Consequences of Misassignment to Treatment: Examining Targeted Policy Interventions in Education

Alec I. Kennedy

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Mark C. Long, Chair

Jacob L. Vigdor

Thomas S. Richardson

Min Sun

Program Authorized to Offer Degree:
Daniel J. Evans School of Public Policy and Governance

University of Washington

Abstract

Consequences of Misassignment to Treatment:
Examining Targeted Policy Interventions in Education

Alec I. Kennedy

Chair of the Supervisory Committee:
Professor Mark C. Long
Daniel J. Evans School of Public Policy and Governance

In education, students are often assigned to specialized programs or receive targeted interventions based on some standards and requirements. The criteria used to select the students for such targeted policies must weigh the trade-off between serving too many students and not serving enough. Under such selection criteria, I identify students who are selected into a treatment category that does not maximize their benefits as *consequences of misassignment to treatment*. As targeted policy interventions are aimed at providing supports for students with specific needs, it is important that we establish a firm understanding and provide guidance on how selection criteria should be designed and used to assure that *all* students are appropriately served. This dissertation explores several aspects of such selection criteria through the examination of three policies in education that seek to select students (or school districts) for targeted interventions and supports. The exploration provides learning and recommendations on how federal, state, and local governments can leverage data and analysis to limit the consequences of misassignment to treatment.

The first chapter of this dissertation introduces a framework to facilitate in the understanding of the consequences to misassignment to treatment. The framework establishes a way of understanding how selection criteria for targeted policy interventions can fail and lead to losses of benefits for both targeted and non-targeted groups.

The second chapter, *Are We Correctly Measuring the Disproportionality of Minority Students in Special Education?*, asks whether the current methods for measuring disproportionality in special education are accurately identifying school districts in need of a change of referral practices. Currently, the U.S. Department of Education's Office of Civil Rights (OCR) monitors whether school districts disproportionately refer minority students into special education services in an effort to prevent overreferral. However, these monitoring methods do not account for student-level factors (notably, socioeconomic status) which have been theorized to be positively associated with the presence of a learning disability. If minority students are more likely to experience low socioeconomic status than their White peers, then there may be good reason for minority students to be overrepresented in special education (i.e., the current measures of disproportionality could be susceptible to bias). In this chapter, I introduce an alternative measure for tracking disproportionality that removes this threat of bias. I then compare the performance of this alternative measure of disproportionality with the one currently in use and find that the currently used measure tends to overstate disproportionality. This finding suggests that current measures used by states to monitor disproportionality may incorrectly label some school districts as having an overrepresentation problem when, in fact, they do not.

The third chapter, *Successfully Transitioning English Language Learners into the Mainstream Classroom*, investigates whether the state-established thresholds on language assessments used to determine whether English language learners (ELLs) are ready to transition from specialized services into the mainstream classroom are set in a way that maximizes the estimated benefits for ELLs in *all* instructional contexts. In this chapter, I find that the cutscores used to determine eligibility for ELL program exit in one state are generally set to promote the successful transition of ELLs into the mainstream classroom. However, I find that the cutscores do not work for all. Specifically, I find that Spanish-speaking students exiting bilingual programs struggle to succeed in English subjects after exiting ELL

services. This finding suggests that the current thresholds are not working for *all* and should be modified in some way. This modification could involve the inclusion of teacher input into transitioning decisions or the close monitoring of recently exited ELL students in order to provide appropriate supports when needed.

The fourth chapter, *Making the Cut: An Optimization Approach for Setting Cutpoints in Targeted Policy Interventions in Education*, provides an introduction to an optimization approach that can be used in Early Warning Indicators (EWI) or Early Warning Systems (EWS) to set *optimal* cutpoints on key leading indicators to determine eligibility for targeted supports. I apply the approach to two district examples of EWI/EWS to identify cutpoints that minimize misclassification rates. In the first example, I identify optimal cutpoints on kindergarten and first grade assessments to identify students who show evidence of having dyslexia and are in need of early literacy interventions. In the next example, I use the approach to produce thresholds that can be used to identify students at risk of becoming chronically absent in order to provide them with additional resources and support during the school year to help them overcome barriers to attendance. The approaches introduced and used in this chapter can offer guidance to districts or states seeking to establish EWI/EWS to identify students in need of targeted interventions or supports.

The final chapter concludes this dissertation by presenting learning from each of the chapters on how selection criteria in targeted policy interventions can be better designed to limit consequences of misassignment to treatment.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I'd first like to thank my committee: Mark, Jake, Thomas, and Min. Without their time, guidance and support, this dissertation would not have been possible. I'd especially like to thank my chair Mark Long, who spent countless hours giving me invaluable advice and challenging me as a researcher. I can honestly say that I would not have made it to this point without him.

I learn best by doing. I received a lot of hands-on experience conducting research during my time at the University of Washington. Maria (Cuky) Perez, Jesse Levin, Mark Long, Min Sun, and Bob Plotnick all allowed me the opportunity to learn from them while collaborating on research projects. I'd also like to thank Amy Li, Min Li, and Eric Anderson for inviting me to contribute to their research. I also want to thank Nathan Adkins, Anna Wheeler, Junmeng Zhu, and Zach LeClair for the conversations about research methods and education policy that greatly benefited my own research. They all have had an influence on the work done in this dissertation.

I didn't come into this program alone. I couldn't have had a better cohort of individuals to learn with than Allison, Ryan, Sarah, and Shane (and, honorary member, Wei). Whether it was long hours of studying for comprehensive exams or chatting about Buffy the Vampire Slayer, it was a joy to share a PhD office with these individuals (even when that office got up to 100 degrees in the summer!).

I met some amazing friends during my time at UW. Their support kept me going. Shout out to: Tyler, Wei, Sarah, Amy, Paul, Linda, Daniel, Ting, Pao, Ti, Dongsheng, and Cricket. It seemed like whenever I felt like I had been knocked down, they were always there to pick me back up. My most cherished memories of my time at UW are with them.

I also have to thank my RPA colleagues who helped me get to the finish line. They lifted my confidence while also adding extra pressure for me to succeed. I look forward to the important work I will do with them in the future.

Finally, I must thank my parents and heroes, Arlene and William Kennedy. They both showed me that you can be passionate about your work. My mother was a teacher for 36 years and my father was a civil rights attorney for over 45 years. They supported me my whole life in searching for my own passion. In writing this dissertation, I believe I have finally found it. It turns out my passion was not too far from their own. I look forward to continuing their work.

DEDICATION

To my parents

Chapter 1

INTRODUCTION

In education, students are often assigned to specialized programs or receive targeted interventions based on some standards and requirements. The criteria used to select the students for such targeted policies must weigh the trade-off between serving too many students and not serving enough. Under such selection criteria, I identify students who are selected into a treatment category that does not maximize their benefits as *consequences of misassignment to treatment*. As targeted policy interventions are aimed at providing supports for students with specific needs, it is important that we establish a firm understanding and provide guidance on how selection criteria should be created and used to assure that *all* students are appropriately served. This dissertation explores several aspects of such selection criteria through the examination of three policies in education that seek to select students (or school districts) for targeted interventions and supports. The exploration provides learning and recommendations on how federal, state, and local governments can leverage data and analysis to limit the consequences of misassignment to treatment.

Students are continuously being evaluated for different interventions in their school. Let's say that a teacher/parent assesses a student for eligibility into a certain treatment offered by the school. We could imagine a teacher/parent assigning a "score" to this student and basing program eligibility on whether this score exceeds or falls below a certain threshold. The goal in this scenario is to create a selection system that assures that the student receives a treatment (or no treatment) that maximizes the benefits he or she receives.

For a concrete example, imagine school discipline as a treatment. The goal of such an intervention may be to correct problem behavior and improve all students' academic achievement. A teacher and principal must collaborate in finding the best intervention for a

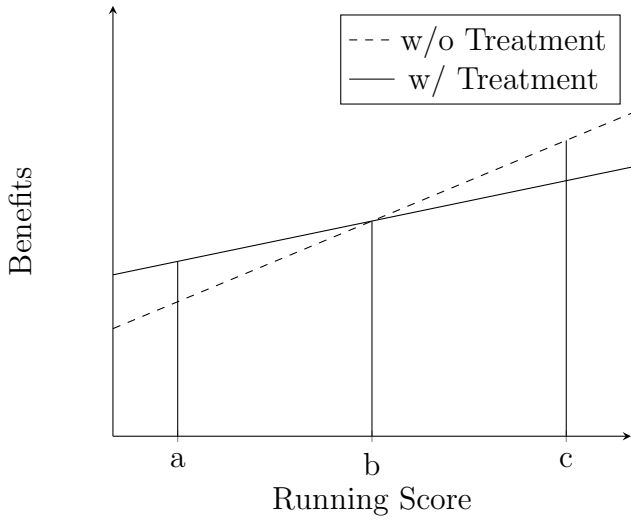
student based on their perception of the student's behavior and their belief in the effectiveness of the disciplinary action. That is, the teacher and principal construct a behavior "score" for a student and use that as a way of deciding what intervention will take place. If there are any errors in the evaluation of the student such as misdiagnosis of the reasons behind the student's behavior, then the student could be given an intervention that neither corrects their behavior nor helps improve their academic achievement. In this case, the student can be thought of as being misassigned to treatment and face potential costs because of it.

A generalization of this scenario is illustrated in Figure 1.1 (this is an adaptation of a figure used in Robinson (2011)).[1] The figure represents the benefits that are received for different values of a running score (i.e., the score assigned to determine eligibility) under two different scenarios: (1) treatment (solid line) and (2) no treatment (dashed line). In this illustration, there is a clear benefit to targeting the treatment to a subset of students. Namely, any student with a running score below b would see higher benefits from receiving the treatment. Of course, this assumes a costless intervention that can be given to every student falling below b.[2] Students with a running score greater than b would benefit more from not receiving the treatment. However, if we were to only treat students with a running score below a, students with a running score between a and b would fare worse given that they are not receiving a beneficial treatment. Conversely, if students with a running score less than c were to be treated, students with a running score between b and c would be taken out of setting (no treatment) that would have led to higher benefits. This toy example emphasizes the role that selection criteria can play in limiting the consequences of misassigning students

[1] It should be noted that Robinson (2011) also includes two more examples of this illustration. A scenario in which treatment is always preferred and a scenario in which no treatment is always preferred. Neither of the scenarios require the establishment of selection criteria. The two scenarios require either a maintenance of the status quo (in the case of no treatment always being preferred) or an overhaul of a system (in the case of treatment always being preferred). In the latter case, cost of the treatment should be considered and expanded up until the point at which marginal benefit is equal to marginal cost. Given the focus of this dissertation on *targeted* policy interventions, it would not make sense to focus on the other two scenarios when illustrating the concept of consequences of misassignment to treatment.

[2] In a scenario where costs are assumed, it would be optimal to apply the intervention where marginal cost is equal to the marginal benefit. This would mean that some students would not get a treatment even though they would benefit from it.

Figure 1.1: Setting Selection Criteria to Maximize Benefits



Source: Adapted from Robinson (2011).

to treatments that offer lower benefits potential.

This dissertation focuses on selection criteria in education. In order to appropriately serve *all* students, selection criteria must be carefully designed to assure that students are placed in a setting that maximizes their benefits. In this dissertation, I examine three targeted policies in education that aim to provide additional supports or services to a subpopulation of students (or districts) that the system believes benefits from such services (and does not serve those who it believes do not benefit). I closely examine the selection criteria used under each policy and provide suggestions on how these criteria can be improved to better balance the trade-off between serving too many and serving too few.

The second chapter, *Are We Correctly Measuring the Disproportionality of Minority Students in Special Education?*, asks whether the current methods for measuring disproportionality in special education are accurately identifying school districts in need of a change of referral practices. Currently, the U.S. Department of Education's Office of Civil Rights (OCR) monitors whether school districts disproportionately refer minority students into special ed-

ucation services in an effort to prevent overreferral. However, these monitoring methods do not account for student-level factors (notably, socioeconomic status) which have been theorized to be positively associated with the presence of a learning disability. If minority students are more likely to experience low socioeconomic status than their White peers, then there may be good reason for minority students to be overrepresented in special education (i.e., the current measures of disproportionality could be susceptible to bias). In this chapter, I introduce an alternative measure for tracking disproportionality that removes this threat of bias. I then compare the performance of this alternative measure of disproportionality with the one currently in use and find that the currently used measure tends to overstate disproportionality. This finding suggests that current measures used by states to monitor disproportionality may incorrectly label some school districts as having an overrepresentation problem when, in fact, they do not.

The third chapter, *Successfully Transitioning English Language Learners into the Mainstream Classroom*, investigates whether the state-established thresholds on language assessments used to determine whether English language learners (ELLs) are ready to transition from specialized services into the mainstream classroom are set in a way that maximizes the estimated benefits for ELLs in *all* instructional contexts. In this chapter, I find that the cutscores used to determine eligibility for ELL program exit in Washington state are generally set to promote the successful transition of ELLs into the mainstream classroom. However, I find that the cutscores do not work for all. Specifically, I find that Spanish-speaking students exiting bilingual programs struggle to succeed in English subjects after exiting ELL services. This finding suggests that the current thresholds are not working for *all* and the system should be modified in some way. This modification could involve the inclusion of teacher input into transitioning decisions or the close monitoring of recently exited ELL students in order to provide appropriate supports when needed.

The fourth chapter, *Making the Cut: An Optimization Approach for Setting Cutpoints in Targeted Policy Interventions in Education*, provides an introduction to an optimization approach that can be used in Early Warning Indicators (EWI) or Early Warning Systems

(EWS) to set *optimal* cutpoints on key leading indicators to determine eligibility for targeted supports. I apply the approach to two district examples of EWI/EWS to identify cutpoints that minimize misclassification rates. In the first example, I identify optimal cutpoints on kindergarten and first grade assessments to identify students who show evidence of having dyslexia and are in need of an early literacy interventions. In the next example, I use the approach to produce thresholds that can be used to identify students at risk of becoming chronically absent in order to provide them with additional resources and support during the school year to help them overcome barriers to attendance. The approaches introduced and used in this chapter can offer guidance to districts or states seeking to establish EWI/EWS to identify students in need of targeted interventions or supports.

The final chapter concludes this dissertation by presenting learning from each of the chapters on how selection criteria in targeted policy interventions can be better designed to limit consequences of misassignment to treatment.

Chapter 2

ARE WE CORRECTLY MEASURING THE DISPROPORTIONALITY OF MINORITY STUDENTS IN SPECIAL EDUCATION?

Abstract

Annually, states are required to monitor whether their school districts disproportionately refer minority students into special education services. However, these monitoring methods do not account for student factors theorized to be associated with the presence of a learning disability. I explore whether conclusions drawn from current measures of overrepresentation remain unchanged once relevant student-level factors are accounted for. Using parameters estimated from logit models, I construct a weighted adjusted risk ratio (WARR) that accounts for student-level characteristics and is comparable across districts within a state. Comparing the WARR with a common and comparable measure of disproportionality, the weighted risk ratio (WRR), I find that the common measures tend to overstate overrepresentation for certain minority groups and underrepresentation for others. Findings from this study point out some of the flaws in the ways states choose to measure disproportionality in districts. I conclude with some suggestions on the ways states can improve their monitoring system to assist in preventing misclassification.

2.1 Introduction

The disproportionate representation of minority students in special education has been one of the more complex problems facing education scholars the past few decades. Lloyd Dunn is often cited as the first to observe this phenomenon. In his 1968 report, he observes that

> about 60 to 80 percent of the pupils taught by [teachers in mild mental retardation or MMR classes] are children from low status backgrounds - including Afro-Americans, American Indians, Mexicans, and Puerto Rican Americans; those from nonstandard English speaking, broken, disorganized, and inadequate homes; and children from other non-middle class environments (Dunn, 1968, p. 6).

One major concern over the findings of Dunn's report was that the disproportionality was indicative of the misplacement of minority students into special education services that they did not need. Because of these civil rights and educational concerns, a number of legal actions were taken to make sure all students were protected from being misidentified into special education (Artiles, Harry, Reschly, & Chinn, 2002; Coutinho & Oswald, 2000; Skiba et al., 2008). The passage of Public Law 94-142 in 1975 – known as the "Education for All Handicapped Children Act" – established protections for students in special education programs, further defining how a child was to be placed into a special education program and requiring schools to keep records and provide information to parents on the child's development. Since 1968, the U.S. Department of Education's Office for Civil Rights (OCR) has been collecting data on key education and civil rights issues at regular intervals. This type of monitoring led to the discovery in the 1980s that the problem of disproportionality was consistent over time. A number of legal battles erupted over these findings. The most well-known being *Larry P. v. Riles* (1972, 1979, 1984, 1986) which focused on the disproportionate representation of African American students in special education programs. The outcome declared disproportionate representation as discriminatory and ordered for the elimination of the overrepresentation problem.

In response to the pressure to resolve the issue, the government reauthorized the Individuals with Disabilities Education Act (IDEA) and included new provisions to prevent schools from exhibiting disproportionate representation. Revisions in the 2004 version of IDEA requires states to monitor disproportionate representation and, if found, to review local policies, practices, and procedures. In addition, districts identified as having significant problems with disproportionality are required to devote 15% of their Part B funds to early intervention strategies directed particularly, but not exclusively, towards the groups identified as being disproportionately represented.[1] Defining "significant disproportionality", until recently, was left to the state. However, a review of state policies regarding "significant disproportionalities" concluded that definitions varied so much across states that there was no assurance of problem districts being correctly identified nationally. For example, Louisiana's definition for "significant disproportionality" led to over half of its districts being required to offer early-intervention services, while no districts were labeled as showing a significant disproportionality under the stricter definition used by Washington state. The report believed that many state thresholds for identifying a "significant disproportionality" were set so high that it was incredibly unlikely for a district to be identified (Government Accountability Office, 2013). The report suggested that the Department of Education (DOE) provide more guidance through the introduction of a standard methodology for measuring disproportionalities. In response, the DOE's Office of Special Education and Rehabilitative Services (OSERS) introduced new regulations that, among other things, limited the types of measures that could be used for monitoring disproportionality (Department of Education, 2016a). While the rule change follows the GAO's recommendation, it is unclear whether the proposed measures are optimal for identifying problem districts. Finding the "correct" methodology for measuring disproportionality has not been rigorously pursued.

Currently, under the recent rule change, disproportionality must be measured using the risk ratio (RR). This measure provides a comparison of the risk of one student subgroup

[1]The Part B funds are federal dollars provided to states for the provision of Special Education and related services to children between the ages of 3 and 21.

being placed in special education and compares it to the risk of all other student groups receiving special education services. It is simply calculated as the ratio of the proportions of one group in special education compared to another group. For example, the risk ratio of a Black student being placed in special education would be calculated as:

$$\text{Risk Ratio} = \frac{\frac{\text{\# of Black students in special education}}{\text{\# of Black students}}}{\frac{\text{\# of comparison group students in special education}}{\text{\# of comparison group students}}} \tag{2.1}$$

A risk ratio of 2 would indicate that Black students are placed in special education at twice the rate of comparison group students. If the comparison group is too small to make a valid comparison, then the state is allowed to use an "alternative" risk ratio which replaces the denominator of the risk ratio with the risk of the comparison group at the state-level.[2]

Recent attention has questioned whether these current measures of disproportionality are accurately capturing the true problems of overrepresentation. Overrepresentation can be indicative of two stories: that (1) there is a biased referral process or (2) there are differential exposures to risk factors across race. Much of the discussion of the disproportionality provisions in IDEA revolved around preventing students from being misidentified and "inappropriately placed in special education" (i.e., the first story) (Samuels, 2004). But the standard measures of overrepresentation do not take into account the second explanation. That is, the risk ratio and alternate risk ratio do not account for the possibility that students (across subgroups) experience different exposures to risk of learning disabilities. In fact, a

[2]There are three important choices a state must make in determining how to define "significant disproportionality." The first is the type of measure to be used. Second, is the threshold that is used to determine when a district begins to show significant disproportionalities and, related, the number of years this threshold must be exceeded. Third, is the minimum cell/n size that must be exceeded for a district's disproportionality measure to be counted. The cell size is the number of students in a subgroup with a particular outcome. For example, in the risk ratio of a Black student being placed in special education, the cell sizes would be the number of Black students/comparison group students in special education. Analogously, the n size refers to the number of students in each subgroup. Again, in the example risk ratio, this would be the number of Black and comparison group students at the school. The risk ratio becomes unreliable under small cell/n sizes and therefore districts are excluded from the disproportionality calculation when they do not reach the minimum cell/n sizes. This chapter focuses on the first decision, but acknowledges that the other decisions are also important for a valid monitoring system.

number of studies find that minorities are no longer found to be overrepresented in special education (and sometimes are found to be underrepresented) when a number of other risk factors – most notably socioeconomic status and primary language – are controlled for (Hibel, Farkas, & Morgan, 2010; Morgan, Farkas, Hillemeier, & Maczuga, 2012; Morgan et al., 2015; Shifrer, Muller, & Callahan, 2011).

In their guidance to states regarding the new regulations for monitoring disproportionality, OSERS acknowledges the possibility of confounding risk factors that could explain the overrepresentation of minority students in special education, yet their newly introduced standard methodology (the common risk ratio) does not account for any of these student characteristics. They argue that requiring districts to review practices after being identified for a significant disproportionality by the risk ratio might help mitigate these alternative risk factors from causing the observed disproportionalities. That is, they argue that potential false positives could still have a benefit in reducing disproportionalities caused by alternative risk factors. While this is possibly true, I contend that this argument ignores an alternative problem: false negatives. If the OSERS acknowledge that they are using potentially biased measures, then they should be worried that some schools that actually have a misplacement problem may not be identified under such measures meaning that interventions are not happening where they should. This chapter proposes and evaluates the utility of an alternative methodology that can be used by the state for monitoring disproportionality in its school districts: the weighted adjusted risk ratio (WARR) which accounts for student-level characteristics commonly collected by school districts.

In this chapter, I find that models that do not include student-level control variables tend to overstate the overrepresentation problem for minority students who are commonly found to be overrepresented (e.g., Black and American Indian students). Asian students, who are typically underrepresented in special education, also experience an exaggeration of their status. However, these differences are not found to be too meaningful when considering the way states typically categorize disproportionality. That is, districts categorized by their measure on the "adjusted" risk ratio are still found to, generally, remain in the

same disproportionality category as defined by the state. This finding brings into question how beneficial a change to an adjusted risk ratio might be and how much a state can do to purge risk factors unrelated to school misidentification from such measures. While WARR does provide a better, and, arguably, more accurate measure of disproportionality caused by a biased referral process, it doesn't appear to provide any difference in the way current disproportionality measures categorize districts as showing overrepresentation. This brings into question the ability of states to monitor overrepresentation in districts accurately and the ways in which they define disproportionality.

2.2 Conceptual Framework

Overrepresentation of minority students in special education may be indicative of a larger problem: misclassification. Being incorrectly placed into a special education classroom can not only take educational opportunities away from students by removing them from the mainstream classroom, but can also limit their future labor market outcomes because of the negative connotations associated with the labels given to them (Artiles et al., 2002). In addition, inappropriate labeling may act as a self-fulfilling prophecy. For instance, a student labeled as having a learning disorder (LD) may have feelings of self-doubt and lose motivation to learn in the classroom (Harry, Klingner, Delpit, & Artiles, 2014).

It is clear that there are serious *potential* costs associated with misclassification, yet, even after nearly fifty years of work focusing on this problem, scholars still seek to understand the causes of overrepresentation in special education. While plenty of descriptive evidence has been offered, little research has made significant progress towards pinpointing the reasons reponsible for these observed disproportionalities (MacMillan & Reschly, 1998; Waitoller, Artiles, & Cheney, 2009). Multiple factors are entangled in the problem and it is impossible to focus on just one when trying to come up with a solution. Coutinho and Oswald (2000) present two hypotheses that generally capture the explanations for observed disproportionalities (p. 147):

1. Special education referral, assessment, and eligibility rely on processes that "work differently" across ethnic groups.

2. Ethnic groups are differentially susceptible to educational disability due to differences in exposure to structural and instructional factors.

The first hypothesis explains that aspects of the referral process lead to the differential treatment across ethnic groups. This suggest that parts of these processes lead to the higher special education referral rate for minority students when they may otherwise not need such services. The possibility of a problem with the referral system has been brought up in a number of papers, but a rigorous examination of this potential bias has been lacking in the literature (Arnold & Lassmann, 2003; Artiles et al., 2002; Harry & Anderson, 1994; Hernandez, Ramanathan, Harr, & Socias, 2008; Ofiesh, 2006; Skiba et al., 2008). In at least one paper using an experiment, evidence of bias in special education referral based on student SES was found (Podell & Soodak, 1993).

Studies have found that the majority of students (70%-90%) referred by parents/teachers for special education are found eligible for services (Klingner & Harry, 2006; Yoshida, 1980). The referrer, the teacher in a majority of the cases, is the initiator of the process and therefore teachers have the most influence in determining the racial makeup of special education programs. "Gatekeepers", such as school psychologists, have been set up to prevent students, once referred, from getting labeled incorrectly. This often does not correct the problem for students placed in "judgmental" or "soft" categories, such as mental retardation (MR), emotional disturbance (ED), or learning disability (LD), which are more subjective in their assessment and do not rely on a clear biological basis (MacMillan & Reschly, 1998; Skiba et al., 2008). The higher disproportionality in these types of categories has been confirmed in studies (Cross, Donovan, et al., 2002; T. Parrish, 2002). Because it is more likely that a student will be misplaced into these fuzzier categories, the fact that we see more minorities in these groups could be suggestive of two scenarios of misclassification. First, minority students could be at a greater risk of being misclassified. And second, non-minority students,

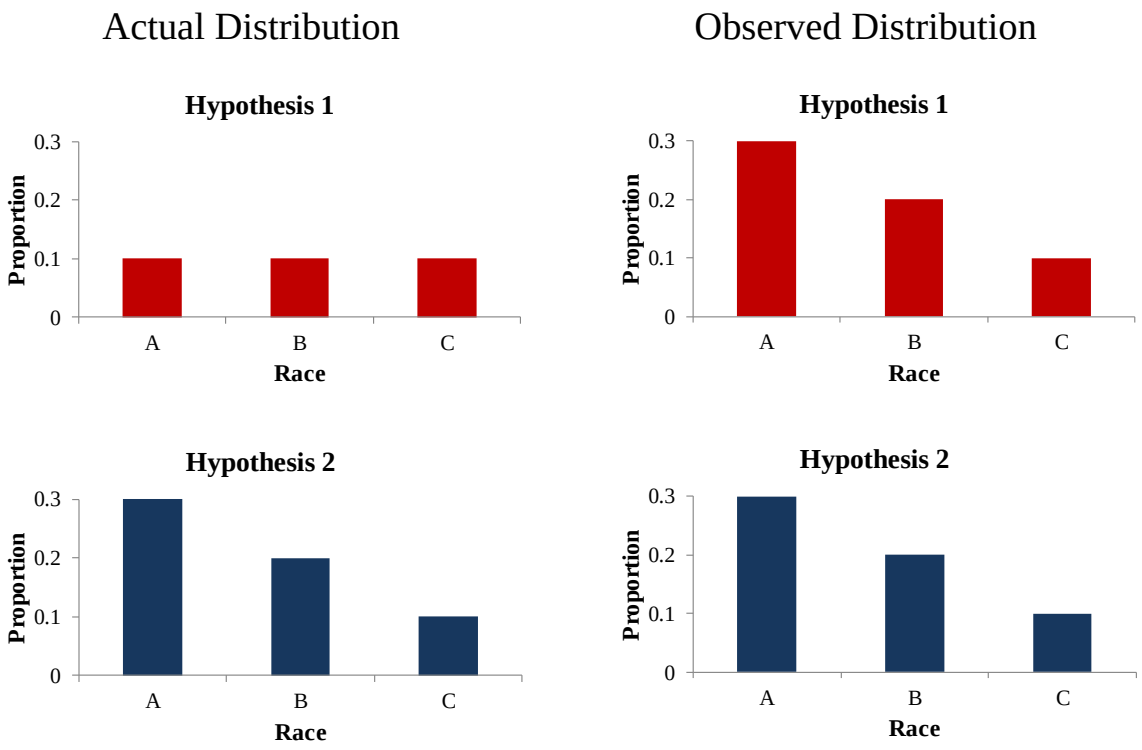
who are underrepresented, could be misclassified as not needing special education services.

The second hypothesis suggests that the underlying distribution of students needing specialized services may vary across different ethnic groups due to the exposure to risk factors out of the control of the student. One well-accepted theory is that the problem has to do with the higher likelihood of minority students to be exposed to poverty and how poverty can lead to academic underachievement and behavioral problems that increase the probability of being placed in special education programs (Coutinho, Oswald, & Best, 2002; MacMillan & Reschly, 1998). A number of studies report that deep and persistent poverty can have negative effects on early cognitive development and school achievement (Duncan, Yeung, Brooks-Gunn, & Smith, 1998; Magnuson & Votruba-Drzal, 2009; McLoyd, 1998). Other factors related with poverty, such as low birthweight, low maternal education, and prematurity were observed as being associated with higher referral rates (Delgado & Scott, 2006). Because of these relationships it is believed that minority students, who often experience poverty, will be more likely to be affected academically and are thus more likely to be placed into special education.

Figure 2.1 presents a visual representation of the two hypotheses. The graphs on the left show the actual distribution of disabilities across races while the bar graphs in the right column show the actual observed distribution of students by race after they are actually placed in special education. The first hypothesis states that the susceptibility of developing a learning disorder is the same across race and that there is differential treatment of certain subgroups in the classification system. That is, the risk of placement in special education is the same across race, but for some reason, current referral or placement processes lead to certain races being placed into special education more often than others. The second hypothesis asserts that the referral process correctly identifies students as needing specialized services, but structural factors are not equally distributed across student subgroups leading some groups to be more likely to need special education services. It is likely that overrepresentation is observed because of some combination of the two explanations.

Each story suggests a different type of solution. The first indicates that the problem can

Figure 2.1: Visual Representation of the Hypotheses



Note: The graphs on the left show the unobserved "true" distribution of disabilities across races. The bar graphs in the right column show the observed distribution of students by race after they are actually placed in special education. Under Hypothesis 1, the risk of having a learning disorder are the same across race, however, placement into special education does not reflect this even distribution. Hypothesis 2 presents a world where the true distribution of learning disabilities is not evenly distributed across races due to different exposures to risk factors. These different exposures explain why we observe different rates of special education placement.

be resolved at the school level. If there is some issue with the referral process, then the district should be able to make adjustments to lower the occurrence of misclassification that removes minority students from mainstream classrooms. This seems to be the current intent of IDEA's monitoring of disproportionate representation in districts. They want to identify the processes under school control that may lead to misidentification. Unfortunately, most states define "significant disproportionality" using measures that do not take into account the second hypothesis that certain minority groups are more susceptible to learning disorders because of structural factors (NASDSE Report, 2007).[3] Recent studies have used more complicated methods for measuring the disproportionalities in special education placement (Hibel et al., 2010; Morgan et al., 2012, 2015; Shifrer et al., 2011). Rather than just using simple population comparisons (i.e., comparing the proportion of a subgroup in special education to their share in the overall population), these studies have used covariate adjustment models with extensive controls collected by nationally representative longitudinal surveys to estimate disproportionalities conditional on other student-, family-, and school-level factors. These studies have found that some subgroups that are commonly seen as being overrepresented in special education no longer show evidence of overrepresentation once these other characteristics are controlled for. In fact, two of the studies find evidence that some groups tend to be underrepresented in special education once other variables (e.g., socioeconomic status) are conditioned on (Hibel et al., 2010; Morgan et al., 2015).

