CONTRIBUTIONS OF STUDENT RESPONSE THEORY TO EVALUATION SYSTEMS: THREE ESSAYS EXTENDING ITEM RESPONSE THEORY PROCEDURES TO MEASURES OF TEACHER PERFORMANCE

By

Tara Kilbride

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Measurement and Quantitative Methods—Doctor of Philosophy

2019

ABSTRACT

CONTRIBUTIONS OF STUDENT RESPONSE THEORY TO EVALUATION SYSTEMS: THREE ESSAYS EXTENDING ITEM RESPONSE THEORY PROCEDURES TO MEASURES OF TEACHER PERFORMANCE

By

Tara Kilbride

This dissertation is a collection of three papers that each use Student Response Theory (SRT), a method for estimating educator effectiveness built upon an analogy of students and teachers to test items and examinees in Item Response Theory (IRT). Prior studies compare SRT to competing methods like value-added and student growth percentile models. This study broadens this literature by shifting the primary focus to comparisons between SRT and IRT and implications of their differences in the context of teacher evaluation and accountability. Using student-teacher linked data from an anonymous large school district in a major U.S. city, unique contributions of IRT concepts and procedures are examined within the SRT context.

The first paper focuses on the construction of a student instructional demand index that operates like an item difficulty parameter in an SRT model. Although prior studies recommend two different methods for estimating this index, they focus primarily on one method (regression analysis) and minimally on the other (IRT calibration). In this paper, I select indicators of instructional demand that are optimal for IRT calibration and compare instructional demand indices across model specifications and estimation methods. I find that the IRT calibration method yields estimates of instructional demand that are more consistent with other indicators of teacher quality than the regression analysis method. However, when students and teachers are evaluated against a difficult standard, neither type of instructional demand index is consistent with these other measures.

In the second paper, I define mathematical functions of SRT model parameters that are comparable to characteristic curves and information functions in IRT. From these functions, I derive several different indicators of teacher performance and compare their respective meanings and the quality of information they each provide. I then identify groups of teachers whose classes comprise equivalent levels of instructional demand, conceptually similar to parallel test forms in IRT. I find high levels of consistency between different effectiveness measures from the same SRT model. The results also suggest that state-determined thresholds for proficiency and advanced proficiency are so difficult for most students in this district that they provide little information about educators. The basic proficiency threshold, however, is informative about most students and teachers.

The third paper applies procedures related to differential item functioning, common item equating, and item selection to the SRT framework as potential tools in a formative teacher evaluation system. Tests of "differential student functioning" find consistent model performance across most subgroups of teachers and classrooms. Although SRT models are successfully equated to a common scale across years, some estimates of teacher growth have unreasonably large standard errors, offering little power to make inferences about changes in performance. Mismatch between the abilities of teachers and demands of their students are the main factor driving this finding. Applications of optimal item selection procedures to SRT, however, may be useful for matching students and teachers with one another in mutually beneficial ways.

These findings are generally encouraging, but also emphasize the importance of setting reasonable goals for students and teachers and thoughtful assignment practices in determining whether this method will be useful in practice.

ACKNOWLEDGEMENTS

I would first like to express my deepest gratitude to my advisor, Dr. Mark Reckase, whose own work inspired me to pursue this project and whose guidance and support was invaluable throughout my graduate school career. I am also profoundly grateful to Dr. Katharine Strunk, who connected me with the resources necessary to begin this work and whose high standards motivate me to grow as a researcher. Thank you to Dr. Joseph Martineau, whose passion for this topic reminds me why it is important and challenges me to think about it in new ways. Thank you to Dr. Ken Frank, who provided thoughtful feedback and fresh perspectives throughout this process. This work would not have been possible without my fantastic dissertation committee.

My sincerest gratitude to Dr. Josh Cowen for taking a chance by trusting me with a challenging project, and for remaining supportive and invested in my success ever since. Thank you to Brooke Quisenberry and Dr. Ting Shen, who have been great classmates and friends to me since we began our doctoral studies together. Thank you to all of the current and former educators in my life, especially those who took the time to share their experiences and perspectives with me as I completed this work.

