidence exists to indicate potential malleable factors, but a few common themes are beginning to emerge such as systemic supports, cultural responsiveness, teacher and administrator behavior and the clarity of behavioral expectations. Looking past 2010 and into the future, the greatest research emphasis has been, and will need to be, on intervention studies and studies examining the role of systemic oppression as bias in school systems.

One such intervention study is Scott, Hirn, and Barber (2012), who conducted an intervention case study in a 1,450-student high school in an urban, Midwestern city in the USA that was not implementing SWPBS. However, most relevant to the current study, the school was engaged in a process to better describe student problem behavior that was similar to the development of operational definitions and of data collection about student problem behavior. The authors supported the school staff, first in the collection and analysis of data on disproportionate disciplinary outcomes and then in the staff's response to their school data on a monthly basis.

Prior to the start of the study, students of color made up 40% of the student body and 73% of all referrals. In three subsequent monthly sessions, the author presented the data to the teaching teams, allowed them to present hypotheses about potential causes, and then provided them with data on their hypotheses. Each of the hypothesized causes posited by the teachers failed to explain the variance in disciplinary outcomes when the data collected by the teachers was evaluated. In the fourth month, the author returned to the staff and they engaged in a data exploration process that revealed a specific time of day as the greatest predictor of differential rates of referral for White and Black students. The staff developed interventions based on their hypothesized explanations for the specific time which included voting to utilize positive behavior support strategies to encourage students to follow expectations during that time. Additionally, the principal provided teachers with professional development on cross-cultural respectful interactions.

At that point, the researcher provided no more support. Prior to the intervention the daily rate of ODRs for Black students was .94 per 100 students, while in the prior year it varied between 2 and 3.64 per 100 students for the same period. However, the daily rate of ODRs for White students post intervention was .24 per 100, down from .39 in the year prior. Given the relatively limited nature of the intervention, the significant reduction in the rate of ODRs is notable, however Black students continued to be disproportionately referred for disciplinary infractions. This study represents the closest to a culturally responsive SWPBS intervention package to date, with locally defined operational definitions of behavior, despite the fact that the school was not implementing the SWPBS framework by name. The extremely promising results demonstrated in such a short time period provide substantial support for the current study.

Finally, in 2017 one of the first examinations of systemic racism's impact on disproportionate disciplinary outcomes was published by Anyon, Lechuga, Ortega, Downing, Greer, and Simmons. They note that qualitative research has indicated that "race-neutral" school discipline policies are most associated with disproportionate disciplinary outcomes, as they are frequently embedded within White, middle-class cultural standards that are presumed universal by the implementers. Given this information, the authors sought to evaluate data from Denver Public Schools disciplinary outcomes in order to assess both school and student level factors that were correlated with the discipline gap, through a critical race theory lens. They chose Denver as the site of this evaluation due to the 8 years of discipline reforms designed to reduce reliance on out-of-school suspension in the district (Anyon et al., 2017). This initiative was focused on the implementation of counseling and universal social emotional programming. District-level discipline policies did not mandate training for educators on implicit bias or culturally responsive pedagogy, eliminate colorblind codes of conduct that criminalize the dress and mannerisms associated with youth of color (e.g., banning hoodies, hats, and particular hairstyles), or address structural concerns such as resource allocation, teacher preparedness, or school segregation

Anyon et al. found that, despite persistent disparities between Black and White students for referrals from classrooms, outside of the classroom Black students were less likely to be referred than their White peers. That is, Black students were much more likely to be over referred by the teachers who they saw on a regular basis than in the spaces were adult contact was more sporadic. This both supports the theory that teachers in classrooms may be a key point of intervention and that systemic solutions that are not specifically designed to address disproportionality but do address aspects of school culture that may contribute to disproportionality may be effective at addressing but insufficient to eliminate disproportionality.

The Anyon et al. research provides the first evidence that despite significant, sustained investment in positive behavior strategies, disproportionality continues to persist without culturally responsive frameworks in place as well. Denver Public Schools represents a wide range of racially, economically and culturally diverse students with a large data set providing an unparalleled opportunity for examining these issues. However, these results have limited application since they focus only on locations where referrals are more or less likely to occur, which is a challenging malleable factor to address. Culturally responsive operational definitions may provide an opportunity to address the concerns raised in the Anyon et al. (2017) article, through the provision of specific operational definitions for each space.

Summary

Across more than three decades of evidence, the impact of bias on disproportionate disciplinary practices is consistent. However, the body of evidence also points to a complex interplay of cultural, school, interpersonal as well as intrapersonal factors impacting the disparity between White and Black disciplinary outcomes in the same schools. There is also an emerging thread of culturally responsive, clear and appropriate disciplinary policies as the next step in intervention planning to address this concern. An evaluation of operational definitions, a key policy decision that impacts teacher training as well as the cultural responsiveness of a behavior system, is a next step to move both research and practice forward.

CHAPTER III

A randomized-control, experimental design was utilized to examine the research questions. Participants were assigned at random to condition for two of the five independent variables: operational definition and race of target student. Time was included as an independent variable to account for pre-test and post-test. For two additional independent variables, participants were exposed to both conditions and randomly assigned to a counterbalanced order to prevent order effects: behavior concern level and behavior class. The remaining independent variable included in the study was cultural context, a qualitative variable that was not manipulated with random assignment and for which no causal conclusions could be drawn. The dependent variable was accuracy of participants ratings of behavior as measured by the Euclidian distance from the expert rating, a quantitative variable where participants' ratings of vignettes were compared with expert ratings to determine total distance, positive or negative, from the experts' ratings.

Participants

Study participants. One-hundred twenty educators with complete data were included in this study. All participants were educators working in K-12 public school settings in the Pacific Northwest or the U. S. Midwest. Thirty-eight of the participants taught in the Pacific Northwest, whereas 82 of the participants reported teaching in the urban Midwest. Participants ranged from 23 to 58 years of age, with a mean age of 38. Participants self-reported their race, with 72 participants or 60% of the sample identifying as White, 30 or 25% identifying as Black and 18 or 15% reporting another racial or ethnic identity such as Hispanic/Latinx, Pacific Islander or Indigenous American.

Participant race by cultural context was examined using a Chi Square analysis. The relation between the race of the participant and their cultural context was significant $\chi^2(2) = 17.21, p < .001$. Participants in the Urban Midwest were significantly more likely to be Black or non-Black persons of color than participants in the Pacific Northwest. A contingency table of participant race by cultural context is presented in Table 1.

Table 1

Contingency Table of Participant Race by Cultural Context

Participant race	Urban Midwest	Pacific Northwest	Total
White	39	33	72
Black	28	2	30
Non-Black person of color	15	3	18
Total	82	38	120

Participants reported their teaching credentials and education as well. A majority of participants reported having an undergraduate teaching degree (*n* = 53, or 44.2%) or a master's degree (*n* = 47, or 39.2%), whereas 13 (10.8%) reported having an alternative teaching certification through their state and 5 (5.8%) reported having obtained a Ph.D. or equivalent doctoral degree in education. Participants reported a wide range in their years' experience teaching, with 34.2% reporting 5 or fewer years-experience, 55.8% of participants had 6 to 15 years' experience, and 10% had more than 16 years of experience.

Participants also provided information about their prior experience with SWPBS. More than 83% of participants (*n* = 100) reported having SWPBS systems in place at their schools. Of those 100 participants, only 4 reported having 1 hour or less of SWPBS training. Thirteen reported having fewer than 3 hours of training, whereas 21 participants had more than 4 but less than 6 hours of training. Sixty-two participants, or more than half of the total sample, had more than 6 hours of training in SWPBS, with 8 reporting having more than 40 hours of SWPBS training.

Experts. Video vignettes were rated by a panel of 4 experts representing both cultural contexts and a variety of professional backgrounds. All experts had substantial training in positive behavior support systems and extensive experience implementing those systems at a district or school level. Expert 1 was a white, male-identifying administrator for a school district in the suburban Pacific Northwest with a graduate degree in special education and more than 25 years' experience as a teacher, principal, and district administrator. Expert 2 was a Black woman who was a Board-Certified Behavior Analyst (BCBA) with more than 5 years' experience working in schools implementing SWPBS. Expert 2 recently relocated to the suburban Pacific Northwest from the suburban Midwest but was originally raised in Nigeria. Expert 3 was a white, non-binary identifying person with a Ph.D. in school psychology and licensure as a BCBA; with more than 10 years' experience implementing behavior programming within SWPBS systems. Finally, Expert 4 was an African American male with a graduate degree in special education from the urban Midwest. He was a former school administrator with more than a decade of experience implementing SWPBS systems in public schools. At the time of the study, he operated a private consultation practice that supported schools' implementation of SWPBS systems nationwide. The expert panel included individuals who represented both cultural contexts and had a wide range of expertise in implementing disciplinary practice and policy within school settings.

Cultural Context

This study was conducted in two different contexts: one large urban school district and a group of smaller, suburban, or rural districts. These districts were geographically and culturally distinct.

Recruitment

Recruitment occurred in one urban school district in the Urban Midwest and in several suburban and rural school districts in the Pacific Northwest. Recruitment was extended to multiple districts in the Pacific Northwest due to difficulty recruiting subjects. Recruitment occurred via outreach directly to schools. The survey was distributed online, and a monetary incentive of $30 was offered in exchange for participation.

Informed Consent

Informed consent was obtained after participants had the opportunity to review a written document explaining the purpose, risks, benefits and procedures of the study on the opening screen of the online module. Informed consent was recorded by selecting the "consent" button on the survey's online interface.

Video Vignettes

The study utilized video vignettes to provide visual depictions of problem behaviors that were realistic to a school context, at multiple grade levels, in order to provide realistic stimuli for decisions about an appropriate disciplinary action. The vignettes, each less than 30 seconds in length, featured one target student misbehaving, surrounded by a diverse cast of fellow students in a realistic school setting. Vignettes were drawn from materials prepared for training or other educational purposes with actors. The video vignettes were rated by an expert panel consisting of 4 individuals with expertise in Positive Behavior Intervention Supports (SWPBS). Experts are described in the following section.

For each video vignette, the experts provided their rating on a 1 to 11 Likert-type Scale on four questions from Okonofua and Eberhart (2015). The questions were (1) “How severe was the student’s misbehavior?” (2) “To what extent is the student hindering you from maintaining order in your class?” (3) “How irritated do you feel by the student?” and (4) “How severely should the student be disciplined?” Ratings were averaged across the panel, for each question individually, for each vignette. This provided a quantitative “expert rating” of the behavior depicted for each question, for a total of 4 expert ratings per vignette.

Experts initially rated a total of 32 videos: 4 potential vignette examples for each behavior class, each level of concern and for each race of the target student (i.e., Black and White objective mild, objective moderate, subjective mild and subjective moderate). The researcher then chose 2 videos from each set of 4 to represent that category and class by comparing videos across levels of concern and race but within class. This was to ensure mild and moderate behaviors were rated distinctly from one another and that videos depicting a Black target student were similarly rated to those depicting a White target student. The researchers selected final video vignettes for the study based on the match to study need based on expert rating. Vignettes were chosen that accurately represented the behavior type, mild and moderate levels of intensity of the behavior, as well as the race of the target student. A total of 16 video vignettes were selected, two for each race of the target student (Black and White) to represent the objective mild, objective moderate, subjective mild and subjective moderate behavior class and severity at Time 1 and Time 2 in a counterbalanced order.

Table 2 depicts the expert mean rating per question and an overall mean expert rating for each vignette included in the study. Average expert ratings across all 4 questions for moderate behaviors ranged from 2.25 to 5.41, while expert average ratings for mild behaviors ranged from 1.28 to 2.00.

Table 2

Expert Average Ratings for Video Vignettes by Question, Behavior Class, Behavior Concern Level and Race

	Subjective mild		Subjective moderate		Objective mild		Objective moderate	
Question	Form 1	Form 2	Form 1	Form 2	Form 1	Form 2	Form 1	Form 2
				White target student				
Q1	2.00	1.00	3.50	4.00	2.00	1.50	4.75	5.50
Q2	1.38	1.25	3.75	3.50	2.00	1.00	2.75	4.25
Q3	1.00	1.25	3.00	3.25	2.00	1.00	1.25	4.00
Q4	1.25	1.00	2.50	2.75	2.00	1.50	4.00	5.50
M	1.41	1.13	3.19	3.38	2.00	1.25	3.19	4.81
				Black target student				
Q1	1.38	2.25	2.25	5.25	1.00	3.00	6.00	5.75
Q2	1.25	2.00	2.25	4.00	1.75	1.00	4.00	3.75
Q3	1.25	1.50	2.00	3.00	2.00	1.00	4.25	3.75
Q4	1.25	1.75	2.50	3.25	1.75	2.25	6.25	5.75
M	1.28	1.88	2.25	3.88	1.63	1.81	5.13	4.75

The mean distance between experts was also calculated to evaluate the stability of experts' ratings. Table 3 depicts the mean distance between expert ratings for each video vignette as well as the minimum and maximum distances between expert ratings. In order to calculate the mean distance, a mean expert rating for each vignette was calculated by taking the average of the 4 question responses for each expert and then calculating the average distance between each expert (E1 to E2, E1 to E3, E1 to E4, E2 to E3, E2 to E4, E3 to E4), which were then averaged again. This final average of all 6 expert pair comparison means is the expert mean distance reported above. Combined mean expert ratings are the expert mean distances for video vignette Form 1 and video vignette Form 2 averaged. Expert ratings were most stable for video vignettes that depicted a White target student demonstrating a mild, subjectively defined misbehavior. Expert ratings were least stable for video vignettes that depicted a Black target student demonstrating a moderate, objectively defined behavior.

Table 3

Mean, Minimum and Maximum Distance between Expert Ratings of Video Vignettes

	Mean distance				
Race, behavior class, and level of concern	Vignette 1	Vignette 2	Combined	Min distance	Max distance
Black objective mild	1.44	3.03	2.23	1	6
Black objective moderate	7.33	7.90	7.61	1	10
Black subjective mild	1.13	3.86	2.49	1	5
Black subjective moderate	2.99	4.84	3.91	1	7
White objective mild	1.89	0.94	1.41	1	3
White objective moderate	4.26	5.57	4.91	1	9
White subjective mild	1.86	0.71	1.28	1	4
White subjective moderate	5.37	4.65	5.01	1	8

Independent Variables

The independent variables that were examined in this study are: operational definition condition, time, race of target student, cultural context, behavior class, and behavior concern level.

Operational definitions. Operational definition condition was a qualitative, between-subjects effect with three levels, control, comparison and experimental, that was experimentally manipulated via randomization to condition. In the control condition they were only provided the video vignette and questions. In the comparison condition participants received the operational definition of the behavior type from the Schoolwide Information System (SWIS). In the experimental condition they received the same definition with a decision-making flow-chart developed by Horner and Nese (2014) to support educators' disciplinary decision making.