The findings of these studies using covariate adjustment models suggests that current federal policy may be singling out districts that do not benefit from making changes to their current identification policies for special education. This also suggests that districts identified as showing significant disproportionalities – commonly determined using simple population comparisons – may intentionally keep their minority students out of special education when the student is actually in need of specialized services.[4] In the only study looking

[3]Commonly used measures of disproportionalities include: the risk ratio, weighted risk ratio, alternate risk ratio, and E-formula none of which adjust for student characteristics (Roy, 2012).

[4]In an alarming example, Houston Independent School District's (HISD) fight against "overidentification" of African-American students led to intentional focus of reducing the number of Black males receiving

quantitatively at the effects of IDEA monitoring of school districts, Enayati (2014) finds that schools do appear to respond to federal monitoring. The results of this chapter indicate that districts singled out for showing a "significant disproportionality" in Michigan subsequently lower their measure of disproportionality measure by 41%. This was achieved by reducing the proportion of Black students being placed in special education rather than increasing the proportion of other students being placed in special education.

Furthermore, from a district perspective, misclassification as showing a "significant disproportionality" is worrisome because they must dedicate a certain amount of their Special Education federal funds towards early intervention programs aimed toward students identified as showing higher risk of receiving special education services. This diversion of funds towards interventions, while potentially beneficial for the identified groups, might take away valuable resources intended for improving general Special Education services in their school.[5]

Students might face even higher costs in districts that are misidentified as exhibiting a "significant disproportionality." Based on the findings by Enayati (2014), the main concern is that student subgroups may be removed from Special Education services when they benefit from being provided such services. The effectiveness of Special Education programs has not been resolved in the literature (Theobald, 2015). In one study, using student-fixed effects to exploit variation across time in special education placement, Hanushek, Kain, and Rivkin (2002) find that being enrolled in special education leads to a 0.1 standard deviation increase in average math scores. In another study, however, Morgan, Frisco, Farkas, and Hibel (2008) use propensity score matching (using a rich set of covariates from a nationally representative longitudinal study) to estimate the effects of special education. They find insignificant effects

---

special education services. African-Americans in the district were less likely to receive special education services than most subgroups in large urban school districts outside of Texas (Rosenthal, 2016).

[5]While districts used to be required to direct the 15% of Part B funds away from Special Education and fully towards early-intervention services for students without disabilities, OSERS, in their recent rule change, has increased the flexibility of the use of these federal funds to include students with disabilities and those of preschool age. However, the use of the funds is still restricted. The money is required to be placed towards coordinated early-intervention services, but could be better used elsewhere to improve the Special Education program within the district if the district is misidentified as showing a "significant disproportionality."

of special education services on student achievement measures on math test scores. They also find, in some models, negative effects of special education on reading scores. In a third paper, Theobald (2015) takes advantage of a special education funding mechanism in Washington that reduces the probability of a student receiving special education services once a school reaches a certain number of students already in special education. The author finds positive, yet insignificant, effects on math test scores for students enrolled in the program. He also finds negative, yet insignificant, effects on reading test scores. None of the studies look at differential effects across race. The results found in these papers might say more about the variability in the effectiveness of different types of special education programs. Alternatively, null findings might be a result of identification strategies used which look at effects on students at the margin (i.e., those students that are deemed eligible for removal from special education services). It is likely that the effects of special education might not be as strong for these students. While the benefits of special education are unclear in the literature, the two strongest methods for identifying causal impacts, student fixed effects and regression discontinuity, find no significant negative impacts of special education placement and possibly positive effects in some programs (What Works Clearinghouse, 2016). The worry is that misclassifying a district as showing a "significant disproportionality" might induce the district to remove students from potentially beneficial services.

Given that there is evidence that districts respond to these state definitions of "significant disproportionality" by reducing their minority enrollment in special education, the question arises if current measures used to monitor disproportionality are appropriate for identifying problem districts or if they may induce schools to remove minority students from the specialized services that they need and are potentially beneficial. This chapter will seek to answer the following questions:

1. **Are racial minorities still overrepresented in special education in Washington state after conditioning on observable characteristics commonly collected by the State?**

2. **How do current measures of disproportionality in special education compare to a measure that takes into account individual student characteristics commonly collected by the State?**

2.3 Data and Methods

2.3.1 Data

To answer the research questions in this chapter I utilize a panel dataset of Washington K-12 students between the 2009-10 and 2014-15 school years. Washington enrolls more than 1 million students with approximately 14% of students receiving some type of special education service. Table 2.1 provides a summary of the types of students that enrolled in Washington public schools during the 2013-14 school year. The table shows the percentage of students by race, disability, and special program category. For sake of space, I only include demographic information on some of the larger disability categories. The full list of disability categories are: developmental delays, emotional/behavioral disability (EBD), orthopedic impairment, health impairment, specific learning disability (SLD), intellectual disability (ID), multiple disabilities, deafness, hearing impairment, visual impairment, deaf-blindness, communication disorders, autism, and traumatic brain injury. As a whole, Washington is demographically quite similar to the nation as a whole. Washington has a slightly higher population of White, multi-race, and Asian students, while enrolling a smaller proportion of Black and Hispanic students. Any generalization of the results of this chapter should take these differences into account. Given Washington's relatively large student population, even with smaller percentages of minority students, my analysis will still be able to answer questions about minority student placement into special education services.

The panel dataset contains information about all students enrolled in public schools in Washington state during the 2009-10 through 2014-15 school years. This includes demographic information such as race, gender, and primary language. It also includes information about the special programs the student is enrolled in: free/reduced price lunch, ELL ser-

Table 2.1: Student Demographics of Washington and US Public Schools (2013-14)

Variable	WA	United States
Race		
American Indian	1.6%	1.0%
Asian/PI	8.1%	5.2%
Black	4.7%	15.6%
Hispanic	21.4%	24.9%
White	57.5%	50.3%
Two or More	6.7%	3.0%
Special Education	14.0%	12.9%
Dev. Delays	2.2%	0.8%
EBD	0.5%	0.7%
Health Imp.	2.5%	0.2%
SLD	4.6%	4.5%
ID	0.5%	0.9%
Autism	1.1%	1.1%
FRPL	49.2%	51.4%
ELL	9.6%	8.9%

Source: Washington State Office of Superintendent of Public Instruction (OSPI), National Center for Education Statistics (NCES)

vices, and special education programs (including disability category). This information is time-varying and is a point-in-time measure of a student's program enrollment for the current school year. It should be noted that this information is commonly collected by all states.

Prior to the recent rule change that set a new standard for measuring disproportionality, Washington used the risk ratio weighted by population proportions as its measure of disproportionality in special education. The weighted risk ratio (WRR) allows for comparison across districts and can be calculated as follows:

$$\text{WRR}_{id} = \frac{(1 - p_i)R_{id}}{\sum_{j \neq i} p_j R_{jd}} \tag{2.2}$$

where R_{id} is the risk of group i being placed in special education. R_{id} is defined as the proportion of group i in special education in district d. The weighted risk ratio for group i in district d (WRR_{id}) shows the risk of group i receiving special education compared to the risk of all other students being placed there in a district (weighted by p_i, the percentage of subgroup i in the state population).

Specifically, Washington defines a "significant disproportionality" in a particular district as "a weighted risk ratio of 4.0 or greater for three consecutive years for any racial/ethnic group."[6] Given that the threshold for determining whether a district has a "significant disproportionality" is relatively high and the definition requires this threshold to be exceeded for at least three consecutive years, it is not surprising that Washington identified no districts as showing a "significant disproportionality" in the past five years (Government Accountability Office, 2013).

Even if few districts in Washington are being identified as showing a "significant disproportionality", they are still monitored annually as part of the state's performance plan (SPP). Among these measures, disproportionality for each school district is reported as well

[6]It is common for states to use multiple years of data when determining districts showing a "significant disproportionality" to account for the unreliability of the estimates when student populations are small. In fact, 23 states, prior to the rule change, use multiple years of data and 13 require the threshold to be exceeded for three straight years (Department of Education, 2016a).

Figure 2.2: Washington State Special Education Disproportionality Categories

Weighted Risk Ratio (WRR)

≤0.5	>0.5 to <0.67	0.67 to 1.5	>1.5 to <2.0	≥2.0 to <3.0 (3 consecutive years)	≥3.0 to <4.0 (3 consecutive years)	≥4.0 (3 consecutive years)
Disproportionate Under-representation	At Risk for Disproportionate Under-representation	No Disproportionate Representation	At Risk for Disproportionate Over-representation	Disproportionate Over-representation*	At Risk for Significant Disproportionality	Significant Disproportionality*

*Note: The results of the calculations will be verified using multiple methods.

Source: WA state OSPI Website

(https://web.archive.org/web/20160731173756/http://www.k12.wa.us/SpecialEd/ProgramReview/Disproportionality.aspx)

as an indicator for whether the WRR exceeds 2.0. While Part B funds are not funneled away for early intervening strategies for being in any category other than "significant disproportionality", the fact that districts are still categorized and this information is publicly released suggests that "significant disproportionality" might not be the only category of disproportionality that matters. Figure 2.2 shows the multiple categories of disproportionality that districts can be contained in.

OSERS' new standard methodology prohibits the use of the weighted risk ratio for the purposes of federal monitoring because it believes the "method fails to provide LEAs and the public with a transparent comparison" (Department of Education, 2016a). While it is true that the WRR is more difficult to interpret, I believe the benefits of the measure outweigh its costs. Bollmer, Bethel, Munk, and Bitterman (2011) provides a nice example of how the calculation of the risk ratio can be different despite there being the same risk across districts if district demographics are different. Imagine that the risk of identification into special education is the same across two districts: 2% for White students, 1% for Hispanic students, and 5% for Black students. If one district with 1000 students has a majority of White students (800) with the rest of the school evenly split between Hispanic and Black students (100 for each group), then the risk ratio for White students compared to other

student groups would be:

$$\begin{aligned}\text{Risk Ratio} =& \frac{\text{Risk of White students}}{\text{Risk of all other student groups}} \\ =& \frac{16/800}{(1+5)/200} \\ =& 0.67\end{aligned}$$

However, in the other district, of the same size and risk of identification, but with a different composition (800 Hispanic, 100 White, and 100 Black) the risk ratio would be calculated as $\frac{2/100}{(8+5)/900} = 1.38$. Now in two districts with the same risk of special education placement, we would come to two different conclusions about their disproportionality using the risk ratio. One district would be seen as having an underrepresentation of White students in special education, while the other would be seen as exhibiting overrepresentation. The use of student population weights eliminates this problem. Now say this state had the following student population: 70% White, 10% Hispanic, and 20% Black students. The weighted risk ratio for *both* schools would be:

$$\frac{(1-0.7)0.02}{(0.1)(0.01)+(0.2)(0.05)} = 0.55$$

The composition of each school no longer is a factor in the calculation. The weighted risk ratio can remove the influence of different racial demographics across districts and provides a measure that is comparable over all districts in the state.

2.3.2 Methods

To answer my first research question, I run the following logistic regression for all students in Washington state:

$$\text{SPED}_{igd} = I(\alpha + \beta\text{RACE}_i + \gamma X_{igd} + \varepsilon_{igd}) \tag{2.3}$$

where SPED_{igd} is a binary indicator for whether student i was offered special education services in grade g enrolled at district d. RACE_i is a categorical variable indicating student i's race/ethnicity. X_{igd} is a vector of student-level control variables (free/reduced price lunch eligibility, primary language, ELL status, gender, homeless status).[7] ε_{igd} is a random error assumed to follow a standard logistic distribution, thus, making this a logistic regression model. The parameters estimated from the logit model can be used to construct conditional probabilities (or risks) of special education placement. The main test is to see whether the coefficients change between the model that includes the X_{igdt} control variables and a model that omits this.

To answer the second research question, I will use the estimated parameters from Equation 2.3 to construct an alternative measure of disproportionality of students in special education that takes into account student-level characteristics that are commonly collected by school districts. From the parameters I obtain an estimate of the conditional probability of special education placement as follows:

$$\Pr\left(\text{SPED}_{igd} = 1 | \text{RACE}_i = j, X_{igd}\right) = \frac{1}{1 + \exp[-(\hat{\beta}_j + \hat{\gamma} X_{igd})]} \tag{2.4}$$

Here, β_j is the estimated coefficient for when $\text{RACE} = j$. The conditional probability can be thought of as the risk of student i being placed into special education based on student i's observable characteristics. Thinking of the conditional probability as a "risk" can help move on to the next step in estimating a risk ratio. Kleinman and Norton (2009) develop a

[7] These are student-characteristics collected by most states in the nation. However, there are two notable omissions: student achievement and school-level aggregates. Student achievement measures are directly linked with Special Education placement as they are used as a part of the evaluation process to determine eligibility. However, most states do not collect annual data on these measures for every grade. For federal accountability purposes, states are only required to annually assess students in grades 3-8 and once during high school. This measure would only be available for these students meaning we would only be able to calculate measures of disproportionality for students in these grade ranges which ignores a large segment of the population. Second, I do not include school-level factors which would compare students in similar academic environments. However, given that I want to produce a measure that captures school-level processes that might be misidentifying students, I do not want to control for measures that might capture that. Therefore, I include only student-level covariates that capture individual student risk and omit school-level factors that might measure institutional factors determining Special Education placement.

way to estimate an adjusted risk ratio (ARR) that takes into account the full characteristics of the population of students. Once again, using the parameters estimated from the logit model, I obtain the regression adjusted risk ratio (ARR_{jkd}) of receiving special education services for a given race/ethnicity j relative to the baseline group k in grade g at district d (of enrollment size n_d) using the following function:

$$\begin{aligned}\widehat{\text{ARR}}_{jkgd} =& \frac{\frac{1}{n_d}\sum_{i=1}^{n_d}\Pr(\text{SPED}_{igd}=1|\text{RACE}=j,X_{igd})}{\frac{1}{n_d}\sum_{i=1}^{n_d}\Pr(\text{SPED}_{igd}=1|\text{RACE}=k,X_{igd})} \\ =& \frac{\sum_{i=1}^{n_d}\frac{1}{1+\exp[-(\hat{\beta}_j+\hat{\gamma}X_{igd})]}}{\sum_{i=1}^{n_d}\frac{1}{1+\exp[-(\hat{\beta}_k+\hat{\gamma}X_{igd})]}}\end{aligned} \tag{2.5}$$

This formula is essentially the ratio of the average predicted probabilities across all students conditional on being of race j relative to being part of group k. It takes into account the full characteristics of the population of students at district d. A further advantage of the ARR approach is that standard errors and confidence intervals of such measures can be obtained using the delta-method.[8].

To produce a measure that is directly analogous with the WRR, I make a modification to Equation 2.5 and add population proportion weights to create the weighted adjusted risk ratio (WARR).

$$\widehat{\text{WARR}}_{jkgd} = \frac{(1-p_j)\sum_{i=1}^{n_d}R_{ijgd}}{\sum_{k\neq j}p_k\left[\sum_{i=1}^{n_d}R_{ikgd}\right]} \tag{2.6}$$

In Equation 2.6, $R_{ijgd} = \Pr(\text{SPED}_{igd}=1|\text{RACE}=j,X_{igd})$. That is, R_{ijgd} is equal to the estimated conditional probability of student i being placed in special education if they were race j and had characteristics X_{igd}. p_j is the state student population proportion of race j. I propose WARR as an alternative measure of disproportionality at the district-level. I assess how this measure compares with the common WRR previously used in Washington (and 24 other states).

[8]A user-written function in Stata command `adjrr` has been created to compute such measures (Norton, Miller, & Kleinman, 2013)

2.4 Results

I start by presenting the estimated odds ratios from the logistic regression presented in Equation 2.3 in Table 2.2. I ran the models separately for each year of data. For sake of space, I only present the results for the 2014-15 school year. Results from other years can be found in appendix tables. The findings are descriptively similar in different years. Across the three panels of Table 2.2, I present pairs of columns for each major disability category along with a general indicator variable for if they received special education services for any disability category.[9] For each pair of columns, I present odds ratios from a model where I do not include any control variables (on the left) and odds ratios from models where I do include additional student-level control variables (the right). All standard errors have been clustered at the district-level.

In general across all models it appears that the estimated odds ratios on race are sensitive to the inclusion of other student characteristics. For example, focusing on general placement into special education services (i.e., they are receiving services for *any* disability category), I find that American Indian, Black, and Hispanic students are significantly overrepresented in special education compared to their White peers. However, once control variables are included, the measure of overrepresentation (in comparison with White students) is greatly reduced for both Black and American Indian students. The measures still indicate overrepresentation, but the magnitude is not as large. So once indicators of possible risk factors are taken into account the overrepresentation problem does not seem as large (but it still remains). A similar reduction can be observed for Black and American Indian students for disability categories in which they are normally found to be overrepresented (e.g., EBD, SLD, intellectual disabilities). Again, it should be noted that these are findings including only those variables commonly collected on all students in the state. Better controls, that are

[9] I omit results for the following disability categories: developmental delays, orthopedic impairment, multiple disabilities, deafness, hearing impairment, visual impairment, deaf-blindness, and traumatic brain injury. These categories are relatively less common than the other disability groups. In Washington's performance indicators, they also remove these categories from the calculation of overrepresentation in their school districts.

Table 2.2: Logistic Regression Results for Special Education Placement (2014-15)

Panel A: Special Education, Emotional/Behavioral Disability, Health Impairment

	SPED		EBD		Health Imp.	
	No Controls	Controls	No Controls	Controls	No Controls	Controls
American Indian	1.663***	1.377***	1.838***	1.238*	1.109	0.909
	(0.074)	(0.063)	(0.200)	(0.136)	(0.079)	(0.068)
Asian	0.510***	0.628***	0.239***	0.417***	0.297***	0.450***
	(0.029)	(0.027)	(0.032)	(0.054)	(0.016)	(0.024)
Black	1.338***	1.151***	2.062***	1.564***	1.167***	1.071*
	(0.037)	(0.032)	(0.240)	(0.199)	(0.051)	(0.038)
Hispanic	1.090***	1.084***	0.558***	0.662***	0.615***	0.785***
	(0.026)	(0.026)	(0.048)	(0.050)	(0.022)	(0.024)
Pacific Islander	0.716***	0.621***	0.435***	0.374***	0.369***	0.372***
	(0.036)	(0.026)	(0.097)	(0.082)	(0.040)	(0.041)
Two or More	1.062**	0.991	1.242***	1.064	0.958	0.893***
	(0.027)	(0.024)	(0.074)	(0.064)	(0.030)	(0.027)
Female		0.488***		0.325***		0.455***
		(0.004)		(0.014)		(0.007)
Non-English		0.535***		0.265***		0.346***
		(0.026)		(0.038)		(0.024)
FRPL		1.745***		3.083***		1.686***
		(0.045)		(0.213)		(0.070)
Homeless		1.192***		2.023***		1.308***
		(0.026)		(0.117)		(0.045)
Migrant		0.750***		0.477***		0.642***
		(0.034)		(0.119)		(0.056)
ELL		1.483***		1.047		1.154*
		(0.078)		(0.145)		(0.099)
N	1,150,081	1,150,081	1,150,081	1,150,081	1,150,081	1,150,081

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$
Odds ratio estimates are presented
Standard errors clustered at the district-level.

Panel B: Specific Learning Disability, Intellectual Disability, Communication Disorder

	SLD		Int. Dis.		Comm. Dis.	
	No Controls	Controls	No Controls	Controls	No Controls	Controls
American Indian	2.314***	1.730***	2.694***	1.933***	1.224***	1.155**
	(0.135)	(0.105)	(0.336)	(0.215)	(0.077)	(0.073)
Asian	0.450***	0.488***	0.782**	0.881	0.610***	0.725***
	(0.048)	(0.036)	(0.075)	(0.071)	(0.047)	(0.050)
Black	1.798***	1.306***	1.894***	1.404***	0.691***	0.667***
	(0.073)	(0.042)	(0.113)	(0.072)	(0.033)	(0.029)
Hispanic	1.675***	1.230***	1.280***	1.006	0.975	1.040
	(0.070)	(0.049)	(0.065)	(0.054)	(0.034)	(0.027)
Pacific Islander	1.164**	0.795***	1.099	0.818	0.617***	0.599***
	(0.082)	(0.054)	(0.146)	(0.105)	(0.087)	(0.081)
Two or More	1.079**	0.963	0.856**	0.755***	1.060	1.041
	(0.038)	(0.030)	(0.063)	(0.054)	(0.050)	(0.048)
Female		0.704***		0.786***		0.578***
		(0.007)		(0.020)		(0.009)
Non-English		0.563***		0.863*		0.466***
		(0.034)		(0.074)		(0.029)
FRPL		2.429***		2.659***		1.190***
		(0.054)		(0.112)		(0.042)
Homeless		1.279***		1.249***		0.935
		(0.033)		(0.095)		(0.055)
Migrant		0.862**		1.239*		0.725***
		(0.052)		(0.142)		(0.058)
ELL		2.297***		0.818		2.286***
		(0.158)		(0.119)		(0.107)
N	1,150,081	1,150,081	1,150,081	1,150,081	1,150,081	1,150,081

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$
Odds ratio estimates are presented
Standard errors clustered at the district-level.

Panel C: Autism

	Autism	
	No Controls	Controls
American Indian	0.561***	0.547***
	(0.059)	(0.059)
Asian	0.984	1.412***
	(0.050)	(0.056)
Black	0.857**	0.936
	(0.061)	(0.070)
Hispanic	0.524***	0.768***
	(0.025)	(0.031)
Pacific Islander	0.405***	0.493***
	(0.069)	(0.084)
Two or More	0.909***	0.908**
	(0.034)	(0.036)
Female		0.207***
		(0.005)
Non-English		0.463***
		(0.046)
FRPL		1.091*
		(0.050)
Homeless		0.679***
		(0.051)
Migrant		0.528***
		(0.095)
ELL		0.785*
		(0.108)
N	1,150,081	1,150,081

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$
Odds ratio estimates are presented
Standard errors clustered at the district-level.

not commonly available to the state, such as parental income, might completely eliminate the apparent overrepresentation for some groups.

The results from Table 2.2 indicate that there would be a problem using current unconditional measures of disproportionality that do not account for other student-characteristics. However, the extent of the differences is unclear just by looking at the odds ratios. In Table 2.3, I present estimated RR and ARR for the 2009-10 through 2014-15 school years based on the parameter estimates from the logit models.[10] Note the difference in interpretation of these risk ratios compared to the odds ratios. The odds ratios compare the odds of being placed in special education compared to the baseline group (White students), whereas the risk ratio compares the probability of being placed in special education and compares it to *all other* groups of students. While the risk ratio is not equal to the odds ratio, when risk levels are relatively small – as is the case in Special Education placement – the odds ratio is approximately equal to the risk ratio. The notable change across the two sets of results is the difference in comparison groups. I present these for the general indicator for special education and only for a couple of the larger disability categories: health impairment and specific learning disability. The top panel presents the WRR which does not condition on other student-characteristics. The bottom panel shows the WARR which does adjust for student-level characteristics. I include estimated standard errors (based on the delta method) in parentheses to the right of each measure.

The first thing to notice about the numbers in Table 2.3 is that they stay roughly consistent over time. There appear to be no huge changes in the measures of disproportionality over time. While this may suggest a high reliability, remember these measures are commonly taken on much smaller populations of students at the district-level. However, a comparison of the top and bottom panels reveals that there are some differences between the two measures. The unconditional measures appear to be larger for American Indian and Black students. In

[10]Note, that I do not apply weights to the risk ratio in this exercise given that they are already being calculated at the state-level. Applying state population weights at the state-level would generate the same value as the unweighted risk ratio.

Table 2.3: Selected RR/ARR For Whole State by Disability Category

Panel A: RR

	2009-10	2010-11	2011-12	2012-13	2013-14	2014-15
Special Education						
American Indian	1.44 (0.04)	1.48 (0.05)	1.51 (0.05)	1.48 (0.05)	1.43 (0.07)	1.52 (0.05)
Asian	0.54 (0.02)	0.54 (0.02)	0.54 (0.02)	0.55 (0.02)	0.54 (0.03)	0.53 (0.03)
Black	1.27 (0.06)	1.32 (0.03)	1.32 (0.02)	1.30 (0.03)	1.30 (0.03)	1.29 (0.03)
Hispanic	1.06 (0.02)	1.07 (0.02)	1.08 (0.02)	1.09 (0.02)	1.10 (0.02)	1.09 (0.02)
White	1.00 (0.02)	1.00 (0.02)	0.99 (0.02)	0.98 (0.02)	0.98 (0.02)	0.99 (0.02)
Health Impairment						
American Indian	1.31 (0.08)	1.33 (0.10)	1.30 (0.10)	1.27 (0.09)	1.21 (0.10)	1.28 (0.09)
Asian	0.35 (0.02)	0.34 (0.02)	0.33 (0.02)	0.35 (0.02)	0.35 (0.02)	0.33 (0.02)
Black	1.36 (0.07)	1.36 (0.04)	1.40 (0.04)	1.38 (0.05)	1.36 (0.05)	1.36 (0.06)
Hispanic	0.55 (0.03)	0.60 (0.02)	0.61 (0.02)	0.63 (0.02)	0.65 (0.02)	0.67 (0.02)
White	1.56 (0.06)	1.50 (0.04)	1.48 (0.04)	1.46 (0.04)	1.45 (0.04)	1.44 (0.03)
Specific Learning Disability						
American Indian	1.75 (0.06)	1.80 (0.07)	1.84 (0.08)	1.83 (0.08)	1.79 (0.11)	1.92 (0.09)
Asian	0.45 (0.04)	0.42 (0.03)	0.42 (0.04)	0.41 (0.04)	0.40 (0.04)	0.38 (0.04)
Black	1.54 (0.11)	1.62 (0.06)	1.60 (0.05)	1.59 (0.06)	1.56 (0.05)	1.54 (0.05)
Hispanic	1.50 (0.06)	1.50 (0.06)	1.52 (0.06)	1.55 (0.05)	1.59 (0.05)	1.59 (0.06)
White	0.77 (0.03)	0.76 (0.03)	0.75 (0.03)	0.74 (0.03)	0.73 (0.03)	0.73 (0.02)

Panel B: ARR

	2009-10	2010-11	2011-12	2012-13	2013-14	2014-15
Special Education						
American Indian	1.25 (0.03)	1.27 (0.04)	1.30 (0.04)	1.28 (0.04)	1.22 (0.06)	1.31 (0.04)
Asian	0.65 (0.02)	0.65 (0.02)	0.65 (0.02)	0.66 (0.02)	0.67 (0.03)	0.65 (0.02)
Black	1.12 (0.06)	1.16 (0.02)	1.16 (0.02)	1.15 (0.04)	1.14 (0.02)	1.13 (0.02)
Hispanic	1.09 (0.03)	1.07 (0.03)	1.06 (0.03)	1.05 (0.02)	1.08 (0.02)	1.09 (0.02)
White	1.01 (0.03)	1.03 (0.02)	1.02 (0.02)	1.02 (0.02)	1.01 (0.02)	1.01 (0.02)
Health Impairment						
American Indian	1.06 (0.06)	1.05 (0.08)	1.04 (0.08)	1.01 (0.08)	0.95 (0.08)	1.01 (0.07)
Asian	0.48 (0.03)	0.49 (0.03)	0.47 (0.03)	0.49 (0.03)	0.52 (0.03)	0.49 (0.03)
Black	1.18 (0.08)	1.18 (0.04)	1.22 (0.05)	1.20 (0.06)	1.18 (0.04)	1.19 (0.04)
Hispanic	0.77 (0.03)	0.78 (0.03)	0.78 (0.03)	0.77 (0.03)	0.82 (0.03)	0.85 (0.02)
White	1.38 (0.05)	1.34 (0.05)	1.34 (0.05)	1.35 (0.05)	1.30 (0.03)	1.29 (0.03)
Specific Learning Disability						
American Indian	1.49 (0.05)	1.51 (0.06)	1.55 (0.06)	1.55 (0.07)	1.51 (0.09)	1.63 (0.09)
Asian	0.53 (0.03)	0.49 (0.03)	0.50 (0.03)	0.49 (0.03)	0.49 (0.03)	0.47 (0.03)
Black	1.29 (0.09)	1.34 (0.04)	1.33 (0.04)	1.32 (0.05)	1.29 (0.04)	1.26 (0.04)
Hispanic	1.22 (0.05)	1.19 (0.05)	1.20 (0.05)	1.18 (0.04)	1.23 (0.04)	1.24 (0.04)
White	0.94 (0.04)	0.95 (0.03)	0.95 (0.03)	0.95 (0.03)	0.94 (0.03)	0.93 (0.03)

fact, it appears that measures of overrepresentation tend to be inflated when not controlling for any control variables and measures of underrepresentation tends to be lower than the conditional measures. However, when thinking about the disproportionality categories for Washington, the differences between the two measures do not appear to be too meaningful. For example, in 2014-15 Black students statewide had a RR of 1.29 which would have put them in the category of "No Disproportionate Representation." The ARR for Black students is significantly lower than the RR at 1.13, but this still places them into the same category of "No Disproportionate Representation." So while there are observable changes in the two measures, consistent with the findings in many other studies, the implications of using the RR rather than the ARR, at least at the state-level, does not appear to have such a large implication as far as the way states categorize disproportionate representation in Special Education. There do appear to be some indications of category switching though (although it is not significant). In 2014-15, American Indian students had a RR of 1.52 placing them in the "At Risk for Disproportionate Overrepresentation" category, but the ARR places them in the "No Disproportionate Representation" category (ARR = 1.31). While this isn't much of a meaningful difference it might cause concern when these measures are applied at the district-level (which is done in practice) where there might be a lot more variation in the results.

Next, I calculate the WARR and WRR for each individual school district in the state. As I mention above, there is a lot more variation in the measures and some cannot be reliably calculated when sample sizes are small for certain groups in categories. That is to say, while I run the analysis for all (approximately) 4,000 district-ethnic group combinations in the state, I cannot actually calculate the WRR or WARR for some of these districts for certain categories of disabilities.[11] Similar to the logistic regression, I run this separately for each year of data, but for sake of space only present the findings from the 2014-15 school year

[11]I set a requirement (same as Washington) of a minimum cell size of 10 before I calculate their WARR or WRR. This is a common practice in many states who do not calculate risk ratios for districts that have too small a number of students in a particular disability category. States vary in their determination of the minimum cell size.

since the findings and conclusions are similar across the different school years. In Figures 2.3 – 2.5, I present scatterplot comparisons of the district-level WARR and WRR measures for general special education placement (Figure 2.3) and the two disability categories: intellectual disability (Figure 2.4) and specific learning disability (Figure 2.5). The dashed lines on each plot represent the 45° line.