My most heartfelt appreciation goes to my family, TG, and Justin for their love, encouragement, and patience. And of course, thank you to Pepper for insisting that I stop working and go for a walk every now and then.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

2PL	Two-Parameter Logistic
AIC	Akaike Information Criterion
CIF	Class Information Function
DIF	Differential Item Functioning
DSF	Differential Student Functioning
E-CIF	Educator-Class Information Function
E-SIF	Educator-Student Information Function
ECC	Educator Characteristic Curve
ELA	English Language Arts
ELL	English Language Learner
E-SCC	Educator-Student Characteristic Curve
ETS	Educational Testing Service
FRL	Free or Reduced-Price Lunch
GRM	Graded Response Model
ICC	Item Characteristic Curve
IIF	Item Information Function
IRF	Item Response Function
IRT	Item Response Theory
LEP	Limited English Proficiency
MGP	Mean or Median Growth Percentile
MH	Mantel-Haenszel

PE	Physical Education
SIDI	Student Instructional Demand Index
SIF	Student Information Function
SCC	Student Characteristic Curve
SD	Standard Deviation
SGP	Student Growth Percentile
SPED	Special Education
SRF	Student Response Function
SRT	Student Response Theory
SWD	Student With Disability
TIF	Test Information Function
URM	Under-Represented Minority
VAM	Value-Added Model
VIF	Variance Inflation Factor

CHAPTER 1. INTRODUCTION

1.1 Background

Objective evidence of performance is critical for making meaningful comparisons of teachers or schools across different contexts. However, the methods used to produce such evidence have historically been both limited and controversial. Statewide school accountability systems implemented after the No Child Left Behind Act of 2001 primarily relied on rates of proficiency on standardized tests as measures of school quality. As research unveiled the biases and unintended impacts of these accountability policies (Booher-Jennings, 2005, Neal & Schanzenbach, 2010, Grissom et al., 2013, Dee et al., 2013), alternative metrics emerged that shifted the focus from achievement to growth (Ladd & Lauen, 2010). By 2016, 40 states required objective evidence of student growth to be a component in teacher evaluations (NCTQ, 2017). Most of these states used statistical growth models (typically a Value-Added Model or "VAM") to generate norm-referenced estimates of teachers' contributions to student growth.

1.1.1 Types and Comparisons of Statistical Growth Models

VAMs estimate the effects of individual teachers on student performance adjusting for prior performance and, in some cases, student demographics and other characteristics (Rothstein, 2008, Hanushek & Rivkin, 2010). The SAS Education Value-Added Assessment System (EVAAS), one of the most widely used VAMs, uses hierarchical linear modeling to estimate these effects (Sanders & Horn, 1994). The Value-Added Research Center (VARC), American Institutes for Research (AIR), and Mathematica Policy Research have each developed similar approaches. Student Growth Percentile (SGP) models, which are sometimes considered a type of VAM, rank students within comparable prior achievement groups using quantile regression, and the mean or median SGP across students with the same teacher (referred to as MGPs) are used as indicators of effective teaching (Betebenner, 2011). Largely due to its use of a familiar percentile rank scale and relatively simple calculation of MGPs from SGPs, this method is often preferred over traditional VAMs for its greater accessibility to stakeholders without statistical backgrounds. However, increased transparency of MGP models and some less sophisticated VAMs may come at the expense of greater bias favoring teachers of more advantaged students (Castellano & McCaffrey, 2017, Walsh & Isenberg, 2013) and greater sensitivity to nonrandom sorting (Guarino et al., 2014).