These three conditions were designed to allow for the comparison of multiple conditions that correspond to plausible real-world scenarios. The control condition was designed to represent schools with no SWPBS systems in place or limited systems without common definitions of problem behaviors, the comparison condition represented schools that provide SWPBS as currently recommended, while the experimental condition represents an approach to SWPBS where operational definitions of behaviors are augmented with more decision-making support. It was not feasible to implement a culturally responsive approach to SWPBS in this study, as it would have necessitated individualization of the operational definitions, examples and non-examples within each school within the study.

The operational definitions for the comparison condition were drawn from the School Wide Information System (SWIS), a web-based program used for recording and charting discipline referral data that has been described as an example of an effective practice in the implementation of SWPBS (Tobin, Lewis-Palmer, & Sugai, 2002). The SWIS system is utilized by over 10,000 schools across the United States and includes operational definitions of problem behaviors to encourage accurate data tracking (May et al., 2000; Todd & Horner, 2001, http://www.swis.org). The SWIS system has broad adoption, which presented the opportunity to examine operational definitions that are already widely in use across the nation, increasing the external validity of study results.

For the experimental condition, the SWIS definitions were expanded to include a decision-making flow chart developed by Horner and Nese (2014). This flow chart was designed by one of the developers of SWPBS to support teacher disciplinary decision making when coupled with the SWIS system and appropriate training.

Differences between conditions were not expected at Time 1, or prior to the presentation of the operational definitions. However, at Time 2, after exposure to operational definition condition, it was expected that participant ratings in the experimental condition would have the greatest congruence with the expert rating for each question. Please note that with Euclidian distance, approaching 0 is desirable as it indicates better alignment between participant and expert ratings. Further, operational definitions were expected to interact with race, level of concern about the behavior, and behavior class to reduce bias.

Race. Race of the target student was a qualitative, between-subjects effect with two levels, White and Black, where participants were randomized to condition.

Participants were exposed to video vignettes that featured a single target student misbehaving, surrounded by a diverse cast of fellow students. In every video in a race condition, the "target student," or the student demonstrating the misbehavior, featured either all White or all Black target students in an attempt to prevent subjects from intentionally manipulating their answers based on race. Only White or Black target students were chosen due to the availability of video vignettes and to limit conditions to help ensure sufficient power to evaluate the results. Race was expected to interact with operational definitions, cultural context, behavior class, behavior concern and operational definition condition.

Cultural context. Cultural context was a qualitative, between-subjects effect with two levels, suburban Pacific Northwest and urban Midwestern United States, that was quasi-experimentally evaluated, with no randomization to condition. Cultural context was determined based on the location of the school the participant teaches in, either the suburban Pacific Northwest or the urban Midwestern United States, as reported by the participant.

Although there were significant limitations to describing cultural context as the location where an individual works, there is no perfect instrument for measuring the complex interplay of geography, history, religion, upbringing, education and other factors that comprise an individual's cultural context. However, this study sought to evaluate common cultural beliefs among groups of individuals, which was best served by the comparison of two locations that differed greatly in geographically, density, history, racial and economic characteristics, allowing for a more holistic evaluation of the impact of cultural context beyond any single individual variable. Therefore, an urban location in the Midwest with a metropolitan population of more than 3 million individuals with a huge diversity of race, economic advantage and religion as well as a suburban location in the Pacific Northwest with a predominantly white population of less than 500,000 were chosen to represent fundamentally different contexts. A possible interaction of cultural context and race of the target student was hypothesized. Differing cultural expectations for behavior, which are less explicit and therefore were expected to be more pronounced in the subjective behavior class. For example, teachers in the urban Midwest may have more familiarity with Black students and therefore may be more or less likely to believe that their behavior is disruptive and requires removal from the classroom than those teachers with less familiarity. Analyses were conducted to evaluate the impact of cultural context on participants' ratings of student behavior, as cultural context was expected to interact with race, behavior class and level of concern about the behavior to impact teachers' perceptions of student behavior.

Behavior class. Behavior class was a qualitative, within-subjects variable with two levels, subjective and objective, where participants were exposed to both levels. Behavior types, or specific behavior infractions such as physical aggression or defiance, were sorted into the subjective or objective behavior classes based on a comprehensive review of the literature on disproportionate disciplinary outcomes by behavior class.

Objective behaviors were defined as behaviors that could be identified through the observation of a discrete event without judgement regarding whether the intensity or quality of the behavior warrants an ODR. Objective behaviors included: theft, vandalism, tardiness, physical aggression, inappropriate location and possession of contraband. Subjective behaviors were defined as "behaviors that require not simply observing a discrete, objective event (e.g., a student smoking), but a significant judgement regarding whether the intensity or quality of the behavior warrants an ODR," by Greflund (2013). Subjective behaviors included defiance, disrespect, disruption, verbal harassment/bullying, inappropriate display of affection, lying, and inappropriate language.

A wider gap in rates of office discipline referrals between White and Black students has been found for behaviors classed as subjective than for those classed as objective (Balderas, 2015; Gion, McIntosh, & Smolkowski, 2017; Girvan, Greflund, McIntosh, Mercer, & May, 2014; Skiba, Michael, Nardo, & Peterson, 2002). Black students are referred to the office and suspended at far higher rates for behaviors that are classified as subjective than for those classified as objective. In fact, some research has indicated that there is not a significant gap between White and Black students in referral rates for behaviors classified as objective (Skiba, Michael, Nardo & Peterson, 2002). It was expected that similar differences in participants ratings of objective and subjective behaviors by race would be found in this study as were found in prior work.

Given the significant differences in referral rates for behaviors classed as subjective and objective, this study evaluated the impact of operational definitions on teachers' ratings of both subjective and objective behavior. Behavior class was expected have a significant effect on teachers' perceptions of the problem behavior, with subjective behaviors expected to be rated less accurately, as demonstrated by a greater Euclidian distance from expert rating, than objective behaviors. Examining both conditions allowed for a better understanding of all of the types of decisions teachers must make during the course of an average school day, with the hopes of providing a better understanding of future intervention directions.

Behavior concern level. The level of concern about the behavior was a qualitative, within-subjects variable with two levels, mild and moderate, where participants were exposed to both levels through video vignettes. The level of the concern, mild or moderate, was derived from the average expert rating across all four questions for each vignette, thereby measuring the level of behavior concerns across 4 separate dimensions of behavior. Groups were based on an analysis of mean, median, and range, with the lowest scoring behaviors rated as mild and the highest scoring behaviors rated as severe, and the remaining behaviors rated as moderate. Severe behaviors were removed from the analysis sample in order to analyze more common levels of classroom misbehavior. However, a range in levels of behavior concern was necessary to assess the efficacy of operational definitions in addressing the "grey areas" of discipline. For example, when a student flagrantly violates the rules, it is much clearer that they should receive at least some consequence. However, for more mild and moderate behaviors the researchers hypothesized that instructors would have less consistent guidelines, both internally and externally imposed, and that therefore unconscious beliefs or implicit biases may have been more likely to impact disciplinary decisions. The current reliance on personal beliefs and values was hypothesized to be one factor that could be contributing to disproportionate disciplinary outcomes for White and Black students. Given that SWPBS systems decrease rates of disproportionality, operational definitions were believed to potentially decrease teachers' reliance on personal beliefs and instead provide a specific standard by which to judge behavior. Therefore, it was hypothesized that operational definitions with decision making supports may further decrease reliance on personal beliefs and values which then may further increase the accuracy of ratings of misbehavior across forms for White and Black students.

Dependent Variables

The dependent variable in this study was accuracy of participant ratings of misbehavior, that is how close participants' ratings were to experts' ratings. It was hypothesized that more accurate ratings may reduce disproportionality in disciplinary outcomes due to more accurate identification of problem behaviors requiring out of classroom disciplinary practices. Disproportionality was not directly measured in this study's outcomes.

Participant accuracy was measured based on the distance between their answers and experts' answers on the following four questions from Okonofua and Eberhart (2015): (1) "How severe was the student's misbehavior?" (2) "To what extent is the student hindering you from maintaining order in your class?" (3) "How irritated do you feel by the student?" and (4) "How severely should the student be disciplined?" Each video vignette was rated on a scale of 1-11 for each question by both participants and experts. Expert ratings from each member of the expert panel were averaged to provide a single overall expert rating for each question. Participants also provided ratings for each question, for each vignette they were presented, through a slider bar that allowed for fractional answers in the Qualtrics web interface.

Participant accuracy was assessed by the Euclidian distance of the participant's rating from the expert rating of the video vignette. Euclidean distance was defined as the Squareroot(sum(Q_i – E_i)^2) where Q_i was the participants' rating of the Question i and E_i is the expert mean rating of the question for that vignette. The Euclidean distance can also be understood as the “true straight line” distance between two points. The Euclidian distance was selected to allow for an examination of the differences between the expert ratings and the participant ratings that accounted for the difference without consideration to whether the difference is positive or negative. For example, if a video vignette had an expert rating of 2.9 on Question 1 while one participant rated it 0.9 and another rated it 4.9, both would have a Euclidian distance of 2 from the expert rating. This allowed for the examination of whether or not an independent variable was related to increases in the accuracy of teacher’s ratings of behavior and their related decisions regardless of the direction of the difference. Examining this difference regardless of the direction is important as both too much and not enough appropriate disciplinary responses are inappropriate and have negative consequences for students.

Procedure

The study was conducted using the Qualtrics Online Survey software (Qualtrics Labs Inc., 2009). This study utilized an online survey interface to ensure all participants were exposed to the conditions with automated blinding and randomization, reducing the chance of human error. This software includes randomization features, can only be accessed with a unique user link to prevent unauthorized access and protects subject confidentiality, and meets human subjects research compliance standards. An online survey allowed the study to be completed either at the school site or at any location convenient to the participant.

Table 4 displays an overall schematic of the participant group allocation for operational definitions and study flow.

Participants began on the informed consent screen, where they had the opportunity to understand the risks and benefits of the study before providing consent. After providing consent, participants provided demographic information. Participants were then randomized to condition based on (a) race of the target student, (b) operational definition condition, and (c) counterbalanced order for viewing vignettes. They then viewed 4 vignettes at Time 1: (a) one mild objective behavior vignette, (b) one moderate objective behavior vignette, (c) one mild subjective behavior vignette, and (d) one moderate subjective behavior vignette. The order of these vignettes was counterbalanced to control for order effects across: (a) form, (b) intensity of the problem behavior, and (c) the behavior class. That is, the order in which the participants were presented the subjective and objective behavior class videos, the levels of concern (either moderate or mild) and the video forms (either 1 or 2) was manipulated in order to ensure that potential order effects were balanced across: (a) objective/subjective behavior, (b) mild/moderate behavior, and (c) form.

Participants were randomized to one of 16 potential conditions/orders. Orders were counterbalanced such that each combination of race, behavior type, concern, and video form was presented first an equal number of times. For example, a participant could have been randomized to either White target student condition 6 or Black target student condition 14 first saw a video depicting a student demonstrating an objective problem behavior at moderate intensity for form 2 at Time 1 while a participant assigned to either condition 3 or 11 first saw a Form 1 video vignette depicting a subjective mild intensity problem behavior. See Tables 5 and 6 for a complete delineation of behavior vignette counterbalancing conditions. Counterbalancing controlled for any order effects that might have occurred as a result of all participants seeing the videos in any one particular order and any potential differences in Form 1 and Form 2 of the vignette. Participants then rated those vignettes on the same questions and with the same Likert scale as the expert panel.

Table 4

Study Procedures by Operational Definition Condition and Time

Group (randomized)	Time 1 (pre-test) (counterbalanced by form, class & concern)	Operational Definition Condition	Time 2 (post-test) (counterbalanced by form, class & concern)
Experimental	Time 1 vignettes (4 vignettes: (a) subjective mild; (b) subjective moderate, (c) objective mild, and (d) objective moderate)	SWIS ODs + decision making framework (Horner & Nese, 2014)	Time 2 Vignettes (4 vignettes: (a) subjective mild; (b) subjective moderate, (c) objective mild, and (d) objective moderate)
Comparison	Time 1 vignettes	SWIS ODs	Time 2 vignettes
Control	Time 1 vignettes	None	Time 2 vignettes

Note. All time 1 ratings were pre-intervention, while time 2 ratings were post- operational definition intervention. Participants were also randomly allocated to either White or Black target students for all video vignettes at both time points.

Participants were randomized to one of 16 potential conditions/orders. Orders were counterbalanced such that each combination of race, behavior type, concern, and video form was presented first an equal number of times. For example, a participant could have been randomized to either White target student condition 6 or Black target student condition 14 first saw a video depicting a student demonstrating an objective problem behavior at moderate intensity for form 2 at Time 1 while a participant assigned to either condition 3 or 11 first saw a Form 1 video vignette depicting a subjective mild intensity problem behavior. See Tables 5 and 6 for a complete delineation of behavior vignette counterbalancing conditions. Counterbalancing controlled for any order effects that might have occurred as a result of all participants seeing the videos in any one particular order and any potential differences in Form 1 and Form 2 of the vignette. Participants then rated those vignettes on the same questions and with the same Likert scale as the expert panel.

Participants who were randomized to the control condition were only exposed to the video vignettes and questions, while those in the comparison condition were exposed to the operational definitions of the behavior types and those in the experimental condition were exposed to the operational definitions of the behavior types as well as a decision-making flow chart. All participants then viewed 4 more vignettes: one mild objective behavior vignette, one moderate objective behavior vignette, one mild subjective behavior vignette and one moderate subjective behavior vignette. The order of these vignettes was counterbalanced to balance order and form effects across both the intensity of the problem behavior and the behavior class (see Table 1 and Table 2 vignette flow above). Participants again rated the vignettes on the same questions and with the same Likert-type scale as previously described

Analyses

The research questions were evaluated with a 6-way, mixed-effects analysis of variance. The between subject effects were race of the target student, operational definition intervention condition, and cultural context. Race of the target student was a two-level (White and Black) between subject effect, as was cultural context (Pacific Northwest and Urban Midwest) while operational definition had three levels: control, comparison and experimental.