The scatterplots show that the differences in the measures are not too large as they all tend to hug the 45° line. However, it appears that once again, Black and American Indian groups receive larger measures of overrepresentation using the unconditional WRR measure as most of their points lie below the line. It also appears that Hispanic students receive larger measures with the WRR for the SLD category. Conversely, Asian and White students tend to see their points lying above the line indicating that the WRR gives them lower values of disproportionality than the conditional WARR. But as I mention previously, these departures do not seem to be large enough to be meaningful, at least under Washington's definition of disproportionality.

I next present *confusion matrices* to show the extent of agreement between the classification patterns of two measures (see Grimmer and Stewart (2013) for an example of a confusion matrix). I present these matrices in Tables 2.4-2.5. In these confusion matrices, the rows show the classification of the WARR measure and the columns show the classification of districts under the WRR measure. I limit the presentation to only the general special education placement indicator (Table 2.4) and the specific learning disability indicator (Table 2.5). I also only present results for Black, Hispanic, and White students for comparison as they are generally larger groups and contain more district measures of disproportionality.

The confusion matrices show that there is general agreement between the two measures in terms of categorizing school districts into disproportionality categories. This can be observed by looking at the diagonal of the matrices. Any disagreements seem to be off by one or two categories, which from a state (and perhaps, Federal) perspective is not too worrisome. It appears that using the WARR as a measure of disproportionality in special education would not lead to too many differing conclusions as far as identifying districts into broad categories

Figure 2.3: Visual Comparison of District-level WARR and WRR Measures: General Special Education Placement (2014-15)

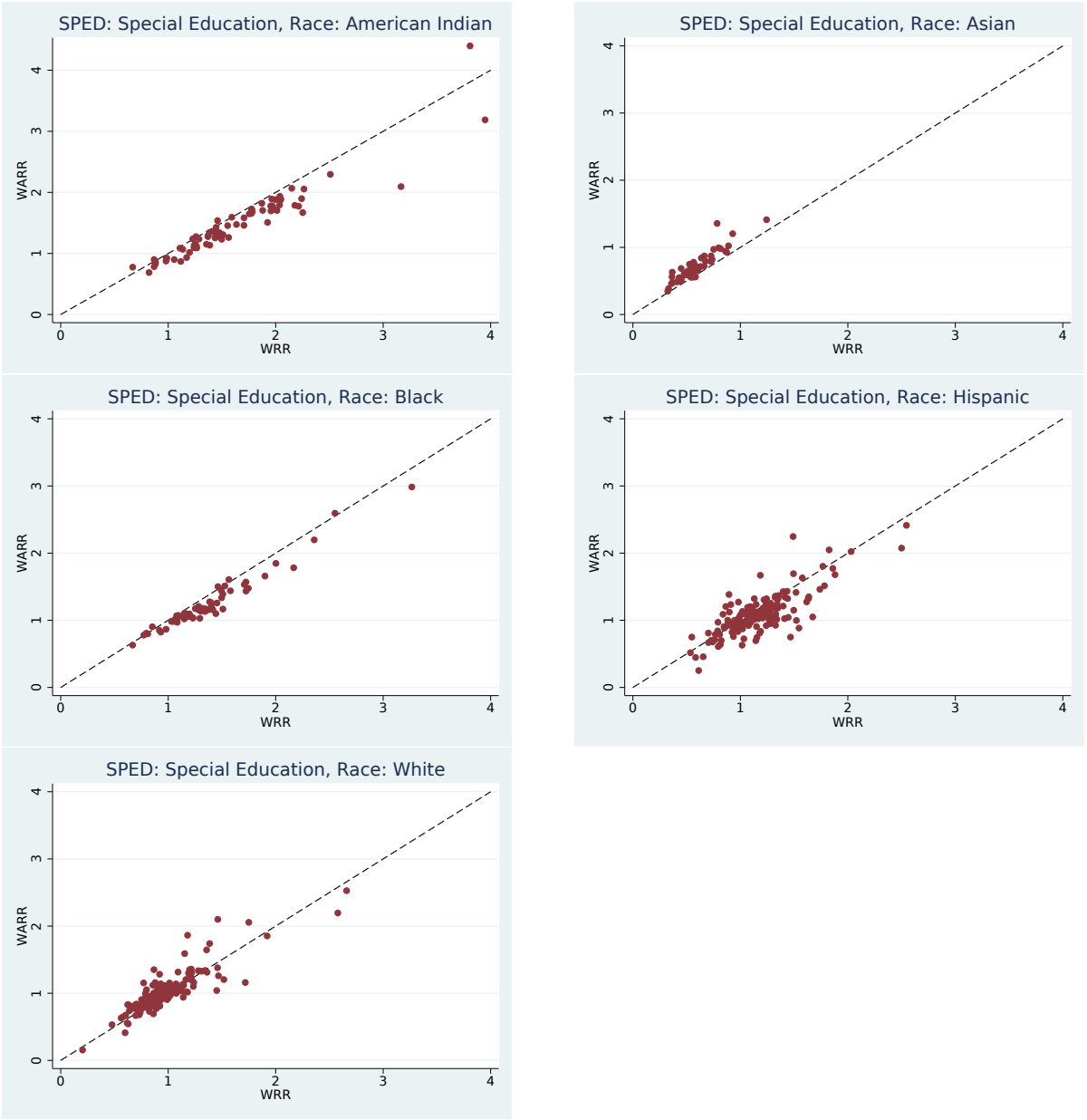


Figure 2.4: Visual Comparison of District-level WARR and WRR Measures: Intellectual Disability

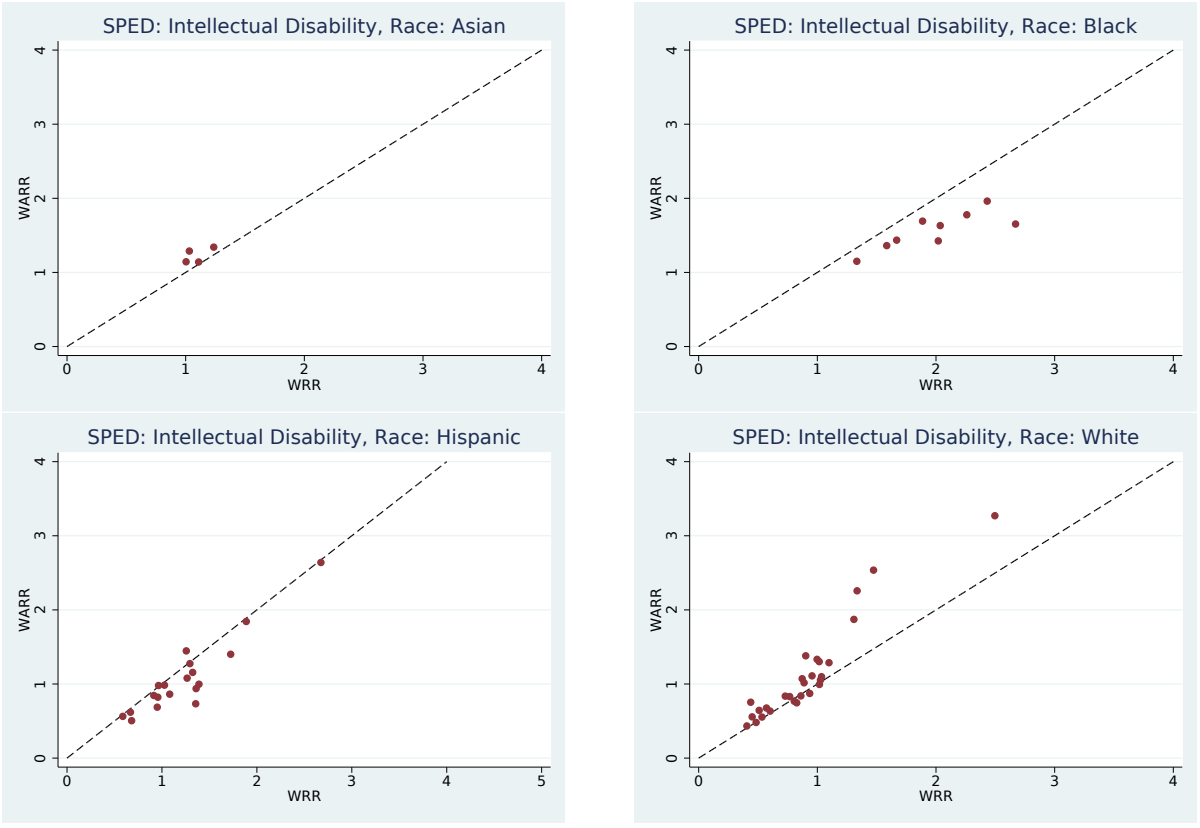


Figure 2.5: Visual Comparison of District-level WARR and WRR Measures: Specific Learning Disability

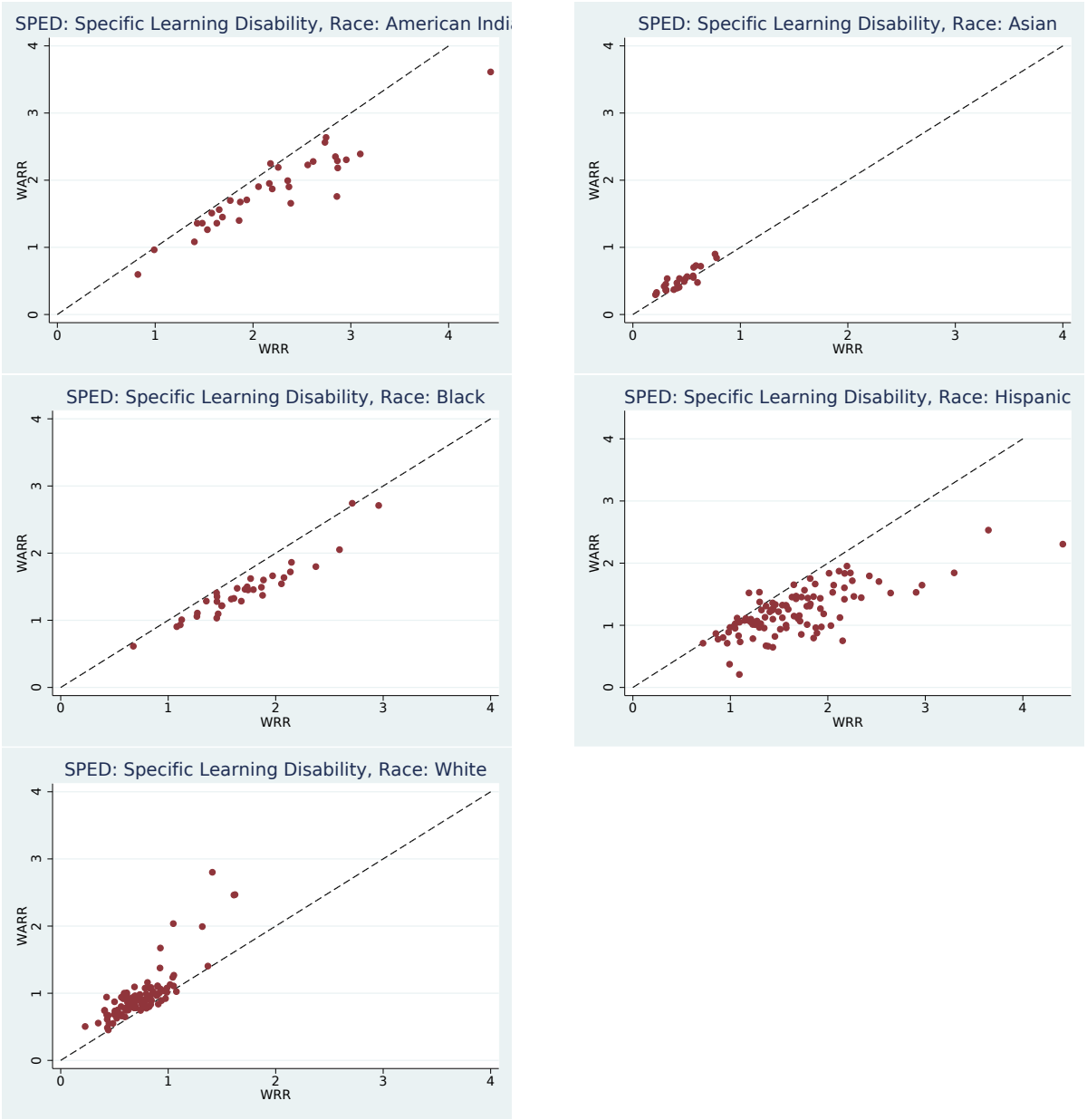


Table 2.4: Confusion Matrix for General Special Education Placement: WARR (Rows) vs WRR (Columns)

Black							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Under (1)	0	0	0	0	0	0	0
At Risk Under (2)	0	1	0	0	0	0	0
No Disprop (3)	0	0	47	5	0	0	0
At Risk Over (4)	0	0	1	5	2	0	0
Over (5)	0	0	0	0	2	1	0
At Risk Significant (6)	0	0	0	0	0	0	0
Significant (7)	0	0	0	0	0	0	0
Hispanic							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Under (1)	0	3	0	0	0	0	0
At Risk Under (2)	0	1	4	0	0	0	0
No Disprop (3)	0	1	136	9	0	0	0
At Risk Over (4)	0	0	2	5	0	0	0
Over (5)	0	0	2	1	3	0	0
At Risk Significant (6)	0	0	0	0	0	0	0
Significant (7)	0	0	0	0	0	0	0
White							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Under (1)	1	1	0	0	0	0	0
At Risk Under (2)	1	4	1	0	0	0	0
No Disprop (3)	0	5	155	2	0	0	0
At Risk Over (4)	0	0	4	1	0	0	0
Over (5)	0	0	1	1	2	0	0
At Risk Significant (6)	0	0	0	0	0	0	0
Significant (7)	0	0	0	0	0	0	0

Table 2.5: Confusion Matrix for Specific Learning Disability: WARR (Rows) vs WRR (Columns)

Black							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Under (1)	0	0	0	0	0	0	0
At Risk Under (2)	0	1	0	0	0	0	0
No Disprop (3)	0	0	12	11	0	0	0
At Risk Over (4)	0	0	0	3	5	0	0
Over (5)	0	0	0	0	3	0	0
At Risk Significant (6)	0	0	0	0	0	0	0
Significant (7)	0	0	0	0	0	0	0
Hispanic							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Under (1)	0	0	2	0	0	0	0
At Risk Under (2)	0	0	2	0	0	0	0
No Disprop (3)	0	0	40	30	6	0	0
At Risk Over (4)	0	0	2	4	14	0	0
Over (5)	0	0	0	0	0	1	1
At Risk Significant (6)	0	0	0	0	0	0	0
Significant (7)	0	0	0	0	0	0	0
White							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Under (1)	2	0	0	0	0	0	0
At Risk Under (2)	5	4	1	0	0	0	0
No Disprop (3)	5	28	55	0	0	0	0
At Risk Over (4)	0	0	2	0	0	0	0
Over (5)	0	0	2	2	0	0	0
At Risk Significant (6)	0	0	0	0	0	0	0
Significant (7)	0	0	0	0	0	0	0

of disproportionality. However, results comparing the measures indicates that there do exist differences between the two ways of assessing disproportionality and these differences could be important for some districts offering special education services to their students.

2.5 Discussion

The results provide evidence that the current RR measure of disproportionality tends to be skewed due to omitted variable bias. However, using information commonly collected by the state on public school students to "adjust" the risk ratio does not appear to alter the conclusions about disproportionality in districts. That is, in the vast majority of cases, a district that shows overrepresentation under the risk ratio continues to show evidence of overrepresentation under the adjusted risk ratio. This would seem to indicate that there is not much to be gained from this exercise. However, I argue that my findings have important policy implications regarding the monitoring of districts for disproportionality.

First, my findings do *not* indicate that the unadjusted risk ratio is a valid measure of disproportionality. I find that the omission of student-level covariates tends to overstate the magnitude of overrepresentation or underrepresentation. The fact that the measure that adjusts for student-level characteristics does not change the disproportionality category that a district falls in could suggest that the risk ratio can continue to be used for monitoring purposes given that the, arguably, improved measure makes no difference in the categorization of districts into disproportionality levels. However, one could interpret my results a different way. Given that my findings provide contradictory evidence to a number of other studies using covariate adjustment models (that student-covariates completely explain the overrepresentation of minorities in Special Education), I must look into the differences between my study and others. The most obvious difference is the set of covariates included in my models. Other studies use information from nationally representative longitudinal datasets which have detailed information on student, family, and school-factors that are theoretically linked to needing Special Education services. I am limited to only what a state commonly collects on its public school students which lacks the depth of these nationwide surveys.

To test whether the set of controls makes a difference in the findings of this study, I construct my measure using the rich set of covariates included in the Early Childhood Longitudinal Study (ECLS-K).[12] In contrast to Morgan et al. (2015), I calculate the risk ratio only for the first wave of the study, rather than treating the data as longitudinal. This way it will be as if I were just calculating these risk ratios for the first cohort during their kindergarten class. There are some obvious comparability issues with this part of the analysis, but I believe it will be sufficient to make the point. Table 2.6 shows the comparison between the risk ratios calculated across three different models: (1) the model with no controls, (2) the model with Morgan et al. (2015) controls, and (3) the model with control variables commonly collected by the state.[13] The numbers appear to indicate that the control variables available to the state are sufficient to explain a high percentage of the bias in the risk ratio. That is, it seems as though the state available controls have similar explanatory power for general special education placement as the relatively richer set of information collected in the national survey. For three of the race groups (Black, Hispanic, and White), I find that the "bias" (defined as the difference between the no control and full control risk ratio) is nearly fully accounted for using state available information. For the *other* race grouping, I find that only 10% of the bias is explained using state-available controls. However, the bias for this category is much smaller compared to the others. This exercise shows that state available controls can capture a substantial portion of the bias in the risk ratio meaning that the WARR could provide a valid measure of overidentification.[14]

Second, one might conclude from my study that the risk ratio and adjusted risk ratio are essentially the same in the spectrum of identifying schools with overrepresentation. I argue

---

[12]I use the publicly available ECLS-K dataset for this part of the analysis. Most of the information used in Morgan et al. (2015)'s study is available with the notable exception of disability category. For the purposes of this exercise, I include only comparisons using the general special education placement category.

[13]I use the equivalent of the control variables included in the models in this chapter. There are two exceptions: (1) homeless status which is suppressed in the public dataset and (2) migrant education status which had no direct analog.

[14]I note that this assumes that the Morgan et al. (2015) control variables yield completely unbiased measures of disproportionality.

Table 2.6: Comparison of Risk Ratios Calculated using ECLS-K Data

Race	No Controls		Full Controls		State Controls		% Explained
Black	1.69	(0.36)	1.09	(0.34)	1.15	(0.33)	91%
Hispanic	0.67	(0.18)	0.51	(0.20)	0.54	(0.20)	85%
Other	0.31	(0.11)	0.30	(0.12)	0.31	(0.12)	10%
White	1.06	(0.18)	1.66	(0.47)	1.57	(0.41)	85%

Full controls refers to models that include controls from Morgan et al. (2015).
State controls refers to models that include controls commonly available to the state
% Explained calculates the % of the bias explained by State controls = $\frac{\text{abs(State}-\text{NoCtrl)}}{\text{abs(Full}-\text{NoCtrl)}}$.
Standard errors in parentheses to the right of the risk ratio

that that would be an incorrect conclusion. It is unclear whether the disproportionality categories were defined by the state in a way that accounted for observable distinctions in the experiences at different school districts. For example, was there work done to identify whether a shift in the risk ratio from 1.9 to 2.0 leads to a substantial shift in the likelihood of benefiting from an intervention? If not, I believe one could argue that any "adjusted" risk ratio *statistically significantly* above one signifies evidence of overrepresentation. More empirical work can be done to better define disproportionality categories that can better identify the school districts who would benefit the most from an intervention.

It is possible that the categories were created with an understanding that there is bias in the risk ratio and therefore it would be incorrect to compare the risk ratio and adjusted risk ratio across these predefined categories. I would say that the adjusted risk ratio provides the state with at least some guidance on understanding where they should set their thresholds in defining overrepresentation. Specifically, it will give them some way to estimate the bias in risk ratios and account for that bias in their categorization of disproportionality.

2.6 Conclusion

This chapter uses data from all of Washington's public school students to better understand the disproportionality of minority students in special education. I find that, when conditioning on student-level characteristics collected by the state that could also contribute to the probability of needing to receive special education services (e.g., FRPL eligibility, ELL

status), measures of overrepresentation for certain minority groups (American Indian, Black, and sometimes Hispanic) are reduced suggesting that current measures of overrepresentation are overstating the problem. That is, current risk ratio measures seem to be exaggerating the extent of the overrepresentation because they do not account for the fact that some minority students are exposed more to risk factors that could lead to their placement into special education. This concern has been noted and studied in a number of papers (Hibel et al., 2010; Morgan et al., 2012, 2015; Shifrer et al., 2011). However, this is the first study to use full administrative data from a state to explore the issue and the findings are roughly consistent, except that overrepresentation does not disappear after student-level factors are included as controls.

The second goal of this chapter was to assess the utility of an alternative measure of disproportionality that accounts for these select risk factors. I propose a weighted adjusted risk ratio – "adjusted" for covariates – to estimate district-level disproportionality. I compare this measure to the analogous weighted risk ratio – not adjusted for covariates – in its ability to detect schools showing disproportionality in special education. While it is not surprising that the two measures differ based on the findings summarized above, the differences might not be meaningful for states monitoring disproportionality. Using disproportionality categories defined by Washington, I find that the WARR provides almost identical classifications as the WRR. So although the WARR can provide an, arguably, better measure of disproportionality in special education, it does not provide estimates that are too far off from the WRR such that states would have to refocus their attention on a different set of districts based on the definitions of overrepresentation in one state. An exercise conducted using ECLS-K data provides evidence that the WARR can reduce a substantial proportion of the bias in the risk ratio caused by omitted variable bias. So this does not suggest that the lack of drastic changes using WARR is due to the relative lack of control variables available to the state.

The results of this chapter gives one clear message: the risk ratio provides a biased measure of disproportionality. What the states do with this information as they plan out new ways to monitor districts for overrepresentation is up to them. I argue that they can

take one of two actions. First, they can replace the risk ratio with the adjusted risk ratio, taking into account student-level characteristics that they are required to collect each year. This would assure that they do not use inflated measures to identify schools with over/underrepresentation problems. The main purpose of this would be to prevent false positives from occurring that might induce districts to take students out of Special Education when it would not benefit these students to do so. However, this path comes at a cost to states. The risk ratios are relatively simple to calculate, while the adjusted risk ratios could be too data intensive to calculate annually. With this cost in mind, I would argue that states could use the information about the bias in the risk ratio from this study to set their thresholds determining significant disproportionalities. Regardless of the choice they make, it is clear there are some issues in the ways in which states monitor disproportionality in Special Education.

APPENDIX

Table 2.A1: Logistic Regression Results for Special Education Placement (2009-10 through 2011-12)

	2009-10		2010-11		2011-12	
	No Controls	Controls	No Controls	Controls	No Controls	Controls
American Indian	1.530***	1.292***	1.588***	1.312***	1.638***	1.356***
	(0.052)	(0.043)	(0.063)	(0.053)	(0.069)	(0.057)
Asian	0.529***	0.622***	0.521***	0.614***	0.527***	0.624***
	(0.022)	(0.023)	(0.020)	(0.022)	(0.025)	(0.025)
Black	1.306***	1.129*	1.369***	1.175***	1.380***	1.180***
	(0.082)	(0.080)	(0.038)	(0.033)	(0.031)	(0.031)
Hispanic	1.063**	1.084**	1.066**	1.053	1.079***	1.053
	(0.030)	(0.037)	(0.028)	(0.035)	(0.027)	(0.035)
Pacific Islander	0.634***	0.572***	0.653***	0.580***	0.708***	0.621***
	(0.036)	(0.032)	(0.036)	(0.030)	(0.042)	(0.035)
Two or More	1.134***	1.077***	1.029	0.968	1.057*	0.993
	(0.036)	(0.029)	(0.022)	(0.020)	(0.033)	(0.030)
Female		0.486***		0.489***		0.489***
		(0.004)		(0.004)		(0.004)
Non-English		0.549***		0.576***		0.572***
		(0.036)		(0.044)		(0.037)
FRPL		1.745***		1.763***		1.746***
		(0.045)		(0.039)		(0.041)
Homeless		1.291***		1.218***		1.261***
		(0.033)		(0.036)		(0.029)
Migrant		0.707***		0.689***		0.742***
		(0.034)		(0.031)		(0.035)
ELL		1.418***		1.372***		1.459***
		(0.090)		(0.102)		(0.086)
N	1,097,389	1,097,389	1,114,283	1,114,283	1,118,809	1,118,809

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$
Odds ratio estimates are presented
Standard errors clustered at the district-level.

Table 2.A2: Logistic Regression Results for Special Education Placement (2012-13 and 2013-14)

	2012-13		2013-14	
	No Controls	Controls	No Controls	Controls
American Indian	1.609***	1.325***	1.545***	1.264***
	(0.075)	(0.059)	(0.104)	(0.079)
Asian	0.535***	0.631***	0.526***	0.642***
	(0.024)	(0.026)	(0.030)	(0.030)
Black	1.364***	1.166***	1.358***	1.160***
	(0.050)	(0.056)	(0.037)	(0.032)
Hispanic	1.094***	1.037	1.097***	1.075***
	(0.026)	(0.033)	(0.027)	(0.029)
Pacific Islander	0.744***	0.643***	0.718***	0.628***
	(0.045)	(0.036)	(0.038)	(0.030)
Two or More	1.083**	1.013	1.073**	0.999
	(0.037)	(0.032)	(0.032)	(0.028)
Female		0.489***		0.488***
		(0.004)		(0.004)
Non-English		0.597***		0.543***
		(0.042)		(0.029)
FRPL		1.735***		1.771***
		(0.044)		(0.040)
Homeless		1.255***		1.220***
		(0.030)		(0.036)
Migrant		0.644***		0.659***
		(0.076)		(0.035)
ELL		1.393***		1.487***
		(0.101)		(0.079)
N	1,123,383	1,123,383	1,128,185	1,128,185

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$
Odds ratio estimates are presented
Standard errors clustered at the district-level.

Chapter 3

SUCCESSFULLY TRANSITIONING ENGLISH LANGUAGE LEARNERS INTO THE MAINSTREAM CLASSROOM

Abstract

English Language Learners (ELLs) are one of the fastest growing student subgroups in the United States. Current law indicates that these students must be offered access to education programs designed to assist them in reaching English proficiency. In response, Washington state has created the Transitional Bilingual Instructional Program (TBIP) to assist in the development of English proficiency "enables meaningful access to grade level curricula and instruction." ELL students take an English proficiency test annually to determine whether they are ready to transition out these specialized services. This study explores the effects of these transitions on students in the first year following program exit. Using a regression discontinuity, I assess whether Washington state's current criteria – used by all public schools in the state – are set in such a way to assure a smooth transition of ELL students *overall* into English-taught classrooms. I find that in most scenarios, ELLs are succeeding in their transition into the mainstream classroom. However, I find that Spanish-speaking ELLs exiting bilingual programs tend to struggle to succeed in English subjects after exiting their TBIP services.

3.1 Introduction

The population of English Language Learners (ELLs) – students who are in the process of learning to both communicate and learn in English, has risen steadily over the past few decades. Approximately 9% of the U.S. student population are enrolled in specialized programs designed for ELL students (NCES, 2013). While the majority of ELL students are in states with large immigrant populations (e.g., California, Nevada, New Mexico, and Texas), all states have seen a rise in the number of students whose primary language is not English (Payán & Nettles, 2006).

States must react to these shifting student demographics by making sure that their schools are suited to meet the unique needs of this growing population of ELL students. On top of needing to learn academic content, ELL students must also develop their English language skills. Due to a number of nation-wide legal decisions, schools throughout the country are required to offer specialized programs for their ELL students to assist them in reaching English proficiency while ensuring equal access to academic content. The Bilingual Education Act of 1968 brought national attention to the special needs of ELL students and allowed for the use of federal dollars to support special programs designed for them. The decision in *Lau v. Nichols* of 1974 required schools to provide services that enable students to overcome language barriers that impeded their equal participation. These services generally place ELL students into a special instructional setting where they are taught in some combination of their native language and English. The programs are designed to develop ELL English proficiency, while making sure these students do not fall behind in other academic subjects. Once an ELL meets a number of benchmarks demonstrating their proficiency in English, the student is removed from specialized services and placed into the mainstream classroom.

Washington state currently enrolls the ninth most ELLs in the nation who represent over 200 different languages. Approximately 9% of Washington state's students participate in specialized ELL services. The state's Transitional Bilingual Instructional Program (TBIP) was created to assist these students develop their language proficiency and enable access to

grade-level academic content. Students in this program are assessed annually to determine whether they have reached an English proficiency level deemed suitable for them to "transition" into a mainstream classroom. There has been recent concern over whether ELLs are transitioning at the correct time. A 2015 Dear Colleague letter released by the Department of Justice/Department of Education (DOJ/ED) advises school districts to monitor the academic progress of recently exited ELLs for up to two years to assure that they have not been prematurely exited from ELL services (DOJ/ED, 2015). Washington state has provided its school districts with additional funds to make sure that these recently exited ELLs are provided the services they need to meaningfully participate in their new instructional settings.[1] This study focuses specifically on ELL students who are making these transitions out of TBIP. This chapter seeks to understand how successful TBIP is in preparing ELLs for English-only classrooms. Given the diversity of ELL instructional programs utilized by Washington's public schools, this chapter also explores whether certain program models are better suited for preparing ELLs of all types for this important transition.

Overall, I find that TBIP is successful in its preparation of ELLs entering the mainstream classroom. When focusing my analysis on *all* ELLs in the state, I do not find any significant negative impacts of ELL exit on their test scores during the first year outside of TBIP services. However, when I focus my analysis on Spanish-speaking ELLs exiting bilingual programs, I estimate that they perform approximately 0.1 SD worse than they would have had they remained in TBIP services. This finding could indicate that these ELLs are being exited from their program too early and would benefit from remaining in these bilingual programs for longer. Or, the finding could suggest that these ELL students face specific challenges in adjusting to a new instructional setting and could use further intervening services to help them make this transition.

---

[1] See http://www.k12.wa.us/MigrantBilingual/pubdocs/ExitedTBIPStudents.pdf for more details on the program.

3.2 Background on Washington State's Transitional Bilingual Instructional Program

There is a multi-step process through which a district identifies eligible students for TBIP services. Upon enrollment in a Washington state school, a home language survey is sent home in order to identify the student's primary language. If the survey is returned with a language other than English identified, then the student is assessed using a state-approved language proficiency placement test. In the majority of cases, ELL students will take the Washington English Language Proficiency Assessment (WELPA) to assess their eligibility for services. Students whose scores fall within Level 1 (Beginning/Advanced Beginning), Level 2 (Intermediate), or Level 3 (Advanced) qualify for the TBIP. Those who score at Level 4 (Transitional) do not qualify for services. Each "Level" encompasses a range on a scale score and the cutoffs between each level are set annually.

