Predictive Utility of a Proficiency Cut Score in a Benchmark Assessment

by

Takeshi Terada

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2021 by the
Graduate Supervisory Committee:

Ying-Chih Chen, Chair
Michael Edwards, Co-chair
David Garcia

ARIZONA STATE UNIVERSITY
August 2021
ABSTRACT

Since the No Child Left Behind (NCLB) Act required classifications of students’ performance levels, test scores have been used to measure students’ achievement; in particular, test scores are used to determine whether students reach a proficiency level in the state assessment. Accordingly, school districts have started using benchmark assessments to complement the state assessment. Unlike state assessments administered at the end of the school year, benchmark assessments, administered multiple times during the school year, measures students’ learning progress toward reaching the proficiency level. Thus, the results of the benchmark assessments can help districts and schools prepare their students for the subsequent state assessments so that their students can reach the proficiency level in the state assessment. If benchmark assessments can predict students’ future performance measured in the state assessments accurately, the assessments can be more useful to facilitate classroom instructions to support students’ improvements. Thus, this study focuses on the predictive accuracy of a proficiency cut score in the benchmark assessment. Specifically, using an econometric research technique, Regression Discontinuity Design, this study assesses whether reaching a proficiency level in the benchmark assessment had a causal impact on increasing the probability of reaching a proficiency level in the state assessment. Finding no causal impact of the cut score, this study alternatively applies a Precision-Recall curve - a useful measure for evaluating predictive performance of binary classification. By using this technique, this study calculates an optimal proficiency cut score in the benchmark assessment that maximizes the accuracy and minimizes the inaccuracy in predicting the proficiency level in the state assessment. Based on the results, this study discusses issues regarding the conventional approaches of establishing cut scores in
large-scale assessments and suggests some potential approaches to increase the predictive accuracy of the cut score in benchmark assessments.
ACKNOWLEDGEMENTS

First, I would like to thank my primary advisor, Dr. Ying-Chih Chen, for accepting me as his advisee and mentoring me during the PhD program. He gave me so many suggestions to help me complete the PhD program. In particular, his advice helped me consider what I should do as a PhD student and how I should proceed in the PhD program. My research experience under his supervision taught me so many lessons on academic research in education. I would also like to thank my co-advisor, Dr. Michael Edwards, for providing me with so much technical and meaningful advice on my dissertation and other research issues. Because he accepted me as his “extra” student, I had great opportunities to discuss and learn about many research issues. His in-depth knowledge in quantitative research helped me deepen my understanding of quantitative research methodology. Dr. David Garcia also provided me with very valuable suggestions from an educational policy perspective as my committee member. His insightful vision helped me broaden my understanding of educational policy and consider my dissertation research from a variety of perspectives.

I really appreciate my best friend, Takeshi Yanagiura. As my close friend, he gave me countless suggestions and advice on my dissertation research and academic life during my graduate career. Since we got to know each other in graduate school at the University of Minnesota, it has always been so valuable and refreshing to chat with him. My dissertation research idea came out of a conversation with him. I cannot thank him enough for his support, insight, and words of encouragement.

I also appreciate the benchmark assessment company, Edmentum, for allowing me to use their data for my dissertation. Without a summer internship opportunity offered by this company,
I could not have completed my dissertation research. The internship was a remarkable experience for me to enrich my psychometric research skills and deepen my understanding of the roles of benchmark assessments. I am very grateful for Edmentum’s psychometric research staff for their valuable research opportunities.

Finally, I deeply appreciate my parents for their constant support. I will never forget that I submitted my application to this PhD program on the same day as the funeral for my mother, and that I accepted the offer of admission and emailed the Mary Lou Fulton Teachers College on her birthday. I hope she is happy about my accomplishment from the Shinnyo spiritual realm.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>List</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Problem Statement</td>
<td>3</td>
</tr>
<tr>
<td>Purpose of Study</td>
<td>5</td>
</tr>
<tr>
<td>Rationales</td>
<td>8</td>
</tr>
<tr>
<td>Predictive</td>
<td>10</td>
</tr>
<tr>
<td>Significance</td>
<td>14</td>
</tr>
<tr>
<td>2 LITERATURE REVIEW</td>
<td>17</td>
</tr>
<tr>
<td>Benchmark Assessment</td>
<td>19</td>
</tr>
<tr>
<td>Proficiency Level</td>
<td>23</td>
</tr>
<tr>
<td>3 METHODS</td>
<td>37</td>
</tr>
<tr>
<td>Evaluation of Proficiency Cut Score</td>
<td>37</td>
</tr>
<tr>
<td>Regression Discontinuity Design</td>
<td>38</td>
</tr>
<tr>
<td>Precision-Recall Curve</td>
<td>45</td>
</tr>
<tr>
<td>4 RESULTS</td>
<td>55</td>
</tr>
<tr>
<td>Results of Descriptive Statistics</td>
<td>55</td>
</tr>
<tr>
<td>Results of Regression Discontinuity Design</td>
<td>66</td>
</tr>
<tr>
<td>Results of Precision-Recall Curve</td>
<td>73</td>
</tr>
<tr>
<td>CHAPTER</td>
<td>Page</td>
</tr>
<tr>
<td>---------------------</td>
<td>------</td>
</tr>
<tr>
<td>5 DISCUSSION</td>
<td>77</td>
</tr>
<tr>
<td>Findings</td>
<td>77</td>
</tr>
<tr>
<td>Limitations</td>
<td>81</td>
</tr>
<tr>
<td>Educational Significance</td>
<td>84</td>
</tr>
<tr>
<td>Future Research</td>
<td>88</td>
</tr>
<tr>
<td>Conclusion</td>
<td>91</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>93</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table                                                                                                                                           Page
1. Precision Recall Matrix..................................................................................................48
2. Precision Recall Matrix in the Benchmark and State Assessment .............................50
3. Precision Recall Matrix in the Example .....................................................................53
4. Results of Overall Benchmark assessment by performance level ..............................56
5. Results of Overall State Assessment by Performance Level ......................................57
6. Cut Score Ranges in Benchmark and State Assessments ...........................................59
7. Numbers and Percentages of Students by Performance Level in Benchmark and State
    Assessments ..................................................................................................................59
8. Descriptive Statistics of Sampled Students near the Cut Score of 55 .......................68
9. Result of Checks for Discontinuities at the Proficiency Cut Score .............................69
10. Average Scores and Percentages of Proficient Students near Cut Scores of 55 and 65 for
    Benchmark and State Assessments ..............................................................................71
11. Descriptive Statistics of Sampled Students near a Score of 65 .................................72
12. Results of Checks for Discontinuities at a Score of 65 ..............................................73
13. Percentages of proficient students accurately and inaccurately predicted to be proficient
    in the state assessment .................................................................................................76
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Linear Regression Lines with/without any Discontinuity</td>
<td>39</td>
</tr>
<tr>
<td>2. Example of Precision Recall Curve</td>
<td>48</td>
</tr>
<tr>
<td>3. Q-Q Plot of Residuals of the Regression</td>
<td>62</td>
</tr>
<tr>
<td>4. Histogram of Residuals of the Regression</td>
<td>63</td>
</tr>
<tr>
<td>5. Plot of Standardized Residuals for State and Benchmark Assessment Scores</td>
<td>63</td>
</tr>
<tr>
<td>6. Scatter Plot of Benchmark and State Assessment Scores</td>
<td>66</td>
</tr>
<tr>
<td>7. Precision-Recall Curve for an Optimal Proficiency Cut Score in the Benchmark Assessment</td>
<td>74</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

Educational assessment has been used for a long time to inform decisions about student learning, curriculum instruction, and educational policy in K-12 education in the U.S. (AERA et al., 2014). When students’ knowledge and skills are measured based on a specific set of academic standards (e.g., state-defined standards), assessment results are even more important in serving a variety of purposes such as improving classroom instructions to meet students’ needs, evaluating curriculum and instruction in the district- or school levels, identifying students, teachers, and schools who need support, and predicting likelihoods of student’s success on a summative assessment (AERA et al., 2014; Herman et al., 2010; Immekus & Atitya, 2016; Lazarín, 2014; Perie et al., 2009). In particular, assessment scores are frequently used to evaluate the status, progress, or accomplishments of student learnings as well as school, school districts, and states (e.g., AERA et al., 2014; Figlio & Loeb, 2011; Ho, 2007; Koretz, 2008; Perie et al., 2009).

In terms of test-based evaluation, state test scores have mainly been used in the educational policy decisions as the enactment of the Improving American’s School Act of 1994, (Perie et al., 2009). Furthermore, since the No Child Left Behind (NCLB) Act required classifications of students’ performance levels, state-wide assessment scores are even more importantly used to measure and monitor students’ academic performance (e.g., Herman, et al., 2010; Immekus & Atitya, 2016; Lazarín, 2014; Perie et al., 2009). When assessment scores are used in such a way as identifying students’ performance levels, assessment results are even more significantly used for other high-stakes policy decisions such as assigning students to educational
programs or courses, including English Language Learner (ELL) course and special education programs, (Carlson & Knowles, 2016; Reyes & Hwang, 2019; Umansky, 2016) or retaining low-performing students in their grades (Greene & Winters, 2007; Greene & Winters, 2009; Özek, 2015; Schwerdt et al., 2017).

On the other hand, many schools and districts have started using benchmark assessments to complement state assessments (e.g., Brown & Coughlin, 2007; Bulkley, Christman, et al., 2010; Diao & Sireci, 2018; Herman et al., 2010; Linn, 2011; Wheadon, 2014). The benchmark assessments are developed to predict students’ performance on the end-of-the-year state assessments (Bulkley, Oláh, et al., 2010; Olson, 2005). Accordingly, districts and local classroom teachers use the results of the benchmark assessments as additional information resources to monitor student learning progress and to identify who needs instructional supports in order to reach higher performance levels of future state assessments (Brown & Coughlin, 2007). If benchmark assessments can do a reliable job of accurately predicting students’ future academic performance levels to be measured in future state assessments, the benchmark assessment results can be used to facilitate classroom instructions to support students’ learning progress measured in the subsequent state assessment.

In terms of the predictive accuracy of benchmark assessments, there is a limited number of studies that focus on statistical properties of benchmark assessments in predicting students’ performance in state assessments (Babo at al., 2014; Brown & Coughlin, 2007; Immekus & Atitya, 2016). If the predictive accuracy of the benchmark assessments is supported, classroom teachers can use benchmark assessment scores to predict their students’ performance levels determined by state assessment scores in advance. To support the use of benchmark assessments
in order to adapt the classroom teachers’ instructional practice and curriculum development to better meet the needs of students, it is necessary to increase the accuracy of benchmark assessments in predicting the performance levels measured in the subsequent state assessments.

In the following sections, I would like to address problem statement, the purpose of my study, and research questions. Then, I will describe the rationales of the study by introducing some technical and theoretical issues that are directly related to predictive accuracy of benchmark assessments. In the last section, I will explain how my research will contribute to policy issues of interest as the significance of the study.

Note that I use both terms, test and assessment in my paper, sometimes interchangeably. Although the latter term encompasses broader sources of information, it is often the case that both terms are used interchangeably in many educational research articles (AERA at al., 2014). Therefore, I use both terms interchangeably in the following sections.

**Problem Statement**

Since A Nation at Risk, was published during the Reagan administration in 1983, the standard-based educational reform movement has been pushed and influenced federal and state educational policy (Papay et al., 2010). In particular, the development of high-stakes tests to monitor the progress of student academic performance to master state-defined standards has been more and more important in the educational reform efforts (Papay at al., 2010). In such test-based accountability systems, states have held their local districts and schools accountable for students’ academic performances. To respond to the accountability purpose, states are required to decide the performance levels to be reported and the clear definitions of each level in the assessment. In particular, since the NCLB started, it has been very important for states to identify
students who are at proficient or higher levels and who are below the proficient level through assessment results (Brown & Coughlin, 2007; Ho, 2008; Zieky et al., 2006). However, in terms of a proficiency level, there are some critical concerns regarding the interpretation of proficiency rates. The biggest concern is about the classification of students who are near the cut score. Classification decisions on educational assessment are typically made by comparing a student’s test score to cut scores set in the process of standard-settings (Diao & Sireci, 2018). If the cut score is too high, some students who deserve being proficient can be classified as below proficient. If the cut score is too low, students who should not be proficient can be classified as proficient. In other words, setting an appropriate cut score is essential in determining students’ proficiency levels. Therefore, to distinguish clearly students who are at a performance level from students who are in the next lower or higher level, it is very critical to verify the appropriateness of the cut scores in the assessment as the validation process (AERA et al., 2014; Brown & Coughlin, 2007; Zieky & Perie, 2006).

Based on the importance of cut scores in the assessment, this study attempts to evaluate the location of a proficiency cut score in the benchmark assessment. If a proficiency cut score is set in the ‘right’ spot, the benchmark assessment can differentiate proficient students from non-proficient students based on the proficiency cut score. Additionally, if students who are classified as proficient in the benchmark assessment are also proficient in the state assessment, it means that the benchmark assessment can be used to predict a proficiency level in the state assessment. That is, students who are proficient in the benchmark assessment should be proficient in the state assessment. If the cut score is in the ‘wrong’ spot, students who are proficient in the benchmark assessment can be classified as non-proficient in the state assessment and vice versa. For
example, if the probability of achieving at a proficiency level in the future state assessment is statistically equivalent between students who barely reach the proficiency level in the benchmark assessment and those who barely fail, the cut score needs to be changed to improve its accuracy of predicting the proficiency level in the state assessment. If this probability of reaching the proficiency level in the state assessment is statistically different between students who are proficient and not proficient in the benchmark assessment, the proficiency cut score differentiates students’ proficiency based on the probabilities of reaching the proficiency level in the state assessment. Therefore, setting the proficiency cut score in the benchmark assessment should be determined based on that probability to increase its predictive accuracy.

However, the current approach of setting a proficiency cut score is not based on statistical evidence of such probabilities, but only on judges’ subjective decisions. That is, the proficiency cut score is established regardless of the probability of reaching the proficiency level in the subsequent state assessment. Now that school districts and schools use the benchmark assessment results to predict students’ future performance in state assessments, setting the proficiency cut score in the benchmark assessment should not be based on the judgment of policymakers, educators, measurement professionals, and other experts. Instead, it should involve students’ probabilities of reaching the proficiency level in the subsequent state assessment. Hence, by focusing on the probability of reaching the proficiency level in the state assessment, this research will investigate the predictive accuracy and the location of the cut score with regard to the proficiency level in the benchmark assessment.

**Purpose of Study and Research Questions**
This study will use the data of 459 students in the 7th grade who took both the English Language Arts (ELA) benchmark and the end-of-year state assessments in the school year of 2017-18. The participants are public charter school students in the northeast region of the U.S. The benchmark assessment has 28 multiple-choice items with two constructed-response items. Because the constructed-response items were rated by local teachers who did not receive any training for grading students’ responses, those scores are excluded in the data analysis. The rest of the 28 multiple-choice items, which have a scale of 0-1 points, will be analyzed. The state assessment has 63 multiple-choice items that have a scale of 0-1 points. Both assessments indicate similar four performance levels (i.e., below basic, basic, proficient, and advanced).

The purpose of this study is twofold: to assess the accuracy of a cut score of a proficiency level in the benchmark assessment in predicting the proficiency level in the state assessment and to find an optimal proficiency cut score in the benchmark assessment in predicting the proficiency level in the state assessment. To achieve these goals, this study will respond to the following questions:

1. Does a proficient level determined by a proficiency cut score in the benchmark assessment predict a proficiency level based on the state assessment?

1.1. Are probabilities of reaching the proficiency level in the state assessment statistically equivalent or different between students who barely reach the proficiency level in the benchmark assessment and students who barely fail?

1.2. Does reaching a proficiency level in the benchmark assessment have any causal impact on the probability of reaching a proficiency level in the state assessment?
2. What is the most optimal proficiency cut score in the benchmark assessment in predicting a proficient level in the state assessment accurately?