Table 5

Vignette Orders for White Target Students Counterbalancing the Effects of Race of Target Student, Vignette Form, Behavior Class, and Behavior Concern Level

Order	Time 1 vignette order				Time 2 vignette order			
	1st	2nd	3rd	4th	1st	2nd	3rd	4th
1	White Form 1 obj mild	White Form 1 obj mod	White Form 1 sub mild	White Form 1 sub mod	White Form 2 obj mild	White Form 2 obj mod	White Form 2 sub mild	White Form 2 sub mod
2	White Form 1 obj mod	White Form 1 obj mild	White Form 1 sub mod	White Form 1 sub mild	White Form 2 obj mod	White Form 2 obj mild	White Form 2 sub mod	White Form 2 sub mild
3	White Form 1 sub mild	White Form 1 sub mod	White Form 1 obj mild	White Form 1 obj mod	White Form 2 sub mild	White Form 2 sub mod	White Form 2 obj mild	White Form 2 obj mod
4	White Form 1 sub mod	White Form 1 sub mild	White Form 1 obj mod	White Form 1 obj mild	White Form 2 sub mod	White Form 2 sub mild	White Form 2 obj mod	White Form 2 obj mild
5	White Form 2 obj mild	White Form 2 obj mod	White Form 2 sub mild	White Form 2 sub mod	White Form 1 obj mild	White Form 1 obj mod	White Form 1 sub mild	White Form 1 sub mod
6	White Form 2 obj mod	White Form 2 obj mild	White Form 2 sub mod	White Form 2 sub mild	White Form 1 obj mod	White Form 1 obj mild	White Form 1 sub mod	White Form 1 sub mild
7	White Form 2 sub mild	White Form 2 sub mod	White Form 2 obj mild	White Form 2 obj mod	White Form 1 sub mild	White Form 1 sub mod	White Form 1 obj mild	White Form 1 obj mod
8	White Form 2 sub mod	White Form 2 sub mild	White Form 2 obj mod	White Form 1 obj mild	White Form 1 sub mod	White Form 1 sub mild	White Form 1 obj mod	White Form 1 obj mild

Note. Participants are randomly assigned to 1 of the 16 counterbalanced orders.
Note. Sub = subjective, obj = objective, mod = moderate

Table 6

Vignette Orders for Black Target Students Counterbalancing the Effects of Race of Target Student, Vignette Form, Behavior Class, and Behavior Concern Level

	Time 1 vignette order				Time 2 vignette order			
Order	1st	2nd	3rd	4th	1st	2nd	3rd	4th
9	Black Form 1 obj mild	Black Form 1 obj mod	Black Form 1 sub mild	Black Form 1 sub mod	Black Form 2 obj mild	Black Form 2 obj mod	Black Form 2 sub mild	Black Form 2 sub mod
10	Black Form 1 obj mod	Black Form 1 obj mild	Black Form 1 sub mod	Black Form 1 sub mild	Black Form 2 obj mod	Black Form 2 obj mild	Black Form 2 sub mod	Black Form 2 sub mild
11	Black Form 1 sub mild	Black Form 1 sub mod	Black Form 1 obj mild	Black Form 1 obj mod	Black Form 2 sub mild	Black Form 2 sub mod	Black Form 2 obj mild	Black Form 2 obj mod
12	Black Form 1 sub mod	Black Form 1 sub mild	Black Form 1 obj mod	Black Form 1 obj mild	Black Form 2 sub mod	Black Form 2 sub mild	Black Form 2 obj mod	Black Form 2 obj mild
13	Black Form 2 obj mild	Black Form 2 obj mod	Black Form 2 sub mild	Black Form 2 sub mod	Black Form 1 obj mild	Black Form 1 obj mod	Black Form 1 sub mild	Black Form 1 sub mod
14	Black Form 2 obj mod	Black Form 2 obj mild	Black Form 2 sub mod	Black Form 2 sub mild	Black Form 1 obj mod	Black Form 1 obj mild	Black Form 1 sub mod	Black Form 1 sub mild
15	Black Form 2 sub mild	Black Form 2 sub mod	Black Form 2 obj mild	Black Form 2 obj mod	Black Form 1 sub mild	Black Form 1 sub mod	Black Form 1 obj mild	Black Form 1 obj mod
16	Black Form 2 sub mod	Black Form 2 sub mild	Black Form 2 obj mod	Black Form 1 obj mild	Black Form 1 sub mod	Black Form 1 sub mild	Black Form 1 obj mod	Black Form 1 obj mild

Note. Participants are randomly assigned to 1 of the 16 counterbalanced orders.

The within-subjects’ effects were behavior class, behavior concern level, and time. Behavior class was a 2-level (subjective and objective) within subjects' effect, as were the level of behavior concern (mild and moderate) and time (Time 1 and Time 2).

Hypotheses

It was expected that at Time 2 there would be an interaction effect between behavior class, behavior concern level, operational definition condition, and race. The experimental operational definitions condition was expected to significantly increase the accuracy of teachers' ratings of student behavior (that is, a smaller distance from experts). However, this effect was expected to depend on the behavior class, the level of behavior concern, and race of the target student. Consistent with prior research, it was expected that teachers within each cultural context would rate White and Black students relatively similarly on behaviors classified as objective; while there would be significant differences in participants' ratings of White and Black students behaviors that were classified as subjective, indicating racial differences that were tied to differences in behavior class. Further, those differences were hypothesized to be more pronounced for behaviors classified as moderate than those classified as mild, as moderate misbehavior was expected to be more universally agreed upon than more mild behaviors. A hypothetical pattern of interaction for each behavior class, level of concern, race of target student and operational definition condition is depicted in Figure 2:

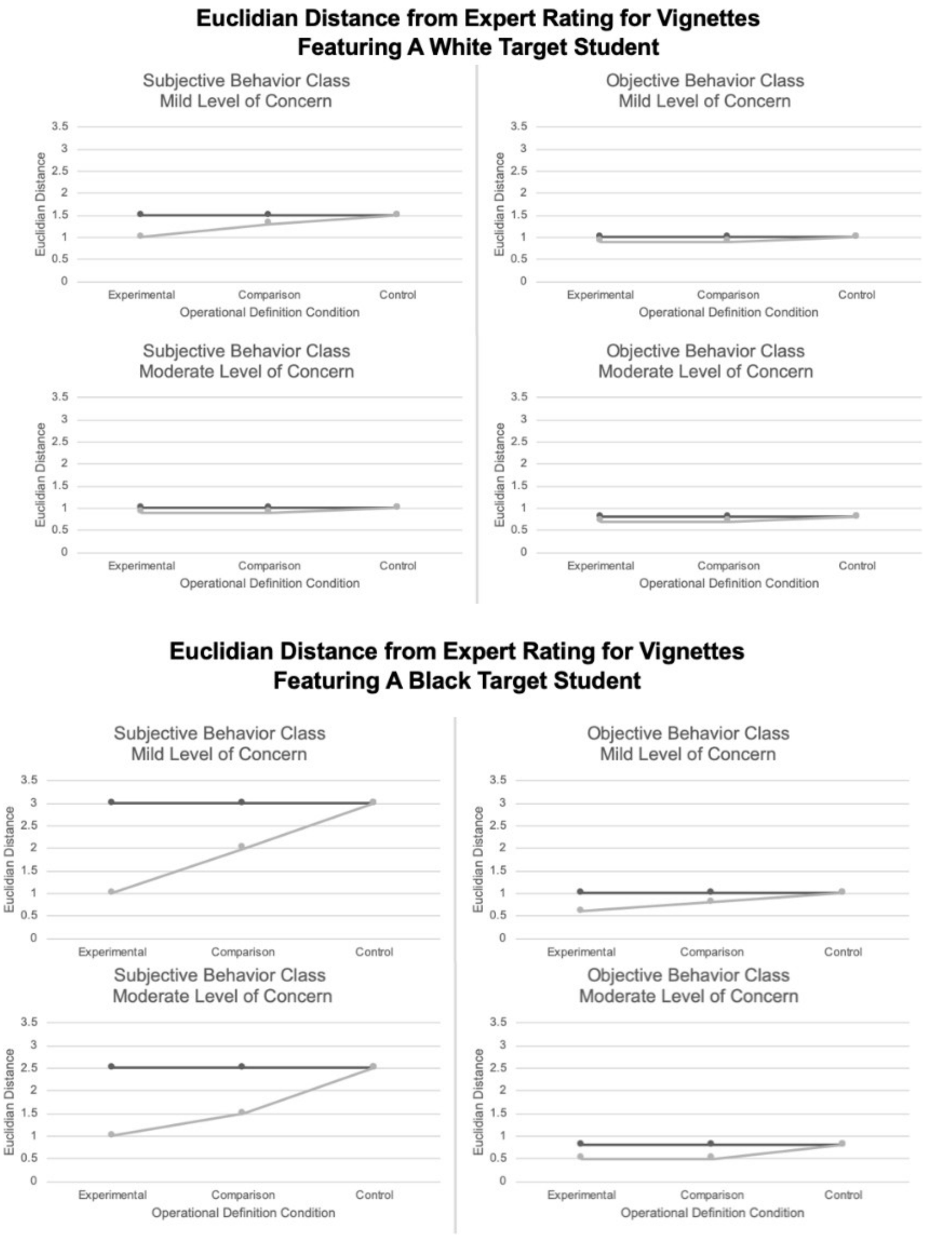


Figure 2. Predicted 5-way interaction of operational definition condition, time, behavior class, level of behavior concern and race.

Research Questions

In the theoretical hypotheses outlined above, the most significant increase in accuracy of participant ratings was predicted for those teachers in the experimental condition, who were exposed to vignettes depicting a Black target student engaging in a mild, subjective behavior. That is, a five-way, time*condition*race*concern*class interaction effect. Further, the researchers expected that the predicted five-way interaction may also depend on the cultural context in a six-way interaction.

1. To what extent does providing operational definitions and decision-making support improve educators' accuracy in rating behavior, as compared to an expert panel?
2. Does the impact of operational definition condition depend on the level of concern about the behavior, the race of the target student, and the behavior class?
3. Does this impact depend on cultural context?

Summary

This study utilized a randomized-control, experimental design to evaluate the impact of six independent variables on participants' accuracy in rating video vignettes of student misbehavior: operational definition condition, race of target student, time, behavior concern level, behavior class, and cultural context in order to answer the following questions: 1) To what extent does providing operational definitions and decision-making support improve educators' accuracy in rating behavior, as compared to an expert panel? 2) does the impact of operational definition condition depend on the level of concern about the behavior, the race of the target student, and the behavior class? and 3) Does this impact depend on cultural context?

The dependent variable was the Euclidian distance from the expert rating, a quantitative variable where participants' ratings of vignettes were compared with expert ratings to determine total distance, positive or negative, from the experts' ratings. It was hypothesized that more accurate ratings may reduce disproportionality in disciplinary outcomes due to more accurate identification of problem behaviors requiring out of classroom disciplinary practices. Disproportionality was not directly measured in this study's outcomes.

One-hundred twenty participants with complete data for analysis across both the Urban Midwest and the Pacific Northwest completed an online survey where they were exposed to video vignettes depicting students misbehaving in both high school and elementary contexts at two different time points. Participants were randomly assigned to an operational definition condition (control, comparison, and experimental) and a race of the target student (White or Black). The video vignette order was counterbalanced to prevent order effects across time, behavior class (subjective and objective), level of behavior concern (mild and moderate) and vignette form (1 and 2). In order to evaluate the impact of the independent variables, a 6-way analysis of variance was conducted.

CHAPTER IV

The current study examined the impact of operational definitions, cultural context, behavior class, level of behavior concern, and race of the target student on the accuracy of teachers' ratings of student behaviors. Effects were examined by evaluating the impact of these factors on participants' rating of video vignettes depicting student misbehavior within an online module. The primary research questions were 1) Does the impact of operational definition condition depend on the behavior concern level, the race of the target student, and the behavior class? 2) Does this impact depend on cultural context?

The impact of the five independent variables, operational definition condition, race, cultural context, behavior class and level of behavior concern, on the dependent variable was evaluated using a six-way, mixed-effects analysis of variance (ANOVA) with time as a within-subjects effect. A mixed-effects ANOVA evaluates main and interaction effects for the independent variables, while accounting for participants repeated measures, in this case, the Time 1 and Time 2 vignettes. When evaluating the effects of the operation definition condition, interactions with time are the primary effects of interest. A main effect of operational condition is generally not interpretable. This design allowed for the analysis of the between-subjects effects of race, operational definition condition, and cultural context, as well as the within-subjects effects of the behavior concern level and class within a single analysis.

The ANOVA includes the main and interaction effects of each variable on the dependent variable, rating accuracy as measured by the Euclidian distance from the expert raters. The primary interest was in the interaction effect of operational definition condition and time, with cultural context, behavior class, race, and behavior concern level. The 5- and 6-way interaction effects were predicted a priori and were prioritized in the reporting of results. The results of the predicted 5- and 6-way interaction effects, including descriptive statistics, are reported in this chapter.

Survey Completion

Data were complete for 120 of the 151 potential respondents at both pre- and post-test, and so participant attrition was examined to determine if there were significant differences between the participants who did and did not complete the survey. Attrition effects were assessed based on cultural context, education level, exposure to SWPBS, age, participant race and race of the target student. Participants who completed the survey were not significantly different from participants who did not complete the survey based on cultural context, education level, exposure to SWPBS, age or participant race, $p > .05$. However, participants were significantly more likely to leave the study with incomplete data if they were initially exposed to a video vignette depicting a Black target student than if they were first exposed to a video vignette depicting a White target student, $\chi^2(1) = 6.97$, $p = .008$. This finding has implications for the remainder of the study, as race is one of the variables of interest and data were not missing at random.

Table 7

Contingency Table of Survey Completion by Race of the Target Student

Participant data	White target student	Black target student	Total
Complete	64	56	120
Not complete	5	17	22
Total	69	73	142

Note. 9 participants quit prior to answering any vignette-based questions, and therefore race could not have affected completion. These participants are not included in the table above.

Given the complex study design, it was not possible to include participants without complete data in the study. Therefore, the researchers eliminated participants without complete data (*n* = 31) from the analysis sample (*n* = 120). As the data were not missing at random, attrition is a limitation to conclusions based on the race of the target student.

Six-Way Interaction/Primary Analysis

In order to evaluate the primary research question on the accuracy of teacher ratings of misbehavior, data were analyzed using a six-way, mixed effects analysis of variance (ANOVA). The between-subjects independent variables were race of the target student, operational definition condition, and cultural context. The within-subjects independent variables were behavior class, behavior concern level, and time.

The quantitative dependent variable was the extent to which participants' ratings of video vignettes of student behavior align with expert ratings on the four questions for each video vignette. The distance between participant ratings and expert ratings at pre- and post-test was calculated using the Euclidian distance. Expert ratings were calculated by taking the average score of each individual behavior rating question for each video vignette across all 4 expert raters. The Euclidean distance score was then calculated by taking the square-root of the sum of distance ratings for each question squared for pre and post-test. These Euclidean distance scores were the dependent variable for the analysis.

The six-way, mixed-effects ANOVA was conducted in SPSS using the general linear model, repeated measures procedure. Descriptive statistics, effect sizes, and estimated marginal means for both hypothesized interaction effects were examined. The primary effect of interest, hypothesized a priori, was the 5-way interaction effect of operational definition condition, race, the level of behavior concern, behavior class, and time. The secondary effect of interest was the corresponding 6-way interaction with cultural context.