Once deemed eligible to receive specialized ELL instruction, the student is placed in one of the state-approved programs. This takes one of two forms: bilingual instructional programs and alternative instructional programs. Bilingual instructional programs utilize two languages for instruction and include the following services:

1. **Dual Language Program (or Two-Way Immersion)**: includes both ELL and native English speakers in a classroom where the priority is to help both groups reach proficiency in their native language as well as their second language while supporting learning of academic content in both languages. These programs typically split instruction evenly across the two languages. ELL students who "transition" out of this program are allowed to remain in this instructional setting.

2. **Developmental Bilingual Education**: is a program designed for ELL students who are on a slower path towards English proficiency. Students begin the program in kindergarten where they are generally taught with a 90/10 split in L1/L2 (native language/English). The use of the native language in the classroom gradually decreases

until there is an even split. Like students enrolled in Dual Language Programs, students can continue in this program after being "transitioned" out.

3. **Transitional Bilingual Education**: like the Developmental Bilingual Education program, ELL students start with a 90/10 split between L1/L2. They continually increase the exposure to L2 until all instruction is provided in English. Students may continue in the program even after "transitioning" out.

Alternative instructional programs, on the other hand, are typically designed such that most of the instruction is provided in English (an immersion approach). All of the following services are considered as alternative instructional programs:

1. **Content-Based Instruction/Sheltered Instruction**: generally involve classrooms made up of a majority of ELLs where the teacher has experience in supporting English language development and teaching academic content. Instruction is in English, where content and language instruction are integrated.

2. **Supportive Mainstream**: ELL students remain in the mainstream English-speaking classroom, yet are removed for special instruction by teachers with training in second language acquisition. Content is delivered in English by the mainstream classroom instructor.

The TBIP proclaims a focus on bilingual education. They cite a study by Collier and Thomas (2004) that finds that bilingual programs are more effective than instructional methods that use only English for instruction.[2] Despite this open support for the use of native language in ELL instruction, during the 2013-14 school year, only 14% of ELL students were

[2]It should be noted that the advantage of bilingual education is not so clear cut. Other research papers have been inconclusive over which instructional method is the most advantageous for ELL students in terms of academic achievement and secondary language proficiency. What is clear from this citation is that the State of Washington believes that the research clearly points to the bilingual model of instruction as the correct way to educate its ELL students.

enrolled in bilingual educational programs. The majority of ELL students were placed in alternative instructional programs. This incongruity between the state's goals and reality is mainly due to the fact that most districts meet criteria allowing them to only offer alternative instructional programs.[3] Most districts do not have the necessary means to offer a bilingual education program to their ELL students.

The goal of TBIP, like many other ELL programs, is to prepare and successfully transition ELL students into an English-speaking classroom. Each year, ELL students are assessed using an English Proficiency Exam. As with the initial WELPA exam, students whose score falls within Level 1, Level 2, and Level 3 remain eligible to receive TBIP services. Parents have the option to waive the TBIP option for their children. ELLs scoring at Level 4 are no longer eligible for TBIP and lose their ELL classification. The state intends for Level 4 ELLs to be proficient enough to participate successfully in the mainstream classroom. All ELLs in the state are tested using the same English proficiency assessment and are subject to the same criteria for determining eligibility for TBIP which is based exclusively on the performance on WELPA.[4] Washington provides additional funding to school districts for ELLs who have exited TBIP within the past two years to provide additional instructional support for those ELLs who are struggling in their new instructional setting.

---

[3]From *WAC 392-160-040*, school districts exhibiting one of the following:

1. Necessary instructional materials in the student's primary language are unavailable and the district has made reasonable efforts to obtain necessary materials without success;

2. The capacity of the district's bilingual instructional program is temporarily exceeded by an unexpected increase in the enrollment of eligible students;

3. Bilingual instruction cannot be provided to students without substantially impairing their basic education because of their distribution throughout many grade levels or schools, or both; or

4. Teachers who are trained in bilingual education methods and sufficiently skilled in the non-English primary language(s) are unavailable, and the district has made reasonable attempts to obtain the services of such teachers.

may offer an alternative instructional program to their ELL students.

[4]In 2015-16, Washington introduced a new English Proficiency assessment: English Language Proficiency Assessment for the 21st Century (ELPA21). This chapter's analysis focuses only on the period prior to the ELPA21. But it should be acknowledged that a new test could alter some of the findings of this chapter.

3.3 Literature Review

ELL students transitioning out of ELL services (commonly referred to as "reclassification") has often been viewed as a goal to work towards. Reclassification is a major milestone in the academic trajectory of an ELL student and has often been used as a proxy for English proficiency to gauge the impact of different ELL instructional programs (Conger, 2010; Greenberg Motamedi, Singh, & Thompson, 2016; Grissom, 2004; Thompson, 2015; Umansky & Reardon, 2014). These studies generally utilize survival analysis to understand the time it takes to reach reclassification. The meaning behind reclassification is ambiguous across states (and sometimes districts) as there are variations in the criteria used to redesignate ELLs (Hill, Weston, & Hayes, 2014; Mahoney & MacSwan, 2005). This suggests that reclassification as an outcome could have different meanings in different contexts.

Other studies have conceptualized reclassification as a cause rather than an effect (Robinson, 2011). Robinson (2011) argues that reclassification is often associated with a change in a "bundle of services" offered to the student. After reclassification, most ELL students no longer receive the specialized services designed to support their English language development. They may be placed into a new classroom with different peers and instructors. A successful ELL program is one that is able to smoothly transition ELL students into these new settings. That is, we do not want to remove ELL students from services too early where they experience a shock of being placed into an entirely different classroom. Robinson (2011) finds negligible impacts on elementary school grade test scores as a results of reclassification. However, negative effects are found in high school grades. This suggests that a smooth transition out of ELL services occurs in elementary grades, but not for high school ELLs. Carlson and Knowles (2016) finds that reclassification in 10^{th} grade has beneficial impacts on student's performance on the ACT, high school graduation, and college enrollment. The authors hypothesize that these positive effects are due to the reclassified students experiencing a stronger emphasis on postsecondary preparation. This suggests that the "ELL" label might prevent ELL students from receiving some the support offered by the high school for

college preparation.

Robinson-Cimpian and Thompson (2015) look at the impact of a change in reclassification criteria that occurred in California. They utilize this exogenous change in policy to examine the difference in the effect of reclassification under two different sets of criteria. They find that once the difficulty of reaching reclassification was raised, there was a significant impact on high school test scores and graduation rates. Findings from this study suggest that the choice of criteria for a program plays an important role in the success of students who are eligible and may be in need of such services.

Other studies have argued that there could be a heterogenous effect of reclassification. R. Callahan, Wilkinson, and Muller (2010) find that the impact of an ELL status varies as a function of English proficiency, number of ELLs in the school, length of time in the United States, and socioeconomic status. Among their results, they find that students with higher English proficiency measures and more time in the US benefit more from reclassification (i.e., they do better outside of ELL services). Umansky (2016) posits that the impact of reclassification differs across different instructional models. She finds that the ELL label (no reclassification) has a general negative impact which is mostly concentrated in immersion type classrooms. The study focuses on a comparison of ELL-Initial Fluent English Proficient (IFEP) students at the initial stages of ELL identification. This is particularly unique in that it is able to identify to some extent the impact of ever receiving an "ELL" label.

As hinted at in Umansky (2016)'s study, it seems likely that the "bundle of services" being taken away by reclassification might differ for students enrolled under different instructional models. ELL programs can generally be categorized as following one of two approaches. The first approach follows a "bilingual model" in which students are initially taught in their primary language and then transition to English-only classrooms. In contrast, the "immersion" method places ELLs into classrooms in which all their instruction is delivered in English with a specifial curriculum designed to meet the needs of ELL students.

The two approaches rest on different theories of second language acquisition. The approach that is taken by English immersion programs is to maximize students' exposure to

English. This approach reflects a "time on task" theory of language learning, where language is thought to be a direct consequence of exposure (Cummins, 1992; Porter, 1996). Students in English immersion classrooms are not only taught in English, but are more likely to share the classroom with native English speaking peers (Grunow, 2011). This allows students more time to practice their English in both social and academic settings. A key motivation for this type of approach is to support rapid English language development and assist ELLs in reaching proficiency at faster rates with the idea that proficiency will lead to academic success (e.g., see discussion by R. M. Callahan (2005)).

On the other side of the debate, many researchers have suggested that the most efficient route to second language proficiency may be through the native language which provides the rationale for bilingual programs (Cummins, 1991, 1992). Researchers argue that native language instruction builds a general language capacity – or "common underlying proficiency" – that supports the development of a second language. Cummins (1979) proposes a "linguistic interdependence hypothesis" that argues that the development of proficiency in a second language is partially a function of the competence previously developed in the primary language. Bilingual programs make it a goal for students to acquire the so-called "common underlying proficiency" with the hopes of making it easier for ELLs to transition to English-only classrooms at a later stage with the necessary academic content knowledge. In contrast to the immersion approach, the theory underlying bilingual programs believes that ELLs should be prioritizing developing academic skills and primary language proficiency in order to make reaching English proficiency faster at a later date.

The two instructional models typically enroll different types of students and therefore the peer exposure differs across the different settings. For example, students in bilingual programs have lower initial English proficiency compared to ELLs in an English immersion program (T. B. Parrish et al., 2006). In addition, schools that offer bilingual education are systematically different from schools that do not. For example, schools that offer bilingual education tend to have a higher proportion of English language learners, higher levels of poverty (based on percentage of students receiving free or reduced price lunch) and lower

percentages of credentialed teachers (T. B. Parrish et al., 2006).

In summary, there is a valid concern about understanding ELL student performance near thresholds for reclassification to understand how smooth their transition is to a different educational setting. It appears that instructional models might play a role in preparing students for success in mainstream English-speaking classrooms. This study seeks to explore this relationship. Specifically, it will seek to answer the following questions: (1) *Do the current score thresholds determining TBIP eligibility smoothly transition Washington's ELLs into English-only classrooms?* and (2) *Do the impacts of reclassification vary across different instructional models?*

3.4 Data and Methods

3.4.1 Data

I utilize student-level information from Washington State to answer my research questions. I look at ELL students in grades 2-7 during the 2011-12 through 2013-14 school year. During these years, I know the specific type of instructional program model they are enrolled in. I have several pieces of demographic information for each student: gender, race, free- or reduced-price lunch eligibility, special education enrollment, primary language, homeless status, and gifted status. Furthermore, I have information on each students' test performance in the subsequent years. For federal accountability, students in Washington state are required to take an assessment during grades 3-8 in reading and math.[5] All scores have been standardized to have mean zero and a standard deviation of one by subject and grade. Finally, I have information on the students' WELPA scores for the 2011-12 through 2013-14 school years. Scale score cutpoints used to determine eligibility for TBIP (along with each other proficiency level) are posted publicly online for each grade level.

[5]Students are also required to take Reading/ELA and Math tests in high school. However, during the sample period the high school grade being assessed changed. Prior to 2014-15, 10^{th} graders took the High School Proficiency Exam (HSPE) in reading, math, and writing. Afterwards, 11^{th} graders took the Smarter Balanced Assessment (SBA) in ELA/Math. In order to avoid complications in the analysis due to this testing change, I focus the analysis only on students assessed in grades 3-8.

Table 3.1 provides summary statistics of the cohort of ELL students that I study in this chapter. I show the summary statistics for three groups: all ELL students, ELL students who receive a Level 3 score and ELL students who receive a Level 4 score on the WELPA. The latter two groups are particularly important for my identification strategy. There are two takeaways I see from this descriptive table. First of all, it appears that achievement (and other variables commonly correlated with achievement, e.g., FRPL, SPED) are also correlated with performance on the WELPA exam. That should not be too surprising given that the WELPA measures student English proficiency which is partially measured in the annual assessments. Level 4 ELL students are 0.2 standard deviations below the average on reading tests, whereas all ELL students are nearing 0.9 standard deviations below average. Second, it is clear a simple comparison of Level 3 and Level 4 ELL students on test score performance would not yield unbiased results. It appears that Level 3 and Level 4 students differ significantly on a number of factors that are also likely correlated with test score performance (e.g., special education enrollment, FRPL eligibility). This highlights the importance of having more detailed information on WELPA test performance to be able to compare ELL students who are close to the eligibility threshold who are likely to be more similar. This leads to the methodology of my chapter.

3.4.2 Methods

Given that there is a well-defined cutpoint each academic year determining a student's eligibility for ELL services, I choose to use the regression discontinuity design (RDD) (Lee & Lemieux, 2010). In this specific case, the RDD uses random variations in student performance around the eligibility threshold to estimate the effect of TBIP placement. Under certain assumptions, RDD provides an unbiased estimate of the local average treatment effect (LATE). That is, rather than identifying the average treatment effect (ATE), output from this RD model will estimate the impact of TBIP on highly proficient Level 3 and lower proficient Level 4 students.

Table 3.1: Summary Statistics for Washington State ELL Students in Grades 2-7

	All ELLs	Level 3	Level 4
Demographic and Program Information			
% Female	45.43%	46.43%	49.19%
% FRPL	89.69%	91.63%	83.44%
% Homeless	3.63%	3.01%	2.33%
% Special Education	17.37%	14.09%	5.24%
% Primary Language: Spanish	69.45%	72.06%	61.51%
% Asian	11.97%	11.05%	18.12%
% Hispanic	70.05%	71.80%	61.34%
% White	9.99%	9.50%	13.60%
Instructional Model Used Prior to WELPA Test			
% Bilingual Programs	11.88%	11.08%	8.04%
% Alternative Programs	86.88%	88.04%	90.25%
Standardized Test Score			
Standardized Math Score	-0.80	-0.77	-0.22
Standardized Reading Score	-0.95	-0.92	-0.30
N	303,544	125,014	28,866

Note: The "Level 4" column shows statistics for ELL students who scored at proficient or above on the WELPA. These ELL students are eligible for TBIP exit. The "Level 3" column shows statistics for ELL students who scored one level below proficient on the WELPA.

In order to obtain the RD estimate, I use the following specification:

$$Y_i = \alpha + \gamma f(x_i - x*) + \delta \text{noTBIP}_i + \beta X_i + \eta X_j + \varepsilon_i \tag{3.1}$$

where $f(x_i - x*)$ is some function of the "distance" between a student's score on the WELPA and the cutpoint needed to be eligible for services (I refer to this as the running score or running variable). For f I run models assuming a local linear and local quadratic relationship between the running score and the outcome to explore how my findings change due to choice of functional form. noTBIP_i is an indicator for whether a student transitions out of TBIP services. X_i is a vector of individual student characteristics. I also aggregate demographic characteristics of all students to the school level and include them as controls in X_j. Y_i is an outcome that measures student achievement. To measure student achievement, I look at both standardized reading and math performance during 2014-15 (the year after ELL students take the 2013-14 WELPA test).

In addition to the model described above, I estimate the effectiveness of different program types offered by running the model described in Equation 3.1 on subsets of students. Specifically, I look at the difference between those students in bilingual programs and those in alternative programs (as defined by Washington state). It should be noted that the assignment of students to different program types is not random and coefficient estimates could be biased due to unobserved school effects that both determine whether a school offers such services and influence student achievement. The inclusion of school aggregate characteristics could potentially eliminate this bias, but the impact could still be contaminated by bias if family or any other unmeasured factors play a role in the placement of ELL students into schools that offer certain ELL services.

Given that I have a relatively large sample, I estimate the RD effect using what are commonly referred to as a "nonparametric/local strategy" (Jacob, Zhu, Somers, & Bloom, 2012, p. 28). Under this approach, the researcher attempts to locate an "optimal" bandwidth around the cutpoint in which the data behave approximately linear (or another specified

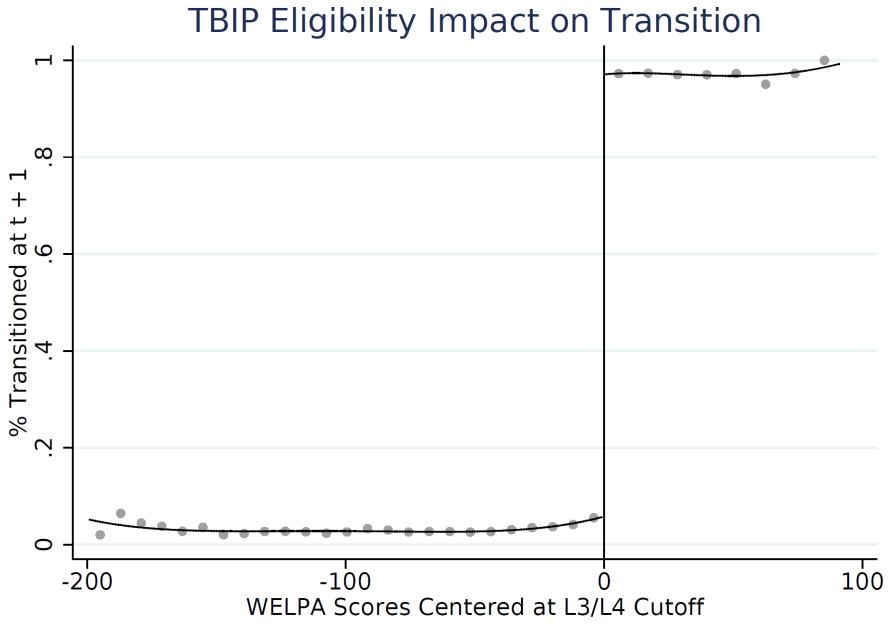
polynomial). Local regression is then conducted and the points on either side of the cutpoint are estimated to produce an estimate of the RD effect. This has an advantage over the alternative "parametric/global strategy" in which the full sample is utilized and a functional form is chosen to fit the data. Under this scenario, choosing the incorrect functional form can lead to a bias in the effect estimate (with the trade-off being an effect estimated with higher efficiency). For the analysis in this chapter, I utilize the `rdrobust` package in Stata to estimate all models (Calonico, Cattaneo, Farrell, & Titiunik, 2017; Calonico, Cattaneo, & Titiunik, 2014).

3.4.3 Assessing the Validity of RDD

Using RDD is appropriate only under certain circumstances (Cattaneo, Idrobo, & Titiunik, 2017; Jacob et al., 2012; Lee & Lemieux, 2010). I start by exploring whether there is an actual discontinuity in treatment (removal from TBIP) probability at the cutoff score. Figure 3.1 shows the percentage of "transitions" that occur for binned groups of ELLs on either side of the treatment threshold. There is a sharp, although not complete, shift in the probability of ELL status at the cutpoint. This noncompliance is not all that surprising given that scoring below the threshold does not guarantee the student will enroll in TBIP services. Even if the student is eligible for TBIP, parents can request to waive these services for their child. On the other side, students who are ready to "transition" out of TBIP, can still continue receive specialized services (yet they are not counted as ELL for funding purposes). Given the noncompliance, I use the *fuzzy* RD design for the analysis (Cattaneo, Idrobo, & Titiunik, 2017; Jacob et al., 2012; Lee & Lemieux, 2010). Under this design, the treatment impact is equal to the ratio of the intent-to-treat effect (from the *sharp* RD) to the effect of assignment on treatment take-up. This again limits my estimated effect to generalize only to those who follow the WELPA program assignment (i.e., those that transition out of TBIP services after reaching Level 4 on WELPA and those who remain who score below Level 4). In other words, I am measuring the local average treatment effect for compliers.

I next explore if there is any evidence of manipulation of the running score. Specifically,

Figure 3.1: Change in ELL Status around WELPA Score Cutpoint



I am worried about students or teachers manipulating scores so that there is an unexpected number of ELLs scoring on one side of the threshold. The worry would be that any effects found on either side of the threshold cannot be attributable to the change in probability of receiving TBIP services, but also to motivations of teachers or students to receive certain services. However, this manipulation seems unlikely in the context of WELPA given that the raw scores needed to pass the test are unknown for each year prior to Level determination. Furthermore, there is less incentive for students/teachers to "cheat" on the exam given that they have some freedom to choose to continue in TBIP or not, regardless of score.[6] Nevertheless, it is still important to check the density of the WELPA score to validate the use of RDD. Panel (a) of Figure 3.2 shows the distribution of WELPA scores within 40 scale score points of the cutoff. Visually, there doesn't seem to be any indication that student

[6]it is possible that funding could be an incentive for schools to "game" the test if they want to receive more resources for their ELL students. However, it is unclear whether the additional funds outweight the costs of the resources needed to serve the additional ELL students.

scores are stacking on either side of the threshold (which could be evidence of manipulation).

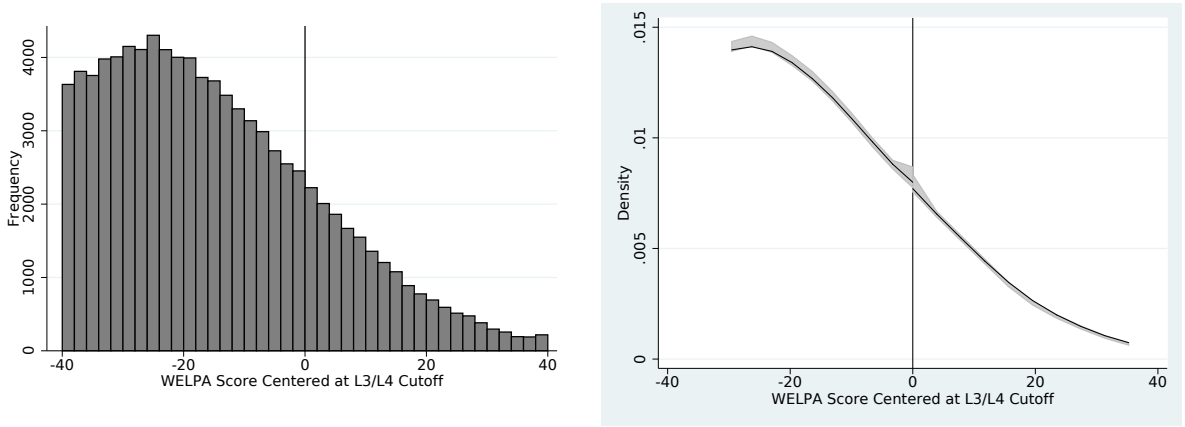
I formally test for manipulation using a nonparametric density estimator designed to, among other goals, test for manipulation near a boundary point (Cattaneo, Jansson, & Ma, 2017b).[7] The test estimates the density function of the running score on both sides of the cutpoint using local polynomial regression and calculates the difference at the cutpoint. Using this test on the WELPA score yields a t-test statistic of -1.50 and a corresponding p-value of 0.13. With a p-value greater than 0.05 I cannot reject the null hypothesis (i.e., no evidence of manipulation at the cutpoint). One might be worried about the marginally significant test-statistic produced by the test, however, a visual inspection of the estimated density shows very little difference in the estimated density functions on either side of the cutpoint (see Panel (b) of Figure 3.2). That is, the difference, although nearly statistically significant, does not appear to be substantially large when viewing the plot. I carry out one more manipulation test designed for discrete running variables as presented in Frandsen (2017). While one could likely assume that the WELPA score is continuous, it can only take integer values. Frandsen's (2017) test fails to reject the null hypothesis of manipulation at the cutpoint (p-value = 0.229), providing further evidence that no manipulation of the running score has taken place.[8]

Finally, no other measures should change significantly at the RD cutoff. That is, like any good experiment, we want to have a balance of observable (and unobservable) characteristics in both the treatment and control groups otherwise we may not be able to attribute the estimated impact only to the treatment. I use a number of observed measures to explore this possibility. I estimate the intent-to-treat effect of TBIP non-eligibility on a number of

[7]This was carried out using a Stata package called `rddensity` (Cattaneo, Jansson, & Ma, 2017a).

[8]McCrary (2008) also provides a formal density test to check for manipulation of the running variable in an RD framework. Surprisingly, the test indicates a significant drop off after the cutpoint (log difference in height estimated at -0.095 with a standard error of 0.014). A number of criticisms of the McCrary (2008) test point out that it assumes that the running variable is continuous (Frandsen, 2017; Jacob et al., 2012). Given that I am using an integer scale score as a running variable, it would suggest that the McCrary (2008) test would not be appropriate for my data. Furthermore, findings in Frandsen (2017) suggests that the McCrary (2008) test overrejects the null hypothesis of no manipulation when the sample size becomes large, another characteristic of my dataset.

Figure 3.2: Distribution of WELPA Scores around the L3/L4 Threshold



(a) Actual Distrubtion

(b) Estimated Distribution w/ 95% CI

program and demographic characteristics: gender, FRPL, disability, Spanish-speaker, race, math score, reading score, and prior WELPA score.[9] For this to be a valid RD, there should be no significant jumps in any of these variables.

Column 3 of Table 3.2 presents treatment impacts on each of these variables. All models were estimated using mean square error (MSE)-optimal bandwidth (column 2 of Table 3.2) and triangular kernel to estimate local linear regressions. The confidence intervals and significance are determined using bias-corrected point estimates and a robust variance accounting for the estimation of the bias (Calonico et al., 2014). As expected, there are no significant estimated shifts in these predetermined covariates at the WELPA score cutoff (no p-value is below 0.249). This finding provides evidence that the treatment effect found is not influenced by the effects of these other covariates. It also assures that there was no manipulation on these observed characteristics in and out of the treatment. Figure 3.3 visually plots the relationships between the WELPA scores and each of these covariates around the cutpoint.

[9]Reading and math scores are for the school year in which the WELPA is taken. These exams were taken simultaneously with the WELPA. While they should be positively correlated, they should not suggest a significant shift at the cutpoint given that these students had not yet received the offer to leave TBIP. The WELPA score is from the previous year's administration, when available.

Table 3.2: Impact of Scoring Above Threshold on Predetermined Covariates

Variable	BW	Estimate	p-value	CI	N
Female	21	0.012	0.249	[-0.01 , 0.037]	47,730
FRPL	19	0.005	0.671	[-0.017 , 0.027]	43,247
Disability	18	-0.002	0.584	[-0.017 , 0.009]	40,826
Primary Language: Spanish	19	0.019	0.318	[-0.021 , 0.063]	43,246
Asian	19	-0.008	0.667	[-0.032 , 0.02]	43,247
Hispanic	22	0.017	0.376	[-0.022 , 0.059]	49,986
White	22	0.004	0.774	[-0.018 , 0.025]	49,986
Prior WELPA Score	17	-0.289	0.996	[-2.532 , 2.518]	25,101
Reading Score	18	0.007	0.647	[-0.038 , 0.062]	24,956
Math Score	19	0.007	0.648	[-0.043 , 0.069]	26,553

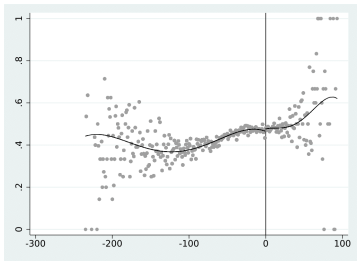
Bandwidths (BW) chosen are MSE-optimal using the `mserd` option in `rdrobust`.
p-values and CIs are constructed after correcting for asymptotic bias in the RD estimate.

The plots show the full range of the relationship rather than limiting the sample to a narrow bandwidth. The fourth order estimated polynomial is fit to visually show the relationship. Again, there does not appear to be any obvious shifts in these outcome variables around the cutpoint.

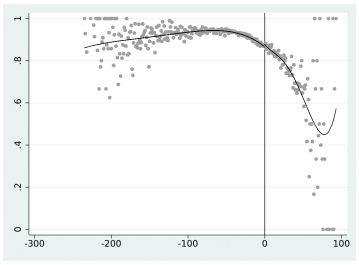
3.5 Results

In this section, I present estimated impacts of transitioning out of TBIP on ELL students in grades 2-7 from a regression discontinuity design. I estimate how this impacts several subsets of students. First, I present results by the type of instructional setting: bilingual approaches vs. alternative programs (i.e., those generally taking the immersion approach). Furthermore, I look at differences in impacts by language: any language vs. Spanish. Spanish is the largest language minority in Washington and provides a sufficiently large subgroup to estimate impacts from. The outcome of interest is the performance on standardized tests (math and reading) in the year following their transition out of ELL services. Put differently, I am estimating the impact of exiting ELL services and entering a mainstream classroom on the performance of that student in the mainstream classroom. Each model is estimated using local polynomial regression within bandwidths determined to minimize the mean-squared error (MSE).

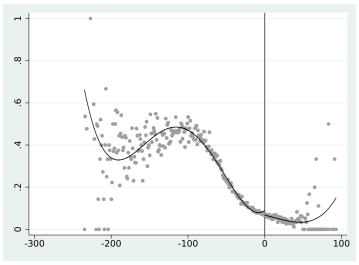
Figure 3.3: Visual Plots of the Impact of TBIP Eligibility on Various Outcomes



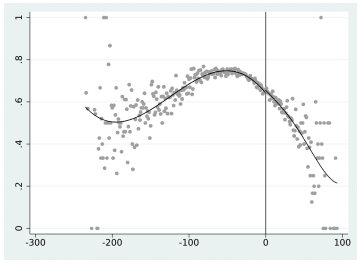
(a) % Female



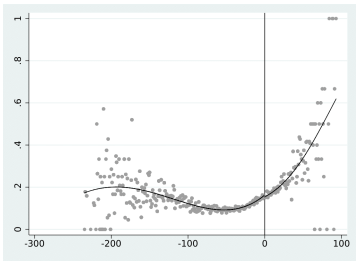
(b) % FRPL



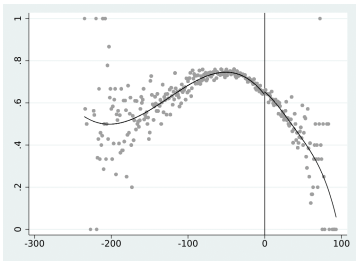
(c) % Disability



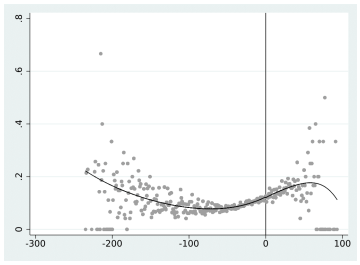
(d) % Primary Lang.: Spanish



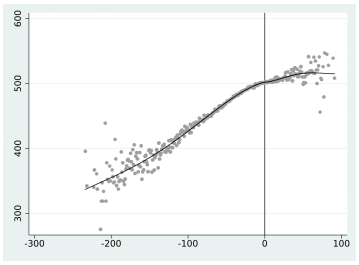
(e) % Asian



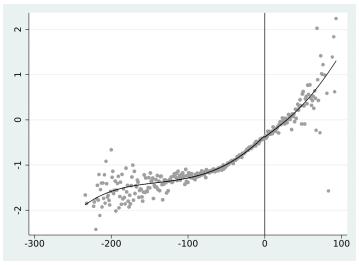
(f) % Hispanic



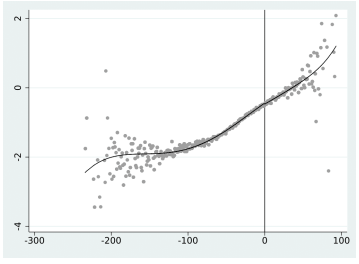
(g) % White



(h) Prior WELPA Score



(i) Math Score



(j) Reading Score

Table 3.3 presents the results of the RD for ELLs in all grades included in my analytical sample (Grades 2-7). The estimated impact of reclassification is presented for a variety of modeling choices to provide evidence of the robustness of my findings. The top panel shows the results using a uniform kernel (similar to running a standard regression) and includes no covariates. The second panel runs the same model, but uses a triangular kernel that reduces the weight of observations within the bandwidth linearly as you move further away from the cutpoint. The third panel adds covariates to the model to aid in the estimation of the optimal bandwidths and local regression. The covariates used in the model include the gender, race, FRPL status, homeless status, and Special Education status of the individual student. In the models that include ELLs of all languages, I include an indicator for whether the language spoken is Spanish. In addition to these individual characteristics, I include school-level covariates to account for differences in school settings. These include the proportion of ELLs, gifted students, females, FRPL eligible, homeless, and Special Education students at the school. I also include information about the racial demographics at the school. Finally, these covariate models include indicators for grade to account for differences in the grade-level test performance. I further include year dummy variables to account for time trends in standardized test performance. Finally, the last panel of the table includes results when the bandwidth is allowed to be of a different size on either side of the cutoff. This specification offers some additional flexibility in determining the bandwidths that minimize MSE. For each specification, I include the estimated effect, standard error (clustered at the school level), optimally determined bandwidth (or averaged bandwidth in the case of the two-sided MSE optimally determined bandwidth), and sample size included within the specified bandwidth. Significance is based on robust bias-corrected p-values.