The first question focuses on the predictive accuracy of a proficiency cut score in the benchmark assessment. Specifically, it aims to test whether the probability of reaching a proficiency level in the state assessment is statistically equivalent between students who barely reach a proficiency level in the benchmark assessment and students who barely fail. If this probability is statistically different between students who reach the level and did not, the proficiency cut score of the benchmark assessment can differentiate proficient students from non-proficient students. If the probability is statistically not different, the proficiency cut score does not differentiate students with regard to a proficiency status. To evaluate this role of distinguishing proficient and non-proficient students in the benchmark assessment, this research will use one of the quasi-experimental design methods, Regression Discontinuity Design (RDD). This method is a very useful tool in assessing the causal impacts of any intervention on the outcome. In this research, RDD is applied to evaluate the causal impact of reaching a proficiency level in the benchmark assessment on the probability of reaching a proficiency level in the state assessment. The second question is about an optimal proficiency cut score. If a proficiency cut score is too high in the benchmark assessment, some students who should be proficient in the state assessment are not predicted to be proficient in the benchmark assessment. If the cut score is too low, students who should not be proficient in the state assessment are predicted to be proficient in the benchmark assessment. To find a “right” cut score that is neither too high nor too low, this research will use Precision-Recall (PR) curves, which is one of the machine learning approaches. The PR curve is a useful measure to evaluate binary classification models (e.g., pass/fail, agree/disagree, and
positive/negative). By using PR curve measures, this research will calculate an optimal proficiency cut score in which a correct classification (i.e., proficiency and non-proficiency) is maximized, and an incorrect one is minimized. For both RDD and PR curves, detailed descriptions of these methods and related past research are provided in the chapter of methodology.

**Rationales**

Educational testing has been used for a long time to inform decisions about student learning, curriculum instruction, and educational policy (AERA et al., 2014). According to standards published by AERA (2014), there are three major purposes of educational tests:

1) To make inferences that inform teaching and learning at the individual or curricular level.
2) To make inferences about outcomes for individual students and groups of students
3) To inform decisions about students, such as certifying students’ acquisition of particular knowledge and skills for promotion, placement in special instructional programs, or graduation (p.184).

In response to these purposes, assessment results are often used to evaluate the status, progress, or accomplishments of students as well as schools, school districts, and states. In particular, when students’ knowledge and skills are relative to a specific set of academic standards (e.g., state-defined standards), assessment results need to serve a variety of purposes such improving classroom instructions to meet students’ needs, evaluating curriculum and instruction, identifying students, teachers, and schools who need support, and predicting the likelihood of student’s success on a summative assessment (AERA et al., 2014).
When assessment results are used not only for evaluating students’ performance levels relative to state- or district-defined standards but also for assigning students to specific educational programs or courses such as ELL courses and special education programs, it is not sufficient to provide the evidence of validity regarding assessment scores. Rather than this, some supportive evidence regarding the classification decisions is necessary to show which categories or programs would be the best fit for students to be assigned to (AERA et al., 2014). If unintended consequences result from such test-based decisions (e.g., students are assigned to the wrong programs or classified as having a higher performance level than the students can reach), it means that the assessment itself or the assessment score interpretation may be considered invalid. To prevent such unintended consequences from happening, it is highly important to provide supporting evidence regarding the validity of classification decisions based on assessment scores through statistical analysis.

When it comes to supporting evidence with respect to test-based classification decisions, it is necessary to provide some statistical evidence to support test-based educational policy decisions using performance levels determined in the assessment. For example, when assessment results of student performance levels are used to determine which performance categories or specific programs students should fit or be assigned to, students need to be accurately classified as the right performance level. To support the classification decisions regarding performance levels in the assessment, it is necessary to specify cut scores used to determine students’ performance levels such as “basic” and “proficient”. If students whose scores are at the cut score and students whose scores are one-point below the cut score have essentially equal academic skills, this cut score needs to be changed. Therefore, this study aims to investigate the utility of a
proficiency cut score to support classification decisions based on assessment scores. By using both benchmark and state assessment scores, this research will present statistical evidence on whether proficiency cut scores are set at the right point to support the classification decisions.

Predictive Accuracy and Validity Evidence

One of the primary roles of benchmark assessments is to help students improve their academic performances in order to reach state-defined standards. In the current standard-based accountability system, students are expected to reach the proficient level in the state assessments. Local school districts and schools use benchmark assessments to help achieve this goal. By monitoring the progress of students’ performance through the benchmark assessments, the districts and schools attempt to identify who is on the right track toward reaching the proficient level in the state assessment and to know how much more the students who are not on the right track need to improve. If the results of the benchmark assessments are not accurate enough to predict the performance measured in the state assessments, the benchmark assessments do not support district and school efforts to help their students reach the proficiency level. In this sense, the predictive accuracy of the benchmark assessments is one of the essential factors in deciding whether benchmark assessments should be used for high-stake policy decisions.

Predictive Accuracy and Validity

With regard to predictive validity, this research considers that a significant correlation between test scores and other test scores provides only circumstantial evidence of validity. Certainly, if the benchmark assessment is closely aligned with the state assessment in the measured content areas, assessment structure, and test formats, this close alignment should help the benchmark assessment to predict accurately how prepared the students are for the state
assessment (Edmentum, 2018a; Edmentum, 2018b). Putting it another way, the estimation of the extent to which test scores predict scores in other measures, such as state assessments, demonstrates how well the test scores predict other subsequent measures.

On the other hand, a correlation between test scores and other targeted outcomes is used as evidence of validity by some K-12 testing companies. Testing whether measurement tools, including achievement tests, can predict some interesting things based on the correlation coefficients, which is known as predictive validity (Borsboom et al., 2004). However, the primary objective of the validity study in validating test scores should not be to establish significant correlations between test scores and subsequent targeted outcomes but to offer a theoretical explanation of the processes that supports the measurement outcomes. Hence, this type of correlation (i.e., the correlation between test score and subsequent targeted outcomes, including state assessments) indicates only circumstantial evidence of validity (Borsboom et al., 2004). Therefore, my research considers such correlations not predictive validity but circumstantial evidence of validity.

The crux of the prediction of students’ performance in the state assessment is not the validity of assessment scores – the extent to which an assessment measures what the assessment intends to measure - but the level of the accuracy (i.e., how accurately the benchmark assessment can predict students’ performance measured in the subsequent state assessment). Correlation analysis and classification accuracy analysis refer to the evaluation of the associations between benchmark and state assessments in the performance levels. Neither of them indicates anything with regard to predictive accuracy. As one piece in a large picture of validity arguments, correlation and classification accuracy analysis can certainly provide some supporting
information. In evaluating the predictive accuracy of the benchmark assessment, it is a probability that needs to be assessed – how likely selected diagnostic outcomes are to be true. Predictive accuracy refers to the probability that the benchmark assessment predicts future state assessment results accurately (Hintze & Silberglitt, 2005; Shapiro et al., 2006). If a student classified as proficient in the benchmark assessment is classified as proficient in the state assessment, this “diagnostic” outcome in the benchmark assessment is accurate. If a student classified as non-proficient in the benchmark assessment is classified as proficient in the state assessment, this benchmark assessment’s diagnosis is incorrect. Both outcomes should therefore be tested to evaluate the predictive accuracy. To put it differently, it refers to the proportion of agreement between the predictor (i.e., benchmark assessment) and criterion measures (i.e., state assessment), that is the accuracy of the benchmark assessment to predict a proficiency status in the subsequent state assessment. Therefore, this research will focus on evaluating the level of the accuracy of predicting a proficiency level in the state assessment by the result of the proficiency level measured in the benchmark assessment.

**Predictive Accuracy of Benchmark Assessment**

In related to predictive accuracy of benchmark assessments, there are some studies that focus on the prediction of state assessment scores. For example, some of the benchmark assessment companies have performed correlation analysis as predictive correlation analysis. They provided estimates of the extent to which their benchmark assessment scores predicted the scores of a state assessment or simple correlations between benchmark and state assessment scores. Although the magnitude of correlation depends on the types and subject of assessments, the correlation coefficients generally ranged from 0.40 to 0.90 in past reports (e.g., Edmentum,
The higher the correlation coefficients are, the more strongly the alignment between benchmark and state assessments are supported. If the alignment is supported by statistical analysis, it suggests that the benchmark assessment is considered predictive of how prepared students could be for their state assessments (Edmentum, 2018b). As such, they reported that the coefficients are high enough to support the validity of their benchmark assessment.

In addition, some of the benchmark assessment companies also performed classification accuracy analysis – the percentages of students who are classified in the same performance levels in the benchmark and state assessments. In this analysis, to evaluate the prediction of students’ proficiency in their state assessments, classification accuracy statistics are reported as one of the predictive studies. Although it depends on grades and tested subjects, overall classification accuracy ranges from 0.80 to 0.92 (Northwest Evaluation Association, 2018; Northwest Evaluation Association, 2019; Renaissance Learning, 2019). Although classification accuracy refers to measures of associations between benchmark and state assessments in the performance levels, the analysis does not include the impact of cut scores on the performance level decisions.

Overall, assessment companies’ technical reports do not include the analysis regarding the location of cut scores. They provide a variety of analyses to evidence the validity of their assessments, including predictive accuracy. However, correlation analysis and classification accuracy analysis refer to the evaluation of the associations between benchmark and state assessments in the performance levels. When predictive accuracy is investigated, it is important
to assess the level of accuracy - how accurate a benchmark assessment result is in predicting a result in the subsequent state assessment. Certainly, there are some factors that might affect the predictive accuracy. Because testing occasions are different between benchmark and state assessments (i.e., students take benchmark assessments before state assessments), these different time points of measuring students’ academic skills can attenuate the predictive accuracy. However, correlation analysis shows the association between benchmark and state assessments, not the level of predictive accuracy. Therefore, this research will not use correlation and classification analysis but focus on the probability that the benchmark assessment predicts future state assessment results correctly.

**Significance of the Study**

This research aims to address multiple important topics in educational measurement and assessment policy. First, this research presents a method for evaluating whether the proficiency cut score is appropriate in differentiating proficient and non-proficient students that future state assessments measure. As already mentioned, the current approach performs correlation analysis and classification accuracy to support the evidence of the validity of the benchmark assessments. However, the validity analyses do not provide the evidence of predictive accuracy in predicting state assessment results. The benchmark assessments are used to predict students’ future performance, but their validity studies do not include evaluation of the cut scores used to determine students’ performance levels. Hence, research findings in this study will be useful for (1) testing whether the benchmark assessment can be used to predict the proficiency level in the state assessment and (2) estimating what the optimal point of a proficiency cut score is in predicting the proficiency level determined by future state assessments.
In addition, to the best of my knowledge, no benchmark assessment exists that uses both RDD and PR curves in assessing its proficiency cut score. By using RDD, this research aims to investigate the impact of classifying students as proficient in the benchmark assessment. Specifically, it aims to test if reaching the proficiency level in the benchmark assessment increases the probability of reaching the proficiency level in the state assessment. To put it differently, RDD helps to understand the consequences of proficiency decisions based on the benchmark assessment – both intended and unintended – through the proficiency decision in the state assessment. The intended consequence is that classifying students as proficient in the benchmark assessment causes the proficient students’ probability of reaching the proficiency level in the state assessment to increase more, compared with the students who fail in the proficiency level in the benchmark assessment. Rather than looking at the proficiency cut score just as the cut-point to divide students into proficient and non-proficient groups, this research aims to conduct a unique investigation of causality - if classifying students as proficient has a causal effect on the probability of reaching the proficiency level in the state assessment. Moreover, a PR curve can help optimize the predictive accuracy of the proficiency cut score in predicting the probability of reaching the proficiency level in the state assessment. In the regular process of setting a proficiency cut score in the assessment, policymakers can influence the classification decisions in students’ performance levels. In particular, in educational measurement, setting cut scores in large-scale assessments involves policymakers, educators, measurement professionals, and other experts. However, the benchmark assessment can avoid issues the state-based assessment cannot. This allows you to choose statistical techniques which maximize predictive accuracy without needing to consider wide swaths of stakeholders’ inputs.
Specifically, by linking the proficiency cut score in the benchmark assessment to the probability of reaching the proficiency level in the state assessment, this benchmark’s cut score is determined by state assessment results, without judgments of policymakers and other test designers.

Therefore, the research’s findings are significant because they will indicate how benchmark assessments should improve. When benchmark assessments do not only measure students’ learning periodically but also predict proficiency levels that future state assessments measure, the needs of benchmark assessment would increase more. If the benchmark assessment can serve both purposes, the benchmark assessment can provide strong incentives and motivate students to take the benchmark assessment more seriously. Schools and teachers also can use the result of the benchmark assessment to encourage students to reach the proficiency level. In this sense, this research will provide both technical support and educational impacts for improving the role of benchmark assessment.
CHAPTER 2
LITERATURE REVIEW

In an era of high stakes accountability, national, state, and district-level assessments are used to respond to the questions of policymakers (Shepard, 2006). Such large-scale assessments have had more impacts on high-stake policy decisions, classroom instructions, and other educational issues. In this test-based accountability era, a variety of assessment results are so widely and frequently used in the educational settings. Classroom teachers use short assessments such as quizzes throughout the school year to grade their students’ classroom performances. School districts use district-wide assessments to monitor their students’ progress and make district-wide policy decisions. States use state assessments to measure and classify students’ performance levels, such as basic and proficient, for responding to accountability requirements. Thus, each of the assessments serves several purposes in various educational settings.

In this section, there are two main issues to be discussed, assessment and proficiency. First, definitions of three main assessments – formative, benchmark, and summative – are provided. Their intended purposes, roles, and use of the results are partly different and overlap with one another. Therefore, it is necessary to distinguish among these three types of assessments. After clarifying the definitions and uses of each assessment, I provide important issues with regard to the proficiency level in the assessments. Next, I introduce the definition of proficiency in the context of educational assessments. Proficiency levels, rates, and its impacts are interconnected with each other in the current K-12 education. Accordingly, I provide in-depth descriptions of proficiency-related issues. Then, I address important issues in terms of the proficiency-based policy decisions.
Note that this research uses the term, benchmark assessment, not interim or diagnostic assessment. Some testing experts call such assessments interim assessments but the term, benchmark, is the most widely used in the context of educational measurement. Accordingly, I use the term, benchmark assessment, in the rest of the sections.

**Formative Assessment**

Formative assessment is performed during the instructional process to improve teaching and learning (Clark, 2011; Shepard, 2006). This type of assessment is usually administered by classroom teachers who need to check their students’ understandings constantly and is the most frequently used assessment with the smallest focus and shortest cycle (Perie, Marion, & Gong, 2009). Because the results of formative assessment help classroom teachers evaluate their students’ understanding of the lessons, formative assessment is very helpful in improving classroom teachers’ instructions and lesson plans (Fancsali, Zheng, Tan, Ritter, Berman, & Galyardt, 2018; Lazarin, 2014).

One of the significant features pertaining to formative assessment is that formative assessment involves formal and informal methods such as observation, traditional quizzes, and portfolios (Shepard, 2006). As a matter of fact, teachers need not only to give feedback about whether their students get correct or wrong answers but also to facilitate students’ learning and give feedback explicitly linked to performance standards. Accordingly, formative assessments play an important role in helping teachers make instructional adjustments as well as promoting students’ own responsibility of their own learning progress (Black & Wiliam, 2010; Sadler, 1989; Shepard, 2006). In addition, formative assessment is developed on various classroom teachings and learnings and does not count on standardized instruments designed outside the
classroom (Shepard, 2006). Unlike summative assessment that is built upon mastery or accountability standards, the main task of formative assessment is to monitor and guide students’ learning progress over time for the purpose of improving student performance (Clark, 2011; Boston, 2002; Shepard, 2006).

**Summative Assessment**

Summative assessment is the assessment that calibrates students’ performances as demonstrating a mastery of academic skills and knowledge (Shepard, 2006). Unlike formative assessment, this type of assessment is usually administered one time at the end of an instructional unit, in particular, the end of the school year in the state-wide level (e.g., Boston, 2002; Fancsali, Zheng, Tan, Ritter, Berman, & Galyardt, 2018; Perie, Marion, & Gong, 2009; Shepard, 2006). Instead of providing useful information to support classroom instructions, summative assessment focuses on student attainment of content knowledge and skills against achievement levels. Because results of summative assessments represent students’ academic performance relative to annual learning goals or a defined set of academic standards and to inform educational policy, summative assessment is an important part of the current accountability system in the U.S. education (Lazarin, 2014; Nhouyvanisvong, 2016; Perie, Marion, & Gong, 2009). Therefore, federally-required, state-wide assessments are considered summative assessments (Perie, Marion, & Gong, 2009).