Transparency and fairness are both central priorities in teacher accountability, but one of these priorities is often compromised for the sake of the other when choosing between the accountability metrics that are typically available to stakeholders. This contributes to an ongoing debate about whether it is appropriate for these metrics to influence decisions about teachers. Specific concerns about some statistical models include effects of non-random sorting of students and teachers (Rothstein, 2009, Harris, 2009), instability of ratings (Koedel & Betts, 2007, Ballou, 2005), effects of omitted variables (Harris, 2009), and the assumption of a linear relationship between prior and expected performance (Lissitz & Doran, 2009). Concerns about transparency include limited interpretability using norm-referenced statistical growth models (Thum, 2003), as there is no direct link between teachers' ratings and concrete performance criteria. Consequently, the results neither deliver explicit guidance for improvement nor provide a basis to assess longitudinal changes in effectiveness. The role of these models in accountability systems, as a result, is generally limited to assigning summative ratings to teachers in order to administer rewards or sanctions and influence personnel decisions, despite experts cautioning against their use in high-stakes decisions (Braun, 2005). This results in apprehension about impacts on instructional climate, educator morale, and student welfare (Lee, 2011).

Reckase and Martineau (2014) proposed an alternate method that offers solutions to many of these concerns: the estimates produced are criterion-referenced and defined in relation to student characteristics, longitudinal changes can be assessed in a meaningful way, random sorting is not assumed, and results may lend well for formative uses in addition to summative. The method, initially titled the Educator Response Function and renamed Student Response Theory (SRT) in subsequent work (Martineau, 2016), is grounded in Item-Response Theory (IRT) methodology. SRT is an extension of techniques typically used to estimate student performance from responses to different types of test questions to the context of estimating educator capacity from outcomes for different types of students. In IRT, examinees' locations on a latent capacity scale are estimated by maximizing the joint probability functions associated with observed strings of responses to test items with specific properties such as item difficulty (Lord, 2012). SRT conceptualizes students as the test items; teachers are tasked with helping students to reach pre-defined performance goals and succeeding to do so with a particular student is analogous to a correct response to a test question. The set of outcomes for all of a teacher's students constitute that teacher's test of capacity.

Just as test questions vary in their levels of difficulty, the difficulty levels of teachers' tasks vary depending on characteristics of their students. For instance, it will likely be easier for teachers to help students who reached an advanced proficiency level in the prior school year to reach a target proficiency level than students who only reached a basic proficiency level. Absences, instances of discipline or behavior problems, and learning disabilities likely increase the level of challenge associated with reaching a performance goal, while high levels of family support, strong work ethic, and good study habits likely decrease it. These factors influence the ease with which students respond to instruction and, therefore, the level of instructional proficiency required for a teacher to successfully help a student reach an academic benchmark. An index of these factors serves as a scale of “instructional demand” analogous to IRT’s difficulty parameter, also referred to as a “student challenge index” in previous studies. The two-parameter logistic (2PL) model also estimates a slope parameter that describes an item’s level of discriminating power. In SRT, a slope parameter is estimated at the teacher level and describes the strength of the relationship between instructional demand and target attainment for a particular teacher (Reckase & Martineau, 2014), also referred to as a “consistency parameter” (Martineau, 2016).

SRT bears some similarities to VAMs and SGPs. Like these statistical growth models, SRT estimates effectiveness of teachers on the basis of test scores, while adjusting for student characteristics that are likely to affect performance. Preliminary studies have demonstrated that rankings of teachers and schools are similar across these different model types. Ham (2014) computed capacity estimates for the same group of teachers using both SRT and VAMs. Rank-order correlations (averaged across grade levels) ranged from 0.69 to 0.89 for reading and from 0.83 to 0.92 for math, depending on the model specifications. Martineau (2016) compared school-level capacity measures using VAMs, SGPs, and SRT. Correlations were highest when SRT performance goals were fixed at a moderate difficulty level (0.81 to 0.84 for reading and 0.82 to 0.76 for math) and with individualized performance goals based on prior performance (0.80 to 0.84 for reading and 0.73 to 0.78 for math). Correlations between estimates from VAMs and SGPs were notably higher (0.91 to 0.99) than the correlations between VAMs or SGPs and

SRT, suggesting that although the capacity estimates produced by SRT are related to those produced by these growth models there may be key differences in the information they capture.