Descriptive statistics by race of the target student, behavior class and behavior concern level for Euclidian distance from expert rating are presented in Table 8. The analysis sample included 120 participants with complete pre- and post-test data. Fifty-six participants viewed vignettes with Black target students and 64 participants viewed vignettes with White target students. All 120 participants viewed vignettes for each behavior class (subjective or objective) and level of behavior concern (mild or moderate) at Time 1 and Time 2. The dependent variable was the Euclidian distance from expert rating for each vignette type at each time point. The minimum Euclidian distance from expert rating for participants was 5.68, for vignettes depicting a Black target student engaging in an objective, moderate misbehavior at Time 1. The maximum distance from expert rating was 9.52 for video vignettes depicting a White target student engaging in an objective, mild misbehavior. Because the video vignette forms were carefully counterbalanced to prevent order effects across time, there should be no significant differences between Time 1 and Time 2 scores due to the vignette form used.

To contextualize these Euclidian distance (ED) scores, the ED between the four experts serving as the comparison point was examined. Experts were, on average, 2.68 points apart across all races, behavior classes and behavior severities, indicating that, although experts did not have complete agreement on video vignette ratings, they were reasonably similar. In contrast, the mean ED for participants all operational definition conditions, races, behavior classes, and behavior severities was 7.42, nearly 3 times the distance between experts. In sum, participants were substantially further away from the expert average ratings than the experts were from each other.

Table 8

Descriptive Statistics for Euclidian Distance by Race, Behavior Class and Behavior Concern Level

	Black target student			White target student		
Behavior	*n*	*M*	*SD*	*n*	*M*	*SD*
Objective mild, time 1	56	8.20	5.55	64	8.85	5.15
Objective mild, time 2	56	8.64	5.06	64	9.52	5.34
Subjective mild, time 1	56	7.31	3.07	64	7.99	3.94
Subjective mild, time 2	56	7.80	3.33	64	8.48	3.61
Objective moderate, time 1	56	5.68	5.26	64	6.76	5.79
Objective moderate, time 2	56	6.22	5.29	64	7.80	6.01
Subjective moderate, time 1	56	5.73	3.70	64	6.16	3.67
Subjective moderate, time 2	56	6.86	4.29	64	6.76	3.80

Pearson correlation coefficients were calculated to better understand the relations between Time 1 and Time 2 participant accuracy. Correlations were computed among the 8 combinations of behavior class, behavior concern level, and time for 120 participants. Overall correlations are reported in Table 9. All correlations were statistically significant and were greater than or equal to $r = .230$, $p < .05$.

Pearson correlations were also computed for the video vignettes by race of the target student to examine possible differences in intercorrelations by race. To evaluate whether the correlations were significantly different for Black and White students, repeated Fisher's *Z* tests of equal correlations for independent samples were computed. To control for Type 1 error, a Bonferroni adjusted *p* value across the 21 tests was used.

Table 9

Pearson Correlations for Euclidian Distance from Expert Rating by Behavior Class, Behavior Concern Level and Time

Variables	1	2	3	4	5	6	7	8
1. Objective mild, time 1	-							
2. Objective moderate, time 1	.421**	-						
3. Subjective mild, time 1	.685**	.462**	-					
4. Subjective moderate, time 1	.662**	.508**	.853**	-				
5. Objective mild, time 2	.568**	.481**	.702**	.608**	-			
6. Objective moderate, time 2	.490**	.230*	.395**	.451**	.490**	-		
7. Subjective mild, time 2	.680**	.598**	.741**	.656**	.684**	.355**	-	
8. Subjective moderate, time 2	.671**	.594**	.704**	.651**	.692**	.465**	.797**	-
M	8.55	7.67	6.25	5.96	9.11	8.16	7.07	6.80
SD	5.33	3.56	5.55	3.67	5.21	3.48	5.72	4.02

Note. n = 120. Means reported are mean Euclidian distances.

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

No correlation coefficients were significantly different between White and Black students except the correlations between objective mild time 1 and objective mild time 2. The correlation between objective mild Time 1 and objective mild Time 2 for Black students (*r* = .80) was significantly larger than the corresponding correlation for White students (*r*

= .36), z = -3.85, p = .002. Overall respondents' accuracy in rating video vignettes was reasonably stable across time, behavior class, as well as level of behavior concern and was generally the same for White and Black target students.

Table 10

Pearson Correlations for Euclidian Distance from Expert Rating by Behavior Class, Behavior Concern Level and Time for Vignettes Depicting a White Target Student

Variables	1	2	3	4	5	6	7	8
1. Objective mild, time 1	-							
2. Objective mild, time 2	.359**	-						
3. Objective moderate, time 1	.503**	.414**	-					
4. Objective moderate, time 2	.392**	.439**	.129	-				
5. Subjective mild, time 1	.657**	.687**	.508**	.416**	-			
6. Subjective mild, time 2	.669**	.672**	.692**	.330**	.841**	-		
7. Subjective moderate, time 1	.643**	.612**	.570**	.487**	.893**	.787**	-	
8. Subjective moderate, time 2	.655**	.638**	.675**	.448**	.765**	.803**	.793**	-
M	8.85	9.52	7.99	8.48	6.76	7.80	6.16	6.76
SD	5.15	5.34	3.94	3.61	5.79	6.01	3.67	3.80

Note. n = 64. Means reported are mean Euclidian distances.

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

Table 11

Pearson Correlations for Euclidian Distance from Expert Rating by Behavior Class, Behavior Concern Level and Time for Vignettes Depicting a Black Target Student

Variables	1	2	3	4	5	6	7	8
1. Objective mild, time 1	-							
2. Objective mild, time 2	.800**	-						
3. Objective moderate, time 1	.313*	.579**	-					
4. Objective moderate, time 2	.600**	.547**	.374**	-				
5. Subjective mild, time 1	.721**	.717**	.378**	.352**	-			
6. Subjective mild, time 2	.698**	.695**	.432**	.370**	.591**	-		
7. Subjective moderate, time 1	.681**	.601**	.422**	.401**	.804**	.488**	-	
8. Subjective moderate, time 2	.691**	.762**	.521**	.496**	.654**	.823**	.514**	-
M	8.20	8.64	7.31	7.80	5.68	6.22	5.73	6.86
SD	5.55	5.06	3.07	3.33	5.26	5.29	3.70	4.29

Note. n = 56.

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

A six-way, mixed-effects analysis of variance (ANOVA) was conducted to evaluate the impact of the 6 independent variables on the dependent variable, Euclidian distance from expert rating. The primary, a priori hypothesis predicted a 5-way interaction between operational definition condition, time, behavior class, behavior concern level, and race. The second, a priori hypothesis predicted the 5-way interaction effect would depend on the cultural context (i.e., a 6-way interaction). These two a priori hypotheses were evaluated using α = .05 per comparison. Results of the remaining main and interaction effects in the six-way analysis of variance were also evaluated to determine if there were other, unanticipated factors significantly impacting the dependent variable. In order to control Type I error at .05 for the set of 61 effects that were not hypothesized a priori, results were evaluated with a Bonferroni adjusted *p* value. The results of the predicted intervention effects represented by the 5- and 6-way interaction effects were evaluated first. Then, other time by operational definition condition interaction effects were evaluated, followed by other race effects, other effects of the level of behavior concern, and other behavior class effects were evaluated using the Bonferroni adjusted *p* value.

Predicted intervention effects. First, the researchers evaluated the predicted intervention effects represented by the 5- and 6-way hypothesized interaction effects for research questions 2 and 3. Following the conventional analysis of variance logic, the 6-way interaction was evaluated first and was not significant, *F*(2, 108) = 0.16, *p* = .850, partial eta-squared = .00. Excluding cultural context, the predicted 5-way interaction between operational definition condition, behavior class, behavior concern level, and race was also not significant, *F*(2, 108) = 0.74, *p* = .481, partial eta-squared = .01. Both the predicted 6-way and 5-way interactions were not only non-significant, they also explained a very small proportion of the variance, 1 percent or less. That is, the predicted pattern of intervention effects, with or without considering cultural context, had little to no impact on the accuracy of behavior ratings of participants with regards to video vignettes of student misbehavior.

Research Question 1. After the theoretical predictions were analyzed, the researchers evaluated the impact of the other operational definition*time interactions next. These interaction effects were analyzed in order to assess any other significant impacts of the operational definition condition intervention on participants' rating accuracy when compared with an expert panel, or Research Question 1. Operational definitions were expected to increase teachers' accuracy in rating student behavior when compared to an expert panel, which in turn was believed to potentially reduce disproportionate disciplinary outcomes for Black students.

Contrary to the expectation that operational definitions would assist teachers in aligning their ratings of student behavior vignettes with experts, the results do not indicate a significant impact of operational definition condition under any conditions examined in this study. None of the 16 potential interactions of time*condition were significant, and all had partial eta-squared values below .02. Thus, operational definition condition did not appear to have any meaningful impact on the accuracy of teacher ratings of behavior, and it explained 2% or less of the variance in participant ratings. Operational definitions did not appear to impact teacher decision making, even when supplemented with a decision-making support such as the flow chart provided in the experimental condition. Further sensitivity analyses with a reduced model also indicated no impact of the operational definition condition by time interaction $F(2, 117)$, $p = .085$, partial eta-squared = .041.

Table 12

Six-Way, Mixed-Effects Analysis of Variance Summary

Source	*df*	*SS*	*MS*	*F*	*p*	Adj. *p*	Partial eta
Between-subjects							
Cultural context (C)	1	791.02	791.02	7.33		.482	.06
Race (R)	1	119.87	119.87	1.11		1.000	.01
Operation definition (D)	2	128.87	64.44	0.60		1.000	.01
C * R	1	0.85	0.85	0.01		1.000	.00
C * D	2	86.96	43.48	0.40		1.000	.01
R * D	2	161.17	80.58	0.75		1.000	.01
C * R* D	2	122.68	61.34	0.57		1.000	.01
Error	108	11,660.53	107.97				
Vithin-subjects							
Behavior class (B)	1	559.78	559.78	52.80		< .001	.33
B * C	1	7.69	7.69	0.73		1.000	.01
B * R	1	3.89	3.89	0.37		1.000	.00
B * D	2	3.30	1.65	0.16		1.000	.00
B * C * R	1	5.31	5.31	0.50		1.000	.00
B * C * D	2	24.25	12.12	1.14		1.000	.02
B * R * D	2	2.84	1.42	0.13		1.000	.00
B * C * R * D	2	5.67	2.84	0.27		1.000	.00
Error (B)	108	1,145.02	10.60				
Behavior concern level (S)	1	92.29	92.29	6.11		.915	.05
S * C	1	15.60	15.60	1.03		1.000	.01
S * R	1	5.93	5.93	0.39		1.000	.00
S * D	2	4.75	2.38	0.16		1.000	.00
S * C * R	1	34.54	34.54	2.29		1.000	.02
S * C * D	2	1.79	0.90	0.06		1.000	.00
S * R * D	2	1.17	0.59	0.04		1.000	.00
S * C * R * D	2	1.98	0.99	0.07		1.000	.00
Error (S)	108	1,632.37	15.12				
Time (T)	1	57.77	57.77	6.41		.781	.06
T * C	1	32.57	32.57	3.61		1.000	.03
T * R	1	1.22	1.22	0.14		1.000	.00
T * D	2	20.62	10.31	1.14		1.000	.02
T * C * R	1	12.89	12.89	1.43		1.000	.01

(continued)

Source	*df*	*SS*	*MS*	*F*	*p*	Adj. *p*	Partial eta
T * C * D	2	16.43	8.21	0.91		1.000	.02
T *R * D	2	1.57	0.79	0.09		1.000	.00
T * C * R * D	2	0.30	0.15	0.02		1.000	.00
Error (T)	108	973.96	9.02				
B * S	1	16.07	16.07	3.52		1.000	.03
B * S * C	1	0.85	0.85	0.19		1.000	.00
B * S * R	1	30.18	30.18	6.61		.702	.06
B * S * D	2	11.19	5.59	1.22		1.000	.02
B * S * C * R	1	19.27	19.27	4.22		1.000	.04
B * S * C * D	2	1.33	0.67	0.15		1.000	.00
B * S * R * D	2	4.92	2.46	0.54		1.000	.01
B * S * C * R * D	2	0.80	0.40	0.09		1.000	.00
Error (B * S)	108	493.49	4.57				
B * T	1	16.25	16.25	1.44		1.000	.01
B * T * C	1	16.33	16.33	1.44		1.000	.01
B * T * R	1	3.16	3.16	0.28		1.000	.00
B * T * D	2	7.48	3.74	0.33		1.000	.01
B * T * C * R	1	22.88	22.88	2.02		1.000	.02
B * T * C * D	2	10.40	5.20	0.46		1.000	.01
B * T * R * D	2	1.14	0.57	0.05		1.000	.00
B * T * C * R * D	2	25.31	12.66	1.12		1.000	.02
Error (B * T)	108	1,223.43	11.33				
S * T	1	0.33	0.33	0.04		1.000	.00
S * T * C	1	16.54	16.54	2.11		1.000	.02
S * T * R	1	6.24	6.24	0.80		1.000	.01
S * T * D	2	0.95	0.47	0.06		1.000	.00
S * T * C * R	1	0.03	0.03	0.00		1.000	.00
S * T * C * D	2	15.98	7.99	1.02		1.000	.02
S * T * R * D	2	4.02	2.01	0.26		1.000	.00
S * T * C * R * D	2	5.66	2.83	0.36		1.000	.01
Error (S * T)	108	847.15	7.84				
B * S * T	1	0.46	0.46	0.06		1.000	.00

(continued)

Source	*df*	*SS*	*MS*	*F*	*p*	Adj. *p*	Partial eta
B * S * T * C	1	3.28	3.28	0.42		1.000	.00
B * S * T * R	1	3.03	3.03	0.39		1.000	.00
B * S * T * D	2	11.92	5.96	0.77		1.000	.01
B * S * T * C * R	1	0.04	0.04	0.01		1.000	.00
B * S * T * C * D	2	0.38	0.19	0.03		1.000	.00
B * S * T * R * D	2	11.40	5.70	0.74	.481		.01
B * S * T * C * R * D	2	2.51	1.26	0.16	.850		.00
Error (B * S * T)	10 8	835.62	7.74				

Note. Adj. *p* is the Bonferroni adjusted *p*-value controlling for Type I error across 61 effects.

Other race effects. The impact of student race on participant ratings was evaluated in determine if race played a significant role in participants' accuracy in rating student problem behavior. Race was generally expected to impact teachers' rating of student behavior, with participants ratings often believed to be less aligned with experts for video vignettes depicting a Black target student than for those depicting a White target student. However, no significant main effect of race was found, *F*(1, 108) = 1.11, *p* = 1.000, partial-eta squared = .01. In addition, race did not interact significantly or meaningfully with any combination of cultural context, behavior class, behavior concern level or time. The partial-eta squared was less than or equal to .04 for all 32 main and interaction effects with race. Thus, the race of the target student did not appear to impact meaningfully participants' accuracy in rating student misbehavior. This result was unexpected, as most other research in the area has found race to be a significant factor in educator disciplinary decision making. This could indicate that racial differences in disciplinary decision making found in real-world educational environments incorporating positive behavior supports may be related to other factors not examined in this study.