Table 3.3 assumes a linear relationship between the running score and the outcome variable within the specified bandwidth.[10] I'll start by commenting on the general patterns

[10]Table 3.A1 shows the same impacts when assuming a quadratic relationship between the WELPA score and each of the outcomes. This provides a more flexible functional fit. The results are very similar to those presented in Table 3.3 showing the robustness of the findings. For this reason, I present only results assuming a linear relationship for the rest of the chapter.

observed in the table. Overall, reclassification out of TBIP services appears to have a null or insignificant impact on test scores following reclassification providing evidence that WELPA's cutpoints are valid for determining an ELL student's preparation for the mainstream classroom. However, there are some notable deviations from this pattern. First, I note that in all cases the estimated impact, regardless of significance, is more negative on reading test scores. This makes sense given that these language minority students are entering a classroom with a, sometimes drastically, different approach to teaching the subject of English. The largest negative impacts are estimated for Spanish-ELL students leaving bilingual programs where I find a negative effect ranging from 0.12-0.14 SD. This effect is marginally significant across all specifications suggesting that Spanish-speaking ELLs would have done better to remain an additional year in their ELL program services. The negative effect, although at a smaller magnitude, is picked up when the looking at ELL students of any language. Results that do not include Spanish-speaking ELL students indicate that the negative effect is primarily driven by Spanish-ELL students. I do not include these results due to there being insufficient non-Spanish ELLs in bilingual programs to estimate such effects in more detailed analyses (i.e., looking at the impacts by grade-level).

The plots presented in Figure 3.4 and Figure 3.5 present visual representations of the estimations carried out in Table 3.3. The bandwidths are MSE-optimal and the local linear regression are estimated with a triangular kernel (i.e., these are the results from the third panel of Table 3.3). Again, in almost all cases, there appears to be little evidence of significant impacts to ELLs exiting TBIP services, with the exception of Spanish-speaking ELL students exiting bilingual program services. However, despite not finding *significant* impacts, it appears that, in the majority of cases, ELL students who exit TBIP services who are close to the L3/L4 threshold tend to struggle upon entering the mainstream classroom. This can be seen in each of the plots by observing the line shifting downwards after the score threshold. While no direct conclusion regarding the validity of the score threshold can be made from such findings, It still seems alarming that students just to the right of the cutscore have a lower average than those directly to the left.

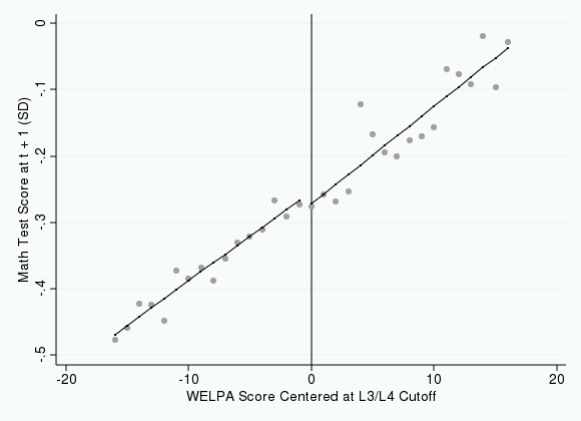
Table 3.3: Regression Discontinuity Estimates for Full Sample (Linear)

	Grades 2-7											
	Any Program				Bilingual Programs				Alternative Programs			
	All Languages		Spanish		All Languages		Spanish		All Languages		Spanish	
	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading
	MSE, Uniform, No Covs											
Effect	-0.022	-0.037	-0.009	-0.043	-0.074	-0.114*	-0.074	-0.131*	-0.013	-0.031	0.007	-0.030
SE	(0.027)	(0.023)	(0.028)	(0.026)	(0.084)	(0.061)	(0.081)	(0.065)	(0.027)	(0.025)	(0.03)	(0.028)
BW	12	14	14	14	23	25	30	29	14	14	14	14
N	22,763	26,384	17,660	17,616	3,877	4,248	4,905	4,657	24,245	24,190	15,611	15,586
	MSE, Triangular, No Covs											
Effect	-0.009	-0.039	-0.014	-0.049	-0.089	-0.116*	-0.097	-0.140**	0.001	-0.031	0.001	-0.035
SE	(0.023)	(0.020)	(0.025)	(0.024)	(0.069)	(0.055)	(0.071)	(0.058)	(0.024)	(0.021)	(0.026)	(0.027)
BW	16	19	19	18	28	23	33	27	14	17	18	15
N	30,237	35,813	24,204	22,737	4,870	3,839	5,428	4,310	24,245	29,328	20,081	16,699
	MSE, Triangular, Covs											
Effect	-0.009	-0.039	-0.014	-0.049	-0.089	-0.116*	-0.097	-0.140**	0.001	-0.031	0.001	-0.035
SE	(0.023)	(0.020)	(0.025)	(0.024)	(0.069)	(0.055)	(0.071)	(0.058)	(0.024)	(0.021)	(0.026)	(0.027)
BW	16	19	19	18	28	23	33	27	14	17	18	15
N	30,237	35,813	24,204	22,737	4,870	3,839	5,428	4,310	24,245	29,328	20,081	16,699
	2-sided MSE, Triangular, Covs											
Effect	-0.015	-0.041*	-0.014	-0.049	-0.087	-0.104	-0.093	-0.122	-0.008	-0.034	-0.002	-0.037
SE	(0.021)	(0.019)	(0.025)	(0.024)	(0.065)	(0.055)	(0.067)	(0.059)	(0.022)	(0.020)	(0.026)	(0.026)
BW	19	21	19	18	31	22	31	23	19	20	19	17
N	41,449	38,117	24,024	21,879	8,060	5,211	7,661	4,912	36,255	34,088	22,264	18,021

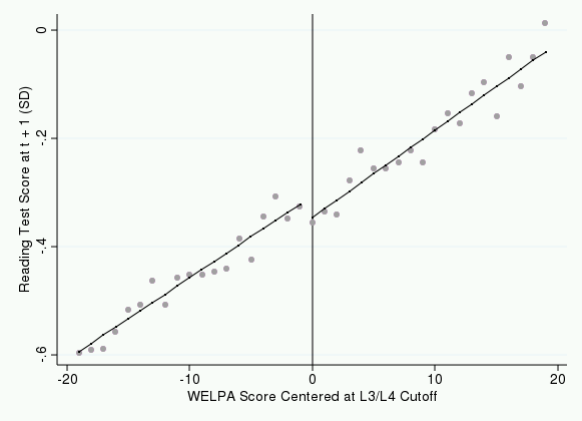
Robust standard errors clustered at the school level are in parentheses below.
Significance based on bias-corrected robust confidence intervals. $^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$
The 2-sided MSE optimal bandwidth allows there to be different bandwidths on either side of the cutpoint. The bandwidth listed is an average of the two.

Figure 3.4: Visual Plots of the Impact of TBIP Eligibility on Next Year's Reading and Math Scores (All ELLs)

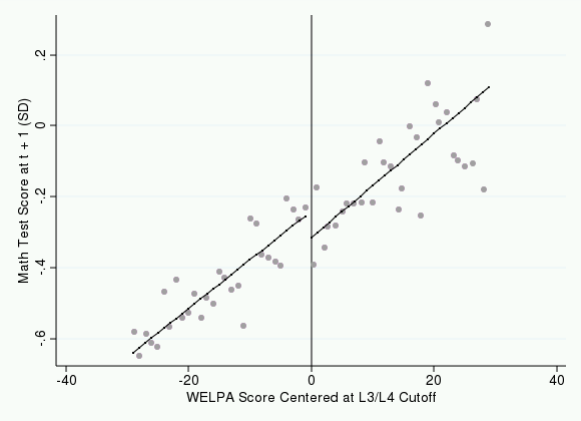
(a) Math Test Score, TBIP



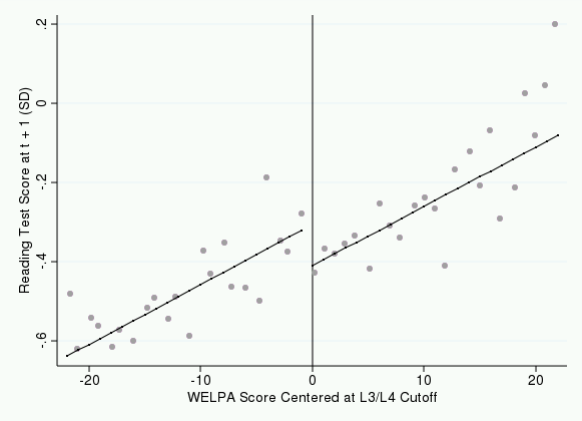
(b) Reading Test Score, TBIP



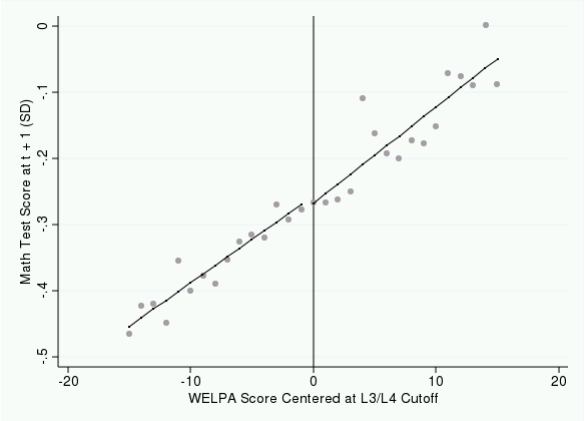
(c) Math Test Score, Bilingual



(d) Reading Test Score, Bilingual



(e) Math Test Score, Alternative



(f) Reading Test Score, Alternative

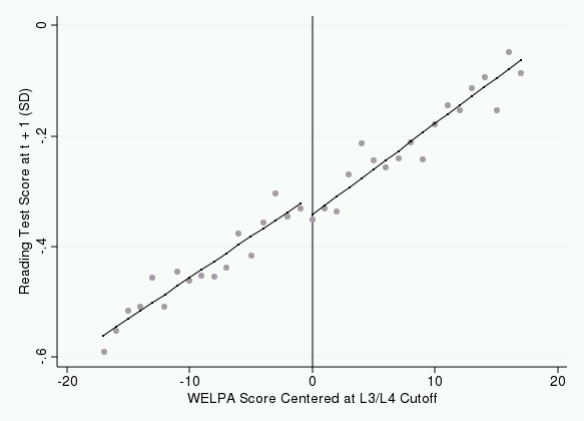
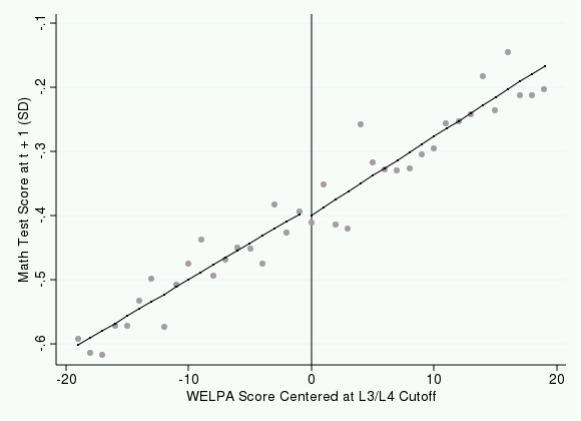
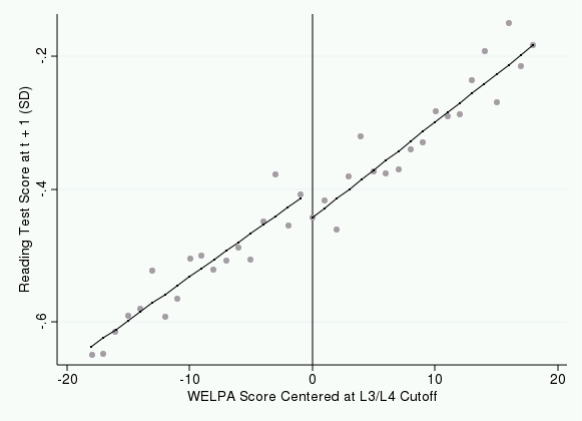


Figure 3.5: Visual Plots of the Impact of TBIP Eligibility on Next Year's Reading and Math Scores (Spanish-speakers)

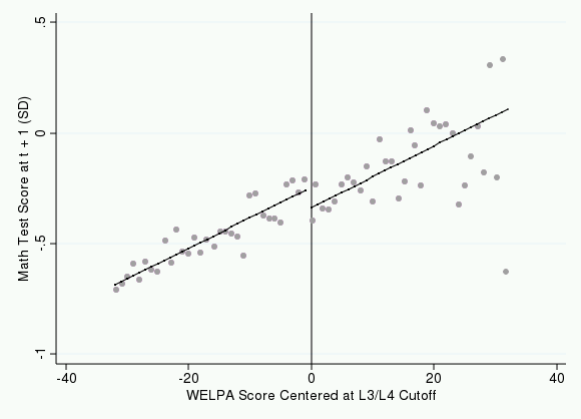
(a) Math Test Score, TBIP



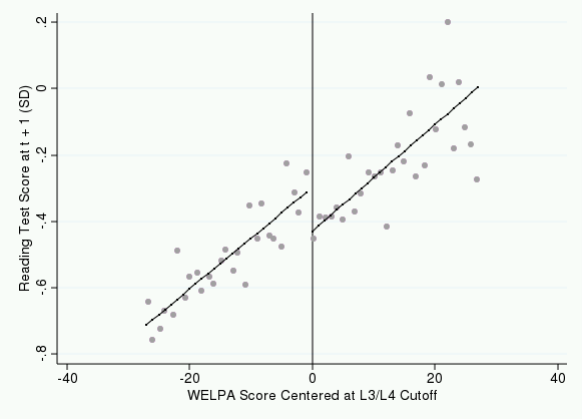
(b) Reading Test Score, TBIP



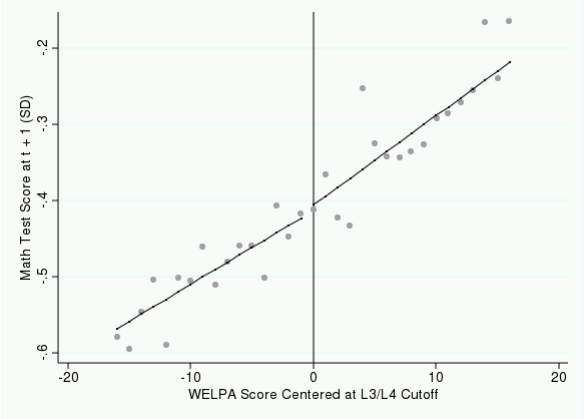
(c) Math Test Score, Bilingual



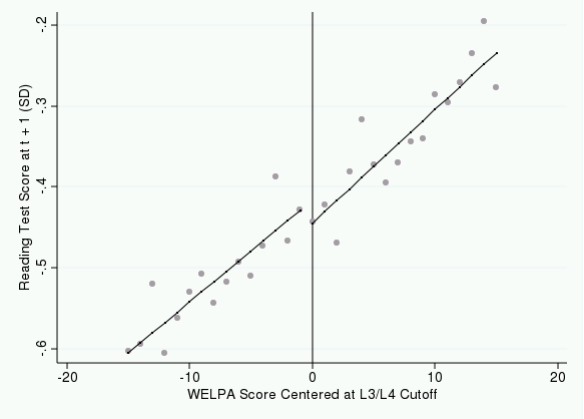
(d) Reading Test Score, Bilingual



(e) Math Test Score, Alternative



(f) Reading Test Score, Alternative



I next present results by grade level groups. I contrast the effects by Elementary (grade 2-5) and non-Elementary grade (grade 6-7) ELLs. WELPA cutpoints for TBIP eligibility are set differently for each grade level and there is sufficient need to assess whether these different cutpoints are validly set.[11] Looking at ELLs who were in elementary school at the time of the WELPA exam (Table 3.4), I find similar patterns to those presented in Table 3.3. There appear to be mostly null findings (although, most are negative and imprecise), with the exception of Spanish-language ELLs in bilingual programs where I find a negative effect of transitioning to the mainstream classroom on reading scores (between 0.11-0.13 SD impact). Again, this provides evidence that the cutpoints for these particular students might not be validly set.

The estimated effects on non-elementary grade ELL students are less consistent across model specifications (see Table 3.5). Again, in general, there appears to be a negative insignificant impact on reading scores for those students exiting TBIP. Although this effect is found to be a significant one for TBIP students in the all language sample. An interesting deviation is that Spanish-speaking bilingual enrolled students appear to be worse off in math after exiting their bilingual services (this is significant for 2 of the 4 specifications, but negative in all). it is possible that the rigor of middle school math courses may prove to be too difficult for these recently exited ELLs. it is unclear if this provides evidence of an incorrectly set cutpoint, or the effect of a changing math curriculum.

3.6 Conclusion

This study explored the impact of TBIP exit on the subsequent academic achievement of exited ELLs. I find that in most cases, ELL students are able to successfully transition out of TBIP and into the mainstream classroom (i.e., I do not find any negative impacts on this

[11]Given that there are different thresholds set for each grade level, it would make more sense to present results by individual grade level rather than these grade groupings. However, the power lost due to breaking the data down in this way leads to imprecise estimates. Therefore, I only present the grouped findings in the main results section of the chapter. Regardless, I present the findings from this analysis in the appendix (Table 3.A2 and 3.A3).

Table 3.4: Regression Discontinuity Estimates for Elementary Grades (Linear)

	Grades 2-5											
	Any Program				Bilingual Programs				Alternative Programs			
	All Languages		Spanish		All Languages		Spanish		All Languages		Spanish	
	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading
	MSE, Uniform, No Covs											
Effect	-0.045	-0.038	-0.029	-0.046	-0.012	-0.077	-0.028	-0.126	-0.039	-0.032	-0.010	-0.026
SE	(0.026)	(0.024)	(0.03)	(0.027)	(0.09)	(0.065)	(0.089)	(0.071)	(0.027)	(0.025)	(0.032)	(0.029)
BW	18	17	14	17	26	29	34	25	17	17	13	15
N	26,684	25,318	13,738	16,814	3,719	4,196	4,675	3,328	23,010	23,042	11,182	12,924
	MSE, Triangular, No Covs											
Effect	-0.019	-0.033	-0.024	-0.047	-0.051	-0.110	-0.054	-0.133*	-0.017	-0.029	-0.023	-0.036
SE	(0.027)	(0.023)	(0.029)	(0.026)	(0.077)	(0.063)	(0.079)	(0.063)	(0.029)	(0.023)	(0.029)	(0.029)
BW	13	17	17	19	22	20	30	29	12	17	19	17
N	19,361	25,318	16,787	18,895	3,067	2,752	4,088	3,924	16,364	23,042	16,442	14,691
	MSE, Triangular, Covs											
Effect	-0.019	-0.033	-0.024	-0.047	-0.051	-0.110	-0.054	-0.133*	-0.017	-0.029	-0.023	-0.036
SE	(0.027)	(0.023)	(0.029)	(0.026)	(0.077)	(0.063)	(0.079)	(0.063)	(0.029)	(0.023)	(0.029)	(0.029)
BW	13	17	17	19	22	20	30	29	12	17	19	17
N	19,361	25,318	16,787	18,895	3,067	2,752	4,088	3,924	16,364	23,042	16,442	14,691
	2-sided MSE, Triangular, Covs											
Effect	-0.030	-0.033	-0.030	-0.047	-0.033	-0.087	-0.040	-0.110	-0.028	-0.028	-0.026	-0.037
SE	(0.024)	(0.022)	(0.028)	(0.026)	(0.071)	(0.062)	(0.075)	(0.067)	(0.025)	(0.023)	(0.029)	(0.028)
BW	17	18	19	19	30	21	30	22	17	18	19	18
N	27,293	25,787	20,402	19,030	6,583	3,891	6,372	3,821	23,540	22,586	17,082	15,509

Robust standard errors clustered at the school level are in parentheses below.
Significance based on bias-corrected robust confidence intervals. $^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$
The 2-sided MSE optimal bandwidth allows there to be different bandwidths on either side of the cutpoint. The bandwidth listed is an average of the two.

Table 3.5: Regression Discontinuity Estimates for Non-Elementary Grades (Linear)

	Grades 6-7											
	Any Program				Bilingual Programs				Alternative Programs			
	All Languages		Spanish		All Languages		Spanish		All Languages		Spanish	
	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading
	MSE, Uniform, No Covs											
Effect	0.037	-0.065	0.017	-0.028	-0.388**	-0.174	-0.425**	-0.236	0.069	-0.062	0.082	-0.020
SE	(0.061)	(0.044)	(0.07)	(0.065)	(0.169)	(0.154)	(0.227)	(0.194)	(0.062)	(0.043)	(0.069)	(0.066)
BW	11	16	8	23	8	9	6	5	12	18	9	9
N	4,447	6,305	2,282	2,665	221	232	154	124	4,509	6,619	2,325	2,295
	MSE, Triangular, No Covs											
Effect	0.038	-0.070*	0.032	-0.049	-0.250	-0.086	-0.239	-0.161	0.073	-0.054	0.077	-0.028
SE	(0.047)	(0.037)	(0.058)	(0.064)	(0.145)	(0.141)	(0.144)	(0.128)	(0.051)	(0.044)	(0.059)	(0.066)
BW	19	25	14	25	11	7	9	7	17	17	13	10
N	7,616	9,837	3,922	2,895	314	187	221	167	6,384	6,286	3,318	2,540
	MSE, Triangular, Covs											
Effect	0.038	-0.070*	0.032	-0.049	-0.250	-0.086	-0.239	-0.161	0.073	-0.054	0.077	-0.028
SE	(0.047)	(0.037)	(0.058)	(0.064)	(0.145)	(0.141)	(0.144)	(0.128)	(0.051)	(0.044)	(0.059)	(0.066)
BW	19	25	14	25	11	7	9	7	17	17	13	10
N	7,616	9,837	3,922	2,895	314	187	221	167	6,384	6,286	3,318	2,540
	2-sided MSE, Triangular, Covs											
Effect	0.052	-0.064	0.033	-0.050	-0.307***	-0.138	-0.283***	-0.159	0.078	-0.050	0.071	-0.055
SE	(0.043)	(0.037)	(0.047)	(0.065)	(0.121)	(0.099)	(0.111)	(0.099)	(0.045)	(0.038)	(0.047)	(0.048)
BW	26	27	25	28	22	16	20	16	24	27	27	23
N	13,162	12,925	9,154	4,007	805	576	704	547	11,512	12,889	9,099	7,882

Robust standard errors clustered at the school level are in parentheses below.
Significance based on bias-corrected robust confidence intervals. ***$p < 0.01$,**$p < 0.05$,*$p < 0.1$
The 2-sided MSE optimal bandwidth allows there to be different bandwidths on either side of the cutpoint. The bandwidth listed is an average of the two.

shift in type of instruction that they receive). This would suggest that the WELPA and cutpoints used to determine TBIP eligibility are adequate at identifying the students who are ready to be moved to the mainstream classroom. There appears to be some exceptions, however. In subsequent analyses, I run the same test on subsamples of the ELL population. In this analysis, I find that ELL students receiving bilingual services tend to struggle more on reading tests once being removed from this instructional setting. This is mostly driven by the effects on Spanish-speaking ELLs leaving these services who make up the majority of the students receiving this type of instructional model (95% of ELLs receiving bilingual services are Spanish-speaking). This seems to indicate that the cutpoints might be set too low for some students (i.e., they are removed too soon from their instructional model). Given that all ELLs in Washington are subject to the same English proficiency test and standards determining eligibility to TBIP, my findings might suggest that a top-down approach to reclassification may not benefit *all* ELL students across language and instructional services they are receiving. With the alternative approach being allowing districts and, even, schools to set up their own reclassification criteria. Or at least letting schools, teachers, or parents provide input into reclassification decisions.

I also argue that the negative effects I find may not indicate that the score thresholds are incorrectly set. The WELPA test may provide an accurate assessment of the preparedness of ELL students ability to learn academic content in English, however, it may not be able to assess their readiness to enter a different instructional environment. Under this hypothesis, it would make sense that students receiving bilingual services would struggle more after exiting these programs. Bilingual programs have a classroom environment and instruction very different from the mainstream classroom. Bilingual programs generally involve being placed in classrooms with all other ELL students (with the exception of dual language programs) where the instruction is in multiple languages. It's entirely possible that these students, after a year of adjusting to a new classroom environment, return to their original academic trajectory after an initial dip. Research would benefit from further research to test this hypothesis. Due to data constraints and changes to Washington's ELL TBIP policies in

2015-16, this hypothesis cannot be tested in this paper.

Washington has updated some of its policies regarding ELL student transitions that might be able to catch any ELLs who may be struggling due to the decisions made under reclassification criteria. The state recently began requiring districts to monitor recently exited ELLs up to two years following their transition. Districts are provided funds to intervene if there is evidence that the student is falling behind in their new instructional setting. This could provide safeguards for students who may be transitioning too early due to the current reclassification criteria or may be struggling due to being placed in a different instructional setting. The findings from this study could highlight the exact students who are in need of these additional supports.

Furthermore, Washington recently made a change to its criteria determining eligibility to TBIP. The new English Language Proficiency Assessment for the 21st Century (ELPA21) was administered for the first time during the 2015-16 school year. The four levels have been removed and replaced with three (Emerging, Progressing, and Proficient). Instead of program eligibility being determined based on an overall scale score, the ELL must perform well across four domains (Reading, Writing, Listening, and Speaking). It will be interesting to see how this change in criteria impacts the success of language minority students transitioning to the mainstream classroom in Washington state. Examination of these changes will be left to future research.

APPENDIX

Table 3.A1: Regression Discontinuity Estimates for Full Sample (Quadratic)

	Grades 2-7											
	Any Program				Bilingual Programs				Alternative Programs			
	All Languages		Spanish		All Languages		Spanish		All Languages		Spanish	
	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading
	MSE, Uniform, No Covs											
Effect	-0.017	-0.041	-0.007	-0.042	-0.128	-0.113	-0.125	-0.145*	-0.009	-0.028	0.007	-0.031
SE	(0.029)	(0.025)	(0.031)	(0.029)	(0.088)	(0.071)	(0.092)	(0.075)	(0.033)	(0.028)	(0.034)	(0.031)
BW	22	26	26	26	22	28	31	30	19	22	23	24
N	41,518	49,018	33,534	33,434	3,700	4,822	5,073	4,857	32,783	37,733	25,795	26,916
	MSE, Triangular, No Covs											
Effect	-0.002	-0.038	0.005	-0.042	-0.110	-0.109	-0.127*	-0.134*	0.004	-0.032	0.019	-0.030
SE	(0.027)	(0.023)	(0.032)	(0.028)	(0.079)	(0.067)	(0.080)	(0.063)	(0.028)	(0.024)	(0.033)	(0.031)
BW	21	27	21	25	34	29	48	44	23	27	22	24
N	39,645	50,823	26,833	32,038	6,030	5,005	7,913	7,240	39,491	46,195	24,671	26,916
	MSE, Triangular, Covs											
Effect	-0.002	-0.038	0.005	-0.042	-0.110	-0.109	-0.127*	-0.134*	0.004	-0.032	0.019	-0.030
SE	(0.027)	(0.023)	(0.032)	(0.028)	(0.079)	(0.067)	(0.080)	(0.063)	(0.028)	(0.024)	(0.033)	(0.031)
BW	21	27	21	25	34	29	48	44	23	27	22	24
N	39,645	50,823	26,833	32,038	6,030	5,005	7,913	7,240	39,491	46,195	24,671	26,916
	2-sided MSE, Triangular, Covs											
Effect	-0.008	-0.037	-0.009	-0.039	-0.094	-0.093	-0.103	-0.115	-0.002	-0.032	0.006	-0.028
SE	(0.023)	(0.021)	(0.028)	(0.027)	(0.071)	(0.063)	(0.076)	(0.068)	(0.024)	(0.022)	(0.029)	(0.029)
BW	31	33	28	27	45	35	45	37	32	32	29	27
N	68,133	68,812	44,984	41,527	10,753	8,454	10,001	8,428	60,489	59,304	39,273	34,187

Robust standard errors clustered at the school level are in parentheses below.
Significance based on bias-corrected robust confidence intervals. $^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$
The 2-sided MSE optimal bandwidth allows there to be different bandwidths on either side of the cutpoint. The bandwidth listed is an average of the two.