**Benchmark Assessment**

Benchmark assessment falls between formative and summative assessment (Herman, Osmundson, & Dietel 2010; Nhouyvanisvong, 2016; Perie, Marion, & Gong, 2009). In terms of cycle, benchmark assessment is typically administered at multiple times throughout the school
year in a district or school level. Unlike state assessments usually administered on the end of the school year, this periodic administration of benchmark assessments provides classroom teachers immediate and actionable information about their students’ learning performance (Bulkley, Christman, et al., 2010). However, unlike formative assessment that is most frequently administered, because the scope of benchmark assessment is not as narrow as the curricular focus of formative one but covers school and district curriculum, benchmark assessment can also inform district- and school-level policy and decision-making (Herman, Osmundson, & Dietel 2010; Nhounyvanisvong, 2016; Perie, Marion, & Gong, 2009). In this sense, benchmark assessment includes the utilities of formative and summative assessments.

In terms of benchmark assessment results, benchmark assessment scores are usually used to evaluate students’ knowledge and skills and monitor the progress of students’ learning toward short- and long-term learning goals. Unlike both summative and formative ones, because benchmark assessment is aligned with district-level curriculum or standards, evaluative data from the benchmark assessments is beyond the classroom level and can be used for district-level policy-making and decision-making (Brown & Coughlin, 2007; Bulkley, Christman, et al., 2010, Herman, Osmundson, & Dietel 2010; Nhounyvanisvong, 2016; Olson, 2005).

At the school level, benchmark assessment scores are used to adapt teachers’ instruction curriculum to improve student academic performance. Unlike summative assessment linked with annual learning goals, benchmark assessment focuses on measuring students’ skills and knowledge needed to reach the annual academic standards. In addition, benchmark assessment is uniform in terms of timing and content across classrooms and schools (Herman, Osmundson, & Dietel 2010; Nhounyvanisvong, 2016). While formative assessment is so embedded in ongoing
classroom instruction as to monitor students’ daily progress and give timely feedback in the classroom level, school districts implement benchmark assessment closely aligned with the district-level curriculum or standards. Hence, benchmark assessments help monitor student progress and check the level of instruction and learning materials to match the students’ abilities beyond the classroom and school levels. The results, accordingly, can be aggregated at the grade, school, or district levels. Therefore, benchmark assessment can provide immediate information to help teachers and other school staff modify and adjust curriculum instruction in the school and district levels (Bulkley, Christman, et al., 2010; Clark, 2011; Olson, 2005; Renaissance Learning, 2016).

Lastly, benchmark assessment can be used for predictive purposes to judge whether students are on the right track toward meeting performance goals. In particular, as state-wide end-of-year assessment scores are used for accountability purposes, benchmark assessment results are used to identify who need more supports and who meet or excel the proficiency level based on the state assessment scores (Bulkley, Oláh, et al., 2010; Brown & Coughlin, 2007; Immekus & Atitya, 2016; Nhoyvanisvong, 2016; Perie, Marion, & Gong, 2009).

**Roles of Benchmark Assessment**

Benchmark assessments serve three primary purposes: instructional, evaluative, and predictive. In terms of instructional purposes, benchmark assessments are designed to provide valuable diagnostic feedback on strengths and weaknesses in students’ learning and to help identify where the students’ learning difficulties come (Herman et al., 2010). In particular, classroom teachers view benchmark assessments as one of the diagnostic tools needed to judge whether students reach the mastery level of context standards (Bulkley, Oláh, et al., 2010;
Immekus & Atitya, 2016). Specifically, benchmark assessments can help classroom teachers to identify instructional strategies that can promote student learning effectively. For the improvement in the instructional practices, classroom teachers use benchmark assessments to monitor students’ learning progress and guide instructional decisions (Immekus & Atitya, 2016; Oláh et al., 2010; Willam et al., 2014).

On the other hand, benchmark assessments are frequently used among various stakeholders, including school and district leaders, for evaluative and predictive purposes. At the school and district levels, benchmark assessment scores are used to inform various educational policies and practical decisions (Bulkley, Christman, et al., 2010 p.177-178; Immekus & Atitya, 2016). For example, when school curriculum or instructions change in the schools or districts after the end of the school year, benchmark assessments can provide various pieces of information about school- or district-wide curriculum development. State assessments also present evaluative data but are designed to cover too many content areas to provide precise information needed to improve the curriculum. On the other hand, because benchmark assessments are designed to diagnosis where each student is in his/her learning and gaps between students in their knowledge and understanding, the major difficulties or misconceptions that occur in student learning are specified relative to content areas (Herman et al., 2010; Perie et al., 2007). In this sense, benchmark assessments can serve as evaluation tools to help enforce the modifications of curriculum and instruction (Perie et al., 2007). Furthermore, benchmark assessments are used to determine students’ likelihood of meeting criterion scores in the end-of-year state assessments (Bulkley, Christman, et al., 2010; Perie et al., 2007). Such predictive purposes are becoming more and more important to respond to the accountability requirements.
Although there are few benchmark assessments that focus solely on the prediction of summative assessments, including state assessments, benchmark assessments are more frequently used to determine weak areas in students’ learning, identify students who are not on track to succeed, and to help students improve their performance measures in the state assessments. In this sense, benchmark assessments play a role in helping identify short-term strategies that support students’ improvement in the end-of-year state assessments (Bulkley, Oláh, et al., 2010; Immekus & Atitya, 2016; Perie et al., 2007).

**Proficiency Level in the Assessment**

As already mentioned, the current U.S. accountability system requires states to decide the performance levels to be reported and the clear definitions of each level in the assessment. In this system, each state defines performance levels differently for various accountability purposes (e.g., Bracey, 2007; Ho, 2008; Hamilton et al., 2012; McClarty, 2016). In particular, since the NCLB started, it has been very important for states to identify students who are at the proficient or higher levels and who are below the proficient level through assessment results (Brown & Coughlin, 2007; Ho, 2008; Zieky et al., 2006). In the NCLB, all students in grades three through eight in any racial, ethnic, socio-economic status were required to be proficient in both math and reading by 2014 (e.g., Cronin et al., 2007; Ho, 2008; Rothstein et al., 2006). Accordingly, states were also required to report the percentage of students who perform above or below a proficiency cut score as the primary score-reporting metric for school accountability decisions (Cronin et al., 2007; Hamilton et al., 2012; Ho, 2008; Rothstein et al., 2006).

In this proficiency-based accountability system, there are some political and technical concerns. The biggest concern is that a proficient level provides some misleading descriptions of
students’ academic performance. With regard to such shortcomings of proficiency level, past studies including Ho (2008) and Rothstein et al. (2006) already reported incorrect indication or statistical limitation of statistical properties of proficiency level. In the next sections, the following questions and related discussions will be addressed:

1. What is proficiency?
2. How is proficiency level determined?
3. What are the impacts of proficiency rates on high-stakes policy decisions?

In the first question, I will provide definitions of proficiency and proficiency level commonly used in the educational assessment. Through these definitions, I will explain the main issues with regard to proficiency. For the second question, I will describe how the proficiency level is established and determined. Although there are a wide variety of discussions of proficiency issues among stakeholders, it is questionable whether they discuss the proficiency issues based on the correct understanding of the process of determining a proficiency level. Some of the proficiency issues might result from the misunderstanding or lack of methods of setting a proficiency level. To put another way, the explanation of how a proficiency level is set will help explain how proficiency-based policy decisions generate misleading policy decisions. Thus, I will provide a broad description of the process of establishing a proficiency level in this section.

For the third question, based on the definitions of proficiency and methods of setting a proficiency level, I will illustrate the impacts of proficiency-based decisions on high-stakes policy decisions. What policy impacts happen in school, district, and state-level if proficiency-based decisions are made? By describing each of the main issues with regard to proficiency-based policy makings and decisions, I will clarify the main impacts of proficiency issues.
What is Proficiency?

Proficiency pertains to specific levels of academic achievement at a specific time point (Bryan, 2017). For example, the National Assessment of Educational Progress (NAEP) defines proficiency as demonstrated competency over challenging subject matter, including subject-matter knowledge, application of said knowledge to real world situations and analytical skills (Zieky & Perie, 2006). NAEP does not connect proficiency to a specific standard defined at each grade level. On the other hand, in the state- or district-level assessments, proficiency is referred to as demonstrating the required level of performance toward academic standards that each state defines and expects its local students to reach. Proficiency simply means that a proficient student has learned enough in his or her grade to be promoted to the next grade. Because states and school districts establish their own unique standards respectively that their local students are required to follow. Hence, they define proficiency differently, too (Cronin et al., 2007; Durant & Dahlin, 2011; McClarty, 2016). Whether it is defined in nationwide, state, or district levels, proficiency generally means that the student has learned enough in his or her grade to be ready to study in the next grade. Therefore, proficiency includes the meaning of the content mastery specified at each grade level (Bryan, 2017; McClarty, 2016).

In terms of the use of proficiency, in the school level, proficiency is generally used to judge whether a student is academically ready to study in his or her next grade. If a student does not reach the proficiency level, the student falls short of the academic standards and need to study, in order to meet the required standards. At the same time, for teachers and other educators, proficiency helps them identify whether students need extra helps (McClarty, 2016). For their parents, this result helps to make sure that their kids are on the right track toward academic
achievement. On the other hand, in the state level, proficiency is usually measured by a single end-of-year assessment. Unlike district-level assessments that are administered multiple times a year, such state-level assessment measures not only simple understandings obtained in earlier grades but also complex applications required in higher grades, proficiency usually indicates content mastery that students are required to reach (Bryan, 2017).

Overall, proficiency is generally considered as a specific level of mastery of standards that is common in nation, state, district, and school levels. When a student reaches a proficiency level, the student is considered as passing the standards defined in his or her grade level or being on the right track toward academic achievement.

**Proficiency and Standard Setting**

Based on the definition of proficiency, what is the proficiency level? If student performance levels including proficiency are determined in the state assessment, how are such performance levels determined in the state assessment? Students’ proficiency level is determined based on a cut score established in the process of standard setting (Hambleton & Pitoniak, 2006; Ho, 2008). In addition, to determine the performance levels of students requires a set of standards of performance. In other words, standard setting is the process of establishing the cut scores required to determine the performance levels based on the standards of performance in the educational assessment.

Originally, the term, standards, has been used in a wide variety of settings. In the licensure or certification testing programs, to set a standard of performance is necessary to step toward specific classifications such as pass or fail and certify or not certify (Cizek, 1996, Cizek & Bunch, 2007). In the context of educational assessment, standard setting involves the process
of determining the boundaries (e.g., cut score) that define more than two levels of performance or differentiate performance levels such as basic, proficient, and advanced (Cizek, 1996). In K-12 education, as already mentioned, a current accountability system mandates that achievement levels for all students should be reported based on performance levels such as below basic, basic, proficient, and advanced. To classify students’ performances into such multiple levels, multiple cut scores are necessary in defining the borderlines between performance levels (Cizek & Bunch, 2007). Put differently, cut scores are outcomes of standard setting and very important in determining the percentages of students at each performance level. In this sense, it is important to choose an appropriate standard setting method for setting cut scores.

Methods of Standard Settings

Choosing a proper standard-setting method is very important in influencing the resulting cut scores (Cizek & Bunch, 2007). The range of student academic skills, the format of the assessment tasks, and the administration protocol are all involved in the choice of the standard-setting method (Ferdous et al., 2011). As a matter of fact, the Standards for Educational and Psychological Testing (2014) say that there can be no single method for determining cut scores for all tests or for all purposes and no single set of procedures (AERA et al., 2014). As new approaches are developed, standard-setting methods that have been used, are becoming more and more important.

In the evolving cycle of various methods, one of the commonly used methods is Angoff’s method (Cizek, 1996; Cizek et al., 2004; Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006). In the Angoff method, the standard-setting panelists review each multiple-choice question and make a judgment of the probability of minimally competent test-takers answering it correctly.
Summed over test items, the probability ratings from all panelists are averaged to determine the suggested cut score (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006). In the multiple-choice items, the probability of getting a correct answer should be higher as the item is easier (Zieky & Perie, 2006). This method can be used not only to set a single cut score to determine the borderline between acceptable and unacceptable levels but also to create multiple cut scores to determine two or more performance levels (Ricker, 2006). In addition, the extended Angoff method is the one that can be used when an assessment includes polytomously scored tasks or constructed-response items. As estimated in the Angoff method that handles dichotomous or proportional judgments, the extended Angoff method is to have the panelists estimate the average number of scale points that minimally competent test-takers would obtain on each polytomously scored items (Cizek et al., 2004; Zieky & Perie, 2006). Cut scores are calculated in the same approach as the normal Angoff method (Cizek et al., 2004). In other words, cut score decisions rely on the standard-setting panelists in both Angoff and extended Angoff methods.

In the educational measurement, the Bookmark method has been a popular one to establish cut scores (Cizek et al., 2004; Cizek & Bunch, 2007; Ferdous et al., 2011; Zieky & Perie, 2006). Introduced by Lewis, Mitzel, and Green in 1996 (Mitzel et al., 2001), this procedure has been widely adopted for setting proficiency levels in K-12 education settings including state assessments (Cizek et al., 2004; Engelhard, 2011). In this method, items in a test are placed in difficulty order from easiest to hardest, determined by the number of students who answer correctly. Judges are required to draw the line between the hardest item that examinees at the borderlines around each of the performance levels are expected to answer correctly and the easiest one that the examinees are expected not to answer correctly. The advantage of the
Bookmark method is to include not only multiple-choice items but also constructed-response items in this type of test booklet ordered by item difficulty (i.e., placed easiest items first and hardest item last), which is called as an ordered item booklet (Cizek & Bunch, 2007; Ferdous et al., 2011; Zieky & Perie, 2013). Hence, as Angoff methods, the Bookmark method also determines a cut score based on the third-party judgments.

The common feature regarding cut scores established in these popular standard-setting methods is that cut scores is based on decisions of judges, not statistical evidence. Regardless of student performance levels, each cut score is determined by standard-setting experts. The cut-point that experts believe is the right location to differentiate students above and below the specific levels is established as a cut score in the current standard setting methods. Therefore, even popular standard-setting methods includes arbitrary decisions of cut scores.

**Cut-off Score Point**

Cut score is the score that separates a test score scale into two or more groups and creates categories of performance or classifications of test-takers. As standard-setting is referred to as the process of establishing one or multiple cut scores on assessments, cut scores are used to divide students’ academic performances into two or more categories of performance (AERA et al., 2014; Cizek & Bunch, 2007; McClarty et al., 2013). In the state assessment, cut scores are used to determine performance levels of students such as “basic”, “proficient”, and “advanced”. Some states use a cut score to specify a minimum passing score. A cut score is also used to identify students who are ready to go to college. In reverse, if the cut scores are inappropriately set, it will affect the assessment results including the determination of students’ performance.
levels (Kaftandjieva, 2010). Therefore, it is very important to provide information on the validity
of the cut scores and the rationales for setting the cut scores (AERA et al., 2014; Cizek, 1996).

In terms of setting cut scores, the way to set cut scores depends on types and purposes of
assessment. According to AERA (2014), there cannot be any single method for determining cut
scores for all tests and for all purposes, nor can there be any single set of procedures for
establishing the defensibility of the cut scores. For example, the Pass/Fail or other classifications
of performance levels resulting from the application of a cut score are essentially inferences
(Cizek & Bunch, 2007). In educational measurement, setting cut scores in the large-scale
assessments including state assessments involves policymakers, educators, measurement
professionals, and other experts in a multiple, judgmental process (AERA et al., 2014; Ho, 2008;
McClarty et al., 2013; Zieky & Perie, 2006). If cut scores are established in such an arbitrary
fashion as choosing a cut score, such as answering 70% of test items correctly, the cut scores
have little meaning or information (AERA et al., 2014). In addition, cut scores established based
on experts’ judgments are reported to be not consistent with their expectations for the impact of
the cut scores (e.g., the percentage of students answering correctly is lower or higher than the
judges’ estimates) (Impara & Plake, 2000; Linn & Shepard, 1997). Because cut scores are
determined in this type of judgmental process, the cut scores could affect the accuracy of
classification of students’ performance levels (Ercikan & Julian, 2002).