Other level of behavior concern effects. The impact of the level of behavior concern was also evaluated in order to better understand the unique impact of this independent variable on participants' accuracy. The main effect of level of behavior concern was not significant, *F*(1, 108) = 6.11, *p* = .915, partial eta squared = .05. That is, neither mild nor moderately concerning behaviors had a significant impact on participants' accuracy in rating problem behavior. In addition, none of the 32 interaction effects with level of behavior concern were significant. The researcher initially hypothesized increased participant accuracy when the behavior depicted was moderately concerning and decreased accuracy when the behavior depicted was only mildly concerning. However, participant accuracy was similar regardless of the level of concern about the behavior depicted in the vignette, even when the race of the target student and the operational definition condition were included in the analysis.

Other behavior class effects. Finally, the impact of the final independent variable, behavior class, on participants' accuracy in rating problem behaviors was evaluated. A main effect of behavior class was significant, *F*(1, 108) = 52.80, *p* < .001, partial eta squared = .33. That is, behavior class explained 33% of the variance on the dependent variable, which would be considered a large effect (Cohen, 1988). Participant ratings on vignettes depicting misbehaviors categorized as subjective (*M* = 7.01) were significantly more accurate (that is, closer to expert ratings) than those categorized as objective (*M* = 8.76). That is, participant ratings more closely matched experts when the behavior shown was subjective, such as defiance and disrespect, by 1.75 points than when the behavior shown was objective, such as physical aggression or smoking.

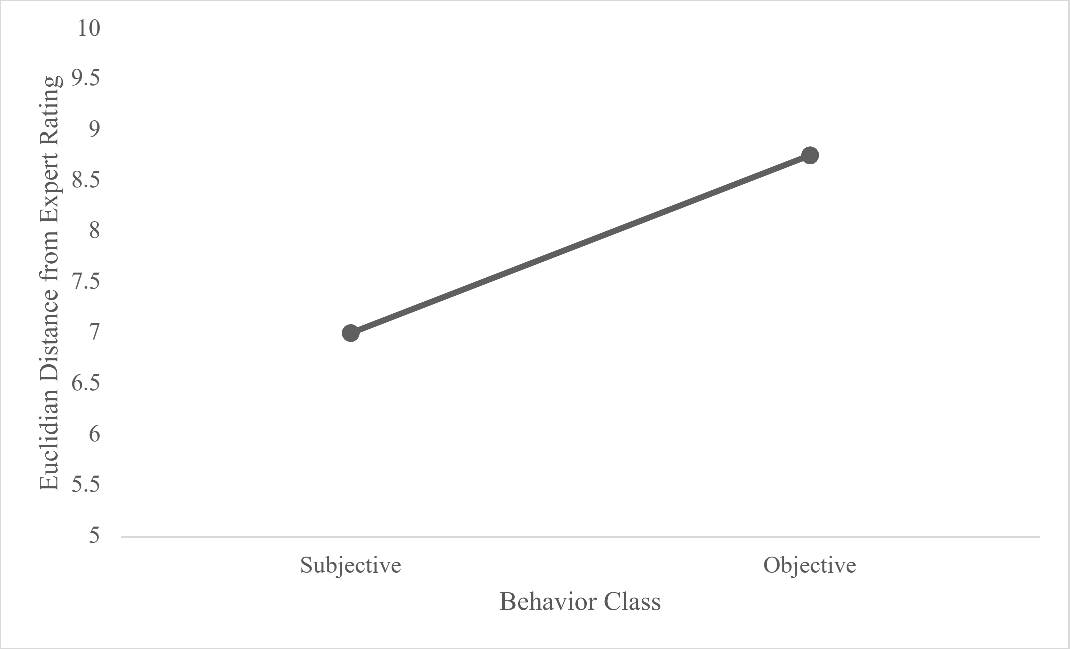


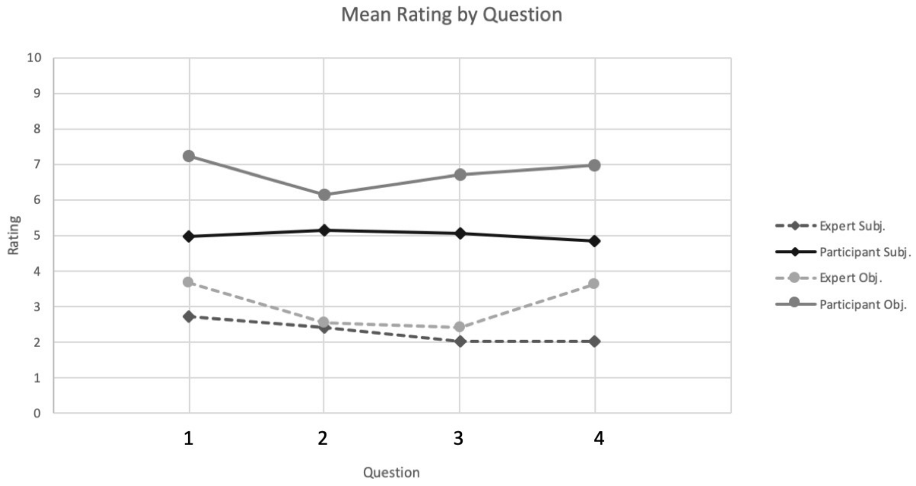
Figure 3: *Estimated Marginal Means for Euclidian Distance from Expert Rating by Behavior Class*

Figure 3 depicts the estimated marginal means for participants' Euclidian distance from expert rating by behavior class. Participants were significantly more accurate in rating behaviors classed as subjective than those classed as objective. This finding is unexpected, as it is commonly believed that objective behaviors, being easily observed visually, would be more likely to result in stable and accurate ratings across both experts and participants. Instead, participants were more accurate in rating behaviors classified as subjective, which included vignettes depicting "defiance" and "disrespect," regardless of the race of target student.

Post hoc analyses. In order to better understand the impact of behavior class, the mean rating, not Euclidian distance, was also examined, disaggregated by question, for both participants and experts. Question one asked, "How severe was the students' misbehavior," and question four asked "How severely should the student be disciplined?"

Question two asked, “to what extent is the student hindering you from maintaining order in your class,” and question three asked, “How irritated do you feel by this student?” Results of that analysis are presented in Figure 4.

Figure 4:



Expert mean ratings for subjective behavior ranged from 2.03 for Questions 3 and 4 to 2.70 for Question 1, whereas the participant mean ratings for vignettes depicting Subjective behaviors ranged from 4.87 for Question 4 to 5.17 for Question 2. The expert mean ratings for objective behaviors ranged from 2.41 for Question 3 to 3.69 for Question 1 and participant mean ratings for objective vignettes ranged from 6.16 for Question 2 to 7.22 for Question 1.

Overall, experts and participants rated the objective behaviors as more concerning than the subjective behaviors. However, both experts and participants rated question 4, how severely students should be disciplined, substantially higher for objective than subjective behaviors.

CHAPTER V

Summary

The purpose of this study was to examine the impact of operational definition condition, race, behavior class, level of behavior concern and cultural context on educators' accuracy in disciplinary decision making using a randomized control, pre-test/post-test intervention study. Participants were randomly assigned to one of three operational definitions conditions, control, comparison and experimental and to one race of a target student: Black or White. Participants' accuracy in rating students' misbehaviors was examined by measuring participant responses pre- and post-test on a series of four questions from Okonofua and Eberhart (2015) regarding their reaction to a series of video vignettes depicting student misbehavior selected by the researcher. Those video vignettes were counterbalanced by form, race, and condition to prevent order effects across time, level of behavior concern and behavior class. Those pre- and post-test responses were then used to calculate a Euclidian distance from expert ratings for use in the analysis.

Intervention effects were hypothesized to be more pronounced for Black, mild, subjective behaviors when exposed to the experimental operational definition condition at Time 2 (i.e. five- or six-way interaction effects). However, the predicted interactions were not significant and explained a trivial amount of variance. Given the lack of significance of the predicted interaction effects, the remaining main and interaction effects within the 6-way analysis of variance were evaluated with an adjusted Bonferroni procedure to better understand the impact of the study's independent variables on the accuracy of participants' ratings.

Other potential intervention effects were evaluated by examining all time*operational definition condition interactions. There were no significant intervention effects because none of the 16 time*operational definition condition interactions were significant. Thus, there was no discernable impact of operational definition condition on participants' accuracy in rating student problem behavior. The impact of race on participants' accuracy was similarly not significant; there were no main or interaction effects of race. This result was surprising given the pervasive impact of race on evaluations of student misbehavior in the literature.

Neither race nor operational definition condition had any impact on participants' distance from expert ratings. They also did not have a significant impact when interacting with the other independent variables: behavior class, level of concern about the behavior or cultural context. Neither variable, alone nor interacting with other variables, explained more than 4% of the variance in Euclidean distance from expert raters, a very small amount of variance explained. Overall, this study does not provide empirical evidence for the efficacy of operational definitions in increasing participants' accuracy in rating student problem behavior and therefore operational definitions are not expected to decrease disproportionate disciplinary outcomes.

The impact of the level of concern about the behavior and participants' cultural context were also assessed. However, neither had a significant main or interaction effect on the accuracy of participants ratings. Therefore, neither the level of concern about the behavior nor the participants' cultural context were a factor in the accuracy of their ratings of video vignettes of student misbehavior when compared to a panel of experts.

The impact of behavior class on participants' rating accuracy was evaluated and a significant main effect of behavior class was found. Behavior class, that is objective vs subjective behaviors, explained 33% of the variance in participants accuracy in rating student misbehavior. Participants were more accurate in rating video vignettes depicting a subjectively defined behavior such as defiance or disrespect than they were in rating videos depicting objectively defined behavior such as physical aggression or smoking.

This study sought to evaluate factors that might increase educators' accuracy in rating problem behaviors when compared to expert ratings. Increased accuracy provides a way to reduce the rate of disproportionality in office discipline referrals between White and Black students. Specifically, the study examined the impact of an operational definition intervention, the impact of the cultural context of the participant, the impact of the level of concern about the behavior and the subjectivity or objectivity of the depicted behavior on the rate of disproportionality Operational definitions are a component of a systemic school policy initiative known as SWPBS.

In order to evaluate these impacts, this study utilized a randomized control, pretest-posttest experimental design. Educators' ratings of video vignettes depicting student misbehaviors were evaluated based on their accuracy, or how similar their ratings were when compared to a panel of experts' ratings of the same video vignette. A 6-way analysis of variance statistical analysis revealed no significant effects of operational definition intervention, race or cultural context. However, a significant effect of behavior class was found. Limitations, study findings, implications, and recommended future directions follow.

Limitations

To most accurately and effectively interpret the findings of this study, it is important to understand the limitations. While some of the considerations are methodological, such as design limitations and differential participant attrition; others are more philosophical, such as the nature of expertise or culture within our society. However, both types of limitations are important to consider when evaluating the study findings.

Methodological Limitations

Methodological limitations of this study are related to design or procedural choices and the results of data collection. Methodological limitations are grounded within a traditional, dominant cultural perspective on scientific research. These include limitations on the types of video vignettes selected for study, the layout of the survey in the online software, participant attrition and the use of experts as a standard for judgement.

Euclidian distance. The dependent variable for this study was accuracy of teacher ratings of behavior, in the sense of distance from expert ratings. The choice of Euclidian distance from expert mean was chosen to account for the theoretical belief that distance from expert mean, either a lack of concern or too much concern, would both be undesirable outcomes. A limitation of this analysis choice is that the study does not address the direction of the distance. Therefore, it is not possible to interpret whether participants were more or less concerned about the behavior displayed by the students depicted in the vignettes. This study examined whether participants rated more like experts. Future studies should address the direction of any differences.

Video vignettes. The video vignettes included in the study came from multiple sources and were not designed explicitly for the study's purposes. All videos in the subjective behavior class came from the same research project, where they were developed for training purposes within a SWPBS system to demonstrate the same misbehavior with both a White and Black target student. The videos utilized the same script, same location, and same clothing. The only difference between videos was the race of the target student (either White or Black) and the target students' 2 friends included in the vignette (White, Black or both). Similar videos were not available for the objective behavior class, so a collection of comparable videos was compiled from a variety of research projects and publicly available videos. While they were rated similarly by the experts, the objective videos did not depict the exact same behaviors, nor were the students in similar outfits or locations.

Thus, while the subjective behaviors were functionally identical across both White and Black target students, the objective behavior class video vignettes were not. Despite the fact that experts rated the behaviors as functionally equivalent, educators may have responded to subtle differences in the video context or in the behavior depicted. Therefore, the behavior class effect may be confounded with of the source of the video vignettes. However, given careful counterbalancing of the design, any impact of these differing videos should have been addressed.

Questions. The questions chosen for this study came from Okonofua & Eberthart's (2015) study, Two Strikes: Race and the Disciplining of Young Students, which evaluated the impact of racial stereotypes or bias on disparities in teacher responses to student behavior. This study provided an important framework to evaluate the impact of bias on teacher disciplinary decision making, they may not have provided sufficient sensitivity to the impact of an operational definition. The four questions, (1) “How severe was the student’s misbehavior?” (2) “To what extent is the student hindering you from maintaining order in your class?” (3) “How irritated do you feel by the student?” (4) “How severely should the student be disciplined?” may be better measures of teachers’ emotional response to the behavior and the severity of the behavior than the constructs impacted by operational definitions. In future studies it may be appropriate to examine alternative or additional questions possibly including “should this student be sent to the office for this behavior?”; “can this behavior be managed in 60 seconds or less?”; “how would you classify this problem behavior?”; and “do you have a strategy to respond to this behavior?”

Behavior class. In this study, behavior class was defined based on a review of the literature evaluating disproportionate disciplinary outcomes. In many studies, subjective and objective behaviors have been defined based on the overall behavior class. That is, whether or not the type of behavior described by the overall behavior class, such as defiance or smoking, can be understood to be binary. For example, a student can be slightly defiant, but they can’t be slightly smoking in a school context – they either are smoking, or they are not. This is a reasonable and defensible way to define subjectivity vs objectivity, but it is not the only way it could be understood. However, with enough specificity, it is possible to develop an objective operational definition of any behavior. In which case, it would be more appropriate to evaluate the relative subjectivity and objectivity of each behavior class through consistency in response to the behavior. Behaviors that were rated more consistently by experts would be considered more objective whereas behaviors that were rated less consistently would have more subjective operational definitions. These questions warrant future research exploration.

Survey layout. In the experimental and comparison conditions, the layout of the survey in Qualtrics when viewed by participants had the operational definition directly below the video vignette. While the directions twice requested that participants refer to the operational definition before rating, it is possible that participants did not review the operational definition and provided ratings not guided by operational definitions regardless of condition. Therefore, the study's attempt to manipulate the operational definition condition may have been compromised. Thus, the lack of effect of operational definition condition may not be attributable to a failure of impact of the operational definition intervention but instead an inadequate implementation of the operational definition intervention. This may be merely technical, where educators did not attend to the operational definition condition due to the survey design. In future studies, it may be appropriate to provide the operational definition for each video vignette on an introductory screen prior to viewing the vignette, which would require an active response from participants' before progressing to the vignette, and again beneath the vignette prior to rating in order to increase the likelihood of attending to the operational definition.