Table 3.A2: Regression Discontinuity Estimates by Grade

	Any Program				Bilingual Programs				Alternative Programs			
	All Languages		Spanish		All Languages		Spanish		All Languages		Spanish	
	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading
	Grade 2											
Effect	-0.094***	-0.041	-0.115**	-0.064	-0.057	-0.091	-0.041	-0.142	-0.093***	-0.097***	-0.129**	-0.073
SE	(0.033)	(0.040)	(0.049)	(0.055)	(0.131)	(0.118)	(0.123)	(0.127)	(0.030)	(0.028)	(0.050)	(0.048)
BW	26	15	16	11	15	16	21	14	39	46	17	18
N	13,119	7,552	5,102	3,451	609	656	853	505	16,958	19,154	4,773	5,040
	Grade 3											
Effect	-0.002	0.005	0.029	0.006	0.107	0.057	0.113	0.017	-0.007	0.017	0.034	0.008
SE	(0.046)	(0.045)	(0.054)	(0.059)	(0.218)	(0.141)	(0.213)	(0.137)	(0.049)	(0.040)	(0.060)	(0.056)
BW	16	13	15	12	10	13	9	14	15	19	13	15
N	6,356	5,204	3,903	3,109	349	448	289	450	5,444	6,871	2,942	3,417
	Grade 4											
Effect	0.038	-0.070*	0.032	-0.049	-0.250	-0.086	-0.239	-0.161	0.073	-0.054	0.077	-0.028
SE	(0.047)	(0.037)	(0.058)	(0.064)	(0.145)	(0.141)	(0.144)	(0.128)	(0.051)	(0.044)	(0.059)	(0.066)
BW	19	25	14	25	11	7	9	7	17	17	13	10
N	7,616	9,837	3,922	2,895	314	187	221	167	6,384	6,286	3,318	2,540

Robust standard errors clustered at the school level are in parentheses below.
All Models Use MSE-optimal bandwidths, triangular kernels, and include covariates.
Significance based on bias-corrected robust confidence intervals. $^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$

Table 3.A3: Regression Discontinuity Estimates by Grade

	Any Program				Bilingual Programs				Alternative Programs			
	All Languages		Spanish		All Languages		Spanish		All Languages		Spanish	
	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading
	Grade 5											
Effect	-0.005	-0.040	0.001	-0.077	-0.020	-0.146	-0.028	-0.180	-0.004	-0.025	0.011	-0.058
SE	(0.041)	(0.037)	(0.051)	(0.047)	(0.118)	(0.106)	(0.126)	(0.106)	(0.040)	(0.039)	(0.054)	(0.051)
BW	22	24	18	22	16	15	16	15	26	25	19	22
N	6,622	7,215	3,886	4,710	517	480	480	443	6,912	6,669	3,550	4,060
	Grade 6											
Effect	0.142*	-0.002	0.119	-0.016	-0.233	-0.169	-0.220	-0.040	0.178**	0.001	0.127***	-0.024
SE	(0.068)	(0.065)	(0.072)	(0.069)	(0.192)	(0.118)	(0.190)	(0.111)	(0.069)	(0.064)	(0.053)	(0.058)
BW	13	12	15	14	18	18	14	20	13	13	27	24
N	3,044	2,794	2,463	2,285	280	268	213	270	2,825	2,820	4,039	3,575
	Grade 7											
Effect	-0.162	-0.150*	-0.128	-0.148	-	-	-	-	-0.133	-0.140	-0.078	-0.129
SE	(0.092)	(0.078)	(0.103)	(0.090)	-	-	-	-	(0.098)	(0.081)	(0.110)	(0.098)
BW	9	10	9	11	-	-	-	-	9	10	9	11
N	1,530	1,650	1,035	1,223	-	-	-	-	1,436	1,543	949	1,121

Robust standard errors clustered at the school level are in parentheses below.
All Models Use MSE-optimal bandwidths, triangular kernels, and include covariates.
Significance based on bias-corrected robust confidence intervals. $^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$
No observations available to estimate the RD effect of transitioning for 7^{th} graders.

Chapter 4

MAKING THE CUT: AN OPTIMIZATION APPROACH FOR SETTING CUTPOINTS IN TARGETED POLICY INTERVENTIONS IN EDUCATION

Abstract

Early warning indicators (EWI) or early warning systems (EWS) are becoming more prevalent in education as ways to inform educators on designing and implementing early interventions to help students before they fall too far off-track. Such systems compile a list of leading indicator variables to predict off-track status and use information from these variables to identify candidate students for interventions. While there is a fast growing literature on the implementation, use, and effects of such systems, there has been little guidance on how to establish thresholds or cutpoints on leading indicators that can *optimally* separate students who are potential candidates for targeted interventions or supports and those who are not. This chapter borrows approaches used in the field of medicine for clinical diagnosis and produces a flexible approach to cutpoint or threshold identification in EWI/EWS that minimizes misclassification error. The approach is illustrated in two examples. First, I identify optimal cutpoints in kindergarten and first grade assessments to identify students who show evidence of having dyslexia and are in need of academic interventions. Next, using detailed district student attendance data, I produce thresholds that can be used to identify students at risk of becoming chronically absent in order to provide them with targeted supports during the school year. The approach has the potential to be useful moving forward as schools continue to adopt and implement EWI/EWS to identify students for targeted supports and interventions.

4.1 Introduction

Targeted interventions or specialized programs are common in education (e.g, individualized academic/behavioral interventions, special education, etc.). While school staff can use informal evaluations to determine student eligibility for such services, formal criteria created by the state or district often guide the decisions for which students to serve. One recent and prevalent example of a formal system for identifying students in need of targeted supports are Early Warning Indicators (EWI) or Early Warning Systems (EWS) that are set up to catch students at risk of not graduating high school (Allensworth & Easton, 2005; Department of Education, 2016b). The purpose of such a system is to identify a small set of leading indicators to assist educators in making informed decisions on designing and implementing early interventions to help students before they get too far "off-track" (Balfanz, 2012; Bruce, Bridgeland, Fox, & Balfanz, 2011; Davis, Herzog, & Legters, 2013; Frazelle & Nagel, 2015; Gurantz & Borsato, 2012; Ikbal et al., 2015; Macfadyen & Dawson, 2010; Mac Iver, 2013; Neild, Balfanz, & Herzog, 2007; Planty, 2010). Early findings from studies of such programs are promising and indicate that they can help students improve predictors of long-term success (Faria et al., 2017).

While the literature on setting up these types of systems is rich, the guidance on determining thresholds on key indicators for such eligibility criteria is not as advanced (Frazelle & Nagel, 2015). While most cutpoints are determined using the advice of national experts, some have used locale-specific data to search for "optimal" thresholds (Balfanz, Wang, & Byrnes, 2010; Frazelle & Nagel, 2015). However, the methods used in these examples focus only on evaluating thresholds *ex post* rather than taking a data-driven approach towards establishing these cutpoints. Further, these examples prioritize establishing thresholds that capture a greater number of "at-risk" students while ignoring false positives. More advanced approaches have been utilized in establishing thresholds in other studies that do account for the tradeoff between true positives and false positives (Stuit et al., 2016; Uekawa, Merola, Fernandez, & Porowski, 2010). Limiting false positives is important in cases where targeted

supports or interventions may have no benefits or, in some cases, even harm students who are incorrectly classified (Robinson, 2011, or see discussion in the Introduction to this dissertation). For example, if a student is incorrectly labeled by the system as a potential candidate for a reading intervention, then we might expect them to gain no benefit or be harmed by this misclassification if they miss class time to be given an intervention that has little benefit to their learning. Educator input is extremely valuable in these situations given that many centralized systems will have errors. Recent research has called for the consideration of both true positives and false positives in evaluating EWS (Bowers & Zhou, 2019).

This chapter develops a data-driven approach for establishing thresholds/cutpoints that borrows strategies used in the medical literature (López-Ratón, Rodríguez-Álvarez, Suarez, & Sampedro, 2014; Williams, Mandrekar, Mandrekar, Cha, & Furth, 2006). While Uekawa et al. (2010) and Stuit et al. (2016) have used similar strategies in determining cutpoints for EWS, the approach developed here provides more flexibility for different scenarios that may arise and is more aligned with an approach that seeks to minimize misclassification errors. After introducing the approach, I apply it to two examples. The first, identifies cutpoints in kindergarten and first grade assessments to identify students who show evidence of having dyslexia and are in need of academic intervention. The second, uses detailed district student attendance data to produce thresholds throughout the year that can be used to identify students at risk of becoming chronically absent in order to provide them with targeted supports.

4.2 Methods for Finding an Optimal Cutpoint

The task of identifying cutpoints on key indicators is necessary in any EWI/EWS. In EWI/EWS, there is an assumption that there are two (or more) populations of students: (1) those who will benefit from targeted interventions and supports ("off-track") and (2) those who will not ("on-track"). The cutpoint on a continuous leading indicator should be selected to best separate on- and off-track students. A realistic goal for setting this cutpoint could be to minimize both the proportion (or number) of off-track students who do not re-

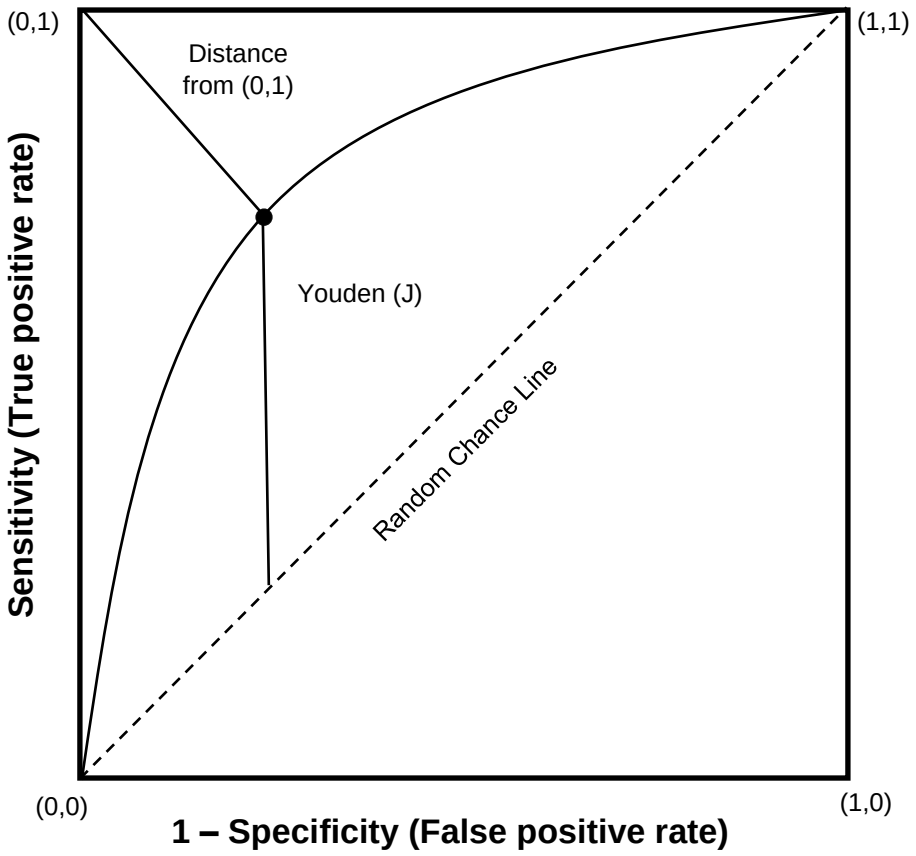
ceive targeted interventions and supports (false-negatives) and the proportion (or number) of on-track students receiving these services (false-positives). That is, the optimal cutpoint is the one that minimizes the misclassification errors (under the assumption that the cost of misclassification errors are symmetric, this assumption is relaxed later).

The clinical medicine literature has generated a wealth of information on the different ways to identify "optimal" cutpoints (López-Ratón et al., 2014). Most methods for obtaining an optimal cutpoint involve analysis of the receiver operating characteristic (ROC) curve (Metz, 1978; Swets, 1986). This curve plots *sensitivity* against (1 - *specificity*) (see Figure 4.1 for an example). In the context of EWS, *sensitivity*, or true-positive rate, is defined as the proportion of off-track students who get labeled as "off-track" ($Se(c) = \frac{TP(c)}{TP(c)+FN(c)}$, where $TP(c) = \#$ of true-positives, $FN(c) = \#$ of false-negatives under cutpoint c). *Specificity*, or the true-negative rate, is defined as the proportion of on-track students who get labeled as "on-track" ($Sp(c) = \frac{TN(c)}{TN(c)+FP(c)}$, where $TN(c) = \#$ of true-negatives, $FP(c) = \#$ of false-positives under cutpoint c). It can easily be derived that 1 - *specificity* is equal to the false-positive rate. Therefore, the ROC curve shows the trade-off between true-positive and false-positive rates. If the cutpoint is set such that all students are labeled "off-track", the true-positive rate would be 100% while the false-positive rate would be 0%. Conversely, if the cutpoint were set such that all students were labeled "on-track", the false-positive rate would be 100% while the true-positive rate would be 0%. The ROC curve shows this balance at all possible cutpoints.

Numerous measures from the ROC curve can be maximized or minimized to generate an "optimal" cutpoint. In determining cutpoints for EWI/EWS, Uekawa et al. (2010) and Stuit et al. (2016) decided to choose a cutpoint that minimized the Euclidean distance to (0, 1). This can be visualized in Figure 4.1 as the point that is closest to the upper-left corner of the ROC plot. Equation 4.1 formalizes the optimization problem.

$$\min_c \sqrt{(1 - Se(c))^2 + (1 - Sp(c))^2} \tag{4.1}$$

Figure 4.1: Illustration of ROC Curve Analysis



Source: Modified from original visualization in Kumar and Indrayan (2011).
Notes: The ROC curve plots sensitivity ($Se(c)$) and 1-specificity ($Sp(c)$) for the range of all possible cutpoints c. The diagonal 45-degree line represents the results of randomly guessing in classification. The larger the area under the curve (AUC), the better the performance of the indicator in differentiating the two populations. The AUC is a common statistic used to measure the performance of predictors in separating target and non-target populations. It is measured as the area under the ROC curve. An optimal cutpoint can be identified at the point on the ROC curve that is closest to the point (0,1) in the plotting area (Distance from (0,1) approach) or the point at which the vertical distance from the "random chance" line (Youden's index) is largest. While the visualization above shows the two optimal cutpoints as being the same, this is not always the case (Perkins & Schisterman, 2006).

However, Perkins and Schisterman (2006) show that the "distance from (0,1)" criterion does *not* produce a cutpoint that minimizes misclassification rates (i.e., the sum of the false-positive and false-negative rates). Another common way of locating an optimal cutpoint is to find the point at which the Youden, or J, index is maximized Youden (1950). The Youden index is the sum of sensitivity and specificity minus one. In looking at Figure 4.1, the cutpoint that maximizes the Youden index is the point where the vertical distance between the diagonal 45-degree line is maximized or, equivalently, the point where the difference between the true-positive rate and false-positive rate is largest. Equation 4.2 formalizes the optimization problem.

$$\begin{aligned} &\max_c\{Se(c) + Sp(c) - 1\} \text{ or} \\ &\max_c\{Se(c) - (1 - Sp(c))\} \end{aligned} \tag{4.2}$$

In contrast to the "distance from (0,1)" approach, using the Youden index yields a cutpoint that can be shown to minimize misclassification rates. Even though the visualization in Figure 4.1 shows that both approaches generate the same cutpoint, the two approaches often disagree (Perkins & Schisterman, 2006). Despite being geometrically intuitive, the "distance from (0,1)" approach does not always match with the goals of minimizing misclassification rates. If the goal were to minimize misclassification errors, the Youden index approach would be preferred.[1]

[1]This is derived in Perkins and Schisterman (2006). Essentially, the Youden index can be rewritten as $\min_c\{(1 - Se(c)) - (1 - Sp(c))\}$ whereas the distance from (0,1) index is rewritten as $\min_c\{(1 - Se(c)) + (1 - Sp(c)) + (Se(c)^2 + Sp(c)^2)/2\}$. Recall that (1-Se(c)) and (1-Sp(c)) are the false-negative and false-positive rate, respectively. So the Youden index can be rewritten as the minimization of the sum of the misclassification rates whereas the distance from (0,1) is equivalent to the minimization of the sum of misclassification rates plus some additional function of selectivity and specificity.

4.3 Application: Identifying Potential Candidates for an Early Literacy Intervention

This section presents an application of the tools from ROC curve analysis in an educational setting.[2] The example uses student-level data from a large urban school district. The district is developing a universal screener to identify students who show signs of dyslexia in response to a state mandate. The hope is to provide these students with phonological skills interventions early so that their learning disability does not prevent them from successfully participating in the classroom. Such early reading interventions have shown promise (Gersten, Fuchs, Williams, & Baker, 2001; Gersten, Newman-Gonchar, Haymond, & Dimino, 2017; What Works Clearinghouse, 2009, 2017). The analysis described in this section seeks to identify the optimal timing and cutoff scores on multiple kindergarten and first grade assessments (Fountas & Pinnell foundational skills test) that can be used in a universal screener to generate a list of students eligible for this targeted tier 2 academic intervention.

In collaboration with the district, I identify a list of 348 students who match the profile of student who would benefit from a phonological skills intervention. These students are in special education settings and had learning profiles consistent with a student struggling with phonological skills based on formative assessments (the district currently does not diagnose students with dyslexia). With this information, I am able to longitudinally track 93 of these students back to kindergarten to observe their scores on the Fountas & Pinnell (F&P) reading assessment. The F&P reading assessment measures nine foundational skills: (1) upper case letter recognition, (2) lower case letter recognition, (3) letter/sound recognition, (4) high-frequency word recognition, (5) early literacy behaviors, (6) initial sounds, (7) blending, (8) segmenting, and (9) rhyming. Kindergarten and first grade students are assessed on these skills three times during the school year. The F&P foundational skills tests work well as leading indicators for students with dyslexia risk. The tests mostly align with predictors of difficulties with phonological skills and it is administered to *all* students in the district

[2]All tables and figures for the application presented in this section are labeled with "(Dyslexia Risk)" in their title.

Table 4.1: Mastery and Maximum Scores for F&P Foundational Skills Tests (Dyslexia Risk)

Foundational Skill	Master	Max
Upper case letter recognition	26	26
Lower case letter recognition	26	26
Letter/sound recognition	26	26
High-frequency word recognition	25	25
Early literacy behaviors	10	10
Initial sounds	8	10
Blending	8	10
Segmenting	8	10
Rhyming	10	10

meaning it can be used as part of a universal screener (Catts, Nielsen, Bridges, Liu, & Bontempo, 2015; Compton et al., 2010; Jenkins & Johnson, 2008).[3] Even for those skills that do not overlap with those that have been shown to be effective predictors of dyslexia, we might expect the signs of dyslexia to also manifest themselves in these other foundational reading skills. Table 4.1 shows the mastery and maximum scores for each foundational skills subtests. It should be noted that the mastery score was recommended by the test publisher as a way of identifying students struggling with foundational skills in reading. It's possible that the mastery score cutpoints could have been used as a way of identifying potential candidates for the Dyslexia intervention.

The segmenting foundational skills test was rarely administered in kindergarten and first grade and was therefore removed from the analysis. Further, testing was inconsistent across schools and testing windows. Therefore, many students were missing scores on some or all foundational skills tests throughout the year. For the purposes of this illustration and analysis, listwise deletion is used to deal with these missing values. While filling in these missing values through multiple imputation could be appropriate in some circumstances, the district was not prepared to use such techniques to assign scores to students who were not assessed (Cameron & Trivedi, 2005). This has implications for the analysis in that not all ROC curves will include the same population of students. The differences in analytical

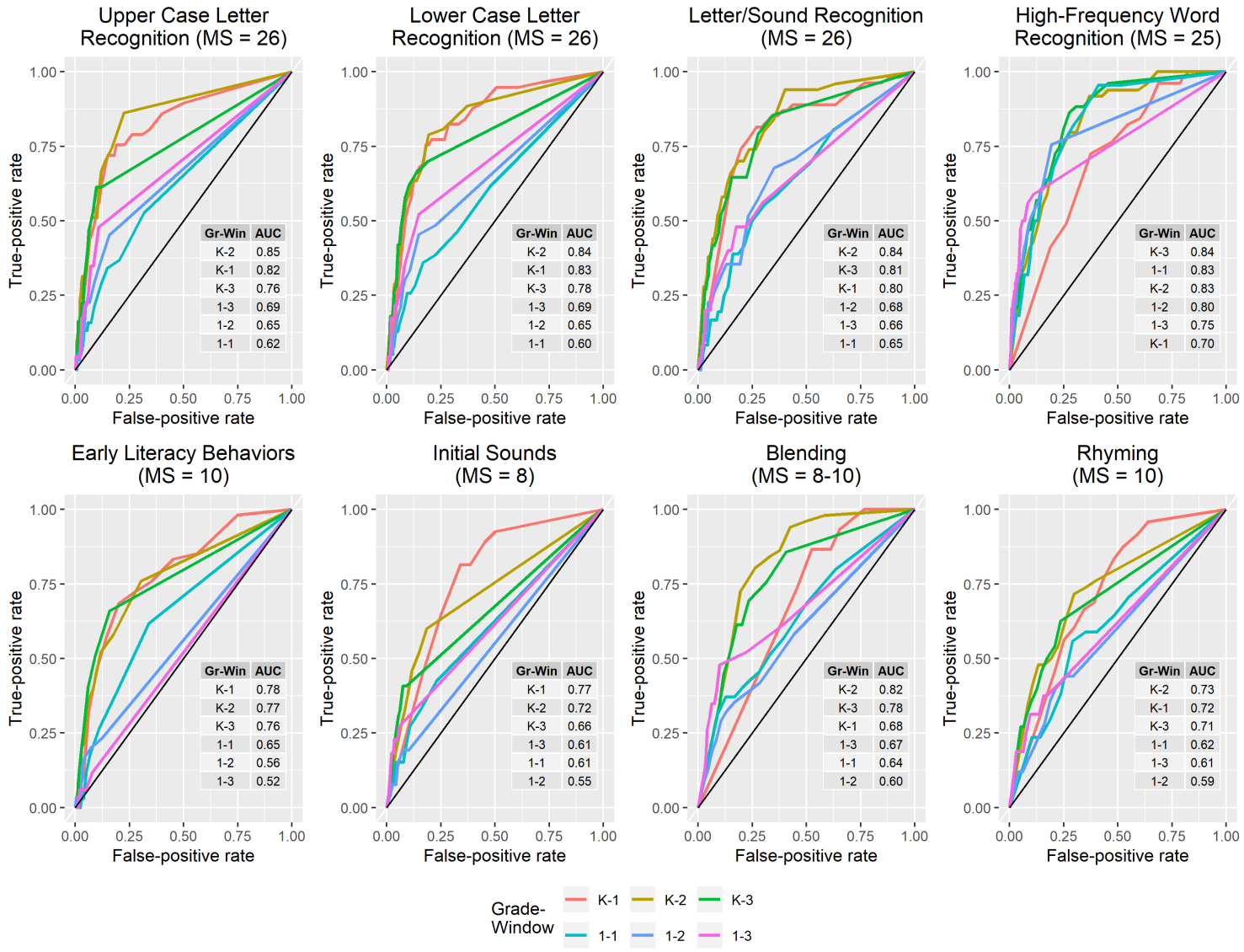
[3]See https://dyslexiaida.org/universal-screening-k-2-reading/ for guidance on what makes a good universal screener for dyslexia.

samples should be considered whenever making comparisons across subtests and testing windows.

The district believes that each of these foundational skills tests can act as a leading indicator of dyslexia risk and can be used to flag students for academic intervention. Given the test is administered multiple times each year, it is important to determine the timing of the test that best differentiates targeted and non-targeted students. This introduces a new measure: area under the curve (AUC). The AUC measures the area that falls beneath the curve on the ROC plot. It ranges from 0.5 (does just as well as random guessing) and 1 (perfectly distinguishes between target and non-target populations). We want to choose the measures – and timing, in this case, – that produce ROC curves with the largest AUC. Figure 4.2 shows ROC curves for each F&P foundational skill subtest and testing window (K-1 = Kindergarten Window 1, K-2 = Kindergarten Window 2, K-3 = Kindergarten Window 3, 1-1 = First Grade Window 1, 1-2 = First Grade Window 2, 1-3 = First Grade Window 3). Upper case letter recognition (max AUC = 0.85), lower case letter recognition (0.84), letter/sound recognition (0.84), high frequency words (0.84), and blending (0.82) all had testing windows that produced AUCs that were greater than 0.8. Relative to these five subtests, early literacy behaviors (max AUC = 0.78), initial sounds (0.77), and rhyming (0.73) did not do as good of a job at separating between the two groups of students.

While variation across the different subtest measures is interesting for selection of leading indicators, the within measure variation can help identify the timing that the measure is most effective. Kindergarten testing windows are almost always preferred to first grade testing windows. it is interesting to see that for all foundational skills tests, kindergarten administrations of the test are better than first grade tests in separating target group students. it is plausible that the foundational skills tests are more effective predictive measures in kindergarten rather than first grade. This would be consistent with research that states that first grade screening measures may differ and be more complex than those measured in kindergarten (Compton et al., 2010). This could suggest that a universal screener for first graders should measure different skills. Another possible explanation of the lack of sepa-

Figure 4.2: Area Under the Curve at Different Testing Cycles for Kindergarten Assessment Subtests (Dyslexia Risk)



Note: Each ROC curve represents a different testing cycle (e.g., "K-1" represents the Kindergarten test during Cycle 1). The area under the curve (AUC) statistic shows the ability of each testing cycle for each subtest to differentiate targeted and non-targeted students.

Table 4.2: AUC Values for Each Foundational Skill Test and Grade-Window (Dyslexia Risk)

Foundational Skill	K-1	K-2	K-3	1-1	1-2	1-3
Upper case letter recognition	0.82	**0.85**	0.76	0.62	0.65	0.69
Lower case letter recognition	0.83	**0.84**	0.78	0.60	0.65	0.69
Letter/sound recognition	0.80	**0.84**	0.81	0.65	0.68	0.66
High-frequency word recognition	0.70	0.83	**0.84**	0.83	0.80	0.75
Early literacy behaviors	**0.78**	0.77	0.76	0.65	0.56	0.52
Initial sounds	**0.77**	0.72	0.66	0.61	0.55	0.61
Blending	0.68	**0.82**	0.78	0.64	0.60	0.67
Rhyming	0.72	**0.73**	0.71	0.62	0.59	0.61

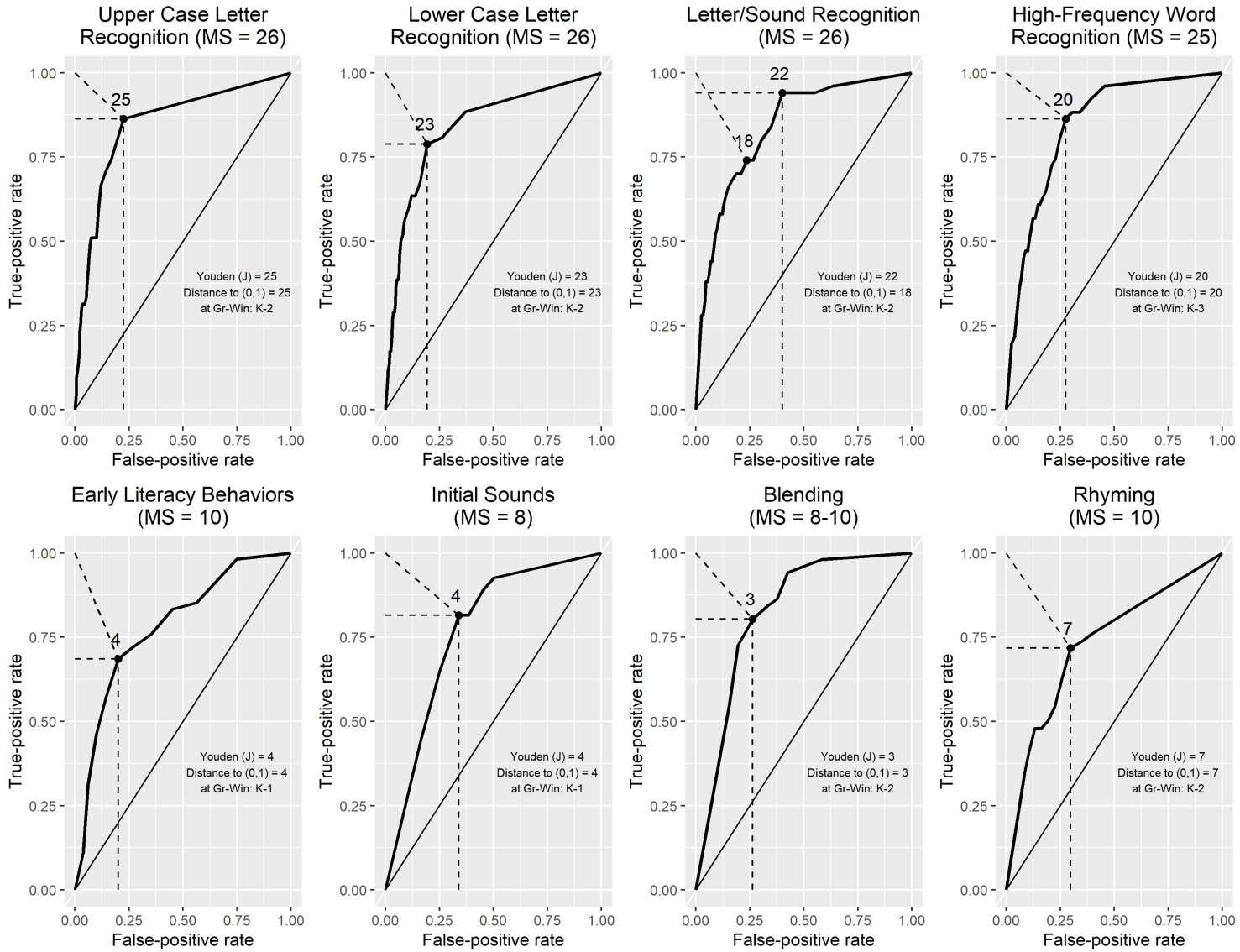
Note: (K-1 = Kindergarten Window 1, K-2 = Kindergarten Window 2, K-3 = Kindergarten Window 3, 1-1 = First Grade Window 1, 1-2 = First Grade Window 2, 1-3 = First Grade Window 3)

ration during first grade testing cycles is that, through repetition of the same foundational skills tests, target students are able to catch up and match scores of non-target students. That is, given that the same test is given during every cycle, target students may be able to succeed on these tests once they have gone through it multiple times. Table 4.2 summarizes the ROC-AUC analysis by highlighting the testing window with the maximum AUC value.

With the ROC-AUC analysis, I was able to identify the measures and timing of those measures that can best isolate target students. I apply the ROC curve analysis to identify the *optimal* cutscores that can be used to generate a list of potential candidates for the targeted academic intervention. I use the two criteria described in the previous section to determine optimal cutpoints – "distance from (0,1)" and the Youden index. The ROC curves presented in Figure 4.3 identify the points where the optimal cutpoints are found under the two different criteria. In all but one foundational skills tests the Youden and "distance from (0,1)" criteria identify the same cutpoint as being optimal. However, the letter/sound recognition subtest has two different cutpoints: 22 under the Youden index and 18 using the distance from (0,1). Table 4.3 summarizes the findings.

This section illustrates how the tools from the ROC curve analysis can be used to set up an

Figure 4.3: ROC Curves with Optimal Cutpoints on Kindergarten Assessment Subtests (Dyslexia Risk)



Note: The optimal cutpoints are presented on ROC curves (true-positive v. false-positive rate) for eight foundational skills subtests on the F&P test. Mastery scores (MS) are shown below each of the eight foundational skills test names. The cutpoints are determined using the minimized *distance from (0,1)* (the diagonal dashed line) and maximized *Youden (J)* (the vertical and horizontal dashed line) approach.