Ham (2014) outlines three key differences that distinguish SRT from other statistical growth models. First, SRT uses student performance data in a different manner than most common growth models. VAMs and SGPs use a continuous student performance variable as an outcome, providing norm-referenced information about the rank of a student score relative to others in the same year and grade level. SRT uses a discrete, criterion-referenced performance outcome, which aligns with the criterion-referenced nature of most state assessments. Second, SRT incorporates student characteristics in a different way. VAMs treat student characteristics as additional educational inputs in the model, while SRT considers these characteristics separately from the outcome variable to determine the different levels of demand posed to educators. Third, the definition of teacher effectiveness differs between the two types of models. VAMs frame teacher effects as the amount of change in student test scores, after accounting for other educational inputs, whereas SRT defines the teacher effect as a latent trait describing the capacity of a teacher to help students reach a particular performance target. As a result, SRT teacher capacity estimates are sample-independent while VAM teacher effects are sample-dependent.

These differences have several implications for the potential role of SRT in an evaluation system. The criterion-referenced capacity scale allows for estimates to be interpreted in terms of the probability that a specific type of student will achieve a specific performance target. The targets typically used in an SRT model correspond to performance level descriptors that explicitly define what students should be able to know or do in order to be considered successful (Martineau, 2016). These ties between observable student characteristics and concrete performance criteria make results more accessible to stakeholders who may be unfamiliar with a particular statistical model but knowledgeable about student performance standards and their relationships to instructional practices. Separate calibration of student parameters and sample-independent properties of capacity estimates lend better to establishing scales that are comparable across contexts and over time, empowering administrators to monitor absolute changes in teacher performance, rather than simply changes in teachers' relative positions within the distribution. Procedures used to examine item difficulty, comparability of test forms, and estimation precision in IRT may add additional value and decision-making power when applied to analogous concepts within an evaluation framework.

1.1.2 The Two-Parameter Logistic Model in SRT

The analogous form of the two-parameter logistic (2PL) model for SRT is shown in Equation 1-1. The variable x_s represents the achievement level of student *s*, while x_t represents the target achievement level for the student. The outcome variable for the SRT model is a binary indicator that equals 1 if the student reaches this target performance level ($x_s \geq x_t$) and equals 0 otherwise ($x_s < x_t$). This is equivalent to a correct response to a particular test item in IRT. The variable θ_{et} represents the latent capacity of educator *e* with respect to performance target *t* and is equivalent to the latent ability level of an examinee in IRT. The level of instructional demand for student *s*, d_s, is equivalent to the item difficulty parameter in IRT and the slope parameter a_e is equivalent to the discrimination parameter in IRT.

There are a few key characteristics that distinguish this model from its IRT counterpart. First, the location parameter is estimated separately from the other model parameters in SRT. This takes place in an earlier step, and pre-calculated instructional demand estimates are treated as constants in the SRT model used to estimate slope and capacity parameters. Second, the slope parameter is estimated at the teacher level (analogous to the examinee level, rather than the item level). This means that the 2PL in SRT includes two second-level (teacher) parameters and only one first-level (student) parameter, whereas the 2PL in IRT includes one second-level (examinee) parameter and two first-level (item) parameters.

$$P_t(x_s \geq x_t \mid \theta_{et},\ a_{et},\ d_s) = \frac{exp(a_{et}(\theta_{et} - d_s)}{1 + exp(a_{et}(\theta_{et} - d_s)} \tag{1-1}$$

where P_t is the probability associated with a particular achievement target *t*;
x_s is the achievement level attained by student *s*;
x_t is the achievement level required to meet target *t*;
d_s is the instructional demand level of student *s*;
a_{et} is the target *t* slope parameter for educator *e*;
and θ_{et} is the latent capacity of educator *e* to help students reach target *t*.