However, it may be that the operational definition intervention did not have a significant impact because simply providing operational definitions and a flow chart may be insufficient training for educators to make use of this intervention. More specific didactic trainings, with examples and non-examples as well as opportunities to practice the difference between a referable and a less concerning student behavior may have a greater impact on teachers' disciplinary decision making. Merely providing an operational definition or even an operational definition with a flow chart may represent an insufficient implementation of the operational definition intervention.

Attrition by race. Participants did not drop out of the study at random. After randomization to condition, the race of the target student was a significant predictor of participant attrition. Participants exposed to video vignettes depicting a Black target student were more likely to quit than those exposed to video vignettes featuring a White target student. Of the 22 people who dropped out of the study after randomization to a target student race, 17 were exposed to a Black target student, whereas 5 were exposed to a White target student. So, 77% of those who quit the study after seeing at least one video vignette were randomized to the Black target student condition. It may be that something about the video vignettes depicting Black target students influenced participants' decision to drop out.

Although it is not possible to know exactly what caused those participants viewing vignettes depicting Black target students to be more likely to attrit, there are some potential explanations. One possible interpretation is that participants who were more biased were more likely to drop out when faced with video vignettes depicting Black students misbehaving. Another possibility is that educators' who were more concerned about appearing biased were more likely to drop out when faced with a Black target student. Regardless of the reason, all results with race of the target student as an independent variable must be interpreted with caution given the significant pattern of attrition moderated by this variable.

Operational Definitions. The operational definitions utilized in this study were selected from the School Wide Information System (SWIS), a web-based program used for recording and charting discipline referral data utilized by over 10,000 schools across the United States. The SWIS operational definitions were selected in ordered to examine operational definitions that were already widely in use across the nation. Despite their widespread adoption however, the choice to utilize SWIS operational definitions was not without limitations. The operational definitions from SWIS are short, generally a single sentence, and designed to encompass broad categories of problem behaviors across a wide range of local contexts. It was not possible to evaluate the impact of culturally responsive or even context specific operational definitions within the scope of this study; however, this is a limitation that impacts the interpretation of the results. Best practice recommends that operational definitions be developed by faculty within a school. Thus, the study results can only be interpreted as valid when considered in the context of externally imposed operational definitions.

Experts. Despite the impressive resumes of our experts, there is no reason to believe experts are any less likely to be biased than the general population. Although it is unlikely this panel of experts holds secret explicitly biased beliefs against any group, implicit bias is endemic to our culture and experts are not immune. Implicit biases may be shared by both Black and White experts across cultural contexts. Through evaluating participants' adherence to expert ratings, this study may be unable to document the bias of both experts and participants if they are similarly biased in the same direction. Experts are human, and therefore are an imperfect standard against which to measure bias given the high likelihood that all humans exhibit some unconscious bias. Given the dependent variable was Euclidian distance from expert mean, if the experts and participants were similarly biased, that could provide an alternative explanation for study results.

Cultural context. In this study, cultural context was a between-subjects factor determined by the geographical location of the school where the participant reported teaching. This broad measure was an attempt to include as many variables as possible within cultural context, and to avoid narrowly defining cultural context based on factors such as race and socio-economic status alone. However, the broad nature of this measure may be too complex to be meaningful. By incorporating variables such as race, socio-economic class, gender, background, and education, the cultural context variable may be too broad to be a measure of cultural context and may instead be a measure of general beliefs within U. S. society.

Further, the panel of experts were from both cultural contexts and their scores were averaged together for each question. So, it is possible that in using the Euclidian distance from expert mean as the dependent variable, that the study design was not sensitive to differences in the cultural context as defined in the study.

Philosophical Limitations

This study was developed within a dominant cultural understanding of science, where truth is discovered through the systematic evaluation of data. In our culture, "science" offers an epistemological framework for the development of knowledge. That is, science is a framework that renders beliefs into truths by the iterative construction of knowledge through the observations and experiments of detached and unbiased observers. This epistemology is founded on the ontological belief that detached and unbiased collection of evidence is not only possible by humans, but necessary to uncover truth.

However, not all epistemological frameworks share the belief that detached and unbiased data collection is either possible or desirable in the development of knowledge. Critical race theory offers a competing ontological framework that contradicts mainstream science's claims of impartiality as a "camouflage for the self-interest, power, and privilege of dominant groups in U. S. society" (Solorzano, 1997, p. 473). Critical race and decolonization studies perspectives argue "epistemologies are a result of social practices where power is being exercised that can reinforce colourblind, 'race' neutral, ahistorical, and apolitical points of view" (Hylton, 2012, p. 3). That instead of being unassailable "gold standards," mainstream epistemological framework's rules dictating subject-researcher distance, emotional detachment, ethics, values, and methods for ascertaining truths instead represent the culturally specific values of the privileged (Carter, 2003). Scholars argue that this practice of universalizing the specific values of the privileged allows the normalization of oppression and inequality. Instead, critical race theorists value the specific, arguing that specificity is necessary for a grounded epistemology (Carter, 2003). Simply stated, critical race theory argues that "truths only exist for this person, in this predicament and this time in history" (Delgado, 1991, p. 111).

To help make these competing frameworks more tangible, they can be understood through a comparison to an orchard with many types of fruit. If you were interested in finding out what fertilizer resulted in the largest average yield most effectively and inexpensively, you would specifically collect data on fruit yield by fertilizer type across the entire orchard. If you were interested in knowing which fertilizer was most effective not only for yield, but flavor profile, for each variety of each type of fruit, you would collect data on multiple outcome variables – sweetness, intensity of flavor, yield, etc. – as well as matching your fertilizer to each variety of fruit and potentially to each tree. One offers the solution most likely to yield a large crop for the largest number of trees, most quickly and inexpensively, while the other offers a time consuming and costly solution that provides the most desirable outcomes for each type of tree. The dominant scientific framework's search for universal truth is akin to the search for the fertilizer that is most likely to work for all trees, while a critical race theory framework asks what fertilizer for what tree in what location within the orchard. Certainly, this is an oversimplification of a complex and diverse set of beliefs. In their 2017 article with Fixsen, two of the developers of SWPBS, Horner and Sugai, argued that interventions should not be able to be declared evidence based without being able to specify, "exactly what the practice involves, where it should be used, by whom and with whom it should be used, and for what purpose." (Horner, Sugai, & Fixsen 2017, p. 26). But these broad strokes illuminate an important distinction between these two frameworks – the relative value of general vs specific applicability and outcome variables.

The philosophical limitation section of this study seeks to explicate both how this study utilized a mainstream scientific perspective and also discuss the implications of this choice on the results.

Cultural context. Cultural context was evaluated using the extremely broad measure of geographic area worked in, with both experts and participants from both cultural contexts. Beyond the methodological limitations these choices imposed on the study's findings, a further limitation of cultural context is a potential failure to account for similarities due to education level, exposure to past school expectations, socio-economic status and career between participants and between participants and experts.

That is, experts and participants' cultural context may be more alike due to their shared experience in an education career than they are different due to any other variables that differentiate them. These significant similar experiences are likely to have an impact on the values and expectations of school appropriate behaviors of both participants and experts. Thus, educator and expert behavior expectations may be measuring a shared outcome goal.

If this is true, this shared outcome goal is likely to be situated within a dominant cultural framework. Schools have historically been an important site of transmission of dominant cultural values. Critical Race Theorist Gloria Ladson-Billings (2000) described this experience as, "Schools and teachers treat the language, prior knowledge and values of African Americans as aberrant and often presume that the teacher's job is to rid African American students of any vestiges of their own culture" (p. 205). Under the guise of objective, universal, color-blind values and beliefs, schools may transmit a "hidden curriculum" aligned with White, middle class, heterosexual, and Christian beliefs (Yosso, 2005).

Experts. The present study utilized Euclidian distance from expert ratings as the dependent variable, which can also be understood as how close participants' ratings were to the expert mean for each of the 4 questions. This allowed for a comparison of participant ratings to an expert standard. The panel of experts represented a diverse array of backgrounds, training and life experiences, in an attempt to capture the diversity of beliefs across differing key characteristics (e.g. race, cultural context and training experiences) between experts. In this way, the combined panel of experts acted as a proxy for a universal standard against which participants could be judged. This choice of standard allowed for the evaluations of differences in their responses to be interpreted as less and more desirable. That is, this study entails the assumption that experts are an appropriate standard by which to judge educators' ratings of student behavior and that having participants' ratings being more like the panel of experts was a desirable outcome. Although this methodology provides a clear and explicit metric by which to measure desired outcomes, it is not without limitations.

From a dominant epistemological framework, an acontextual and universal set of beliefs around student behavior is not only possible, but desired. That is, experts can be understood as a "yield" metric, a single, universally equivalent metric by which to judge success. However, as in the orchard example, simply measuring one metric of success inherently excludes other metrics. In this instance, compliance with an externally imposed standard of appropriate behavior was the "yield" for this crop of participants. The more compliance, the more desirable the outcome.

It is possible that by utilizing expert beliefs as the standard against which responses were evaluated, this study may be tacitly endorsing a hegemonic cultural norm of whiteness. Gloria Ladson-Billings (1998) argued that "stories provide the necessary context for understanding, feeling and interpreting. The ahistorical and acontextual nature of much of law and other 'science' renders the voices of dispossessed and marginalized group members mute" (p. 13). Positing that there can be a single, objective, and universal understanding of a behavior best measured by expert opinions may itself be a capitulation to the dominant culture that increases the likelihood of disproportionate disciplinary outcomes. Or as Carter (2003) explained, “research methods that focus on external perspectives [i.e. expert opinion] do not offer much hope for counternarratives. Instead they validate what we seek to deconstruct in the first place.” (p. 33).

Operational definitions. If experts are the yield metric of our analogy, operational definitions are the all-purpose fertilizer intervention. Indeed, operational definitions within SWPBS are by design intended to be universal explanations of student misbehavior that are sufficiently clear for any observer to apply in the moment to behavior in a desired way; comparable to the all-purpose fertilizer that will successfully ensure a majority of the trees in the orchard yield a large amount of fruit. This acontextual, universal evaluation is absent any significant information about either the experts or educators’ political and moral realities that may influence their decision making.

Current cultural and methodological demands insist on solutions that can be applied for the largest number of students and whose result is compliance with externally imposed criteria. Operational definitions as they are currently implemented certainly fit these metrics of success. This study sought to evaluate whether this potential solution, operational definitions, was likely to help or hinder rates of disproportionality for a preponderance of educators across geographic diversity because of its value within a dominant cultural paradigm. Despite the defensibility of that choice, it necessarily excluded the opportunity to ask, “for what outcome?” Given the realities of both the demands of dominant scientific methodology and logistics, it was not possible to evaluate contextually specific operational definitions within the current study, that is, to search for the right fertilizer, for the right tree. But operational definitions as they exist are a simple compliance metric, adherence to them fails to provide information about the student's, their family's, or even their teacher's, desired behavioral outcomes because they were not included in the development of those operational definitions.

The results of this study may be best interpreted as increased adherence to dominant cultural norms within schools. This is an important point to consider when evaluating the results. Is increased compliance to a colorblind, externally imposed behavioral standard a desirable outcome? Especially for those students who come from non-dominant family and/or community backgrounds that may not share any important facets of culture with either educators or experts, regardless of their race? These outside behavioral expectations may require students to have the ability to effectively "code switch" between home and community appropriate behaviors and school appropriate behaviors. The failure to appropriately differentiate the cultural needs of students from non-dominant backgrounds from the cultural needs of those students who share their same race, but a dominant cultural perspective may be contributing to our failure to address disproportionate disciplinary outcomes. Instead of moving towards a diversity of solutions and appropriate outcome measures, this study narrowed the field of possible solutions.

Study Findings

Theoretical predictions. The researchers' a priori hypothesis predicted an interaction effect of operational definition condition with time that depends upon the behavior class, level of concern about the behavior, race of the target student, and time. The most significant increase in congruence with expert ratings was expected for those teachers in the experimental condition who were exposed to vignettes depicting a Black target student engaging in a mild, subjective behavior. This prediction was evaluated with a five-way, time*condition*race*concern*class interaction effect.

The five-way interaction was not significant, and the partial eta squared indicated that this interaction was trivial in magnitude. In addition, the five-way interaction was considered to possibly depend on cultural context, however that six-way interaction was similarly nonsignificant and was also trivial in magnitude.

Certainly, the methodological and philosophical limitations may have impacted the ability to identify the expected pattern of results. It could also be that teachers' behavior may no longer be biased based on race but may instead have a bias against behaviors that are considered "objectively wrong." Indeed, the objective behaviors depicted were physical aggression and smoking, behaviors frequently associated with "troublemaker students," which may have led teachers to rate those vignettes less accurately than those depicting subjective behaviors like defiance and disrespect. However, the dependent variable, Euclidian distance, does not allow for the examination of the direction of the difference, so theories about direction of the difference need to be evaluated in subsequent research.

Operational definitions and rating accuracy. The efficacy of the operational definition intervention at increasing participants' accuracy when compared with experts, or Research Question 1, was evaluated via a randomized control experimental design. Operational definitions are designed to provide an objective, universal understanding of the topography of a behavior that is sufficiently clear any observer can use it to immediately identify the target. The operational definitions utilized in this study were obtained from a nationwide behavior tracking data base that is frequently utilized by schools implementing SWPBS.

Participants were randomly assigned to one of three conditions, a control condition with no decision-making guidance, a comparison condition with just operational definitions and an experimental condition where participants were provided with both operational definitions for the video vignette but also a decision-making flow chart for disciplinary decisions. Participants were first provided a pre-test of 4 randomized and counter-balanced video vignettes, and then at an immediate post-test were exposed to either the same conditions for the control group or the conditions described above for the experimental and comparison groups in a carefully counterbalanced order. The experimental operational definition condition was expected to significantly increase the accuracy of teachers' ratings of student behavior (that is, increase their similarity to experts). This would be indicated by a time*condition interaction effect. That is, we would expect to see an impact of operational definition condition at time 2 if the experimental operational definition condition affected participant accuracy.

However, none of the 16 interactions that included operational definition condition*time were significant. In this study, neither operational definitions as presented in SWIS, nor operational definitions coupled with a decision-making flow chart had a significant or non-trivial impact on participants' accuracy in rating video vignettes when evaluated by the time by condition interaction. Participants ratings were similarly accurate with respect to experts' ratings across all three conditions, including between the control condition where no guidance was offered and the conditions where operational definitions were provided. That is, there was no significant difference in the accuracy of participants' ratings regardless of the presence or absence of operational definitions or supplementary decision-making aids.