Table 4.3: Optimal Cutpoints for Grade-Window with Largest AUC (Dyslexia Risk)

Foundational Skill	Window	Mastery	Cutpoint	Sensitivity	Specificity
Upper case letter recognition	K-2	26	25	0.86	0.77
Lower case letter recognition	K-2	26	23	0.79	0.81
Letter/sound recognition	K-2	26	22 (18)	0.94 (0.74)	0.60 (0.77)
High-frequency word recognition	K-3	25	20	0.86	0.73
Early literacy behaviors	K-1	10	4	0.69	0.80
Initial sounds	K-1	8	4	0.81	0.66
Blending	K-2	8	3	0.80	0.74
Rhyming	K-2	8	7	0.72	0.70

Note: The cutpoints are set such that any student scoring at or below ($\leq$) the cutpoint would be eligible for a targeted intervention. The cutpoints and their corresponding statistics are determined using the Youden index criterion. The same statistics for the "distance from (0,1)" criterion are presented in parentheses when they differ from the Youden index criterion.

EWI/EWS for identifying students in need of an academic intervention. With the cutpoints identified in this analysis, the district can identify a list of students whose score profiles on the F&P reading tests are consistent with those students who have already been identified as needing an academic intervention. In setting up this system the school district may decide to identify students who score below the majority *or* at least one of these cutscores.[4] Different decision rules can be applied using these cutscores and analyzed in a similar ROC framework. Namely, using the number of foundational skills tests that fall below the cutoff scores, I can plot the different separations onto an ROC curve and identify the decision rule that best identifies students. This analysis is done in Table 4.4. The analysis here finds that the minimum number of foundational skills tests that a student must score below cutscores on is three out of the eight possible foundational skills tests. This decision rule captures approximately 52% of students who show signs of having dyslexia (AUC = 0.65). This is surprisingly low given how the values were when looking at scores on individual tests.

[4]It was the choice of the district to identify separate cutpoints for each individual foundational skills test. An alternative could have been to generate a predicted probability of needing an intervention for each student based on the eight foundational skills tests and then generating an associated "optimal" cutpoint on that predicted probability. However, the district decided to keep separate cutpoints for each test to help assist teachers to identify the specific areas in which they want to focus on with each individual student.

Table 4.4: Sensitivity and Specificity for each Decision Rule (number of tests scored below cutscore) (Dyslexia Risk)

$\geq$ # of tests	Sensitivity	Specificity	Youden	Dist. from (0,1)
0	1.00	0.00	0.00	1.00
1	0.67	0.50	0.17	0.60
2	0.57	0.68	0.24	0.54
3	**0.52**	**0.77**	**0.29**	**0.53**
4	0.41	0.84	0.25	0.61
5	0.36	0.89	0.26	0.65
6	0.28	0.93	0.21	0.73
7	0.17	0.96	0.13	0.83
8	0.10	0.98	0.08	0.90

Note: AUC = 0.65. Sensitivity and specificity are calculated for the scenario where a student would be eligible for a targeted intervention if they scored below the identified cutscores on $\geq$ the # of tests listed on the table.

However, we should recall that the tests were administered inconsistently and analyses on the tests were conducted on different subpopulations of students whereas the analysis of the decision rule includes all students with any information during the identified optimal testing windows. It should be noted that this analysis ignores specifically *which* tests they are scoring below cutpoints on, which likely matters in this context. A similar analysis can be done for any decision rule that generates a continuous measures (e.g., a weighted sum of the tests that one scores below cutpoint on).

4.4 Adding Flexibility to the Youden Index

The Youden index makes two assumptions: (1) the true on- and off-track student populations are roughly equivalent in size and (2) the costs of a false-positive (i.e., the cost of providing a targeted intervention to an "on-track" student) is equal to the cost of a false-negative (i.e., the cost of not providing a targeted intervention to an "off-track" student). In many cases of EWI/EWS, these assumptions will not be met.[5] Prevalence, as it is referred to

[5]I will note that in the previous application, interventions were assumed to be costless. In this case, the high false-positive rates were not as much of a concern. In a world where an intervention is costly, then

in the medical diagnosis literature, is not accounted for by the Youden index. Both $Se(c)$ and $Sp(c)$ do not account for the differing sizes of the on- and off-track student populations and are, therefore, weighted equally in the calculation. If the off-track student population is much smaller than the on-track student population, then the equal weighting could lead to a greater number of false-positives. For targeted interventions in education, it is unlikely that the cost of a false-positive is equal to the cost of a false-negative. An off-track student who does not receive any targeted intervention is likely to see larger consequences than an on-track student who receives the targeted intervention. Weights can be added to provide more flexibility to the optimization problem.

$$\min_c\{a\pi(1 - Se(c)) + (1 - \pi)(1 - Sp(c))\} \tag{4.3}$$

Equation 4.3 shows the minimization of misclassification rates problem that accounts for these different weights. Given the cost of false-negatives (C_{fn}) and false-positives (C_{fp}), a is equal to the relative cost of false-negatives to false-positives ($a = \frac{C_{fn}}{C_{fp}}$). π is the proportion of the student population that are "off-track." Equation 4.3 can be transformed into a maximization problem to provide a generalized Youden index that provides added flexibility to the common Youden index to account for differing subpopulation sizes and costs (Geisser, 1998; López-Ratón et al., 2014; Rücker & Schumacher, 2010; Schisterman, Perkins, Liu, & Bondell, 2005).

$$\max_c\{Se(c) + rSp(c) - 1\} \tag{4.4}$$

Where $r = \frac{1-\pi}{a\pi} = \frac{1-\pi}{\pi}\frac{C_{fp}}{C_{fn}}$. The maximization of the generalized Youden index provides an "optimal" cutpoint accounting for cost and prevalence. The generalized Youden index identifies a cutpoint that minimizes the expected costs arising from misclassifications.

However, the relative costs and prevalence are often hard to define in practice (Greiner, Pfeiffer, & Smith, 2000). While prevalence can be estimated as the proportion of the pop-

---

the high false-positive rates would be more of a concern.

ulation identified as "off-track", the relative costs of false-positives and false-negatives may be difficult to measure. Further, it may be difficult for schools or districts to disentangle the concepts of relative costs and prevalence in certain situations. For example, in the case of the dyslexia risk screener described above, the prevalence is approximately 1%. A district may estimate the relative cost ratio (a) to be around ten (the cost of a false-negative is ten times that of a false-positive). However, they may be confused when the optimal cutpoint using the generalized Youden index ends up favoring $Sp(c)$ instead of $Se(c)$. That is, if someone says that false-negatives are 10 times the cost of false-positives it could be confusing that the false-positive rate is still higher than the false-negative rate with the optimal cutpoint. It may be hard to understand that the prevalence is bringing this down. It may be difficult for school districts to implement this approach if the concepts of prevalence and costs are hard to quantify.

If minimizing total costs is not desired, then a simpler weighting of the Youden index is needed. Galen and Gambino (1975) provides some guidelines on what conditions warrant the favoring of sensitivity and specificity: (1) prefer sensitivity when the condition (e.g., being "off-track") is serious and can be easily corrected and false-positive results are not stigmatizing and do not cost too much academic instructional time (2) prefer specificity when the condition is not serious or serious with an intervention that has a low probability of working and false-positive results are neither stigmatizing nor do they lead to a loss of academic instructional time (3) specificity and sensitivity are equally important when the condition is serious and can be easily corrected and misclassifications are equally damaging. With these guidelines in mind, Li, Shen, Yin, Peng, and Chen (2013) introduce a weighted Youden index that takes a generic weight w and applies it to sensitivity and specificity (Equation 4.5).

$$\max_c \{2(wSe(c) + (1-w)Sp(c)) - 1\} \tag{4.5}$$

The weight parameter w is required to be between 0 and 1. The weighted Youden index (J_w) ranges from -1 to 1. With fewer parameters, the weighted Youden index keeps some

of the flexibility of the generalized Youden index while eliminating some of the complexity. If schools and districts have good information on the relative costs of misclassifications and the prevalence of "off-track" status within the district, then it would be advised to use the generalized Youden index. However, in the absence of this information, the weighted Youden index can provide an intuitive alternative while offering some control over the optimal cutpoint that is chosen.[6]

4.5 Application: Early Detection of Students who are Chronically Absent

This section applies the flexible Youden approaches to a different example.[7] The data for this example also come from a large urban school district. The district wants to create an early detection system for students who are at-risk of being chronically absent (missing greater than 10% of the school year). Many states and districts have begun to see the importance of reducing the number of students who are chronically absent in their schools and have even included it in their accountability plans for the Every Student Succeeds Act (ESSA) (Bauer, Liu, Schanzenbach, & Shambaugh, 2018). Attendance in school has been positively linked with student achievement and success (Gottfried, 2009, 2014, 2019; Gottfried & Kirksey, 2017; Morrissey, Hutchison, & Winsler, 2014). Given the connection between attendance and school success as well as the use of chronic absenteeism as an accountability measure, a system that can identify students early on during the school year as at-risk for being chronically absent could provide schools with enough time to intervene and hopefully prevent

[6]Given that the *identified* prevalence of dyslexia risk using our indicator is so low in the district (approximately 1%), it seems more appropriate to apply the generalized Youden index to a scenario where the prevalence is higher and less influential on the criterion. Further, the leading indicators used in the dyslexia risk example are discrete – so the ROC curve is piecewise linear – and may not show the subtle changes that arise through different assumptions of prevalence, cost, and relative importance of sensitivity and specificity.

[7]All tables and figures for the application presented in this section are labeled with "(Early Detection of Chronic Absence)" in their title.

a student from missing too much class time during a school year.[8]

Currently at this district, schools are given attendance summary reports in non-overlapping six-week cycles: (1) August - September, (2) October - November, (3) November - December, (4) January - February, (5) February - March, (6) April - May (the months are approximate). For each report, an indicator for students who have been chronically absent through that six-week cycle is provided. The purpose of this indicator is to help the schools identify students who may be in need of an attendance intervention, although, currently, no formal action is required of schools. In the hopes of setting up an early detection system which can identify students for evidence-based attendance interventions, the district is seeking ways to identify candidates for such services. While one candidate for identifying these students throughout the year is to use the 90% attendance rate theshold that is currently being used to flag students, the system could benefit from using more optimized cutpoints to flag students. I compare misclassification patterns using the 90% threshold with other "optimal" cutpoints that are generated using the generalized and weighted Youden approach under a variety of different assumptions over the relative importance of sensitivity and specificity.

Table 4.5 relates cutpoints (and associated statistics) identified using different criteria: (1) 90% threshold (currently in use), (2) Youden index, (3) generalized Youden index (assuming cost-ratio (a) = 1, 10, and 20), and (4) weighted Youden index (assuming weight (w) = 0.3, 0.5, 0.7). The cutpoints are determined using data that come from six-week attendance reports for the 2016-17 and 2017-18 school years. Separate cutpoints are identified for the three different grade levels: elementary (grade K-5), middle (grade 6-8), and high (grade 9-12). In all cases of the generalized Youden index, the prevalence of chronic absenteeism is estimated using the proportion of chronically absent students at each grade-level: (1) elementary (11.1%), (2) middle (8.4%), (3) high (15.2%). The choices for the cost-ratio of false-negatives to false-positives (a) are set for illustrative purposes; value judgments would

[8]While it is beyond the scope of this chapter to identify strategies for improving attendance for those students who have shown to be at-risk for being chronically absent, there are many resources available to schools, districts, and states: https://www.attendanceworks.org/.

have to be made by district administrators to set more appropriate cost-ratios. Similarly, the weights (w) for the weighted Youden index are set for illustrative purposes to show how the optimal cutpoints change in three different scenarios: (1) specificity is favored ($w = 0.3$), (2) specificity and sensitivity are equally favored ($w = 0.5$, or the same result as the unweighted Youden index), and (3) sensitivity is favored ($w = 0.7$).

Comparing the predictive ability of the different cycles, it is not surprising that the AUC rises as the attendance rate encompasses the majority of the year.[9] Cycle 6 is excluded from the table since it produces an AUC of 1 given that the final indicator for being chronically absent is based on the cumulative Cycle 6 attendance report. The AUC values also show very similar patterns across the different grade levels, suggesting that the early detection system does not work better for one grade level compared to another. This similarity, along with the similar values of the cutpoints provides evidence that a universal early detection system may be appropriate for the district. That is, a single set of cutpoints would be set and applied across all grade-levels.

The optimal cutpoints (as shown in Table 4.5) under all criteria are above the currently used 90% threshold with the exception of the generalized Youden index where the costs of false-negatives are assumed equal to the cost of false-positives. The 90% threshold misses identifying more students who end up being chronically absent by the end of the year compared to the other criteria. The optimal cutpoints suggest setting the threshold a bit higher during the early part of the year in order to make sure more students at-risk of missing too much school can get the proper attendance intervention. For the first six-week cycle, the Youden index identifies the cutoff value to be 96-98 compared to the 90 currently used. Under the "optimal" cutpoint, 71-82% of students who end up being chronically absent by the end of the year are captured compared to 48-52% under the current 90% threshold. The optimal cutpoint under Youden index does sacrifice some specificity however: 74-85%

[9]This application, in particular, illustrates a trade-off between accuracy and timing. While we might feel that accuracy increases as we collect more data throughout the school year, there is less time to implement interventions once the additional data is collected. In this scenario, it may be more favorable to choose less accurate optimal cutpoints in order to allow time to implement an intervention.

Table 4.5: Optimal Cutpoints for Early Detection of Students who are Chronically Absent based on Different Criteria (Early Detection of Chronic Absence)

		Elementary School					Middle School					High School				
		C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5
$\leq 90\%$	Cut	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
	Se(c)	(.52)	(.60)	(.66)	(.80)	(.87)	(.48)	(.56)	(.65)	(.79)	(.89)	(.52)	(.62)	(.70)	(.83)	(.93)
	Sp(c)	(.94)	(.97)	(.98)	(.98)	(.99)	(.96)	(.98)	(.98)	(.98)	(.99)	(.97)	(.98)	(.98)	(.98)	(.99)
Youden (J)	Cut	96	95	94	92	92	97	96	95	93	92	98	96	94	93	92
	Se(c)	(.71)	(.85)	(.88)	(.91)	(.95)	(.82)	(.85)	(.89)	(.92)	(.95)	(.81)	(.86)	(.88)	(.93)	(.96)
	Sp(c)	(.85)	(.85)	(.90)	(.93)	(.96)	(.74)	(.87)	(.90)	(.94)	(.96)	(.77)	(.88)	(.91)	(.94)	(.96)
Gen. J (a = 1)	Cut	83	89	89	89	90	88	89	90	89	89	92	91	90	90	90
	Se(c)	(.25)	(.49)	(.62)	(.73)	(.87)	(.36)	(.47)	(.64)	(.74)	(.84)	(.52)	(.62)	(.73)	(.83)	(.93)
	Sp(c)	(.99)	(.98)	(.98)	(.99)	(.99)	(.98)	(.99)	(.98)	(.99)	(1.0)	(.97)	(.98)	(.98)	(.98)	(.99)
Gen. J (a = 10)	Cut	97	95	94	92	92	96	96	95	93	91	98	96	95	93	92
	Se(c)	(.88)	(.89)	(.88)	(.93)	(.96)	(.66)	(.85)	(.89)	(.92)	(.95)	(.81)	(.89)	(.92)	(.94)	(.97)
	Sp(c)	(.65)	(.80)	(.90)	(.91)	(.95)	(.90)	(.87)	(.90)	(.94)	(.97)	(.77)	(.83)	(.86)	(.93)	(.95)
Gen. J (a = 20)	Cut	97	96	95	94	93	97	98	96	95	93	98	98	96	95	93
	Se(c)	(.88)	(.89)	(.94)	(.95)	(.97)	(.82)	(.92)	(.93)	(.95)	(.97)	(.81)	(.91)	(.94)	(.97)	(.98)
	Sp(c)	(.65)	(.80)	(.80)	(.87)	(.92)	(.74)	(.76)	(.84)	(.89)	(.95)	(.77)	(.78)	(.79)	(.86)	(.94)
Wgt. J (w = .3)	Cut	93	93	92	91	91	96	95	93	91	91	94	94	93	91	90
	Se(c)	(.52)	(.71)	(.82)	(.85)	(.92)	(.66)	(.77)	(.81)	(.89)	(.92)	(.66)	(.78)	(.81)	(.89)	(.93)
	Sp(c)	(.94)	(.94)	(.94)	(.96)	(.98)	(.90)	(.92)	(.94)	(.96)	(.98)	(.92)	(.93)	(.95)	(.97)	(.99)
Wgt. J (w = .5)	Cut	96	95	94	92	92	97	96	95	93	92	98	96	94	93	92
	Se(c)	(.71)	(.85)	(.88)	(.91)	(.95)	(.82)	(.85)	(.89)	(.92)	(.95)	(.81)	(.86)	(.88)	(.93)	(.96)
	Sp(c)	(.85)	(.85)	(.90)	(.93)	(.96)	(.74)	(.87)	(.90)	(.94)	(.96)	(.77)	(.88)	(.91)	(.94)	(.96)
Wgt. J (w = .7)	Cut	97	96	95	94	93	97	98	96	95	93	98	98	95	94	92
	Se(c)	(.88)	(.89)	(.93)	(.95)	(.97)	(.82)	(.92)	(.93)	(.95)	(.97)	(.81)	(.91)	(.92)	(.95)	(.97)
	Sp(c)	(.65)	(.80)	(.84)	(.87)	(.92)	(.74)	(.76)	(.84)	(.88)	(.95)	(.77)	(.78)	(.86)	(.90)	(.95)
AUC		0.85	0.92	0.95	0.97	0.99	0.84	0.92	0.95	0.97	0.99	0.85	0.93	0.95	0.98	0.99

Note: The cutpoints are for the daily attendance rate (% of days at school) and are set such that students with an attendance rate $\leq$ to the cutpoint are identified as "at-risk" for being chronically absent. The leading indicators of chronic absenteeism are the daily attendance rate of a student through the Xth six-week cycle ("CX"). The $\leq 90\%$ criteria sets the cutpoint at 90% for all cycles. Youden (J) identifies the cutpoint which maximizes the sum of $Se(c)$ (sensitivity, true-positive rate) and $Sp(c)$ (specificity, true-negative rate). Gen. J is the generalized Youden index which accounts for the prevalence of chronic absenteeism and the assumed relative costs of false-negatives to false-positives ($a = \frac{C_{fn}}{C_{fp}}$). Wgt. J is the weighted Youden index which accounts for the relative importance of $Se(c)$ (w) and $Sp(c)$ ($1 - w$). Cutpoints are rounded to the nearest whole number.

versus 94%-97% for the Youden index and 90% threshold, respectively. By the end of Cycle 5, though, the Youden index still suggests setting the cutpoint above the 90% threshold. The sensitivity (95-96%) and specificity (96%) of the cutpoints determined based on the Youden index provide a greater balance of the two measures than the 90% threshold (sensitivity: 87-93%, specificity: 99%). The main takeaways from the comparisons are that the currently used 90% threshold may not be optimal when considering both false-positives and false-negatives.

The generalized Youden index adds more flexibility to the Youden index by accounting for prevalence and costs. When prevalence is accounted for and the costs of misclassification are assumed to be equal, the Youden index tends to favor specificity (i.e., reducing the number of false-positives). Many of the cutpoints fall below the currently used 90% threshold. However, given that the system is designed to be able to catch students at-risk of being chronically absent, a system that catches 25-52% of chronically absent students in Cycle 1 may not be viewed as a success. Therefore, it may make sense to assume that the cost of false-negatives is much higher than false-positives to increase the percentage of chronically absent students captured. Increasing the cost-ratio ($\frac{C_{fn}}{C_{fp}}$) to 10 and then 20 makes very similar adjustments to the cutpoints, both of which increase selectivity over the generalized Youden index approach which assumes the cost-ratio is one. When the cost-ratio is assumed to be ten, the generalized Youden index gives cutpoints that are very similar to the Youden index due to the prevalence of chronic absenteeism being approximately 10% in the district. Even when the cost ratio is increased to twenty, the cutpoints are still very similar to those set by the Youden index, however, the sensitivity is higher than specificity in all cases when the cost of false-negatives outweights the prevalence.

While it was easy to come up with hypothetical costs for the generalized Youden index, it would be hard to determine these in practice. The weighted Youden index allows one to determine the relative preference for sensitivity and specificity without having to come up with estimates of costs for misclassification that balance our own assumptions over the prevalence. Having only a single parameter that needs to be assumed (even if it accounts

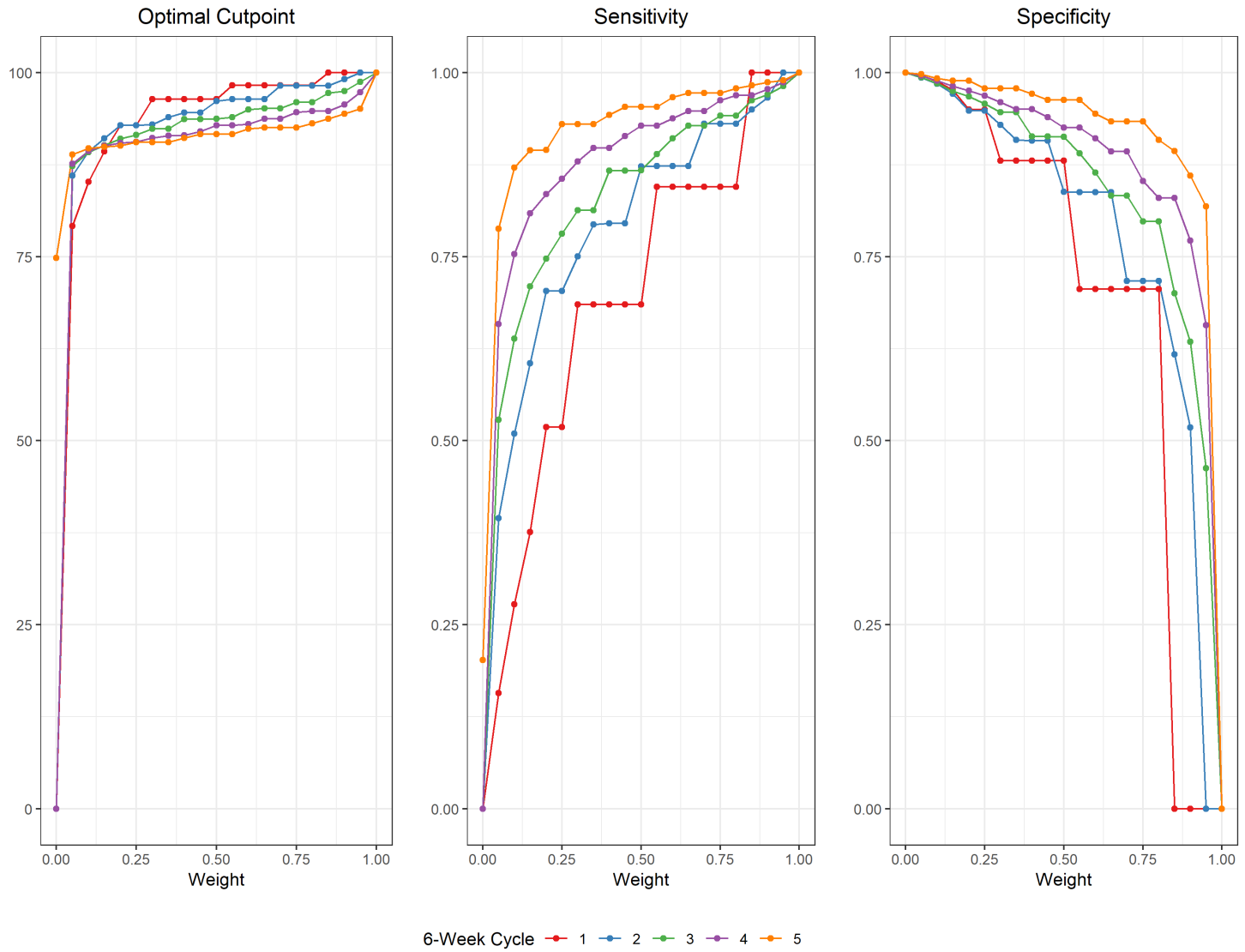
for multiple assumptions) gives the weighted Youden index an advantage in interpretatbility. Setting a weight (w) of 0.5, the weighted Youden index becomes just the normal Youden index (both sensitivity and specificity are of equal importance). Weighting in favor of specificity ($w = 0.3$), as you would imagine, leads to identifying cutpoints that lead to higher specificity. With a weight of 0.3, the weighted Youden index still suggests setting thresholds above the usual 90% attendance rate. Choosing to favor sensitivity ($w = 0.7$), leads to cutpoints that are slightly higher than those set by the Youden index as a way of capturing more students who end up being chronically absent by the end of the year.

The weighted Youden index has the capacity to be useful in this district's context. Given the large amount of stakeholders who are having input into the early detection system, having to come to agreement on just a single parameter, rather than the three required for the generalized Youden index, is a major advantage. In addition, the influence for the weight in the weighted Youden index is easier to understand than the complex relationships between the cost-ratio and prevalence assumptions. Given the utility that the weighted Youden index has in this scenario, I move my attention to focus on the characteristics of the weighted Youden index.

While the purpose of the early detection system is to try and identify *all* students who are at-risk of missing too much school (defined as missing 10% or more of the school year), there is an obvious downside to identifying too many students. First, schools face a constraint on resources, and can rarely provide high-leverage targeted supports to all of their students. Therefore, a balance must be struck to make sure schools are not overburdened with too large of a caseload. Second, the label given to students who fall below the cutpoints determined by the analysis may be stigmatizing. Depending on the type of interventions that are undertaken after a student is identified as "at-risk", the student and family may perceive the label as negative and punitive. While this could be remedied by carefully crafting responses that are positive, supportive, and asset-based, these are real costs to be considered in weighing the importance of selectivity and sensitivity.

Figure 4.4 shows the main statistics related to optimal cutpoints coming out of an ROC

Figure 4.4: Optimal Cutpoint, Sensitivity, and Specificity as a Function of the Weight (w) in the Weighted Youden Index (Early Detection of Chronic Absence)



Note: Weight = the weight placed on sensitivity in the weighted Youden index ($2(wSe(c) + (1-w)Sp(c)) - 1$). The cutpoint is determined such that a student with an attendance rate $\leq$ that point is labeled as "at-risk."

analysis (the cutpoint, sensitivity, and specificity) as a function of the choice of weight (w) used in the weighted Youden index. The optimal cutpoint moves higher as the weight gets larger. That is, as the relative importance of selectivity moves higher, the cutpoint separating target from non-target students gets more inclusive. We can also observe how the values for selectivity and sensitivity change. The trends in sensitivity and specificity over all the different weights are as expected. These graphics can facilitate in the discussion on how to set the cutpoint that creates the desired selectivity and specificity while also minimizing misclassification errors.

4.6 Conclusion

In the education literature, the guidance on setting cutoff scores for districts and states developing EWI/EWS is lacking. The field of medicine has spent more time developing tools for such tasks. This chapter introduces the concepts and tools developed by the medical literature for disease diagnosis into the realm of education. The tools use an optimization procedure that locates a cutpoint that minimizes misclassification errors for both false-negatives and false-positives while accounting for prevalence and costs. I then apply these tools to identify "optimal" cutpoints in two district examples. In the first example, I identify optimal cutscores on a kindergarten and first grade assessment that is being used as a universal screener for dyslexia risk. In the second example, I construct optimal thresholds on mid-year attendance reports to identify students who are at-risk of being chronically absent (missing more than 10% of the school year).

This is only a first attempt at integrating learning from the medical literature into education. More work is needed to guide the appropriate selection of weights for the weighted Youden index to account for the different situations that arise in education. Further, if relevant for a situation, diagnostics on the statistical properties of the cutpoints can be integrated to allow for ranges of possible cutpoints (Schisterman & Perkins, 2007). There is clearly more work to be done to understand how these tools from ROC curve analysis fit in the specific context of education and EWI/EWS. This is left to future research.

Targeted interventions are generally aimed at helping the most vulnerable student populations. Until recently, most of the decisions made on which students to serve were made through non-standardized decisions made by the classroom teachers with few checks to assure the accuracy of such systems. The rise of EWI/EWS in education help standardize this process and are following a trend of evidence-based decision-making coming out of ESSA (Balfanz, 2012; Bruce et al., 2011; Davis et al., 2013; Frazelle & Nagel, 2015; Gurantz & Borsato, 2012; Ikbal et al., 2015; Macfadyen & Dawson, 2010; Mac Iver, 2013; Neild et al., 2007; Planty, 2010). The optimization approaches covered in this chapter help schools and districts leverage information available to them in order to make the most informed decisions of which students to (and not to) serve with targeted interventions.

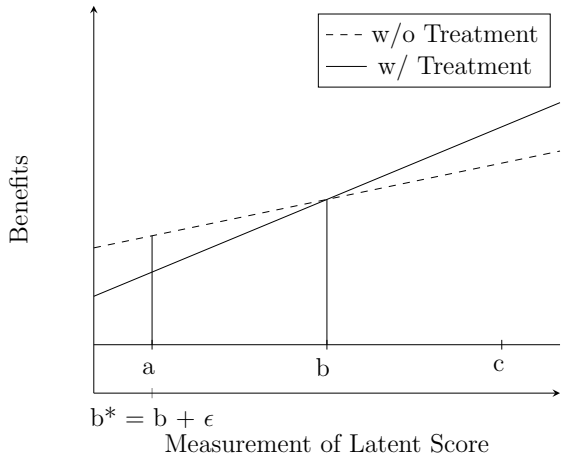
Chapter 5
CONCLUSION

Targeted policy interventions can potentially carry consequences if individuals are deemed ineligible for beneficial services or incorrectly targeted. I call these the *consequences of misassignment to treatment*. These consequences arise when the system set up to select individuals for targeted policy efforts fail. In Chapter 1, I introduce a framework for understanding how these consequences can arise as a result of errors in selection criteria. In this dissertation, I examine ways in which data can be leveraged to validate the selection criteria of the system to limit the consequences of misassignment to treatment. Through this exploration, I learn several practical lessons regarding the design of systems used for determining eligibility for targeted policy interventions. I summarize these lessons below.

1. *The criteria or leading indicators should be checked for validity. Invalid measures have the potential to target students (or districts) incorrectly for treatment.*

   In the second chapter, I examine a federal policy that tracks special education disproportionality in our nation's local education agencies (LEA). When evidence of disproportionality is found, the LEA is required to devote part of their federal funds towards early intervention efforts. However, the research literature has begun to question the ways in which disproportionality is commonly measured. Specifically, there are concerns that measures that do not account for student-level characteristics (e.g., socioeconomic status) will be biased. After creating a covariate-adjusted measure of disproportionality, I find that disproportionality tends to be overstated under commonly used measures. While most state thresholds are set in such a way to mostly avoid LEAs being incorrectly identified as needing interventions, utilization of the incorrect indicators in targeted policy systems could very much lead to consequences.

Figure 5.1: Consequences of Running Score Measured with Error



I cite an alarming example of Houston Independent School District's (HISD) fight against "overidentification" of African-American students. In response to measures citing an overrepresentation problem in their schools, the district intentionally under-referred African-American students, likely taking many students away from the services that they need (Rosenthal, 2016).