The three main assumptions underlying the 2PL in IRT (De Ayala, 2009) also apply within the SRT framework. The first assumption is that the student response function (SRF) increases monotonically with teacher capacity. This makes sense intuitively: the greater the capacity of a teacher, the more likely a student is to reach a performance target. The second assumption is unidimensionality of teacher capacity. Satisfying this assumption hinges on proper specification of the student instructional demand index, as omitting important indicators from this index is a threat to unidimensionality (Ham, 2014). The third assumption, conditional independence of student outcomes, is perhaps the most concerning of the three within the SRT context. Although the average characteristics of students' peers generally have little effect on teacher value-added estimates (Harris & Sass, 2006), the presence of a disruptive peer has been shown to impact both student achievement and teacher value-added (Horoi & Ost, 2015) and could affect SRT capacity estimates in similar ways.

1.2 The Present Study

The purpose of this study is to examine the unique capabilities of SRT models as tools for evaluating the performance of teachers, making appropriate decisions based on this evaluative information, and ultimately driving improvement in instructional quality. These tasks are organized into three main research questions and addressed in three corresponding standalone papers. The first paper focuses on the construction of the student instructional demand index (SIDI). The second paper explores contributions of IRT-specific procedures and their SRT analogs within a summative evaluation framework. The third paper examines the contributions of these types of procedures within a formative evaluation framework.

1.2.1 Research Questions

The first paper broadly addresses the following question: how does the estimation method and set of indicators used to construct the SIDI affect estimates of teacher capacity and student demand? Proper specification of this index is critical in order to satisfy the unidimensionality assumption. Choices between different estimation methods and indicator sets also have practical implications for the role of an SRT model in an evaluation system; these decisions may affect the point within the school year when information about instructional demand will be available and the way standards and expectations are set for teachers with different types of students. These practical implications are considered alongside the statistical properties of indicators, instructional demand estimates, and SRT models, in order to evaluate the consequences of various decisions in the SIDI construction process.

The second paper explores the following question: can mathematical functions of SRT parameters, comparable to item response and test information functions in IRT, be used to generate more comprehensive summative reports of teacher performance? These mathematical functions are used to frame SRT results in multiple ways, creating several indicators of teacher performance with somewhat different meanings. The statistical validity and practical value of these indicators hinge both on the monotonicity assumption and on the levels of variation and concordance among the different indicators. Taking these factors into consideration, this paper explores whether a combination of the indicators can offer non-redundant, non-contradictory information about educators and can therefore provide more nuanced summaries of their performance than any one of the measures on its own.

The third paper addresses a similar question while shifting outside of the summative evaluation framework: can extensions of IRT procedures to the SRT context contribute meaningfully to a formative evaluation system? IRT procedures related to differential item functioning, test form equating, and optimal item selection are extended to SRT as potential tools for identifying differences in performance across subgroups of teachers and classrooms, monitoring changes in the performance of a teacher over time, and making decisions about how best to match students and teachers with one another. The conditional independence assumption is critical to these applications. The extent to which this assumption is violated may impact whether it is feasible or appropriate to equate the instructional demand index across years, under what circumstances performance is expected to differ from model-based predictions, and whether predictions of future performance can be trusted.

1.2.2 Significance and Implications

The primary purpose of this study is to establish whether, how, and to what extent SRT can contribute to an educator evaluation system in ways that a traditional VAM cannot. Earlier studies demonstrate that SRT can perform the same tasks as a VAM and produce similar results. By leveraging the aspects of SRT that distinguish it from other statistical growth models, arising from the vast body of research and well-studied technology within the IRT framework, this study homes in on the practical relevance of SRT and its unique benefits to an evaluation system.

The proposed adaptations of IRT concepts, procedures, and technology to SRT that are explored in the study each aim to align a statistical growth model more closely with the priorities of educators, administrators, and other stakeholders. The first paper considers several decisions about how to measure instructional demand that determine when this information can be made available and how it relates to educational equity. The second paper expands beyond a simple ranking of educators to construct multifaceted reports of teaching performance that highlight strengths and weaknesses of an individual, consider contextual factors for different comparisons, and frame results in terms of concrete performance criteria. The third paper connects SRT measures across time and contexts, creating a framework for monitoring the performance of individual educators, groups of educators, and the entire population of educators over time, and using this information to better meet the needs of students.