It is possible that participants simply did not attend to the operational definitions and/or the flow-chart despite the directions to do so. However, it is also possible that operational definitions, as an externally imposed and one-size fits all approach, did not have a significant impact on teachers' behavior in the moment of actual decision making. It is easy to imagine an educator receiving a list of operational definitions and a decision-making flow-chart at the beginning of the term during a professional development, returning to their classroom and tacking it on the wall near their desk with the best intentions of using it for disciplinary decision making throughout the year. It is equally easy to imagine that individual never again remembering to look at the form, certainly not in the heated moments when it would be most useful. Similarly, within the study it is plausible that educators reviewed the provided operational definitions and decision-making flow-chart and they simply did not utilize that information in their actual decision-making moment.

Impact of level of behavior concern. The effect of operational definition condition was expected to depend on the level of concern about the behavior, because prior research has found a significant impact of more concerning behaviors on disproportionate disciplinary outcomes. However, in this study there was no significant difference in the accuracy of participants' ratings between mild and moderate behaviors across all 32 of the main and interaction effects with level of behavior concern. That is, participants were similarly distant from experts regardless of whether the behavior depicted was of mild or moderate level of concern. This result was unexpected as it was predicted that raters would be more likely to be accurate when the behavior was more moderate, based on the understanding that more moderately concerning behaviors were more likely to have shared agreement while behaviors that were only mildly concerning would be likely to have greater variation and be susceptible to bias. This may suggest that evaluations of a broader range of levels of behavior concern may be necessary in order to better understand the impact of level of concern about a behavior on teacher disciplinary decision making.

Behavior class and rating accuracy. Behavior class did have a significant effect on participants' rating accuracy. It was hypothesized that participants would be less accurate when rating subjective behaviors than objective behaviors, based on prior research that indicated the rate of disproportionality was greater for behaviors classed as subjective rather than objective. However, in this study participants were more like experts when rating behaviors classed as objective than those classed as subjective. The subjective behavior classes represented in this study were defiance and disrespect, whereas the objective behavior class were physical aggression and smoking. Prior research with real-world, extant data sets has indicated that Black student are significantly more likely to be disciplined for behaviors classified as subjective than those classified as objective, with defiance and disrespect specifically being repeatedly mentioned in the literature as significant contributors to disproportionate disciplinary outcomes. In this study, raters were more closely aligned with expert ratings when the behavior depicted was subjective rather than objective.

The participants rated questions most highly for objective behaviors. Their ratings may indicate how concerning or how troubling the behavior was to the participant. Overall, the expert mean rating for objective behavior vignettes for each of the four questions was only slightly higher than their rating for the subjective behavior vignettes. Participants, however, rated objective behavior vignettes as or much more concerning than the subjective behavior video vignettes. This provides initial support for the hypothesis that participants were rating objective behaviors as more troubling than vignettes depicting subjective behaviors. It could be that the behaviors selected to represent objective behaviors – physical aggression and smoking – were not representative of the broader class of objective behaviors included in other analyses of the impact of behavior class on disproportional disciplinary outcomes. That broader class includes tardiness, vandalism, and other less stigmatized behaviors. Physical aggression and smoking may be perceived as more serious behaviors within the objective class and teachers may not regularly observe them, whereas it is likely that teachers regularly experience instances of defiance and disrespect in the course of their work. It could be that familiarity decreases the perception of the seriousness of the behavior.

Racial differences in rating accuracy. Perhaps the most surprising finding in this study was the failure to find a significant main or interaction effect of student race on participants' accuracy in rating student behaviors. That is, the accuracy of the ratings of participants exposed to video vignettes depicting a White target student was not significantly different from the accuracy of the ratings of the participants exposed to video vignettes depicting a Black target student. Prior studies in real world contexts have consistently shown a significant impact of race on nearly all disciplinary decisions, with teachers and administrators persistently over-referring Black students for disciplinary consequences.

The current sample may have failed to identify a significant effect of race due to the attrition of participants who would be more likely to rate a Black student differently from a White student. Participants were not blind to the study purpose. Consent materials explicitly stated that the purpose of the study was to evaluate differences in responses based on race and to address implicit bias in disciplinary decision making. This framing of the study may have impacted participants' choice to drop out when initially presented with a Black target student, with the possibility that more biased participants were more likely to drop out given the desire not to appear biased**.**

Conclusions

This study sought to evaluate the role of behavior class, race, cultural context, level of concern about the behavior and operational definition condition play in increasing the accuracy of participant ratings of behavior as measured by increased agreement with expert ratings, particularly within a SWPBS framework. SWPBS's statistically significant impact on disproportionate disciplinary outcomes but failure to eliminate them led the researchers to hypothesize that operational definitions, when interacting with the other factors described above, may significantly increase teachers' accuracy in identifying true problem behaviors for which disciplinary action is the appropriate response. Increased accuracy may then decrease disproportionate disciplinary outcomes. Despite the evidence to support the theoretical interaction of these variables from prior research, neither of the hypothesized interaction effects were significant. These findings have implications for practice, theory and future research.

Implications for practice. Each of the variables, race of the target student, behavior class, level of concern about the behavior and operational definition condition, provide some direction for practitioners considering these results. For example, the lack of impact of the race of the target student was surprising but does not negate the salience of race in evaluations of real-world disciplinary outcomes. Practitioners must continue to track rates of disproportionality by race within educational contexts.

However, it is not enough to admire the problem, proactive measures to address disproportionality must also be adopted. Administrators and other disciplinarians should note that educators were no more accurate when rating mild than moderate behaviors. In practice, this indicates that educators may be no more likely to accurately refer major or minor infractions for disciplinary consequences. In schools, minor disciplinary infractions are frequently not investigated, indeed sometimes major disciplinary infractions are not investigated beyond a brief interview with the teacher, which may increase the likelihood of inappropriate disciplinary referrals.

The significance of behavior class also suggests the need for a particular focus on more serious objective behavioral violations in addition to the more traditionally concerning subjective. Educators may be more concerned about a student when the behavior was physical aggression or smoking as opposed to defiance and disrespect. Therefore, those imposing disciplinary consequences should be particularly mindful of the over-response to objective behavioral violations. When it comes to disciplinary infractions, it may be more prudent to treat the referral like being charged with a crime – as innocent until proven guilty. Although this may not be the most efficient method for responding to disciplinary infractions, more efficient methods where an educator’s word is taken as fact with no due process may actually be leading to inappropriate disciplinary consequences for both White and Black students.

Lastly, despite the experimental manipulation of the operational definition condition, this study did not identify any impact of the intervention on the accuracy of participants' ratings of problem behavior. Given this outcome, the role of operational definitions within a SWPBS model must be evaluated. SWPBS has been described as a necessary but not sufficient methodology for addressing disproportionate disciplinary outcomes for Black students, as it has not been associated with the elimination of disproportionality, only its reduction. This evaluation of operational definitions may indicate that operational definitions, as applied in this study, are not a significant factor in increasing the accuracy of educators' ratings of behavior within SWPBS, at least in the sense of becoming closer to expert ratings.

Further, the current study suggests it is may not be possible to assign the decreases in rates of disproportionality seen in prior studies of the implementation of SWPBS to the "universal expectations of behavior" created by operational definitions, but instead those decreases may be attributable to other factors within the SWPBS model such as the focus on positive reinforcement and student-staff relationships. Instead of focusing on a universal understanding of what it means to be an appropriate student within a school, schools may see the greatest decreases in disproportionality when they work to improve the connections students have to the staff as well as increase the quality and quantity of positive feedback students receive. It is possible that through a renewed focus on reducing the number of inappropriate referrals through targeted investments in positive reinforcement and student-staff relationships, a decrease in racial disproportionality in discipline referrals will occur.

Philosophical implications and future directions. Although the practical applications of this study are certainly important, the philosophical limitations also have potentially far reaching implications for theory, research, and practice. SWPBS currently operates from a dominant cultural framework where universal expectations for student behavior are considered not only possible but desirable and can be imposed by an external authority. Indeed, most interventions to address student behavioral concerns operate from the premise that compliance with an outside authority is both desirable and attainable. CRT frameworks call on us to reject this as a form of systemic oppression that obscures white hegemonic cultural norms as universal truths.

Perhaps instead of conceptualizing universal behavioral expectations as the goal, instead we can recognize the shared right to define appropriate student behavior within each unique school context. Recognizing students', families' and community members' right to a voice in the definition of appropriate behavior may provide the theoretical foundation for the future of research in disproportionality. Indeed, Smolkowski, Girvan, McIntosh, Nese and Horner (2016) already make a similar recommendation for addressing Vulnerable Decision Points (VDPs) within SWPBS. They state, "school personnel can decrease (but not eliminate) subjectivity by creating and using operational definitions of each behavior, as well as the thresholds for no ODR, a minor ODR, and a major ODR."

Although the suggestion that school personnel develop these operational definitions and supporting materials is an excellent beginning, the inclusion of students, parents and community members may be necessary in order to develop operational definitions of problem behaviors that are not simply replicating the dominant cultural expectations. Bal's promising Culturally Responsive SWPBS model includes a "learning lab" component where constituent groups engage in ongoing, collaborative discussion to not only define the behaviors of interest but to engage in an ongoing process that allows for the refinement of all disciplinary procedures over time.

Future research should also evaluate the role of interventions not based on compliance with externally imposed behavior standards on disproportionality. A component analysis of the aspects of SWPBS related to positive student-staff relationships as well as positive reinforcement, decoupled from the universal, objective definitions of appropriate student behavior may provide valuable evidence about the mechanism behind the efficacy of SWPBS. It may be that relationships have a significant impact on not just students' experience of school expectations and disciplinary procedures but also educators' experiences of student behavior. Future research should not only evaluate the role of positive student-staff relationships but also consider potential interventions to assist students and teachers in repairing challenging relationships. This line of research may provide both key information about disproportionality as well as potential intervention to address this concern.

Regardless of the specific mechanism through which researchers attempt to address disproportionate disciplinary outcomes, we must continue to wrestle with not only the practical aspects of addressing racial disproportionality in disciplinary outcomes but also the ontological framework our work is situated within. Researchers' choices have the power to shape students' reality with interventions. Certainly, SWPBS has already shaped the daily reality of hundreds of thousands of school children. We now have the opportunity to end the unconscious perpetuation of systemic oppression and instead share our power to shape reality with the students, families and communities we seek to help. Instead of being saviors, we can be the co-authors of a shared reality that supports each student, in each school achieving their own unique goals.

REFERENCES CITED

Abikoff, H., Courtney, M., Pelham, W. E., & Koplewicz, H. S. (1993). Teachers' ratings of disruptive behaviors: The influence of halo effects. *Journal of Abnormal Child Psychology*, *21*(5), 519-533.

Anyon, Y., Jenson, J. M., Altschul, I., Farrar, J., McQueen, J., Greer, E., ... & Simmons, J. (2014). The persistent effect of race and the promise of alternatives to suspension in school discipline outcomes. *Children and Youth Services Review*, *44*, 379-386.

APA Zero Tolerance Task Force. (2008). Are zero tolerance policies effective in the schools? An evidentiary review and recommendations. *American Psychologist, 63*, 852-862.

Bates, L. A., & Glick, J. E. (2013). Does it matter if teachers and schools match the student? Racial and ethnic disparities in problem behaviors. *Social science research*, *42*(5), 1180-1190.

Beaver, K. M., Wright, J. P., & DeLisi, M. (2011). The racist teacher revisited: Race and social skills in a nationally representative sample of American children. *Classrooms: Management, Effectiveness, and Challenges. New York, NY: Nova Science Publishers*.

Bennett, C., & Harris, J. J. (1982). Suspensions and Expulsions of Male and Black Students: A Study of the Causes of Disproportionality. *Urban Education*, 16(4), 399-423.

Bradshaw, C. P., Koth, C. W., Thornton, L. A., & Leaf, P. J. (2009). Altering school climate through school-wide Positive Behavioral Interventions and Supports: Findings from a group-randomized effectiveness trial. *Prevention science*, *10*(2), 100-115.

Bradshaw, C. P., Mitchell, M. M., & Leaf, P. J. (2010). Examining the effects of Schoolwide Positive Behavioral Interventions and Supports on student outcomes results from a randomized controlled effectiveness trial in elementary schools. *Journal of Positive Behavior Interventions*, *12*(3), 133- 148.

Bradshaw, C. P., Mitchell, M. M., O'Brennan, L. M., & Leaf, P. J. (2010). Multilevel exploration of factors contributing to the overrepresentation of black students in office disciplinary referrals. *Journal of Educational Psychology, 102*(2), 508-520. doi:10.1037/a0018450

Bradshaw, C. P., Reinke, W. M., Brown, L. D., Bevans, K. B., & Leaf, P. J. (2008). Implementation of School-wide Positive Behavioral Interventions and Supports (PBIS) in elementary schools: Observations from a randomized trial. *Education and Treatment of Children*, *31*(1), 1-26.

Bradshaw, C. P., Waasdorp, T. E., & Leaf, P. J. (2015). Examining variation in the impact of School-wide Positive Behavioral Interventions and Supports: Findings from a randomized controlled effectiveness trial. *Journal of Educational Psychology*, *107*(2), 546.

Carter, M. (2003). Telling tales out of school: "What's the fate of a black story in a White world of stories?" *Counterpoints*, Vol 195, 29 – 48

Carter, P. (2008). Teaching students' fluency in multiple cultural codes. In M. Pollock (Ed.), *Every day anti-racism: Getting real about race in school.* New York: New Press.

Children's Defense Fund. (1975). *School Suspensions: Are They Helping Children?* Cambridge, MA: Washington Research Project.

Childs, K. E., Kincaid, D., George, H. P., & Gage, N. A. (2015). The relationship between School-Wide Implementation of Positive Behavior Intervention and Supports and student discipline outcomes. *Journal of Positive Behavior Interventions*, *18*(2), 89-99.

Delgado, R. (1988) Critical legal studies and the realities of race: does the fundamental contradiction have a corollary?, *Harvard Civil Rights–Civil Liberties Law Review*, 23, 407–413.

Delgado, R. (1989) Storytelling for oppositionists and others: a plea for narrative, *Michigan Law Review*, 87, 2411–2441.

Delgado, R. (1992) The imperial scholar revisited: how to marginalize outsider writing, ten years later, *University of Pennsylvania Law Review*, 140, 1349–1372.

Delgado, R. (1993) On telling stories in school: a reply to Farber and Sherry, *Vanderbilt Law Review*, 46, 665–676.

Delgado, R. & Stefancic, J. (1993) Critical race theory: an annotated bibliography, *Virginia Law Review,* 79, 461–516.

Delgado, R. & Stefancic, J. (Eds) (1997) Critical white studies: looking behind the mirror(Philadelphia, *Temple University Press*).

Delgado, R. & Stefancic, J. (2001) Critical race theory: an introduction (New York, *New York University Press*).

Delgado Bernal, D. (1998) Using a Chicana feminist epistemology in educational research, *Harvard Educational Review*, 68(4), 555–582.

Delgado Bernal, D. (2001) Living and learning pedagogies of the home: the mestiza consciousness of Chicana students, *International Journal of Qualitative Studies in Education*, 14(5), 623–639.