In terms of the effects of cut scores on classification accuracy, the biggest concern is
about the classification of students who are near cut scores. Classification decisions on
educational assessment are typically made by comparing a student’s assessment score to cut
scores set in the process of standard setting (Diao & Sireci, 2018). In this decision, ideally, test-
takers should be classified into “true” performance levels based on the cut scores (Diao & Sireci, 2018; Lee, 2010) but in reality, this is more challenging. For example, if the proficiency cut score is too high, some students who should be rated as proficient can be classified as below proficient. If the cut score is set too low, students can be falsely classified as proficient. In other words, changing cut scores can increase or decrease this type of classification errors (i.e., misclassification) (Zieky & Perie, 2006). If scores of students are not close to cut scores, misclassification would rarely happen. However, if students’ scores are close to the cut score, the students are more likely to be misclassified than students whose scores are not close (Cizek, et al., 2004; Haertel, 2006). Therefore, to clearly distinguish students who are at a performance level compared to students who are in the next lower level or in the next higher level, it is very critical to verify the appropriateness of the cut scores in the assessment as the validation process (AERA et al., 2014; Brown & Coughlin, 2007; Zieky & Perie, 2006).

As for setting cut scores, it is important to be careful about the interpretation of the proficiency rates and performance levels during the process of standard setting. Specifically, the establishment of the cut score involves not only statistical analysis, but also the input of policymakers, measurement experts, and other stakeholders. Thus, the implications of cut scores and performance levels based on these scores are different between benchmark and state assessments. For example, in Florida’s statewide assessment, when the cut scores were renewed, their state board of education reviewed and determined the cut scores based on public input and various recommendations, including local educators and the commissioner of the Florida Department of Education (Verges, 2015). In a nation-wide multi-state assessment, the Smarter Balanced Assessment Consortium (SBAC) assessment, the process of establishing cut scores
also includes a wide variety of experts’ and policymakers’ approval and recommendations. For the initial design of the SBAC assessments, after the cut scores were tentatively established by the standard-setting panelists, the cut scores are determined based on recommendations from K-12 and higher education representatives from states that adopted SBAC assessments and a vote of SBAC governing states (Smarter Balanced Assessment Consortium [SBAC], 2013; SBAC, 2017). When cut scores were reviewed in 2017, their technical report addressed that ELA and math educators from states that used SBAC assessments discussed and modified cut scores (SBAC, 2017). In their interim comprehensive assessment for grades 9 and 10, cut scores were established based on reviews and recommendations from K-12 educators and representatives from higher education (SBAC, n.d.). In setting cut scores in state-wide assessments, the main use of the cut scores is the categorization of performance levels for each student. This is different from the role of benchmark assessments; this is the empirical prediction of scores and performance levels measured in future assessments. Because state-level assessments do not have any specific future scores or performance levels that the assessment needs to predict, the establishment of the cut score is based at least partially on judgment of key stakeholders. Therefore, a wide variety of experts’ and policymakers’ input are necessary to determine the cut scores in state-level assessments.

**Impact of Proficiency-based Policy-decisions**

One of the critical issues regarding proficiency-based policy decisions is that proficiency rates misrepresent student performance and improvement. Specifically, proficiency rates, the percentages of students who reach a proficiency level in the assessment, provide a limited and misleading description of test scores, achievement gaps, and their trends (Ho, 2008). As
described in the previous section, setting a proficiency cut score includes arbitrary and judgmental characteristics. Proficiency rates are calculated based on the number of proficient students determined based on such judgmentally and arbitrarily established cut scores (Hambleton, & Pitoniak, 2006; Ho, 2008). In particular, since the NCLB emphasized measuring proficiency rates, states and school districts are required to report rates of students classified as each of performance levels. In reporting categorical performance levels, schools are expected to increase their proficiency rates. If their proficiency rates are very low, those schools face the sanctions of turn-around (Ho and Reardon, 2012). Therefore, arbitrarily and judgmentally established proficiency cut scores can mislead school evaluation as misrepresentation of schools’ proficiency rates.

Furthermore, proficiency is not equivalent across subjects and grades (Bracey, 2007; Cronin et al., 2007; Durant & Dahlin, 2011). In most states, math cut scores are generally higher than reading or ELA cut scores. If it is the case, a proficiency rate in math is lower than the rate in ELA when student performance is at the same level in both math and ELA. This result might give misinformation that the school needs to focus on improving their students’ math program more than ELA. Thus, non-equivalent proficiency cut scores in math and ELA may mislead local educators.

In addition, when achievement gaps are calculated based on percentages of students’ performance levels determined based on that proficiency cut scores, the gaps significantly vary based on the location of cut scores (Ho, 2008; Ho & Reardon, 2012; Holland, 2002). As already mentioned, the current proficiency-based accountability system requires reporting proficiency rates and not the assessment scores. The school’s performance is also evaluated based on
percentages of students attaining a proficiency level. Achievement gaps are also calculated based on the proficiency rates, not assessment scores. Proficiency-based achievement gap is, accordingly, subject to flaws resulting from the arbitrary selection of a proficiency cut score. As a result, because achievement gaps are calculated based on such a proficiency cut score, the result of achievement gaps can also vary substantially (Bracey, 2006; Ho, 2008; Ho & Reardon, 2012; Holland, 2002; Koretz & Hamilton, 2006; Linn, 2007).

Student growth also changes substantially due to such proficiency cut scores. Because currently used measures to evaluate student growth and school performance are based on categorical models (i.e., calculating the number of students moving from lower performance levels to higher performance levels or just the percentage of students who achieve a proficiency level in the assessment), proficiency cut score can lead to manipulation of student growth. As already mentioned, because the establishment of proficiency levels is not consistent across the grades, student growth based on proficiency rates also vary so significantly across grades (Cronin et al., 2007; Durant & Dahlin, 2011; Ho, 2008; Ho et al., 2009). Most of the differences in student growth based on proficiency rates are due to the difficulty of assessment or location of the cut scores and not the result of differences of student performance (Cronin et al., 2007; Durant & Dahlin, 2011). In other words, the change in the location of proficiency cut scores leads to the change of proficiency rates (Ho, 2007). Therefore, unless assessments are designed to ensure that proficiency in the early grades is aligned with proficiency in the upper grades, student growth measures based on proficiency rates across grades might be misleading.

In the nation-wide level, there are also some critical concerns regarding proficiency-based policy. First of all, the definition of proficiency varies greatly from state to state (Bracey, 2007;
Cronin et al., 2007; Durant & Dahlin, 2011; McClarty, 2016). As mentioned previously, cut scores established in the standard-setting process are used to determine whether students are at the proficient level or not. This proficiency cut score also vary greatly from state to state. Thus, proficiency levels and rates also vary greatly for each state. According to Cronin, Dahlin, Adkins, and Kingsbury (2007), proficiency cut scores ranged from 6th percentile to the 77th percentiles in the state level. This means that the difficulty of reaching the proficiency level varies so significantly. Moreover, states can set performance levels low in their state assessment to make proficiency rates high (Figlio & Loeb, 2011; Ho, 2008; Peterson & Hess, 2006). Such arbitrary choices of the proficiency cut score leads to various policy issues such as manipulating students’ growth trends and minimizing achievement gaps arbitrarily (Ho, 2008). Because there is not any common proficiency standard used in the nation-wide level, it is very difficult to compare proficiency rates in multiple states (Hamilton et al., 2012; Rothstein et al., 2006).

Overall, it is necessary to be cautious about making any high-stakes policy decisions based on proficiency levels. Because the current proficiency-based accountability system counts on the number and rate of students who reach the proficiency level in the assessment, this metric can lead to some misleading, limited, and incorrect results. Certainly, this accountability system has a positive effect that pushes districts and their local schools to make more efforts to improve their student academic performances. If proficiency rates are very low, the states or districts need to invest more money, workforce, or any other necessary things to increase their proficiency rates. However, if this low proficiency rate results from a very high proficiency cut score, this proficiency rate shows misinformation (Durant & Dahlin, 2011). In the current assessment system, proficiency cut score is established in a more or less arbitrary and judgmental way.
Hence, student trends and growth measures based on proficiency level/rates might be incorrect or misleading. Furthermore, students near the proficiency cut score might be misclassified as proficient or not proficiency. This can contribute to incorrect proficiency rates by including students falsely classified as proficient or not proficient. In the nation-level, different definitions and decisions of proficiency cause incomparable proficiency rates among multiple states or districts in different states. In the next section, I will address a plausible solution with regard to setting a proficiency cut score so as to prevent proficiency rates from misleading and misrepresenting the students’ performance.
CHAPTER 3

METHODS

My research focuses on the utility of a proficiency cut score of the benchmark assessment to predict a proficient level in the state assessment. To evaluate the proficiency cut score of the benchmark assessment in predicting the proficient level in the state assessment, I will use one of the quasi-experimental methods, Regression Discontinuity Design (RDD), and one of the machine learning approaches, Precision-Recall (PR) curves. In this research, I will perform RDD to examine whether reaching the proficient level in the benchmark assessment causes students to reach the proficiency level in the state assessment. I will apply PR curves to find an optimal proficiency cut score to accurately predict the proficiency level in the state assessment.

**Evaluation of Proficiency Cut Score**

As already mentioned, in the current test-based accountability system, students are assigned to one of the multiple performance levels (e.g., below basic, basic, proficient, and advanced) based on cut scores in the assessment. I will use RDD to evaluate the proficiency cut score of the benchmark assessment in predicting the similar proficiency level in the state assessment. Both benchmark and state assessments are basically designed to measure students’ academic performances and assign each student to performance levels. However, equally skilled students are not necessarily rated as similar performance levels in these two types of assessments (Papay, Murnane, & Willett, 2010). For example, if a student is classified as proficient in the benchmark assessment, the student would not necessarily be rated as proficient in the future state assessment. The student might be classified as one level lower (i.e., basic level). If a proficiency level determined in the benchmark assessment is reliable and accurate enough to predict whether
a student will reach the similar proficient level in the future state assessment, the benchmark assessments would be a very useful tool for classroom teachers to prepare their students to reach the proficiency level classified in the state assessment.

As for predicting a proficiency level in the state assessment, it is important to evaluate whether a proficiency cut score in the benchmark assessment is appropriately set. For assessing the predictive accuracy of the proficiency cut score, RDD is one of the econometric techniques that can be used to assess unbiased effects of classification of performance levels based on the assessment scores. In investigating the predictive accuracy of the proficiency cut score in the benchmark assessment, this research focuses on students who barely reached the proficiency level and those who barely did not reach. If the proficiency cut score is accurate enough to predict the proficiency level in the state assessment, it should be statistically significant for the proficiency cut score of the benchmark assessment to differentiate students who reach the proficiency level in the state assessment and those who do not. Accordingly, in the following sections, the details on the usage and rationales for using RDD for this research purpose are described.

**Regression Discontinuity Design**

Regression Discontinuity Design (RDD) is a rigorous quasi-experimental design used to assess the effects of interventions in the cases in which an assignment variable is used to divide subjects into a treatment group and a control group (Jacob et al., 2012; Lee & Munk, 2008; Pearson, 2016). In many educational and other social interventions, it has been more and more commonly used to evaluate program or policy effects (e.g., Jacob et al., 2012; Lee & Munk, 2008; Robinson, 2011; Smith, 2014). As the name of RDD shows, this design is used to estimate
the impact of the discontinuity or displacement of the regression line that happens at a specific point, which is called the cutoff point or the cut score.

In RDD, the subjects are assigned to treatment or control groups based on whether their values exceed this cutoff point for the intervention at a rating or assignment variable (Jacob et al., 2012; Smith, 2014). This rating or assignment variable is any continuous variable (e.g., test score) measured before treatment. For example, students are assigned to a remedial course (e.g., writing courses and after-school programs) when their test scores are below a specific cutoff point. Students scoring below this cutoff point are included in the treatment group (i.e., a group of students who participate in writing courses or after-school programs), and students scoring at or above the cutoff point are included in the comparison group (i.e., a group of students who do not participate in writing courses or after-school programs). If any interventions have a causal effect on the outcome variables such as test scores, there is a discontinuity in the regression line (Lee & Munk, 2008). Assuming that an assignment variable with a cut point of 18 and an outcome variable has a linear relationship and a treatment group who receives any intervention is coded as 1 and a control group who does not as 0, Figures 1 is created. The left figure shows the regression line without any break, suggesting that there is no effect of the intervention on the outcome variable, regardless of the position of a cut point. The figure on the right illustrates the regression line with a discontinuity at the cut point of 18. In this case, an estimated value of this discontinuity is interpreted as an estimate of the unbiased intervention effect on those who are barely above the cut point relative to those who are barely below the cut point (Lee & Munk, 2008; Shadish et al., 2002; Smith, 2014).

Figure 1
To describe the main idea of RDD, a regression model with a cut score in this example is the following equation:

\[ Y_i = \beta_0 + \beta_1 T_i + \beta_2 S_i + \varepsilon_i \]

where \( Y_i \) is the outcome score for student \( i \), \( \beta_0 \) is an intercept, \( \beta_1 \) is an effect of intervention on the outcome variable (i.e., the effect of reaching the cut score on the outcome), \( T_i \) is the assignment variable (i.e., 1 if a subject reaches the cut score and 0 if not), \( \beta_2 \) is the regression coefficient of \( S_i \), \( S_i \) is the assignment score (e.g., benchmark assessment scores) and \( \varepsilon_i \) is an error term. In this equation, \( \beta_1 \) is the parameter of interest in the RDD model. If \( \beta_1 \) is not statistically significant, it indicates that reaching the cut score has no causal impact on the outcome. It means that the outcome is statistically equivalent between students who reach the cut score and students who do not as shown in the left figure. If \( \beta_1 \) is significant, it suggests that reaching the cut score has a causal impact on the outcome and accordingly causes a discontinuity as shown in the figure on the right. It means that the outcome of students who reach the cut score is significantly higher than students who do not. To put it differently, if the regression line has
any discontinuity, the intercept of the line changes only for the treatment group (i.e., a group of subjects who reach the cut score) as the right figure shows.

In the interpretation of the above results, the discontinuity starts at the point of 18 on an x-axis. This means that the “right” cut score should be 18 in differentiating two different groups with a different regression line. When this idea is applied to the determination of the proficiency cut score, there should be statistically significant discontinuity if the proficiency cut score is designed to differentiate a group of proficient students and the other group of non-proficient students. Hence, RDD can be applied to evaluate the appropriateness of the cut score, that is the “right” location of the cut score.

Please note that in practical research, RDD is more complicated than the simple straight line shown in the example (Jacob et al., 2012; Lee & Munk, 2008). For example, the constant effect of the interventions on the outcome might not hold. In some cases, the regression line might not be linear but quadratic. Accordingly, in this study, two regression lines, linear and quadratic, will be applied by using RDD. The regression models, accordingly, are as follows:

Linear line: \( Y_i = \beta_0 + \beta_1 T_i + \beta_2 S_i + \varepsilon_i \)

Quadratic line: \( Y_i = \beta_0 + \beta_1 T_i + \beta_2 S_i + \beta_3 S_i^2 + \varepsilon_i \)

In both regression lines, if there is any statistically significant discontinuity in either of the lines, it suggests that reaching the cut score has a statistically significant causal impact on the outcome.

**Application of Regression Discontinuity Design**

In terms of the application of RDD in evaluating cut scores in the assessment, RDD is a rigorous tool to estimate the causal inference even when a randomized experimental design is not feasible due to any practical and ethical issues (e.g., Lee & Munk, 2008; Pearson, 2016; Smith,
When the research focuses on simple predictions, standard regression analysis might be able to offer estimates of proficiency decisions based on benchmark assessment scores. However, without controlling for all observable and unobservable variables related to the proficiency classification process, the standard regression analysis cannot indicate the causal impacts of the proficiency cut score of the benchmark assessment. By contrast, RDD can be applied in the situations in which each subject is randomly assigned into a treatment group and the other group based on the cut score of the assessment (Papay et al., 2010; Pearson, 2016; Robinson, 2011; Umansky, 2016). For the randomization of samples, randomized control trial (RCT) might be a better approach because subjects are randomly assigned to the treatment and control group in RCT. Theoretically, such random assignment can nullify any confounding effects on the estimation of causal impacts (Lee & Munk, 2008; Smith, 2014). However, in practical educational research settings, random assignment of subjects (i.e., dividing subjects randomly into treatment group who receives any intervention such as rewards and control group who does not) is not feasible due to ethical or practical issues. In this sense, RDD entails the assignment of subjects if they are above or below any cut-off point (Jacob et al., 2012; Lee & Munk, 2008; Smith, 2014). RDD does not use completely random assignment but the assignment is based on an assignment variable with a cut-score, local randomization (Jacob et al., 2012; Lee, 2008). If any intervention has some effects on the outcome, as already described, there should be a noticeable discontinuity (e.g., any jump or drop) in the regression line at the cut score point. Thus, despite the lack of random assignment of subjects, RDD provides unbiased estimates of treatment effects as a similar randomized experiment (Smith, 2014; Thoemmes et al., 2017).