While this study takes critical steps towards refining SRT methods for use in a practical setting, it also identifies areas of concern. Violations of model assumptions and fundamental differences between the IRT and SRT frameworks that were discussed in earlier studies are examined through a different lens, in a different setting, using different data. This contributes to larger conversations about limitations of these measures, contextual factors that contribute to model performance, ways in which IRT and SRT are not truly analogous, and directions for future work.

CHAPTER 2. MEASURING STUDENT DEMAND AND EDUCATOR CAPACITY

2.1 Introduction

Student response theory (SRT) is a method for estimating the effectiveness of an educator using a statistical model analogous to an item response theory (IRT) model (Reckase & Martineau, 2014). One of the primary differences between the two-parameter logistic (2PL) model in IRT and the analogous version of this model in SRT is the manner in which the location parameter is estimated. In IRT, the item location (or item difficulty) parameter is typically estimated concurrently with the item discrimination (or slope) parameter and the latent variable (Lord, 2012). In SRT, the student location (or instructional demand) parameter is estimated independently in an earlier step, and then treated as a constant in the model used to estimate the teacher slope and capacity parameters.

Instructional demand is a construct describing the varying levels of challenge to an educator that are associated with helping different students to reach a performance target, given each student's prerequisite knowledge, behavior, attitude, and different needs as a learner. In order to estimate instructional demand, an index is constructed using a combination of indicators describing the academic performance, behavior, attendance, disabilities, and support structures of students. The inclusion of demographic indicators like ethnicity and socioeconomic status can be controversial: on one hand, there are well-documented relationships between these variables and academic performance (Baker et al., 2016). On the other hand, incorporating these factors can inappropriately imply that these relationships are causal or inadvertently set lower standards for educators of students from disadvantaged backgrounds[1]. This necessitates careful consideration of not only the statistical properties of instructional demand indicators, but also the social consequences of their use in an accountability framework.

Reckase and Martineau (2014) discuss two approaches for estimating the student instructional demand index (SIDI): regression analysis and IRT calibration. The two methods produce SIDIs with fundamentally different meanings, and function best with different types of instructional demand indicators. With the regression analysis approach, student performance at the end of a school year is regressed on performance in the prior school year and a set of student-level covariates believed to influence performance. The SIDI is then defined by determining the predicted performance for each student, and reverse-scaling these values so that lower predicted performance corresponds to a higher level of instructional demand. The ideal covariates for regression analysis are highly-correlated with the outcome variable (future performance) but uncorrelated with each other (Chatterjee & Hadi, 2015). The second approach uses IRT calibration to estimate instructional demand as a latent variable with student-level indicators operating as test items. This is equivalent to the first principal component of a set of risk factors (Martineau, 2016). Because this approach does not rely on current year performance in the estimation process, instructional demand can be calculated before the school year begins. The ideal indicators for this method are highly-correlated with future achievement and also highly-correlated with each other.

Preliminary studies of SRT utilize both estimation methods. Ham (2014) constructed both types of SIDIs using prior performance, attendance, economic disadvantage, FRL eligibility,

[1] A set of guidelines issued by the federal government in 2009 explicitly discouraged this practice on the grounds that expectations of growth and standards of achievement should be the same regardless of these characteristics (U.S. Department of Education, 2009). Researchers have argued that this advice is misguided (Ehlert et al., 2013).

targeted assistant eligibility, special education placement, and limited English proficiency as indicators. While the regression method produced an approximately normal distribution of estimated demand levels, the IRT calibration method resulted in very few possible demand values, and ultimately only the regression-based SIDI was used for the SRT analysis. Martineau (2016) used a similar set of indicators to construct both types of SIDIs and the resulting capacity estimates were markedly different from each other. Correlations with estimates from other SRT models and VAMs were weaker, estimates were more stable across grade levels and years, and correlations with demographic variables were stronger for models with the IRT-calibrated SIDI. The reason for these differences was not entirely clear, and analyses focused primarily on models that used the regression-based SIDI.