Delgado Bernal, D. (2002) Critical race theory, LatCrit theory and critical raced-gendered epistemologies: recognizing Students of Color as holders and creators of knowledge, *Qualitative Inquiry*, 8(1), 105–126.

Delgado-Gaitan, C. (1992) School matters in the Mexican American home: socializing children to education, *American Educational Research Journal*, 29(3), 495–513.

Downey, D., & Pribesh, S. (2004). When race matters: Teachers' evaluations of students' classroom behavior. *Sociology of Education, 77*, 267-282.

Finn, J. D., & Servoss, T. J. (2013). Misbehavior, suspensions, and security measures in high school: Racial/ethnic and gender differences. In Losen, D. (Ed.), *Closing the school discipline gap: Research for policymakers.* New York: Teachers College Press.

Fordham, S., & Ogbu, J. U. (1986). Black students' school success: Coping with the "burden of 'acting white'". *The Urban Review*, *18*(3), 176-206.

Freire, P. (2000). *Pedagogy of the oppressed*. Bloomsbury Publishing.

Gibson, P. A., Wilson, R., Haight, W., Kayama, M., & Marshall, J. M. (2014). The role of race in the out- of-school suspensions of black students: The perspectives of students with suspensions, their parents and educators. *Children and Youth Services Review*, *47*, 274-282.

Gregory, A., & Weinstein, R. S. (2008). The discipline gap and African Americans: Defiance or
cooperation in the high school classroom. *Journal of School Psychology*, *46*(4), 455-475.

Gregory, A., Skiba, R. J., & Noguera, P. A. (2010). The achievement gap and the discipline gap: Two sides of the same coin? *Educational Researcher, 39*(1), 59-68. doi:10.3102/0013189X09357621

Horner, R. H. (2013, June 12). *Implementing SWPBIS with quality, equity and efficiency.* Available from www.pbis.org

Horner, R. H., Sugai, G., & Anderson, C. M. (2010). Examining the evidence base for school-wide positive behavior support. Focus on Exceptionality, 42, 1–14.

Horner, R. H., Kincaid, D., Sugai, G., Lewis, T., Eber, L., Barrett, S., et al. (2014). Scaling up school-wide positive behavioral interventions and supports: The experiences of seven states with documented success. Journal of Positive behavioral Interventions, 16, 197–208. doi:10.1177/1098300713503685.

Horner, R. H., Sugai, G., Smolkowski, K., Eber, L., Nakasato, J., Todd, A. W., & Esperanza, J. (2009). A randomized, wait-list controlled effectiveness trial assessing School-wide Positive Behavior Support in elementary schools. *Journal of Positive Behavior Interventions*. *11*(3), 133-144.

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

Education Act, 20 U.S.C. § 1400 (2004).

Ladson-Billings, G. (1998). Just what is critical race theory and what's it doing in a nice field like education?, *International Journal of Qualitative Studies in Education*, 11:1, 7-24, DOI: 10.1080/095183998236863

Ladson-Billings, G. (2000). Fighting for our lives: Preparing Teachers to Teach African American Students. *Journal of Teacher Education*, v51 n3 p206-14 May-Jun 2000.

Ladson-Billings, G. (2005). From the achievement gap to the education debt: Understanding achievement in U.S. schools. *Educational Researcher, 35*(7), 3-12.

Lewin, K. (1946). Action research and minority problems. *Journal of Social Issues, 2*(4), 34-46.

Losen, D. J., & Gillespie, J. (2012). Opportunities suspended: The disparate impact of disciplinary exclusion from school. *UCLA: The Civil Rights Project.* Retrieved from http://escholarship.org/uc/item/3g36n0c3

Losen, D. J., Hodson, C. L., Keith, I. I., Michael, A., Morrison, K., & Belway, S. (2015). Are we closing the school discipline gap? *K-12 Racial Disparities in School Discipline*. Retrieved from https://www.civilrightsproject.ucla.edu/resources/projects/center-for-civil-rights-remedies/school- to-prison-folder/federal-reports/are-we-closing-the-school-discipline- gap/AreWeClosingTheSchoolDisciplineGap_FINAL221.pdf

Martinez, A., McMahon, S. D., & Treger, S. (2016). Individual-and school-level predictors of student office disciplinary referrals. *Journal of Emotional and Behavioral Disorders, 24*(1), 30-41.

May, S., Ard, W. I., Todd, A. W., Horner, R. H., Glasgow, A., Sugai, G., & Sprague, J. (2003). School- wide information system. *Eugene: Educational and Community Supports, University of Oregon*.

McCarthy, J. D., & Hoge, D. R. (1987). The social construction of school punishment: Racial disadvantage out of universalistic process. *Social Forces*,*65*(4), 1101-1120.

McElderry, C. G., & Cheng, T. C. (2014). Understanding the discipline gap from an ecological perspective. *Children & Schools*, *36*(4), 241-249.

McIntosh, K., Barnes, A., Eliason, B., & Morris, K. (2014). *Using discipline data within SWPBIS to identify and address disproportionality: A guide for school teams. OSEP Technical Assistance Center on Positive Behavioral Interventions and Supports.* www.pbis.org.

McIntosh, K., Girvan, E. J., Horner, R. H., Smolkowski, K., & Sugai, G. (2014). *Recommendations for addressing discipline disproportionality in education.* OSEP Technical Assistance Center on Positive Behavioral Interventions and Supports. www.pbis.org.

McKinney, E., Bartholomew, C., & Gray, L. (2010). RTI and SWPBIS: Confronting the problem of disproportionality. *NASP Communiqué*, *38*(6), 1-26.

Nelson, J. R., Benner, G. J., Lane, K., & Smith, B. W. (2004). Academic achievement of K-12 students with emotional and behavioral disorders. *Exceptional Children*, *71*(1), 59-73.

Nelson, J. R., Martella, R. M., & Marchand-Martella, N. (2002). Maximizing student learning: The effects of a comprehensive school-based program for preventing problem behaviors. *Journal of Emotional and Behavioral Disorders*,*10*(3), 136-148.

Nelson, J.L., Palonsky, S.B., & McCarthy, M.R. (2004). The academic achievement gap: Old remedies or new. *Critical Issues in Education: Dialogues and Dialectics, 5th Edition* (pp.103-126). Boston: McGraw-Hill.

Nie, N. H., Bent, D. H., & Hull, C. H. (1970) *SPSS: Statistical package for the social sciences.* New York: McGraw-Hill.

Noguera, P. A. (2003). Schools, prisons, and social implications of punishment: Rethinking disciplinary practices. *Theory into Practice*, *42*(4), 341-350.

Payne, A. A., & Welch, K. (2015). Restorative justice in schools: The influence of race on

Pigott, R. L., & Cowen, E.L. (2000). Teacher race, child race, racial congruence, and teacher ratings of children's school adjustment. *Journal of School Psychology, 38*(2), 177-195.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879.

Raffaele Mendez, L. M., & Knoff, H. M. (2003). Who gets suspended from school and why: A demographic analysis of schools and disciplinary infractions in a large school district. *Education and Treatment of Children, 26*, 30-51.

Raines, T. C., Dever, B. V., Kamphaus, R. W., & Roach, A. T. (2012). Universal screening for behavioral and emotional risk: A promising method for reducing disproportionate placement in special education. *The Journal of Negro Education*, *81*(3), 283-296.

Rocque, M. (2010). Office discipline and student behavior: Does race matter? *American Journal of Education, 116*(4), 557-581.

Skiba, R. J., & Rausch, M. K. (2006). Zero tolerance, suspension, and expulsion: Questions of equity and effectiveness. *Handbook of classroom management: Research, practice, and contemporary issues*, 1063-1089.

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.

Skiba, R. J., Horner, R. H., Chung, C., Rausch, M. K., May, S. L., & Tobin, T. (2011). Race is not neutral: A national investigation of Black and Latino disproportionality in school discipline. *School Psychology Review, 40*(1), 85-107.

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.

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

Skiba, R., & Leone, P. (2001). Zero tolerance and school security measures: A failed experiment. *Racial Profiling and Punishment in US Public Schools*, 13.

Skiba, R., & Peterson, R. (1999). The dark side of zero tolerance: Can punishment lead to safe schools?. *The Phi Delta Kappan*, *80*(5), 372-382.

Solórzano, D. (1989) Teaching and social change: reflections on a Freirean approach in a college classroom, Teaching Sociology, 17, 218–225.

Solórzano, D. (1992) Chicano mobility aspirations: a theoretical and empirical note,Latino StudiesJournal, 3, 48–66.Solórzano, D. (1997) Images and words that wound: critical race theory, racial stereotyping andteacher education, Teacher Education Quarterly, 24, 5–19.

Solórzano, D. (1998) Critical race theory, racial and gender microaggressions, and the experiencesof Chicana and Chicano Scholars,International Journal of Qualitative Studies in Education,11,121–136.

Solórzano, D., Ceja, M. & Yosso, T. (2000) Critical race theory, racial microaggressions andcampus racial climate: the experiences of African-American college students, Journal of NegroEducation, 69(1/2), 60–73.

Solórzano, D. & Delgado Bernal, D. (2001) Critical race theory, transformational resistance and social justice: Chicana and Chicano students in an urban context,Urban Education,36,308–342.

Solórzano, D. & Solórzano, R. (1995) The Chicano educational experience: a proposed frame-work for effective schools in Chicano communities, Educational Policy, 9, 293–314.

Solórzano, D. & Yosso, T. (2000) Toward a critical race theory of Chicana and Chicano education, in: C. Tejeda, C. Martinez, Z. Leonardo & P. McLaren (Eds)Charting new terrains of Chicana(o)/Latina(o) education (Cresskill, NJ, Hampton Press), 35–65.

Solórzano, D. & Yosso, T. (2002b) Maintaining social justice hopes within academic realities: a Freirean approach to critical race/LatCrit pedagogy, Denver Law Review, 78(4), 595–621.

Sprague, J. & Nelson, C. (2012). School-Wide Positive Behavior Interventions and Supports and Restorative Discipline in Schools. Retrieved April 29, 2017 from http://pages.uoregon.edu/ivdb/documents/RJ%20and%20PBIS%20Monograph%20for%20OSEP %2010.11.12.pdf.

Stanfield, J. H., II. (1993). Methodological reflections: An introduction. In R. M. Dennis & J. H. Stanfield (Eds.), *Race and ethnicity in research methods*. Newbury Park, CA: Sag, 3-15.

Sugai, G., & Horner, R. (2002). The evolution of discipline practices: School-wide Positive Behavior Supports. *Child & Family Behavior Therapy, 24*(1-2), 23-50.

Sugai, G., & Horner, R. H. (2006). A promising approach for expanding and sustaining School-wide Positive Behavior Support. *School Psychology Review, 35*, 245–259.

Sugai, G., & Horner, R. H. (2009). Responsiveness-to-intervention and School-wide Positive Behavior Supports: Integration of multi-tiered system approaches. *Exceptionality, 17*(4), 223-237.

Sugai, G., Horner, R. H., & Lewis-Palmer, T. L. (2001). Team implementation checklist (TIC). *Eugene, OR: Educational and Community Supports. Available at www. pbis. org.*

Sugai, G., Lewis-Palmer, T. L., Todd, A. W., & Horner, R. H. (2001). *Effective Behavior Support (EBS) Survey: Assessing and planning behavior supports in schools.* Eugene, OR: Center for Positive Behavioral Supports, University of Oregon.

Sugai, G., Lewis-Palmer, T. L., Todd, A. W., & Horner, R. H. (2001). *School-wide evaluation tool (SET).* Eugene, OR: Center for Positive Behavioral Supports, University of Oregon.

Sugai, G., Sprague, J. R., Horner, R. H., & Walker, H. M. (2000). Preventing school violence: The use of office discipline referrals to assess and monitor school-wide discipline interventions. *Journal of Emotional and Behavioral Disorders, 8*(2), 94-101.

Swain-Bradway, J., Loman, S. L., Vincent, C. G. (2014). Systematically addressing discipline disproportionality through the application of a school-wide framework. *Multiple Voices for Ethnically Diverse Exceptional Learners*, *14*(1), 3-17.

Theriot, M. T., Craun, S. W., & Dupper, D. R. (2010). Multilevel evaluation of factors predicting school exclusion among middle and high school students. *Children and Youth Services Review*, *32*(1), 13-19.

Tobin, T. J., & Vincent, C. G. (2011). Strategies for preventing disproportionate exclusions of Black students. *Preventing School Failure: Alternative Education for Children and Youth*, *55*(4), 192- 201.

Vincent, C. G., & Tobin, T. J. (2011). The relationship between implementation of School-wide Positive Behavior Support (SWPBS) and disciplinary exclusion of students from various ethnic backgrounds with and without disabilities. *Journal of Emotional and Behavioral Disorders, 19*(4), 217-232.

Vincent, C. G., Randall, C., Cartledge, G., Tobin, T. J., & Swain-Bradway, J. (2011). Toward a conceptual integration of cultural responsiveness and schoolwide positive behavior support. *Journal of Positive Behavior Interventions, 13*(4), 219-229.

Vincent, C. G., Sprague, J. R., & Gau, J. M. (2013, July). *The effectiveness of School-wide Positive Behavior Interventions and Supports for reducing racially inequitable disciplinary exclusions in middle schools*. Paper presented at the Closing the School Discipline Gap: Research to Practice Conference, Washington, DC.

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.

Vincent, C. G., Swain-Bradway, J., Tobin, T. J., & May, S. (2011). Disciplinary referrals for culturally and linguistically diverse students with and without disabilities: Patterns resulting from School- wide Positive Behavior Support. *Exceptionality*, *19*(3), 175-190.

Wallace Jr, J. M., Goodkind, S., Wallace, C. M., & Bachman, J. G. (2008). Racial, ethnic, and gender differences in school discipline among US high school students: 1991-2005. *The Negro Educational Review*, *59*(1-2), 47.

Wehlage, G., & Rutter, R. (1986). Dropping out: How much do schools contribute to the problem? *The Teachers College Record*, *87*(3), 374-392.

Welch, K., & Payne, A. A. (2012). Exclusionary school punishment: The effect of racial threat on expulsion and suspension. *Youth Violence and Juvenile Justice, 10*(2), 155-171. doi:10.1177/1541204011423766

Whyte, W. F. E. (1991). *Participatory action research.* Sage Publications, Inc.

Wright, J. P., Morgan, M. A., Coyne, M. A., Beaver, K. M., & Barnes, J. C. (2014). Prior problem behavior accounts for the racial gap in school suspensions. *Journal of Criminal Justice*, *42*(3), 257-266.

Wu, S. C., Pink, W., Crain, R., & Moles, O. (1982). Student suspension: A critical reappraisal. *The Urban Review*, *14*(4), 245-303.

Yosso, T.J. (2005) Whose culture has capital? A critical race theory discussion of community cultural wealth, *Race Ethnicity and Education*, 8:1, 69-91, DOI:10.1080/1361332052000341006