Past RDD Research on Classification based on Cut Scores
When it comes to the practical application of RDD to policy-decisions and evaluation based on assessment scores, there are past studies that focused on a cut-score of pass/fail examinations. These examinations are used for high-stake policy decisions such as classification of ELL students, determination of whether high-school students can graduate or not, and whether students with low-performance should be retained in their grades or promoted to a higher grade. These studies are highly useful in that they focused on the effects of dividing a continuous measure of students’ proficiency into two categories such as pass/fail based on the cut score.

**ELL Reclassification Decisions.** One of the studies in which RDD is applied to the evaluation of cut scores of K-12 achievement tests in educational policy settings is English Language learners (ELLS) reclassification research. In the U.S., more and more students who are not proficient in English are classified as English learners (ELs) and assigned to ELL programs (e.g., Carlson & Knowles, 2016; Reyes & Hwang, 2019; Umansky, 2016). Accordingly, ELL assessments have a big influence on their ELL instruction (Reyes & Hwang, 2019; Robinson, 2011). In the design of ELL assessments, a threshold, a cut score at an appropriate level, is set for ELL reclassification purposes. When ELL students reach the cut score of the proficient level in the ELL assessment, the ELL students are considered as attaining fluent English proficiency and exempt from extra coursework and services (e.g., Robinson, 2011). In this reclassification, it is important to set the appropriate cut score of fluent English proficiency in the ELL assessment. If the cut score is set as a relatively low score and more ELL students reach the fluent level, some reclassified students who should study in the ELL program but are in mainstream courses, might have a hard time keeping up with the courses. In reverse, if the cut score is relatively high and a limited number of ELL students can attain a fluent level, ELL students need a higher level
of ELL service to reach the cut score of fluent English proficiency (Robinson, 2011; Robinson-Cimpian & Thompson, 2016; Umansky, 2016). In the past studies, ELL reclassification studies focused on the classification decision on whether ELL students should be transferred from ELL to regular English courses based on ELL assessment scores (e.g., Carlson & Knowles, 2016; Robinson, 2011). Specifically, RDD is performed to evaluate the appropriateness of the cut score used to determine whether ELL students should be promoted from the ELL program to the mainstream classroom. If the cut score in the ELL assessment is appropriately established (i.e., the cut score is neither too high nor too low), there should be a statistically significant effect on future academic performances of reclassified students. The results evidence that ELL assessments can differentiate ELL students who deserve attending the mainstream classrooms and ELL students who do not (Pearson, 2016; Robinson, 2011).

**High School Exit Examination.** There are some recent studies that apply RDD to estimate the causal effects of the high school exit examination on high school graduation. In these studies, as a nation-wide trend, more and more states have adopted high school exit examinations that students must pass to earn high school diplomas (e.g., Papay, Murnane, & Willett, 2010; Zabala Minnici, McMurrer, & Briggs, 2008). These states aim to confirm that students have mastered the state-defined academic standards through exit examinations. In the middle of test-based educational reforms, past studies tested if the examinations can distinguish students who have a sufficient level of academic skills and those who do not. As a matter of fact, because states make a relatively arbitrary or judgmental decision about setting the pass/fail cut score that high school students must pass, it is relatively judgmental to determine whether the students pass or fail in the exit examinations. To make non-arbitrary decisions, past studies
applied RDD to investigate the difference in the likelihood of high school graduation between students who barely passed and who barely failed. Specifically, RDD is performed to evaluate any causal effects of barely passing or failing the exit examinations on the likelihood of graduation. If there is no significant discontinuity in the probability of graduation between students who barely passed and those who barely failed (i.e., the probability of graduation is statistically same between those two groups of students), it suggests that the cut score does not differentiate students who have academic skills required to graduate from students who do not. It means that the cut score is not in the right location in differentiating these two groups of students.

Overall, some past studies used RDD to evaluate the effects of test-based classification decisions on academic outcomes. In particular, the main discussion of RDD focused on the “right” location of the cut score. In the general approach of setting test-based thresholds, policymakers can significantly influence classification decisions (Robinson, 2011). With regard to this conventional process, RDD calls the validity of test-based thresholds into question. That is, there are some unintended effects of test-based decisions. In the ELL classification, only ELL students with a sufficient level of English skills should be promoted to mainstream courses but some students with limited English skills might also be promoted. To maximize the intended consequences and minimize the unintended ones, it is even more important to set a cut score appropriately. Hence, it is highly necessary to provide statistical evidence to support the location of the cut score used to make highly political decisions. Therefore, RDD can be a powerful tool to test whether any unintended effects are caused regarding the cut score.

**Precision-Recall Curve**
Precision-Recall (PR) curve is a useful measure to evaluate binary classification models that visualize the performance at a range of thresholds (e.g., Boyd et al., 2013; Flach & Kull, 2015; Tharwat, 2018). In machine learning and other related fields, it is often the case that multiple performance measures need to be optimized (Flach & Kull, 2015, p.1). In such cases, it would be more plausible to predict probabilities of a subject belonging to each class in the classification process than predicting the classes directly (Brownlee, 2018). In predicting these probabilities, a PR curve can be a useful diagnostic approach that facilitates the interpretation of probabilistic forecasts for predictive models of binary classification.

In terms of machine learning approaches including PR curves, to the best of my knowledge, machine learning techniques have not been widely used in educational measurement research. This might be because it is not so apparent to calculate the values or to illustrate the mechanism of the analysis (Boyd et al., 2013). However, I believe that a PR curve is useful to calculate an optimal proficiency cut score in the assessment, that is to optimize the level of accuracy regarding the proficiency cut score of the benchmark assessment in predicting the proficient level in the state assessment. If an optimal proficiency cut score is calculated by the PR curves to improve the predictive accuracy of the benchmark assessment, the benchmark assessment would be a more useful tool to prepare the students in reaching a similar proficient level in the future state assessment. To improve the predictive accuracy, a PR curve is a promising approach that is applicable to establishing the cut score in the benchmark assessment.

**Precision and Recall**

Precision is defined as “the proportion of true positives among positive predictions, as performance metric instead of false positive rate” (Flach & Kull, 2015, p. 3). It shows the ratio of
the number of true positives divided by the sum of the true positives and false positives (the name does not seem to match the meaning or definition, but precision means the ratio). It pertains to how good the model is at predicting the positive class. On the other hand, recall indicates how accurately the model predicts the positive class. Precision is generally called the positive predictive value and recall as sensitivity (Brownlee, 2018; Flach & Kull, 2015). In the formal definition, precision, $P$, is defined as the number of true positives, $TP$, divided by the number of true positives plus false negatives, $FP$.

$$P = \frac{TP}{TP + FP}$$

Recall, $R$, is defined as the number of $TP$ divided by the number of $TP$ and $FN$.

$$R = \frac{TP}{TP + FN}$$

PR analysis is useful, in particular, in cases of imbalance in the observations between two classes. In the case in which there are many subjects of no event (i.e., the class is 0) and only a few subjects of the event (i.e., class is 1). If this is the case, it is not so meaningful to examine the capability of predicting the class of 0 accurately, which is high true negatives (Brownlee, 2018). Calculations of PR analysis does not use true negatives and focuses on predicting correctly the class of 1.

Table 1 shows the mechanics of precision recall matrix. If the predicted class is the same as the actual class, this means that the benchmark assessment predicts future assessment status accurately. In this case, values of true positive (i.e., both a predicted and actual status is positive) and true negative (i.e., both a predicted and actual status is negative) is high. If the predicted status is different the actual status, values of false positive (i.e., a predicted status is positive but
the actual one is negative) or false negative (i.e., a predicted status is negative but the actual one is positive) should be higher than values of true positive and true negative. In this way, precision recall curve shows how accurate predicted classes are, compared with inaccurately predicted class.

**Table 1**

*Precision Recall Matrix*

<table>
<thead>
<tr>
<th>Actual class (Proficiency status in the state assessment)</th>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted class (Proficiency status predicted by benchmark Assessment)</td>
<td>True positive (TP)</td>
<td>False positive (FP)</td>
</tr>
<tr>
<td></td>
<td>False negative (FN)</td>
<td>True negative (TN)</td>
</tr>
<tr>
<td>All proficient students = TP + FN</td>
<td>All non-proficient students = FP + TN</td>
<td></td>
</tr>
<tr>
<td>Sensitivity = TP / (TP + FN)</td>
<td>Specificity = TN / (TN + FP)</td>
<td></td>
</tr>
</tbody>
</table>

**Precision-Recall Curve**

To facilitate the interpretation of the results of PR analysis, a PR curve is usually created to visualize the relationship between values of precision and recall. This curve is the line connecting all PR points of classifications by the change of the threshold. The curve indicates the trade-off between true positive rate and positive predictive value for a predictive model using a different probability threshold (Brownlee, 2018). Figure 2 shows an example of the PR plot.

**Figure 2**

*Example of Precision Recall Curve*
In the PR curve in the plot, the x-axis shows the recall and the y-axis, the precision. Every point shown on the line of the PR curve represents a cut-off point. In this example, the score ranges from 0 to 30 with the optimal cut score of 20. If the cut score is too high, 28 or higher in the example, some subjects who should be able to reach the cut-off point do not reach the point. If the cut score is too low, 9 or less in the example, some subjects who should not be able to reach the cut-off point reach the point.

In the interpretation of the result of PR curve, the closer the PR curve is to the upper right corner, which means a larger area under the PR curve, the more accurate is the cut score in maximizing the correct classification and minimizing the misclassification. In other words, the point closest to the higher right corner is considered as the optimal score. As such, PR curves are very useful in visualizing the classification results and showing which score should be the optimal one.

When PR curves are applied to the PR matrix in my research, the PR matrix is shown in Table 2. In my research, TP means that a student is classified as proficient in the benchmark
assessment and is actually classified as proficient in the state assessment, TN means that a student is classified as not proficient in the benchmark assessment is actually classified as not proficient in the state assessment, FP means that a student is classified as proficient in the benchmark assessment is classified as not proficient in the state assessment, and FN means that a student is classified as not proficient in the benchmark assessment is classified as proficient in the state assessment.

Table 2

*Precision Recall Matrix in the Benchmark and State Assessments*

<table>
<thead>
<tr>
<th>Benchmark assessment</th>
<th>State assessment</th>
<th>Proficient (+)</th>
<th>Non-proficient (-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proficient (+)</td>
<td>Students classified as proficient in the benchmark and state assessments</td>
<td>Students classified as proficient in the benchmark but non-proficient in the state assessment.</td>
<td></td>
</tr>
<tr>
<td>Non-proficient (-)</td>
<td>Students classified as non-proficient in the benchmark but proficient in the state assessment</td>
<td>Students classified as non-proficient in the benchmark and state assessments</td>
<td></td>
</tr>
</tbody>
</table>

**Precision Recall and Receiver Operating Characteristic Curves**

In addition to PR curves, a Receiver Operating Characteristic (ROC) curve is also a useful tool to predict the probabilities of a binary outcome. Both ROC and PR curves are similar in providing a diagnostic tool for binary classification (Brownlee, 2020; Davis & Goadrich, 2006). In the ROC curve, the x-axis shows false positive rates (i.e., the number of $FP$ divided by the number of $FP$ and $TN$) and the y-axis shows true positive rates (i.e., the number of $TP$ divided by the number of $TP$ and $FN$). In general, when the datasets are balanced ones in which
true positive and false positive rates are almost equal, ROC curves are appropriate (Brownlee, 2018; Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015).

Compared with ROC curves, PR curves have some unique features that ROC curves do not have. First, in the case of highly skewed datasets, PR curves provide more precise information of the classification performance (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015). When the datasets are strongly imbalanced ones in which the number of negatives or positives is a lot higher than the other, ROC curves can misrepresent the outcome. However, PR curves can provide an accurate picture of the classification performance. Another characteristic regarding PR and ROC curves is about whether linear interpolation is achievable or not (Boyd et al., 2013; Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015). The interpolation between points making curves is different between PR and ROC curves. Specifically, ROC curves use linear and PR curves use non-linear interpolation. Hence, when the classification results show non-linear interpolation between two points, PR curves should be applicable. Lastly, in terms of the interpretation of the results, PR curves represent the rate of accuracy directly (Powers, 2015; Saito & Rehmsmeier, 2015). The value of precision (i.e., the total number of true positive predictions divided by the sum of the true positives and false negatives) can be simply interpreted as the rate of correct prediction. This indication directly translates to the application of the classification to large datasets, in particular the cases in which an estimate of the number of correct classifications among the positively classified outcomes is critical (Saito & Rehmsmeier, 2015). Therefore, PR curves have an advantage in evaluating the rate of making correct predictions.

Past Research on Classification based on Precision-Recall Curves
When it comes to the practical application of PR curves to policy decisions and evaluation based on assessment scores, there are few studies that used PR curves in the contexts of educational measurement or policy studies. In particular, because PR curves have been long used in computer science fields including machine learning and other related scientific fields, PR curves are relatively new approaches in educational research. Accordingly, there are a very limited number of studies using PR curves that are available in educational research fields.

In terms of using PR curves for binary classification purposes, how PR curves should be applied in future educational measurement and policy studies is important. In practical research settings outside educational research, one of the common characteristics in the studies is to apply PR curves to determine the threshold values for practical decision-making. For example, a study by Schwarm (2005) applied PR curves to evaluate the reading level of topical texts. In his research, assuming that existing measures are not so useful in evaluating reading levels, PR curves are used to determine classifications of appropriate levels of topical texts for English language learners. Specifically, using numerical data with regard to topical texts such as average sentence length and average number of syllables per word, PR curves are used to determine at which thresholds each topical text should be classified into an appropriate reading level. For another example, PR curves are used to identify the threshold point used to determine the borderline between mild cognitive impairment and dementia. In the research by Facal et al. (2019), a sample of participants completed a baseline assessment and attended an evaluation session for a long period. Their research focused on the threshold point to determine whether a patient should be diagnosed as having dementia or not. By using machine learning algorithms, they tried to find the best performing models in terms of predictive values of conversion from
mild cognitive impairment to dementia. In both studies, PR curves are used as an indication of classification quality. To facilitate binary classification decisions, PR curves can be applied to provide evidence supporting the validity of a threshold point.

On the other hand, within the context of educational research, PR curves are used to evaluate the appropriateness of the cut score used to predict whether college students drop out of schools or not. According to Sivakumar (2016), to reduce student dropouts, a decision matrix is developed to predict student dropout in advance by using student data including students’ parents’ education levels and occupations, student grade, and educational status. To improve the predictive accuracy of student dropout, PR curves are applied to at what point college students should be flagged as potential students who are expected to drop out. In this study, PR matrix is shown in Table 3. If the accuracy of predicting students’ dropout improves, the decision matrix can work better as a precautionary and advisory measure to prevent student dropout.

**Table 3**

*Precision Recall Matrix in the Example*

<table>
<thead>
<tr>
<th>Predicated outcome</th>
<th>Actual outcome</th>
<th>Actual dropout (+)</th>
<th>Actual non-dropout</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted dropout</td>
<td>Students predicted as dropout actually drop out</td>
<td></td>
<td>Students predicted as dropout do not actually drop out</td>
</tr>
<tr>
<td>Predicted Non-dropout</td>
<td>Students predicted as non-dropout actually drop out</td>
<td></td>
<td>Students predicted as non-dropout do not actually drop out</td>
</tr>
</tbody>
</table>
In this study, PR curves are used to enhance the ability of predicting student dropout based on educational datasets. The results of PR curves help with setting up the guideline and policies related to reducing student dropout rate. In this way, PR curves show the optimal cut points at which policy decisions should be made.