I can illustrate the failure of the selection criteria that was found in Chapter 2 in Figure 5.1. In the example, the system's attempt to measure the running score using the risk ratio did not accurately capture the latent score: bias in special education referral practices. The error in the measurement, led states to mistake noise (or error) for a signal. In the illustration, the value b on the latent score separates the school districts (or students) that benefit from the intervention from those that do not. However, when the running score is being measured by the system, the value on the measure that corresponds with b actually yields a lower value of the true latent running score. That is, $b*$, the system's measurement that it believes represents the point b is actually equal to $b + \epsilon$ which corresponds to point a on the true latent running score. Choosing a as the point of selection due to a belief that $b* = b$ subjects all districts (or students)

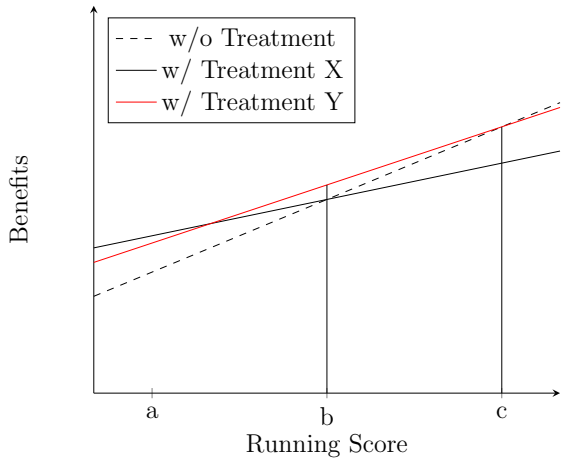
with a running score of a to b to a treatment that does not benefit them. If a system does not accurately measure the latent score, then consequences can arise in the system.

Researchers and psychometricians will play an important role in assuring that the measures used in selection criteria accurately reflect the true measure the system is attempting to gauge. The measures used in such systems should limit, to the extent possible, any error or bias that could arise. Research should provide the theoretical frameworks to help in the identification of possible sources of omitted variable bias or measurement error to ensure the validity of the system. In the event that biased or error-prone measures are the only option (likely the case), the system should attempt to adjust thresholds to account for the expected direction of error to avoid misidentification. Exercises, similar to the one carried out in Chapter 2, can assist in estimating the direction and magnitude of the bias in the proposed measures for the selection criteria. Checking the validity of the indicators used to determine the eligibility for treatment remains an important task in research.

2. *Pre-established thresholds should be assessed for accuracy with consideration to different contexts. Flexibilities should be built into any system to account for these different contexts.*

   Often, targeted policy systems have pre-established thresholds set by higher-level governments or test publishers. While methods have already been introduced to examine whether the thresholds are set to maximize benefits for both targeted and non-targeted individuals, the lesson I take away from the third chapter is that these sorts of analyses need to account for different contexts. In the third chapter, I examine the state-established thresholds that are used to determine which students should remain in English Language Learner (ELL) programs and which are ready to be exited. While the analysis revealed that the cutscores were set to maximize the benefits for the majority of students making these transitions, I did find that students under certain circumstances (Spanish-speakers leaving bilingual programs, $\tilde{1}5\%$ of the ELL population

Figure 5.2: Consequences of Heterogeneity in Treatment Groups



in the state) might have benefited from a modification to the cutpoint. Specifically, this particular subgroup of ELLs appeared to be exiting too early and struggled when removed from ELL services and placed in a mainstream classroom.

Figure 5.2 shows how heterogeneity in treatment (or student) groups might lead to problems when a single threshold is set for the entire population. Whereas students with treatment X are able to smoothly transition from treatment X to no treatment at point b, students with treatment Y would benefit more from a cutpoint set at c. In this illustration, without any constraints on costs, it would be most preferable for students at very low ranges of the running score to receive treatment X until they reach the point of intersection with the treatment Y line. At that point students would benefit most from being placed in treatment Y until they reach level c on the running score. However, in the example presented in Chapter 3, the type of treatment was mostly determined by the amount of resources available to a school and, therefore, the type of transition just described could not be offered in the majority of schools. Extending this example, one could imagine how additional groupings of students could complicate the scenario. For example, if, rather than a treatment group, each student

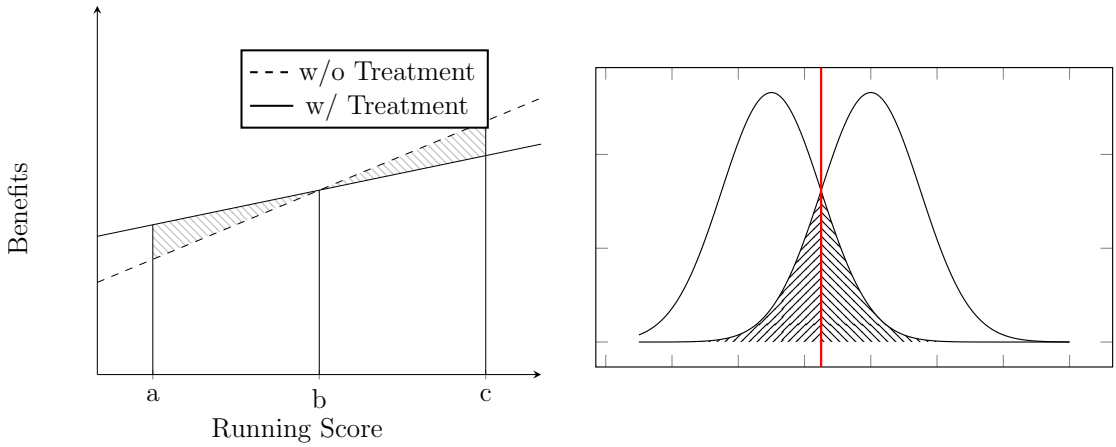
had a slightly different relationship between the running score and benefits potential, it would be nearly impossible to set a single cutpoint from a central level to make sure that all students find a smooth transition from treatment to no treatment. In this case, with heterogeneity in the relationship between the running score and benefits, it is necessary to make some modifications to the system to account for any students who do not benefit from the centrally decided threshold.

A top-down approach to threshold setting may be desired for targeted policy systems when consistency and standardization are desired. However, not all situations will require such rigidity. It should not be expected that the state- or federal-governments should account for *all* possible scenarios in their guidance in setting thresholds for determining targeted individuals. However, built-in flexibilities to the system can help alleviate any consequences that arise from the "one-size-fits-all" approach. Including parent/teacher feedback in such decisions can further ensure the accuracy of student placement. Further, a monitoring system for students who have recently exited treatment or those who scored just below the eligibility criteria can be put in place to make sure that these students are not falling behind. Under such a monitoring system, schools would be given additional resources to provide supports to a student if he or she is found to be struggling after recently exiting treatment or just missing eligibility criteria. The additional supports could be used to reverse the consequences that would have arisen due to misassignment of any students due to treatment or student heterogeneity.

3. *Costs and benefits associated with all possible thresholds can and should be considered in finding "optimal" separation.*

   In the event that the thresholds are not coming from a higher-level of the government, local governments can have control over where to set their thresholds that are specific for their context. While the concept of "optimal" cutpoints is not new, there is currently little guidance on setting cutpoints for targeted interventions at the lo-

Figure 5.3: Minimizing Misclassification Errors



cal level. The fourth chapter takes tools from the medical literature to show how one can account for the consequences of misassignment to treatment (false-positives and false-negatives) when developing targeting systems. The approach requires that a school district or state have longitudinal data that can be used in predictive modeling of certain events that are being targeted (e.g., graduation status). The optimization approach outlined in the fourth chapter not only is specifically aimed at limiting misclassification rates (false-positives and false-negatives), but also has built-in flexibility to account for different scenarios (e.g., prevalence and relative costs).

The misclassification errors that are being minimized can be illustrated in Figure 5.3. In the left panel, I show where misclassification would arise on the framework model introduced in Chapter 1. The shaded areas represent the range of students and costs of setting the thresholds at a point other than b. If the threshold were to be set at point a, the students with a running score between a and b would all be counted as misclassifications. While the optimization approach introduced in Chapter 4 cannot account for costs that vary as a function of the running score, average costs can be applied to attempt to minimize the misclassification errors. In the illustration, the point that would minimize misclassification would be at point b where the number of

misclassifications would be zero.

As brought up in the previous point, education systems are often dealing with heterogeneous populations where a single cutpoint may not work for all. One of the key takeaways of Chapter 4 is that the point b is not always the point where misclassification rates are zero. In fact, the right panel of Figure 5.3 presents a more realistic example in which the heterogeneity of student groups creates an overlap between target and non-target group populations. In this scenario, the minimization of misclassification errors becomes a less trivial exercise. The optimal cutpoint is not always set at the point that perfectly distinguishes between the two populations. In most cases, some misclassification errors will exist and should be expected even when using an optimized threshold. Various tools have been added to adjust for different assumptions on the relative costs of misclassification errors to yield a threshold that minimizes misclassification costs (when known). The extensive literature on finding optimal cutpoints on key diagnostic measures can and should be integrated into targeting systems. Chapter 4 begins to make this integration with the application to two examples in education.

As targeted policy interventions become more prevalent in education (e.g., EWI/EWS) it is crucial that we have a firm understanding of the consequences that could arise from such policies and how to limit them. This dissertation provides a start in developing this understanding. Across the chapters, I examine the ways in which selection criteria are set currently in education and provide suggestions for modifications in order to prevent misclassification. The analytic tools used can provide examples of how data and analysis can be used to improve targeting systems and limit the consequences of misassignment to treatment. While it is suggested that these data tools can be used to limit the consequences, educators will play a crucial role in making sure that all students are appropriately served.

BIBLIOGRAPHY

Allensworth, E. M., & Easton, J. Q. (2005). *The on-track indicator as a predictor of high school graduation*. Chicago, IL: Consortium on Chicago School Research, University of Chicago.

Arnold, M., & Lassmann, M. E. (2003). Overrepresentation of minority students in Special Education. *Education*, *124*(2), 230–236.

Artiles, A. J., Harry, B., Reschly, D. J., & Chinn, P. C. (2002). Over-identification of students of color in Special Education: A critical overview. *Multicultural Perspectives*, *4*(1), 3–10.

Balfanz, R. (2012). A path forward: Evidence-based approaches to educational policy and practice. *Journal of Applied Research on Children: Informing Policy for Children at Risk*, *3*(2), 1.

Balfanz, R., Wang, A., & Byrnes, V. (2010). Early warning indicator analysis: Tennessee. *Baltimore, MD: Johns Hopkins University*.

Bauer, L., Liu, P., Schanzenbach, D. W., & Shambaugh, J. (2018). Reducing chronic absenteeism under the Every Student Succeeds Act. *Strategy Paper for the Hamilton Project*. Retrieved March 15, 2019, from `https://www.attendanceworks.org/wp-content/uploads/2018/04/Hamilton_project_-reducing_chronic_absenteeism_under_the_every_student_succeeds_act.pdf`

Bollmer, J., Bethel, J., Munk, T., & Bitterman, A. (2011, October). *Methods for assessing racial/ethnic disproportionality in Special Education: A technical assistance guide* (Tech. Rep.). Westat Rockville, MD: Data Accountability Center. Retrieved February 10, 2016, from `http://www.isbe.net/spec-ed/pdfs/disproportionality_ta-guide.pdf`

Bowers, A. J., & Zhou, X. (2019). Receiver operating characteristic (ROC) area under the curve (AUC): A diagnostic measure for evaluating the accuracy of predictors of education outcomes. *Journal of Education for Students Placed at Risk*, *24*(1), 20–46.

Bruce, M., Bridgeland, J. M., Fox, J. H., & Balfanz, R. (2011). *On track for success: The use of early warning indicator and intervention systems to build a grad nation.* Washington, DC: Civic Enterprises.

Callahan, R., Wilkinson, L., & Muller, C. (2010). Academic achievement and course taking among language minority youth in US schools: Effects of ESL placement. *Educational Evaluation and Policy Analysis*, *32*(1), 84–117.

Callahan, R. M. (2005). Tracking and high school English Learners: Limiting opportunity to learn. *American Educational Research Journal*, *42*(2), 305–328.

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2017). rdrobust: Software for regression discontinuity designs. *Stata Journal*, *17*(12), 372–404.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust data-driven inference in the regression-discontinuity design. *Stata Journal*, *14*(4), 909–946.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications.* Cambridge: Cambridge University Press.

Carlson, D., & Knowles, J. E. (2016). The effect of English Language Learner reclassification on student ACT scores, high school graduation, and postsecondary enrollment: Regression discontinuity evidence from Wisconsin. *Journal of Policy Analysis and Management*, *35*(3), 559–586.

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2017). A practical introduction to regression discontinuity designs: Volume i. *Cambridge Elements: Quantitative and Computational Methods for Social Science (forthcoming)*.

Cattaneo, M. D., Jansson, M., & Ma, X. (2017a). rddensity: Manipulation testing based on density discontinuity. *Stata Journal*. Retrieved June 10, 2017, from `http://www-personal.umich.edu/~cattaneo/papers/Cattaneo-Jansson-Ma_2017_Stata.pdf`

Cattaneo, M. D., Jansson, M., & Ma, X. (2017b). Simple local regression distribution

estimators. *Working Paper*. Retrieved June 10, 2017, from `http://www-personal.umich.edu/~cattaneo/papers/Cattaneo-Jansson-Ma_2017_LocPolDensity.pdf`

Catts, H. W., Nielsen, D. C., Bridges, M. S., Liu, Y. S., & Bontempo, D. E. (2015). Early identification of reading disabilities within an RTI framework. *Journal of Learning Disabilities*, *48*(3), 281–297.

Collier, V. P., & Thomas, W. P. (2004). The astounding effectiveness of dual language education for all. *NABE Journal of Research and Practice*, *2*(1), 1–20.

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., ... Crouch, R. C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology*, *102*(2), 327.

Conger, D. (2010). Does Bilingual Education interfere with English-language acquisition? *Social Science Quarterly*, *91*(4), 1103–1122.

Coutinho, M. J., & Oswald, D. P. (2000). Disproportionate representation in Special Education: A synthesis and recommendations. *Journal of Child and Family Studies*, *9*(2), 135–156.

Coutinho, M. J., Oswald, D. P., & Best, A. M. (2002). The influence of sociodemographics and gender on the disproportionate identification of minority students as having learning disabilities. *Remedial and Special Education*, *23*(1), 49–59.

Cross, C. T., Donovan, M. S., et al. (2002). *Minority students in special and gifted education*. National Academies Press.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, *49*(2), 222–251.

Cummins, J. (1991). Interdependence of first-and second-language proficiency in bilingual children. In E. Bialystok (Ed.), *Language processing in bilingual children* (pp. 70–89). New York, NY: Cambridge University Press.

Cummins, J. (1992). Bilingual Education and English Immersion: The Ramirez report in theoretical perspective. *Bilingual Research Journal*, *16*(1-2), 91–104.

Davis, M., Herzog, L., & Legters, N. (2013). Organizing schools to address early warning indicators (EWIs): Common practices and challenges. *Journal of Education for Students Placed at Risk*, *18*(1), 84–100.

Delgado, C. E., & Scott, K. G. (2006). Comparison of referral rates for preschool children at risk for disabilities using information obtained from birth certificate records. *The Journal of Special Education*, *40*(1), 28–35.

Department of Education. (2016a). *Assistance to States for the Education of Children with Disabilities; Preschool Grants for Children With Disabilities* (Tech. Rep.). Retrieved May 8, 2017, from `https://www.gpo.gov/fdsys/pkg/FR-2016-12-19/pdf/2016-30190.pdf`

Department of Education. (2016b). *Issue brief: Early warning systems.* Retrieved March 12, 2019, from `https://www2.ed.gov/rschstat/eval/high-school/early-warning-systems-brief.pdf`

Duncan, G. J., Yeung, W. J., Brooks-Gunn, J., & Smith, J. R. (1998). How much does childhood poverty affect the life chances of children? *American Sociological Review*, *63*(3), 406–423.

Dunn, L. M. (1968). Special Education for the mildly retarded: Is much of it justifiable? *Exceptional Children*, *35*(1), 5–22.

Enayati, H. A. (2014). The impact of disproportionality regulations on identification into Special Education programs. *Job Market Paper*. Retrieved January 25, 2016, from `https://www.msu.edu/~enayatih/Enayati_Job_Market_Paper.pdf`

Faria, A.-M., Sorensen, N., Heppen, J., Bowdon, J., Taylor, S., Eisner, R., & Foster, S. (2017). Getting students on track for graduation: Impacts of the early warning intervention and monitoring system after one year. REL 2017-272. *Regional Educational Laboratory Midwest*.

Frandsen, B. R. (2017). Party bias in union representation elections: Testing for manipulation in the regression discontinuity design when the running variable is discrete. In M. D. Cattaneo & J. C. Escanciano (Eds.), *Regression discontinuity designs: Theory*

and applications (pp. 281–315). Bingley: Emerald.

Frazelle, S., & Nagel, A. (2015). A practitioner's guide to implementing early warning systems. REL 2015-056. *Regional Educational Laboratory Northwest*.

Galen, R., & Gambino, S. (1975). *Beyond Normality: The Predictive Value and Efficiency of Medical Diagnoses*. John Wiley & Sons. Retrieved March 15, 2019, from `https://books.google.com/books?id=bGhrAAAAMAAJ`

Geisser, S. (1998). Comparing two tests used for diagnostic or screening purposes. *Statistics & Probability Letters*, *40*(2), 113–119.

Gersten, R., Fuchs, L. S., Williams, J. P., & Baker, S. (2001). Teaching reading comprehension strategies to students with learning disabilities: A review of research. *Review of Educational Research*, *71*(2), 279–320.

Gersten, R., Newman-Gonchar, R., Haymond, K. S., & Dimino, J. (2017). What is the evidence base to support reading interventions for improving student outcomes in grades 1-3? REL 2017-271. *Regional Educational Laboratory Southeast*.

Gottfried, M. A. (2009). Excused versus unexcused: How student absences in elementary school affect academic achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 392–415.

Gottfried, M. A. (2014). Chronic absenteeism and its effects on students' academic and socioemotional outcomes. *Journal of Education for Students Placed at Risk*, *19*(2), 53–75.

Gottfried, M. A. (2019). Chronic absenteeism in the classroom context: Effects on achievement. *Urban Education*, *54*(1), 3–34.

Gottfried, M. A., & Kirksey, J. J. (2017). "When" students miss school: The role of timing of absenteeism on students' test performance. *Educational Researcher*, *46*(3), 119–130.

Government Accountability Office. (2013). *Individuals with Disabilities Education Act: Standards Needed to Improve Identification of Racial and Ethnic Overrepresentation in Special Education* (Tech. Rep.). Retrieved July 10, 2016, from `http://www.gao.gov/assets/660/652437.pdf`

Greenberg Motamedi, J., Singh, M., & Thompson, K. D. (2016). English Learner student characteristics and time to reclassification: An example from Washington state. REL 2016-128. *Regional Educational Laboratory Northwest*.

Greiner, M., Pfeiffer, D., & Smith, R. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine*, *45*(1-2), 23–41.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297.

Grissom, J. B. (2004). Reclassification of English Learners. *Education Policy Analysis Archives*, *12*(36). Retrieved June 11, 2017, from `https://epaa.asu.edu/ojs/article/view/191`

Grunow, A. (2011). *Access to English: Resources for Spanish-speaking children's academic development* (Doctoral dissertation, Stanford University). Retrieved February 24, 2013, from `https://searchworks.stanford.edu/view/9238377`

Gurantz, O., & Borsato, G. (2012). Building and implementing a college readiness indicator system: Lessons from the first two years of the CRIS Initiative. *Voices in Urban Education*, *35*, 5–15.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2002). Inferring program effects for special populations: Does Special Education raise achievement for students with disabilities? *Review of Economics and Statistics*, *84*(4), 584–599.

Harry, B., & Anderson, M. G. (1994). The disproportionate placement of African American males in Special Education programs: A critique of the process. *Journal of Negro Education*, *63*(4), 602–619.

Harry, B., Klingner, J., Delpit, L. D., & Artiles, A. J. (2014). *Why are so many minority students in Special Education? : Understanding race and disability in schools (2nd ed.)*. New York: Teachers College Press.

Hernandez, J. E., Ramanathan, A. K., Harr, J., & Socias, M. (2008). Study of the effects of an intervention to reduce the disproportionate identification in the category of emotional

disturbance in the Los Angeles Unified School District. *Journal of Special Education Leadership*, *21*(2), 64–74.

Hibel, J., Farkas, G., & Morgan, P. L. (2010). Who is placed into Special Education? *Sociology of Education*, *83*(4), 312–332.

Hill, L. E., Weston, M., & Hayes, J. M. (2014). *Reclassification of English Learner students in California*. San Francisco, CA: Public Policy Institute of California (PPIC).

Ikbal, S., Tamhane, A., Sengupta, B., Chetlur, M., Ghosh, S., & Appleton, J. (2015). On early prediction of risks in academic performance for students. *IBM Journal of Research and Development*, *59*(6), 5:1–5:14.

Jacob, R., Zhu, P., Somers, M.-A., & Bloom, H. (2012). *A practical guide to regression discontinuity*. New York, NY: MDRC.

Jenkins, J. R., & Johnson, E. (2008). Universal screening for reading problems: Why and how should we do this. *RTI Action Network*. Retrieved March 15, 2019, from `http://www.rtinetwork.org/essential/assessment/screening/readingproblems`

Kleinman, L. C., & Norton, E. C. (2009). What's the risk? a simple approach for estimating adjusted risk measures from nonlinear models including logistic regression. *Health Services Research*, *44*(1), 288–302.

Klingner, J., & Harry, B. (2006). The Special Education referral and decision-making process for English Language Learners: Child study team meetings and placement conferences. *The Teachers College Record*, *108*(11), 2247–2281.

Kumar, R., & Indrayan, A. (2011). Receiver operating characteristic (ROC) curve for medical researchers. *Indian Pediatrics*, *48*(4), 277–287.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(2), 281–355.

Li, D. L., Shen, F., Yin, Y., Peng, J. X., & Chen, P. Y. (2013). Weighted Youden index and its two-independent-sample comparison based on weighted sensitivity and specificity. *Chinese Medical Journal*, *126*(6), 1150–1154.

López-Ratón, M., Rodríguez-Álvarez, M. X., Suarez, C. C., & Sampedro, F. (2014). Optimal-

Cutpoints: An R package for selecting optimal cutpoints in diagnostic tests. *Journal of Statistical Software*, *61*(8), 1–36.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, *54*(2), 588–599.

Mac Iver, M. A. (2013). Early warning indicators of high school outcomes. *Journal of Education for Students Placed at Risk*, *18*(1), 1–6.

MacMillan, D. L., & Reschly, D. J. (1998). Overrepresentation of minority students the case for greater specificity or reconsideration of the variables examined. *The Journal of Special Education*, *32*(1), 15–24.

Magnuson, K. A., & Votruba-Drzal, E. (2009). Enduring influences of childhood poverty. In M. Cancian & S. Danziger (Eds.), *Changing poverty, changing policies* (p. 153-179). New York, NY: Russell Sage Foundation.

Mahoney, K. S., & MacSwan, J. (2005). Reexamining identification and reclassification of English Language Learners: A critical discussion of select state practices. *Bilingual Research Journal*, *29*(1), 31–42.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*(2), 698–714.

McLoyd, V. C. (1998). Socioeconomic disadvantage and child development. *American Psychologist*, *53*(2), 185–204.

Metz, C. E. (1978). Basic principles of ROC analysis. In *Seminars in nuclear medicine* (Vol. 8, pp. 283–298).

Morgan, P. L., Farkas, G., Hillemeier, M. M., & Maczuga, S. (2012). Are minority children disproportionately represented in early intervention and early childhood Special Education? *Educational Researcher*, *41*(9), 339–351.

Morgan, P. L., Farkas, G., Hillemeier, M. M., Mattison, R., Maczuga, S., Li, H., & Cook, M. (2015). Minorities are disproportionately underrepresented in Special Education: Longitudinal evidence across five disability conditions. *Educational Researcher*, *44*(5), 278–292.

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2008). A propensity score matching analysis of the effects of Special Education services. *The Journal of Special Education*, *43*(4), 236–254.

Morrissey, T. W., Hutchison, L., & Winsler, A. (2014). Family income, school attendance, and academic achievement in elementary school. *Developmental Psychology*, *50*(3), 741–753.

NCES. (2013). Digest of education statistics, 2012. NCES 2014-015. *National Center for Education Statistics*.

Neild, R. C., Balfanz, R., & Herzog, L. (2007). An early warning system. *Educational Leadership*, *65*(2), 28–33.

Norton, E. C., Miller, M. M., & Kleinman, L. C. (2013). Computing adjusted risk ratios and risk differences in Stata. *Stata Journal*, *13*(3), 492–509.

Ofiesh, N. (2006). Response to intervention and the identification of Specific Learning Disabilities: Why we need comprehensive evaluations as part of the process. *Psychology in the Schools*, *43*(8), 883–888.

Parrish, T. (2002). Racial disparities in the identification, funding, and provision of Special Education. In D. J. Losen & G. Orfield (Eds.), *Racial inequity in special education* (pp. 15–37). Cambridge, MA: Harvard Education Publishing Group.

Parrish, T. B., Merickel, A., Pérez, M., Linquanti, R., Socias, M., Spain, A., ... Delancey, D. (2006). Effects of the implementation of Proposition 227 on the education of English learners, K-12: Findings from a five-year evaluation: Final report. *American Institutes For Research*.

Payán, R. M., & Nettles, M. T. (2006). *Current state of English-Language Learners in the US K-12 student population*. Princeton, NJ: Educational Testing Service.

Perkins, N. J., & Schisterman, E. F. (2006). The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology*, *163*(7), 670–675.

Planty, M. (2010). *Understanding education indicators: A practical primer for research and*

policy. New York: Teachers College Press.

Podell, D. M., & Soodak, L. C. (1993). Teacher efficacy and bias in Special Education referrals. *The Journal of Educational Research*, *86*(4), 247–253.

Porter, R. P. (1996). *Forked tongue: The politics of bilingual education*. New Brunswick, NJ: Transaction Publishers.

Robinson, J. P. (2011). Evaluating criteria for English Learner reclassification: A causal-effects approach using a binding-score regression discontinuity design with instrumental variables. *Educational Evaluation and Policy Analysis*, *33*(3), 267–292.

Robinson-Cimpian, J. P., & Thompson, K. D. (2015). The effects of changing test-based policies for reclassifying English Learners. *Journal of Policy Analysis and Management*, *35*(2), 279–305.

Rosenthal, B. M. (2016, December). HISD's focus on "over-identification" of Black students backfires. *Houston Chronicle*. Retrieved May 9, 2016, from `http://www.houstonchronicle.com/news/houston-texas/houston/article/HISD-s-focus-on-over-identification-of-10821607.php?t=feb29e3481`

Roy, L. (2012). On measures of racial/ethnic disproportionality in Special Education: An analysis of selected measures, a joint measures approach, and significant disproportionality. *California Department of Education Report*. Retrieved January 26, 2016, from `ftp://ftp.cde.ca.gov/sp/se/ds/Disproportionality%20Paper%20June%202012.pdf`

Rücker, G., & Schumacher, M. (2010). Summary ROC curve based on a weighted Youden index for selecting an optimal cutpoint in meta-analysis of diagnostic accuracy. *Statistics in Medicine*, *29*(30), 3069–3078.

Samuels, C. A. (2004, December). Renewed IDEA Targets Minority Overrepresentation. *Education Week*. Retrieved May 9, 2016, from `http://www.edweek.org/ew/articles/2004/12/08/15idea.h24.html`

Schisterman, E. F., & Perkins, N. (2007). Confidence intervals for the Youden index and corresponding optimal cut-point. *Communications in Statistics – Simulation and Com-*

putation, *36*(3), 549–563.

Schisterman, E. F., Perkins, N. J., Liu, A., & Bondell, H. (2005). Optimal cut-point and its corresponding Youden index to discriminate individuals using pooled blood samples. *Epidemiology*, *16*(1), 73–81.

Shifrer, D., Muller, C., & Callahan, R. (2011). Disproportionality and learning disabilities: Parsing apart race, socioeconomic status, and language. *Journal of Learning Disabilities*, *44*(3), 246–257.

Skiba, R. J., Simmons, A. B., Ritter, S., Gibb, A. C., Rausch, M. K., Cuadrado, J., & Chung, C. G. (2008). Achieving equity in Special Education: History, status, and current challenges. *Exceptional Children*, *74*(3), 264–288.

Stuit, D., O'Cummings, M., Norbury, H., Heppen, J., Dhillon, S., Lindsay, J., & Zhu, B. (2016). *Identifying early warning indicators in three Ohio school districts.* Washington, DC: US Department of Education, Institute of Education Sciences.

Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psychological Bulletin*, *99*(2), 181–198.

Theobald, R. (2015). *Response to intervention? Estimating the causal effect of Special Education services on student performance* (Doctoral dissertation, University of Washington). Retrieved December 15, 2015, from `https://digital.lib.washington.edu/researchworks/bitstream/handle/1773/34191/Theobald_washington_0250E_15181.pdf?sequence=1`

Thompson, K. D. (2015). English Learners' time to reclassification: An analysis. *Educational Policy*, *31*(3), 330-363.

Uekawa, K., Merola, S., Fernandez, F., & Porowski, A. (2010). Creating an early warning system: Predictors of dropout in Delaware. REL Mid-Atlantic technical assistance brief. REL MA 1.2. 75-10. *Regional Educational Laboratory Mid-Atlantic*.

Umansky, I. M. (2016). To be or not to be EL: An examination of the impact of classifying students as English Learners. *Educational Evaluation and Policy Analysis*, *38*(4), 714–

737.

Umansky, I. M., & Reardon, S. F. (2014). Reclassification patterns among Latino English Learner students in Bilingual, Dual Immersion, and English Immersion classrooms. *American Educational Research Journal*, *51*(5), 879–912.

Waitoller, F. R., Artiles, A. J., & Cheney, D. A. (2009). The miner's canary: A review of overrepresentation research and explanations. *The Journal of Special Education*, *44*(1), 29–49.

What Works Clearinghouse. (2009). *Lexia reading: Intervention report* (Tech. Rep.). Washington, DC. Retrieved March 15, 2019, from `https://ies.ed.gov/ncee/wwc/Docs/InterventionReports/wwc_lexia_063009.pdf`

What Works Clearinghouse. (2016). *Procedures and Standards Handbook Version 3.0* (Tech. Rep.). Retrieved December 15, 2016, from `http://ies.ed.gov/ncee/wwc/Handbooks`

What Works Clearinghouse. (2017). *Leveled literacy intervention: Intervention report* (Tech. Rep.). Washington, DC. Retrieved March 15, 2019, from `https://ies.ed.gov/ncee/wwc/Docs/InterventionReports/wwc_leveledliteracy_091917.pdf`

Williams, B. A., Mandrekar, J. N., Mandrekar, S. J., Cha, S. S., & Furth, A. F. (2006). *Finding optimal cutpoints for continuous covariates with binary and time-to-event outcomes* (Tech. Rep.). Rochester, MN. Retrieved March 15, 2019, from `https://www.mayo.edu/research/documents/biostat-79pdf/doc-10027230`

Yoshida, R. K. (1980). Multidisciplinary decision making in Special Education: A review of issues. *School Psychology Review*, *9*(3), 221–227.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35.