Overall, past studies present some practical applications of PR curves to evaluate the appropriateness of the cut score to improve predictive accuracy. As a common feature in past studies, PR curves are used as an indication of classification quality. To support binary classification decisions, PR curves can be applied to provide evidence supporting the validity of a threshold point. A graphical measure of PR curves also helps to demonstrate how PR curves serve as a classification assessment measure. Hence, this research will provide a practical use of PR curves to determine the appropriateness of the cut score in the educational measurement context.
CHAPTER FOUR

RESULTS

In this chapter, the results of descriptive statistics, RDD, and PR curves are presented respectively. First, the descriptive statistics of sampled students are shown in terms of test scores, demographics, and other student characteristics. In the descriptive statistics, overall test results in both benchmark and state assessments are provided. The results of the specific students whose scores are near the proficiency cut scores in the benchmark and state assessments are also included. Next, the results of the diagnostic analysis are presented to check the assumptions of RDD. The test of the assumptions is necessary in ensuring that RDD can be performed in the study. After the assumptions are checked, the results of the RDD are presented. Focusing on the proficiency cut scores determined in the benchmark assessment company, the results are shown to attend to the first research question regarding the causal impact of the benchmark assessment on the state assessment. Lastly, to address the second research question, the result of the PR curve is presented to investigate the optimal proficiency cut score in the benchmark assessment needed to maximize the predictive accuracy of the cut score.

Results of Descriptive Statistics

Overall Results of Benchmark and State Assessments

The following tables show the summaries of overall benchmark assessment scores: Table 4-1 by performance level, Table 4-2 by performance levels of male students, and Table 4-3 by performance levels of female students. All tables include numbers of students, mean, median, minimum, and maximum scores. Overall, about a half of the students (n=230; 50.3%) were rated as proficient or higher. In both male and female students, the largest percentages of students
were classified as proficient (n=89; 47.8% for males and n=141; 52.0% for females). Table 5-1 presents the summary of overall state assessment scores by performance levels, with Table 5-2 focusing on male students, and Table 5-3 female students. All tables include numbers of students, mean, median, minimum, and maximum scores. As the results of the benchmark assessment show, more female students (n=271; 59.3%) participated in the state assessment than males (n=186; 40.7%).

Table 4-1

*Results of Overall Benchmark Assessment by Performance Level*

<table>
<thead>
<tr>
<th>Performance Level</th>
<th>Number of students</th>
<th>Mean</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below Basic</td>
<td>33</td>
<td>23.05</td>
<td>25</td>
<td>10.71</td>
<td>28.57</td>
</tr>
<tr>
<td>Basic</td>
<td>118</td>
<td>44.98</td>
<td>46.43</td>
<td>32.14</td>
<td>53.57</td>
</tr>
<tr>
<td>Proficient</td>
<td>230</td>
<td>69.83</td>
<td>71.43</td>
<td>57.14</td>
<td>82.14</td>
</tr>
<tr>
<td>Advanced</td>
<td>76</td>
<td>89.94</td>
<td>89.29</td>
<td>85.71</td>
<td>100</td>
</tr>
<tr>
<td>Overall</td>
<td>457</td>
<td>63.38</td>
<td>64.29</td>
<td>10.71</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 4-2

*Results of Overall Benchmark Assessment by Performance Level for Male Students*

<table>
<thead>
<tr>
<th>Performance Level</th>
<th>Number of students</th>
<th>Mean</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below Basic</td>
<td>18</td>
<td>21.23</td>
<td>25</td>
<td>10.71</td>
<td>28.57</td>
</tr>
<tr>
<td>Basic</td>
<td>51</td>
<td>45.24</td>
<td>46.43</td>
<td>32.14</td>
<td>53.57</td>
</tr>
<tr>
<td>Proficient</td>
<td>89</td>
<td>69.26</td>
<td>71.43</td>
<td>57.14</td>
<td>82.14</td>
</tr>
<tr>
<td>Advanced</td>
<td>28</td>
<td>89.67</td>
<td>89.29</td>
<td>85.71</td>
<td>100</td>
</tr>
</tbody>
</table>
### Table 4-3

*Results of Overall Benchmark Assessment by Performance Level for Female Students*

<table>
<thead>
<tr>
<th>Performance Level</th>
<th>Number of students</th>
<th>Mean</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below Basic</td>
<td>15</td>
<td>25.24</td>
<td>25</td>
<td>17.86</td>
<td>28.57</td>
</tr>
<tr>
<td>Basic</td>
<td>67</td>
<td>44.78</td>
<td>46.43</td>
<td>32.14</td>
<td>53.57</td>
</tr>
<tr>
<td>Proficient</td>
<td>141</td>
<td>70.19</td>
<td>71.43</td>
<td>57.14</td>
<td>82.14</td>
</tr>
<tr>
<td>Advanced</td>
<td>48</td>
<td>90.1</td>
<td>89.29</td>
<td>85.71</td>
<td>100</td>
</tr>
<tr>
<td>All females</td>
<td>271</td>
<td>64.94</td>
<td>67.86</td>
<td>17.86</td>
<td>100</td>
</tr>
</tbody>
</table>

### Table 5-1

*Results of Overall State Assessment by Performance Level*

<table>
<thead>
<tr>
<th>Performance Level</th>
<th>Number of students</th>
<th>Mean</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below Basic</td>
<td>9</td>
<td>812.4</td>
<td>822</td>
<td>744</td>
<td>834</td>
</tr>
<tr>
<td>Basic</td>
<td>216</td>
<td>938.6</td>
<td>940</td>
<td>846</td>
<td>995</td>
</tr>
<tr>
<td>Proficient</td>
<td>209</td>
<td>1053</td>
<td>1051</td>
<td>1003</td>
<td>1125</td>
</tr>
<tr>
<td>Advanced</td>
<td>23</td>
<td>1179</td>
<td>1172</td>
<td>1136</td>
<td>1349</td>
</tr>
<tr>
<td>Overall</td>
<td>457</td>
<td>1001</td>
<td>1003</td>
<td>744</td>
<td>1349</td>
</tr>
</tbody>
</table>

### Table 5-2

*Results of State Assessment of Male Students by Performance Level*
### Table 5-3

**Results of State Assessment of Female Students by Performance Level**

<table>
<thead>
<tr>
<th>Performance Level</th>
<th>Number of students</th>
<th>Mean</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below Basic</td>
<td>6</td>
<td>808</td>
<td>821</td>
<td>744</td>
<td>834</td>
</tr>
<tr>
<td>Basic</td>
<td>93</td>
<td>937.3</td>
<td>940</td>
<td>846</td>
<td>995</td>
</tr>
<tr>
<td>Proficient</td>
<td>81</td>
<td>1054</td>
<td>1042</td>
<td>1003</td>
<td>1125</td>
</tr>
<tr>
<td>Advanced</td>
<td>6</td>
<td>1152</td>
<td>1147</td>
<td>1136</td>
<td>1172</td>
</tr>
<tr>
<td>Overall</td>
<td>186</td>
<td>990.7</td>
<td>995</td>
<td>744</td>
<td>1172</td>
</tr>
</tbody>
</table>

### Comparison of Benchmark and State assessments

Table 6 shows cut score ranges of benchmark and state assessments. As already mentioned, the benchmark assessment scores range from 0 to 100, while the state assessment scores range from 600 to 1641. Accordingly, cut score ranges are very different in state and benchmark assessments. Table 7-1 shows the numbers and percentages of each performance level in both benchmark and state assessments overall, with Table 7-2 by male, and Table 7-3 by
female. From these results, the largest number of students (n=138; 30.1%) is classified as proficient in both assessments overall, and this trend applies to both male and female groups. The benchmark assessment has a higher proportion of students rated as advanced, but the state assessment has a higher proportion of students classified as basic.

In terms of proficiency status, 62.2% (n=143/230) of the students who are rated as proficient in the benchmark assessment reached the proficiency level in the state assessment, while 37.8% (n=85/230) of them did not. On the other hand, 79.7% (n=94/118) of the students who are rated as basic in the benchmark assessment are classified as basic in the state assessment, while 20.3% (n=24/118) of them reached the proficiency level in the state assessment. The results show that the benchmark assessment predicts the basic level of the students more accurately (i.e., about 80% of basic students were accurately predicted as basic) than the proficient level (i.e., about 62% of proficient students were accurately predicted as proficient).

**Table 6**

*Cut Score Ranges in Benchmark and State Assessments*

<table>
<thead>
<tr>
<th>Assessment type</th>
<th>Below Basic</th>
<th>Basic</th>
<th>Proficient</th>
<th>Advanced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmark</td>
<td>0 - 29</td>
<td>30 - 54</td>
<td>55 - 84</td>
<td>85 or above</td>
</tr>
<tr>
<td>State Assessment</td>
<td>600 - 844</td>
<td>845 - 999</td>
<td>1000 - 1129</td>
<td>1130 or above</td>
</tr>
</tbody>
</table>

**Table 7-1**

*Numbers and Percentages of Students by Performance Level in Benchmark and State Assessments*
## Table 7-2

*Numbers and Percentages of Male Students by Performance Level in Benchmark and State Assessments*

<table>
<thead>
<tr>
<th>Performance Level</th>
<th>Benchmark assessment</th>
<th>State assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Below Basic</td>
<td>Basic</td>
</tr>
<tr>
<td>Below Basic</td>
<td>5 (1%)</td>
<td>2 (0.4%)</td>
</tr>
<tr>
<td>Basic</td>
<td>29 (6%)</td>
<td>92 (20%)</td>
</tr>
<tr>
<td>Proficient</td>
<td>1 (0.2%)</td>
<td>24 (5.2%)</td>
</tr>
<tr>
<td>Advanced</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Overall</td>
<td>35 (7.5%)</td>
<td>118</td>
</tr>
<tr>
<td>Performance Level</td>
<td>Below Basic</td>
<td>Basic</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------</td>
<td>-------</td>
</tr>
<tr>
<td>Below Basic</td>
<td>2 (0.7%)</td>
<td>1 (0.4%)</td>
</tr>
<tr>
<td></td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>Basic</td>
<td>13 (4.8%)</td>
<td>50 (21.0%)</td>
</tr>
<tr>
<td></td>
<td>81</td>
<td>31</td>
</tr>
<tr>
<td>Proficient</td>
<td>0</td>
<td>16 (29.9%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>State assessment</td>
<td>Advanced</td>
<td>0 (1.1%)</td>
</tr>
<tr>
<td></td>
<td>67</td>
<td>141</td>
</tr>
<tr>
<td>Overall</td>
<td>15 (5.5%)</td>
<td>(24.7%)</td>
</tr>
<tr>
<td></td>
<td>271</td>
<td></td>
</tr>
</tbody>
</table>

**Table 7-3**

*Numbers and Percentages of Female Students by Performance Level in Benchmark and State Assessments*

<table>
<thead>
<tr>
<th>Performance Level</th>
<th>Below Basic</th>
<th>Basic</th>
<th>Proficient</th>
<th>Advanced</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below Basic</td>
<td>2 (0.7%)</td>
<td>1 (0.4%)</td>
<td>0</td>
<td>0</td>
<td>3 (1.1%)</td>
</tr>
<tr>
<td></td>
<td>57</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic</td>
<td>13 (4.8%)</td>
<td>50 (21.0%)</td>
<td>3 (1.1%)</td>
<td>(45.4%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>81</td>
<td>31</td>
<td>128</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proficient</td>
<td>0</td>
<td>16 (29.9%)</td>
<td>(11.4%)</td>
<td>(47.2%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>State assessment</td>
<td>Advanced</td>
<td>0 (1.1%)</td>
<td>14 (5.2%)</td>
<td>(6.3%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>67</td>
<td>141</td>
<td>48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>15 (5.5%)</td>
<td>(24.7%)</td>
<td>(52.0%)</td>
<td>(17.7%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>271</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Model Assumptions

Prior to applying RDD, I checked the assumptions of the regression model. The first one is about multivariate normality. To check the normality of the residuals, Figure 3 shows the normal Q-Q plots of the residuals. The Q-Q plot shows the normality of the residuals save for ones around very low and high scores. Figure 4 also shows the histogram to check the distribution of the residuals. The histogram clearly shows that the residuals are normally distributed.

The other assumption is about linear fit. Figure 5 shows the plots of standardized residuals. The plot shows that in this data the residuals do not gather at the specific area but are more or less random and scattered below and above the line of zero, although residuals at very low and high scores are not around the standardized zero line. This means that the linear regression did not fit completely but can be considered as the moderately acceptable level.

Figure 3

Q-Q Plot of Residuals of the Regression
**Figure 4**

_Histogram of Residuals of the Regression_

**Figure 5**

_Plot of Standardized Residuals for State and Benchmark Assessment Scores_
In addition to checking the assumptions of standard regression models, this study also tested the assumptions of RDD. First, the assignment variable cannot be impacted by any intervention (Jacob & Zhu, 2012). This means that the assignment variable of benchmark assessment scores cannot change prior to the start of the intervention. In this study, benchmark assessment scores could never change after students took the assessment.

Second, the process of assigning students to intervention and comparison groups must be exogenous and strictly applied to all of the sampled students (Papay et al., 2010). The term “exogenous” literally means being external to the design or study, and an exogenous variable is one that is not impacted by any variable within a study (Jacob & Zhu, 2012). In this study, exogenous means that the proficiency cut score in the benchmark assessment is determined independently of any other variables. All of the samples are assigned to either the intervention group (i.e., a group of proficient students) or the comparison group (i.e., a group of non-proficient students) entirely based on their benchmark assessment scores, and not any other variables. The assignment of students to either of the groups depends only on whether a student’s assessment score reaches the proficiency cut score. In other words, students are randomly assigned to either proficiency or non-proficiency groups. It should be impossible for the students to either reach or fail to reach the proficiency cut score intentionally, because the proficiency cut score was calculated after students took the benchmark assessment. Thus, the proficiency cut score is exogenously established in this study.

Third, when any significant discontinuity is found at the cut-off point, this discontinuity is caused only by the intervention effect; no other factors contribute to the discontinuity (Jacob & Zhu, 2012; Lee & Munk, 2008; Smith, 2014). In this study, the assignment of students to either
proficient or non-proficient groups is only based on students’ benchmark assessment scores. If students received any intervention such as extra class activities or teacher supports, this would have depended entirely on their proficiency status. Thus, if there is any significant discontinuity found at the proficiency cut score in the benchmark assessment, the discontinuity would also result from the decision of proficiency status.

The fourth assumption is that the intervention must be restricted to the samples in the intervention group (Lee & Munk, 2008; Smith, 2014). When any intervention is provided, samples only in the intervention group receive the intervention. Otherwise, the effect of the intervention cannot be estimated for examining any causal effect. In this study, students rated as proficient are expected to be different from those rated as non-proficient in terms of receiving any learning support. Thus, the intervention should not have been confounding to the outcome variable: the state assessment scores.

Lastly, an important condition required to support the validity of the result in RDD is a large sample size, relative to RCT (Shadish et al., 2002; Smith, 2014). Unlike the traditional regression analysis or other econometric analysis using the full sample of data, RDD uses solely the sample of subjects near the cut score (e.g., Flaster & DesJardins, 2014; Jacob & Zhu, 2012; Lee & Munk, 2008; Papay et al., 2010; Smith, 2014). Because of the limited number of samples used in the RDD analysis, RDD needs a sample size 2.5-3 times larger if the study attempts to ensure the same level of statistical power as RCT or regular regression analysis (Jacob & Zhu, 2012; Lee & Munk, 2008; Smith, 2014). In other words, if the sample size is not large enough, this limitation “makes it challenging to use RDD to evaluate interventions targeting only those most at need or those with the greatest merit” (Smith, 2014, p.5). In this study, about two
hundred students are near the proficiency cut score in the benchmark assessment. Accordingly, the issues regarding sample size are discussed in later sections.

**Figure 6**

*Scatter Plot of Benchmark and State Assessment Scores*

*Note.* Red lines show proficiency cut scores in benchmark and state assessments, respectively.

**Results of Regression Discontinuity Design**

Prior to examining the discontinuity in the RDD results, it is necessary to determine the bandwidth on either side of the proficiency cut score. The bandwidth is in close proximity to the cut score within which the sampled students to be analyzed lie. The smaller the bandwidth is, the less sampled students are included in the RDD analysis. As already described, one of the important differences between traditional regression analysis and RDD is that the analysis in
RDD includes only sampled subjects near the cut score, unlike the traditional regression analysis using the full sample of data in the analysis (Flaster & DesJardins, 2014; Umansky, 2016). By determining the bandwidth, the score range with the middle point of the proficiency cut score, it is necessary to identify a group of sampled students in the comparison group - a group of non-proficient students - who can be analyzed as counterfactuals for the intervention group – a group of proficient students. In other words, the sample of students in the comparison group need to be identical to the students in the intervention group (i.e., students in both intervention and comparison groups should have essentially equal learning ability). This means that students whose scores are not near the cut score should not be included in both intervention and comparison groups (i.e., the students whose scores at the end of the score range in both sides should not have identical learning abilities). Including students who have scores too far from the cut score might cause bias of the estimation of the causal impact (Umanskye, 2016). Therefore, in this study, some variations of the bandwidth (e.g., 9 or 12 scores away from the proficiency cut scores) are examined in the analysis of RDD.

Impacts of Proficiency Status with a Cut Score of 55

I conducted the RDD analysis of two bandwidths (i.e., 9 or 13 scores away from the cut scores) with a proficiency cut score of 55, which is originally established by the benchmark assessment company. Table 8 presents the numbers of students included in the RDD analysis, numbers of proficient students in the analyses, average scores in benchmark and state assessments for analyzed students, and standard deviations of their scores, respectively. There are 146 students included in this analysis with benchmark scores that are within nine points of the cut score of 55 (i.e., the score range is from 46.43 to 64.29). When this range is expanded to
scores within 13 points of the potential cut score (i.e., the score ranges from 42.86 to 67.86), the sample size increases to 186. The average benchmark assessment score for the bandwidth of 9 scores is 56.09 and the average state assessment score is 974.19, while the average benchmark assessment score for the score range of 13 scores is 56.74 and the average state assessment score is 980.44.

Table 9 shows the results of the RDD for two bandwidth cases. The results indicate that the estimated effect of the proficiency status in the benchmark assessment on the probability of reaching the proficiency level in the subsequent state assessment is not statistically significant in both cases. In the bandwidth of 9 scores, the coefficient of the discontinuity is -0.05 (p-value is 0.745) in the linear model, while the coefficient in the quadratic model is 0.44 (p-value is 0.824). In the bandwidth of 13 scores, the coefficient of the discontinuity is -0.03 (p-value is 0.824) in the linear model, while the coefficient in the quadratic model is 0.08 (p-value is 0.724). These results demonstrate that proficient and non-proficient students are not significantly different in terms of the probability of reaching the proficiency level in the state assessment. Thus, the RDD suggests that the proficiency status of the benchmark assessment based on the cut score of 55 had no causal impacts on the probability of reaching the proficiency level in the state assessment.

Table 8

Descriptive Statistics of Sampled Students near the Cut Score of 55

<table>
<thead>
<tr>
<th></th>
<th>55 with a bandwidth of 9 scores (±3 correct/wrong answers)</th>
<th>55 with a bandwidth of 13 scores (±4 correct/wrong answers)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmark assessment</td>
<td>State assessment</td>
<td>Benchmark assessment</td>
</tr>
<tr>
<td>assessment</td>
<td>assessment</td>
<td>assessment</td>
</tr>
</tbody>
</table>
Table 9

Result of Checks for Discontinuities at the Proficiency Cut Score

<table>
<thead>
<tr>
<th></th>
<th>55 with a bandwidth of 9 scores</th>
<th>55 with a bandwidth of 13 scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(±3 correct/wrong answers)</td>
<td>(±4 correct/wrong answers)</td>
</tr>
<tr>
<td>Coefficient</td>
<td>SE</td>
<td>Coefficient</td>
</tr>
<tr>
<td>Linear</td>
<td>-0.05</td>
<td>0.16</td>
</tr>
<tr>
<td>Quadratic</td>
<td>0.44</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Note: all coefficients are not statistically significant.

Impacts of Proficiency Status with a Cut Score of 65

I applied the same RDD analysis to two bandwidths with a different cut score of 65.

Although the score of 65 is higher than the original proficiency score of 55 by 10 points, the average score on the state assessment for the students whose benchmark assessment scores are near the score of 65 is higher than a score of 1000, a proficiency cut score in the state assessment. The average state assessment score of the students whose benchmark assessment scores is near the proficiency cut score of 55 is below 1000 (i.e., 974 and 980 for the groups of 
the bandwidth of 9 and 13 scores, respectively) as shown in Table 9. In addition, Table 10 shows the average state assessment scores for each of the benchmark assessment scores near the cut scores of 55 and 66. From the table, students whose benchmark assessment scores 67.86 or higher has an average state assessment score that is higher than the proficiency cut score on the state assessment, 1000. Thus, although the result already indicated that there is no causal impact of the proficiency status with a cut score of 55, the RDD is applied to a score of 65 with the bandwidths of 9 or 13 scores which might be a more plausible proficiency cut score in the benchmark assessment.

Table 11 represents the numbers of students included in this second RDD analysis, numbers of proficient students in the analyses, average scores in benchmark and state assessments for analyzed students, and standard deviations of their scores. There are 168 students included in this analysis with benchmark scores that are within nine points of the potential cut score of 65 (i.e., the score range is from 57.14 to 75.00). When this range is expanded to scores within 13 points of the potential cut score (i.e., the score ranges from 53.57 to 78.57), the sample size increases to 234. The average benchmark assessment score for the bandwidth of 9 scores is 66.14 and the average state assessment score is 1003.95, while the average benchmark assessment score for the bandwidth of 13 scores is 66.87 and the average state assessment score is 1006.72. In both cases, because the average state assessment scores are barely above the proficiency cut score of the state assessment, I used the RDD again to test whether there is any causal impact of the score of 65 on the probability of reaching the proficiency level in the state assessment, assuming that a score of 65 is set as a proficiency cut score.
Table 12 shows the results of the second RDD analysis for two bandwidth cases. The results indicate that the estimated effect of the proficiency status in the benchmark assessment on the probability of reaching the proficiency level in the subsequent state assessment is also not statistically significant in both cases with the cut score of 65. In the bandwidth of 9 scores, the coefficient of the discontinuity is 0.16 (p-value is 0.340) in the linear model, while the coefficient in the quadratic model is 0.26 (p-value is 0.434). In the bandwidth of 13 scores, the coefficient of the discontinuity is 0.07 (p-value is 0.600) in the linear model, while the coefficient in the quadratic model is 0.27 (p-value is 0.260). These results also illustrate that proficient and non-proficient students are not significantly different in terms of the probability of reaching the proficiency level in the state assessment, even when the cut score is set at 65. Thus, the RDD suggests that the proficiency status of the benchmark assessment even based on the cut score of 65 had no causal impacts on the probability of reaching the proficiency level in the state assessment.

Table 10

*Average Scores and Percentages of Proficient Students near Cut Scores of 55 and 65 for Benchmark and State Assessments*

<table>
<thead>
<tr>
<th>Benchmark score</th>
<th>Average state assessment score</th>
<th>Number of students</th>
<th>Number of proficient students in the state assessment</th>
<th>Percentage of proficient students</th>
</tr>
</thead>
<tbody>
<tr>
<td>39.29</td>
<td>928.2</td>
<td>15</td>
<td>1</td>
<td>6.67%</td>
</tr>
<tr>
<td>42.86</td>
<td>966.86</td>
<td>14</td>
<td>4</td>
<td>28.57%</td>
</tr>
</tbody>
</table>
Note. The benchmark assessment scores are calculated by dividing the number of correct answers with the total of 28 items.

Table 11

Descriptive Statistics of Sampled Students near a Score of 65

<table>
<thead>
<tr>
<th></th>
<th>65 with a bandwidth of 9 scores</th>
<th>65 with a bandwidth of 13 scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(±3 correct/wrong answers)</td>
<td>(±4 correct/wrong answers)</td>
</tr>
<tr>
<td>Benchmark</td>
<td>State</td>
<td>Benchmark</td>
</tr>
<tr>
<td>assessment</td>
<td>assessment</td>
<td>assessment</td>
</tr>
<tr>
<td>N</td>
<td>168</td>
<td>168</td>
</tr>
<tr>
<td>N of proficient students</td>
<td>87 (51.8%)</td>
<td>95 (56.6%)</td>
</tr>
<tr>
<td>Average score</td>
<td>66.14</td>
<td>1003.95</td>
</tr>
<tr>
<td>SD</td>
<td>6.27</td>
<td>56.61</td>
</tr>
</tbody>
</table>
Table 12

Results of Checks for Discontinuities at a Score of 65

<table>
<thead>
<tr>
<th></th>
<th>65 with a bandwidth of 9 scores (±3 correct/wrong answers)</th>
<th>65 with a bandwidth of 13 scores (±4 correct/wrong answers)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
<td>SE</td>
<td>Coefficient</td>
</tr>
<tr>
<td>Linear</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>Quadratic</td>
<td>0.26</td>
<td>0.34</td>
</tr>
</tbody>
</table>

Note: all coefficients are not statistically significant.

Result of a Precision Recall Curve

To find an optimal proficiency cut score that maximizes the accuracy of predicting the proficiency/non-proficiency, a PR curve is applied. Figure 7 represents the PR curve in which each dot shows a benchmark score. As a reminder, the y-axis in Figure 7 (precision) represents the proportion of true positives among positive predictions (i.e., the proportion of the students who are classified as proficient in both benchmark and state assessments). On the other hand, the x-axis (recall) represents how accurately the model predicts the positive class, which is considered as a sensitivity (i.e., the proportion of students who are accurately included in the prediction of students who reach the proficiency level in the state assessment). For example, the two dots in the left upper corner (i.e., scores of over 90 points) are high enough to accurately predict students who reached the proficiency level in both benchmark and state assessments. However, because of very high scores, students who did not have a high benchmark assessment score but reached the proficiency level in the state assessments are not included in the prediction.
if over 90 points are used as a proficiency cut score. By contrast, the dots in the lower right corner (i.e., scores of less than 50 points) are low enough to include all students who reached the proficiency level in the state assessment in the prediction if these scores are used as a proficiency cut scores. However, due to very low scores, students who did not reach the proficiency level are inaccurately included in the prediction of who would subsequently reach the proficiency level in the state assessment.

In these two cases, the first case shows that the proportion of students who are classified as proficient in both benchmark and state assessments goes up by increasing the cut-off point but the inaccuracy of failing to predict proficient students who reached the proficiency cut score also goes up. The other case indicates that the inaccuracy of failing to predict proficient students who reached the proficiency cut score also goes down considerably by decreasing the cut-off point but the predictive accuracy of not predicting students who did not reach the proficiency level as non-proficient. In the result of the PR curve, because the optimal cut score that maximizes the predictive accuracy and minimizes the inaccuracy, is considered the dot of the score nearest the upper right corner, the proficiency cut score should be 65 as shown in Figure 7. As already discussed, average state assessment scores of students who have a score of over 65 points in the benchmark assessment are over 1,000, which is the proficiency cut score in the state assessment. Thus, the result of the PR curve, which shows 65 as an optimal cut score, matches the result of the descriptive statistics.

**Figure 7**

*A Precision-recall Curve for an Optimal Proficiency Cut Score in the Benchmark Assessment*
Comparison of two cut scores based on the predictive accuracy

In terms of the accuracy and inaccuracy rates, Table 13 shows percentages of students who were rated as proficient in the benchmark assessment accurately and inaccurately predicted as proficient in the state assessment for two cut scores of 55 and 65. When the proficiency cut score is set as 55, an original cut score established by the benchmark company, 306 students were classified as proficient in the benchmark assessment and 67.7% (n=207/306) of them are accurately predicted as proficient in the state assessment. On the other hand, when the cut score is 65, the number of students rated as proficient goes down to 225, while the accuracy rate of predicting proficient students in the state assessment goes up to 75.6% (n=170/225). This suggests that the cut score of 65 improves the predictive accuracy of the benchmark assessment by about 8%, which also means the predictive inaccuracy decreased by about 8%. This
improvement proves that the PR curve helps determine a proficiency cut score to improve the predictive accuracy. At the same time, this evidence also demonstrates that this method of setting a cut score is based solely on statistical evidence, including the result of a PR curve, rather than on subjective inputs of various stakeholders included in the conventional method.

Table 13

Percentages of proficient students accurately and inaccurately predicted to be proficient in the state assessment

<table>
<thead>
<tr>
<th>Proficiency cut score</th>
<th>Number of students accurately predicted</th>
<th>Number of students inaccurately predicted</th>
<th>Total number of proficient students in the benchmark assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>207 (67.7%)</td>
<td>99 (32.3%)</td>
<td>306</td>
</tr>
<tr>
<td>65</td>
<td>170 (75.6%)</td>
<td>55 (24.4%)</td>
<td>225</td>
</tr>
</tbody>
</table>
CHAPTER 5

DISCUSSION

The purpose of this study was twofold: 1) to evaluate the accuracy of a proficiency cut score in the benchmark assessment in predicting the proficiency level in the subsequent state assessment, and 2) to find an optimal proficiency cut score in the benchmark assessment that maximizes the accuracy and minimizes the inaccuracy in predicting the proficiency level in the state assessment. In this chapter, I summarize the main findings from the study. Then, I describe the limitations of the study. Lastly, I discuss the educational policy issues with the educational significance of the study and conclude with the implementation of future research.

Findings

Impacts of a Proficiency Cut Score

To evaluate the accuracy of the benchmark assessment in predicting the proficiency level in the state assessment as addressed in the first research question, the study applied RDD to two proficiency cut scores (i.e., 55 and 66 points). Specifically, the analysis tested whether there was any statistically significant difference between proficient and non-proficient students near the proficiency cut scores in the probability of reaching the proficiency level in the state assessment. The study also examined whether the proficiency level in the benchmark assessment had any significant impact on the probability of reaching a proficiency level in the state assessment. The results indicate that there were no significant impacts on the proficiency status in the state assessment for both of the two cut scores. This means that the probability of reaching the proficiency level in the state assessment was not significantly different between proficient and non-proficient students, as identified by the benchmark assessment, near the cut score. It also
suggests that even when a student reaches the proficiency level in the benchmark assessment, the proficiency status in the benchmark assessment does not always increase the probability of reaching the proficiency level in the subsequent state assessment.

The non-significant result indicates some potential issues or questions on benchmark assessments. First of all, one of the biggest concerns is about whether the use of benchmark assessments is necessary or not in terms of predicting student future performance measured in end-of-year state assessments. If students rated as proficient in the benchmark assessment have the same level of academic performance as students rated as non-proficient, this means that the benchmark assessment does not predict student future performance levels very accurately. In other words, the non-significant results might make school districts revisit whether benchmark assessments should be used or not for predicting student future performance measured in the end-of-year state assessments.

The other concern is whether proficient students and non-proficient students received the same classroom instructions or not between administrations of benchmark and end-of-year assessments. If the students rated as non-proficient in the benchmark assessments received any additional learning supports that proficient students did not, this extra supports would confound the estimation of the impacts of proficiency status in the benchmark assessment on the likelihood of reaching the proficiency level in the state assessment. In other words, whether students received any extra supports based on the proficiency status in the benchmark assessment could affect the predictive utility of benchmark assessments. Even if students were rated as non-proficient in the benchmark assessment, students could reach the proficiency level in the subsequent state assessment by receiving extra supports before taking the state assessment. By
contrast, even if students reached the proficiency level in the benchmark assessment, students could fail to reach the proficiency level in the state assessment by thinking that they could reach the proficiency level in the benchmark assessment without studying so hard as they had done. In other words, how proficient and non-proficient students studied respectively between two administrations of benchmark and state assessments can influence the ability of benchmark assessments to predict student future performance in the state assessment. Therefore, the research results present that the RDD did not provide strong evidence on where a proficiency cut score is located.

**Optimal Cut Score**

To respond to the other research question (i.e., finding an optimal proficiency cut score in the benchmark assessment), the study used a novel method, PR curves, to find the location of the proficiency cut score that supports the accuracy of the benchmark assessment in predicting the future proficiency status in the subsequent assessment. The PR curve is different from other conventional standard-setting methods of establishing a proficiency cut score, such as Angoff and Bookmark methods in that it seeks to maximize the accuracy and minimize the inaccuracy in predicting a proficiency status in future assessments. To improve the predictive accuracy of the assessment, the PR curve is created to explore what points the proficiency cut score should be. Specifically, the PR curve is constructed by calculating values of precision and recall respectively for each of the benchmark assessment scores. The result represents that a point of 65 is the optimal cut score in terms of predictive accuracy of the benchmark assessment. This means that when a proficiency cut score is set at a point of 65, the predictive accuracy in the benchmark assessment can be maximized and the inaccuracy minimized. The result thereby indicates that
even if the proficiency status of the benchmark assessment is found to have no significant
impacts on the probability of reaching the proficiency level in the state assessment in the RDD
analysis, the PR curve is useful to enhance the ability of the benchmark assessment to predict the
proficiency status in the state assessment.

Originally, the analysis of PR curves is used to evaluate how accurately diagnostic
assessments predict actual classes (e.g., positive or negative). In this research, the PR curve is
used to how accurately the benchmark assessment predicts whether students are proficient or
non-proficient in the future state assessments. If the student rated as proficient in the benchmark
assessment reached the proficiency level in the subsequent state assessment, this is true positive
in the analysis PR curve. If the student rated as non-proficient in the benchmark assessment did
not reach the proficiency level in the subsequent state assessment, the student is predicted
accurately (i.e., true negative). The higher values of true positive and negatives are, the
accurately the benchmark assessment predicts student future performance. On the other hand, if
the student rated as proficient in the benchmark assessment did not reach the proficiency level in
the subsequent state assessment, this is false positive in the analysis PR curve. In addition, if the
student rated as non-proficient in the benchmark assessment reached the proficiency level in the
subsequent state assessment, this is false negative in the analysis PR curve. As values of those
false positive and false negative go up, the benchmark assessment predicts student future
performance inaccurately.

In terms of four categories in the PR curve, the information shown in each category is not
necessarily simple. The two categories, true positive and true negative, can be interpreted simply.
These categories show that the benchmark assessment predicts student performance levels
measured in the state assessment accurately (i.e., proficient students in the benchmark assessment reach the proficiency level in the state assessment or non-proficient students fail to reach the proficiency level). However, the other two categories (i.e., false positive and false negative) present different results. If the result shows a high value of false positive, this means that the benchmark assessment predicts more proficient students by mistake than the assessment should predict. Because some students who should be rated as non-proficient are mistakenly classified as proficient in the benchmark assessment, the result of false positive would mislead these students and their teachers into the judgement that the students can reach the proficiency level in the subsequent state assessment. On the other hand, the category of false negative present a different information. In this category, students who should be rated as proficient are mistakenly classified as non-proficient in the benchmark assessment, this false negative would make the students and their teachers think that the students have to study harder to reach the proficiency level in the state assessment. Thus, two categories, true positive and true negative, present the predictive accuracy of the benchmark assessment, while false negative and false positive affect the predictive accuracy and requires careful interpretations.

**Limitations**

There are several limitations in this study that should be considered. In particular, because the limitations in this study pertain to RDD or educational measurement, all of them need to be noted for the implementation of future research.

One of the limitations with regard to the use of RDD is that the results in estimating the impact of the benchmark assessment might be limited by unobserved variables. Because the available data were benchmark and state assessment scores and student demographics such as
gender, this study could not control for any other factors that might have some effects on the probability of reaching the proficiency level in the state assessment. For example, because the benchmark assessments occur a little earlier than the state assessment, the outcome of the RDD should indicate the proximal impacts of the benchmark assessment. Specifically, after the benchmark assessment, a school district and its schools should provide any extra efforts to help students reach the proficiency level in the state assessment. A possible scenario is that the instructional programs change for proficient and non-proficient students respectively based on the results of the benchmark assessment. One possible scenario is that while proficient students studied in the same classroom curriculum, non-proficient students studied in the same curriculum and received any extra learning support outside their regular classroom. At the same time, however, it may be also possible that proficient students do not study as much as they used to, expecting that they can reach the proficiency level in the subsequent state assessment, while non-proficient students study harder to reach the proficiency level because of not reaching the proficiency level in the benchmark assessment. Whatever scenarios happen, those cases can have some effects on the probability of reaching the proficiency level in the state assessment. If only students near the cut score on the benchmark assessment received any extra learning support, unlike students far away from the cut score, the impact of the benchmark assessment can be estimated with regard to the probability of reaching the proficiency level in the state assessment. As this study did not have the data required to evaluate whether this was the case, the most that can be done is to investigate whether there is any significant difference in the probability of reaching the proficiency level in the state assessment based on the benchmark score. The study found from the analysis of the RDD that students near the proficiency cut score on the
benchmark assessment were not necessarily near the proficiency cut score on the state assessment. Therefore, if any future studies attempt to estimate any causal impacts of the benchmark assessment on the state assessment, it is essential to have the data on whether students near the cut scores receive any extra learning support in their schools in order to reach the proficiency level in the state assessment.

Another limitation is the sample size in this study. In general, the analysis of RDD requires a large sample size to accurately estimate any causal effects (e.g., Cappelleri et al., 1994; Lee & Munk, 2008; Shadish et al., 2004; Smith, 2014). In particular, results of RDD depend on the sample size determined by the location of cut-off points and the bandwidth. When the sample size is small, it might be necessary to increase the bandwidth (i.e., the score range with the middle point of the cut-off point) for including more samples in the study. The non-significant results in the analysis of RDD in this study might be due to the small sample size. Due to the relatively small sample size used in this study, the generalizability of the results would be limited. Thus, it would be necessary to perform the same analysis with a larger sample size so that the result of RDD can be generalized to other assessments.

The last limitation pertains to benchmark assessment scores used in this study. The assessment scores are not derived via psychometric techniques, such as Item Response Theory (IRT). In the current educational measurement, large-scale assessments, including benchmark and state-level assessments, often use IRT for accurate test scoring and the development of assessment items. In IRT models, student ability, traits, or behavior characteristics are accurately measured to improve scoring accuracy. However, the benchmark assessment scores used in the study are raw or percent-correct scores which are calculated simply by counting the number of
correct answers in multiple-choice typed dichotomously scored (i.e., 0 or 1) questions. This means that the item difficulty is not scaled on the same metric as person ability and students’ abilities are not scored based on the estimated item parameters. Because the assessment scores are not converted to IRT-based scores to account for differences in item difficulty and person ability, the inferences or information represented based on percent-correct scores and the precision of the scores are limited in this study.

**Educational Significance**

The goals of this study were originally twofold. First, I argued that because the result of benchmark assessments is used to predict students’ future proficiency status in the subsequent state assessment, the benchmark assessment can have causal impacts on the probability of reaching the proficiency level in the state assessment. If benchmark assessment scores precipitate any kind of intervention, then this may affect their ability to predict the state assessment scores. Based on the results of the benchmark assessment, schools and their teachers might intervene with students who almost reached the proficiency level. However, this study found that there is no evidence of a causal impact of the benchmark assessment on increasing the probability of reaching the proficiency level in the state assessment. This might be due to the data that were available - any unobservable factors that might affect the estimation of impacts of the benchmark assessment were not precisely controlled for. As a matter of fact, there was no available data on whether extra instructional supports were provided to non-proficient students who barely failed to reach the proficiency level in the benchmark assessment. It was also unknown whether students who were rated as proficient in the benchmark assessment were less motivated to study hard to reach the proficiency level in the state assessment. What is desirable is that students who
are proficient in the benchmark assessment are more likely to reach the proficiency level in the state assessment than students who are not proficient. Past literature reported prior classification of students’ performance levels based on assessment scores has been found to impact future academic performance in various past studies (e.g., Papay et al., 2010; Robinson, 2011; Umansky, 2016). In this sense, if a study finds evidence supporting causal impact of the benchmark assessment after controlling for factors affecting state assessment scores – including any interventions that might be given to non-proficient students – this can be viewed as evidence supporting the predictive power of the benchmark assessment.

As the other goal, this study outlined a statistical approach to determine an optimal proficiency cut-off point that district policymakers can use to increase the predictive accuracy of benchmark assessments. Prior to presenting this approach, I discussed why conventional methods of determining the proficiency cut score are arbitrary, lacking the statistical evidence to support the accuracy of predicting the proficiency status in future assessments. In the existing approaches, cut scores are not always designed to improve the predictive accuracy of the assessment, because the cut scores are determined based on various inputs, including the judgments of policymakers, educators, and other experts. In particular, state-wide end-of-year assessments are designed to identify what performance levels each student is at or the level of his or her mastery of state-defined standards. This means that such state-level assessments are not designed to predict any specific data. Thus, various inputs from policymakers, educators, and other experts are necessary in determining cut scores. However, in the benchmark assessment, setting cut scores does not need such inputs but can be based on statistical optimization. Therefore, a statistical method for the setup of cut scores presented in this study can contribute to
the improvement of benchmark assessment - maximizing the predictive accuracy and increasing the value of the benchmark assessment. This method can also be useful even when there are no significant impacts of the benchmark assessment on the probability of reaching the proficiency level in the state assessment. In addition, as long as both benchmark and state assessment data are available, any school district can replicate this method of mixing RDD and PR curves to evaluate the predictive utility of the proficiency cut score in the benchmark assessment. In this sense, the finding of this study suggests that by using both benchmark and state assessment scores, the proficiency cut score can be determined based on statistical evidence, not on arbitrary and subjective judgment, to increase the predictive accuracy of the benchmark assessment.

Third, the findings from this study provide a different insight into the use of the benchmark assessment. As already discussed repeatedly, one of the important roles in the benchmark assessments is to predict students’ future academic performances measured in the end-of-year state assessment. Certainly, it is important to improve the predictive accuracy of the benchmark assessment for school districts, schools, and teachers who use the results. As a matter of fact, school districts need an appropriate cut score for proficiency determination to ensure students reach the proficiency level in the subsequent state assessment (Robinson, 2011). The districts can accordingly establish a corresponding curriculum and their teachers can make classroom instructions for students to reach the proficiency level in the state assessments. However, the results in this study indicated that students who were rated as proficient on the benchmark assessment did not necessarily reach the proficiency level on the state assessment. The study also found that some of the students who were rated as basic in the benchmark assessment reached the proficiency level in the state assessment. It would be a positive result for
the district and its schools to have any of the basic students reach the proficiency level in the state assessment. The result of the benchmark assessment might cause some basic students to study harder to reach the proficiency level on the state assessment. By contrast, it would be a serious issue if some of the students rated as proficient in the benchmark assessment failed to reach the proficiency level in the state assessment. Thus, the results of the benchmark assessment might have some unintended effects on these proficient students.

These unintended effects have some implications. From the perspective of curriculum and learning instruction, students who are rated as proficient in the benchmark assessment are supposed not to need extra learning support, as they are expected to reach the proficiency level in the subsequent state assessment. On the other hand, non-proficiency levels would suggest that the students need extra learning supports to reach the proficiency level. If local schools and their classroom teachers need to decide on whether they should provide extra learning opportunities to help students to reach the proficiency level, the benchmark assessment is expected to provide decisive information. Thus, the more the predictive accuracy of the benchmark assessment increase, the more likely classroom teachers are to use the benchmark assessment results to identify students who need extra learning supports. In addition, when benchmark assessments can predict student’s future proficiency status accurately, teachers can count on the results of the assessment as instructional benefits. When professional supports given to classroom teachers are eliminated or reduced, teachers tend to have fewer opportunities and harder times in interpreting the benchmark assessment results for their improvement in the classroom instructions (Bulkley, Christman, et al., 2010). By contrast, the more the predictive accuracy increases, the more easily individual teachers can use benchmark assessment results as one of the additional tools to
improve their students’ performance. Therefore, the increase in the predictive accuracy of the benchmark assessments has some positive impacts on school curriculum and classroom instructions as a decisive factor to determine which students need extra supports and as additional aids to improve teachers’ classroom instructions.

On the other hand, from a district-wide policy perspective, it is important to know whether school districts have any test-based requirements for students rated as non-proficient in the benchmark assessment. This study found no evidence to support the causal impact of the benchmark assessment on students’ proficiency status in the subsequent state assessment. However, prior to assessing the causal impact of the benchmark assessment, it is necessary to determine whether the sampled school district has any decision-making system based on benchmark assessment scores. For example, if the district has a test-based policy that requires non-proficient students to take extra classes, this means that all of the non-proficient students in the sample should have taken additional classes that proficient students did not take. In such a test-based system, non-proficient students should be more likely to reach the proficiency level in the state assessment than districts which do not have the requirements. This means that in such districts, not only proficient but also non-proficient students should reach the proficiency level in the state assessment. The data in this study cannot speak to such test-based district policy but at least if the school district has such a policy, this can explain why the benchmark assessment was found to have no causal impact on the probability of reaching the proficiency level. In this sense, the improvement in the predictive accuracy of the benchmark assessment is important to the districts which make policy decisions based on benchmark assessment scores.

**Future Research**
This study can make several important contributions to future research with regard to the establishment of the proficiency cut scores. First, the approach discussed in this study presents a statistically supported model that can be used to find a proficiency cut score. The literature review shows that the conventional models in determining cut scores are influenced by policymakers, educators, and other measurement experts. That model does not necessarily improve the accuracy of the assessments to predict students’ future performance levels. However, as already described, benchmark assessment, instead of including the inputs from various experts, can focus on the accuracy of predicting the proficiency levels in future assessments in favor of statistical optimization. In this sense, the model presented in this study can be used to revisit the existing cut scores and to improve the predictive accuracy of benchmark assessments.

Second, this study demonstrates that the proficiency status rated in the benchmark assessment had no causal impact on the probability of reaching the proficiency level in the subsequent state assessment. However, the non-significant results are partly due to no available data that show whether students receive extra learning support based on the benchmark assessment results. If such data is available, it is theoretically possible to examine whether the benchmark assessment has any indirect causal impacts on the probability of reaching the proficiency level in the subsequent state assessment. In other words, observable or unobservable factors, which might affect the estimation of the causal effects, need to be controlled for if the data is available. This issue also applies to the estimation of the causal impacts of the benchmark assessment to cause non-proficient students (e.g., students in the basic level) to study harder to reach higher levels. This study reports that some of the students classified as basic in the
benchmark assessment reached the proficiency level in the state assessment. Because the districts attempt to increase the number of proficient students in the end-of-year state assessment, benchmark assessments can encourage local schools to give more support for students rated as basic to reach the proficiency level. If data is available on whether students in each performance level in the benchmark assessment receive any extra support from their schools or teachers, future studies can re-examine the impacts of the benchmark assessment on the state assessment.

Moreover, if the sample size is bigger, the results might be different. In particular, RDD usually requires a large sample size to estimate any causal impacts of interventions on outcomes. In addition to controlling for influential variables that may affect the estimation of the causal impacts, a large sample size might produce different results in the estimation of causal impacts. Therefore, it is necessary to re-examine the causality of the benchmark assessment with a large sample size for proficient and non-proficient students, respectively, when the data on factors that might affect the probability of reaching the proficiency level in the state assessment can be included in the analysis.

Lastly, future studies should be conducted to check whether the results in this study are applied to any other grades. This study used a sample of 7th graders, but it is uncertain if there is no causality of the proficiency status in lower or higher grades. Furthermore, it is also necessary to test if the optimal cut score is still the same or not in other grades. This study demonstrated that, even when there are no causal impacts found in the analysis, the alternative option, the use of PR curves, can be used to maximize the accuracy and minimize the inaccuracy in predicting the proficiency status in future state assessments. If any future research finds causal impacts of the benchmark assessment on the proficiency status in future assessments, the PR curves can be
used to support the location of the best proficiency cut score that maximizes the predictive accuracy. In any case, it would be necessary to test whether there are any causal impacts of the benchmark assessments on the likelihood of reaching the proficiency level in subsequent assessments for any other grades.

**Conclusion**

Prior ratings of students’ performance levels based on assessment scores to predict future performance levels in state assessments are still highly necessary to prepare students for subsequent assessments. This helps local districts and schools to improve their instructions for their students. In particular, school districts inform their local schools of the scores and numbers of students who reach the performance levels in the benchmark assessments. Responding to the assessment reports delivered from the districts, local schools provide their students with more learning opportunities through specialized services so that more and more students can reach the proficiency level in the end-of-year state assessments. To promote school district and school activities to increase the proficiency rates in the state assessment, benchmark assessments should play an important role in measuring students’ prior academic performance, that is how students tend to reach the proficiency level in the state assessment.

In this study, to make benchmark assessments predict students’ future proficiency status more accurately, I focused on the proficiency cut score used to determine the proficiency status – where is the right cut-off point to predict a proficiency level in the future assessment? Accordingly, I added to the existing literature by suggesting a novel method for evaluating the predictive accuracy of the proficiency cut score – the use of RDD and PR curves. I described a general methodology for assessing the causal effect of a proficiency status at two different cut-
off points and for finding an optimal cut score to improve the predictive accuracy. Although applied to a group of middle school students in a specific school district, the method can be used to improve the accuracy of predicting the proficiency status in any large-scale assessments. I believe that this new approach will provide a more objective way to determine students’ performance levels based on assessment scores.
REFERENCES


Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one, 10*(3). https://doi.org/10.1371/journal.pone.0118432