One possible explanation for these results is that the sets of instructional demand indicators used in these studies are better suited for the regression analysis method, and that the IRT calibration method would be more successful with a different set of indicators. If this is the case, IRT calibration could offer three noteworthy advantages over regression analysis in this context. First, the regression analysis method is subject to bias from regression to the mean while the IRT calibration method is not (Reckase & Martineau, 2014). Second, the IRT calibration method does not rely on performance outcomes from the current year and can therefore be estimated in advance of an upcoming school year and potentially be used for planning purposes. Third, IRT equating procedures could be applied to SIDIs from different grade levels and years to establish consistent scales across time and across contexts. This study compares different sets of instructional demand indicators, methods for estimating the SIDI, and the resulting estimates of demand and capacity from SRT models under each combination of an indicator set and estimation method. This is the first study to employ such a rich set of instructional demand indicators, as well as the first to select instructional demand indicators based on optimality for IRT calibration, and the first to analyze the impact of omitting demographic variables from the SIDI construction process.

2.2 Data and Methodology

2.2.1 Source Data

This study draws on data from a large, anonymous, urban school district in a major U.S. city. The district serves a diverse population with a large proportion of high-needs students, providing a strong framework for studying the varying levels of demand faced by educators. Table 2-1 provides descriptive information about students and teachers in the sample. The majority of students in the district are nonwhite and eligible for free or reduced-price lunch (FRL), and the percentage of students classified as English language learners (ELL) is far greater than the 2015 national average of 9.5%. Students in this district tend to perform poorly on standardized math and ELA assessments, relative to other students throughout the state. Scores in both subjects are farthest below state averages in earlier grade levels and move closer to the state average for higher grades. Nearly 90% of teachers in the sample are tenured, and only 1-2% are first-year teachers in each grade level each year. Fewer 4th grade students in the first cohort are eligible for gifted programs than 4th grade students in the second cohort. Within cohorts, eligibility in gifted and special education programs increases between the two years in the study. Aside from these exceptions, characteristics of students and teachers are fairly consistent across cohorts, grade levels, and years.

Table 2-1. Summary of student and teacher characteristics by cohort.

	Cohort 1		**Cohort 2**	
School year	2014-2015	2015-2016	2014-2015	2015-2016
Grade level	3rd	4th	4th	5th
Students				
Percent female	48.8	48.8	48.8	49.0
Percent nonwhite	85.2	85.5	85.7	85.6
Percent FRL eligible	83.4	80.0	83.3	79.4
Percent ELL	32.0	29.1	26.9	24.0
Percent special education	11.3	12.7	12.6	13.8
Percent gifted	7.5	9.7	14.1	14.7
Mean number of absences	6.9	6.5	6.7	6.2
Mean standardized math score	-0.61	-0.26	-0.27	-0.09
Mean standardized ELA score	-0.77	-0.41	-0.43	-0.07
Teachers				
Mean number of students	17.5	20.4	20.2	20.7
Percent female	79.9	75.7	75.9	74.2
Percent tenured	88.9	86.3	86.1	86.2
Percent first-year teachers	1.9	1.5	2.1	1.2

Student-level indicators of instructional demand include performance level categories on state standardized math and ELA assessments, course grades for achievement and effort in each of ten subject areas (math, reading, writing, speaking, listening, science, history, health, art, and physical education), number of days enrolled in the district during the year, number of days absent during the year, number of suspensions, length of suspensions, parent education level, eligibility for specific gifted and special education programs, ethnicity, eligibility for free or reduced-price lunch (FRL), gender, language spoken at home, English language learner (ELL) status, and previous ELL status. SIDI estimates are linked to performance levels for the same students on state standardized assessments in the following year (as 4th or 5th graders in 2015-2016) and identifiers for their 2015-2016 classroom teachers. Teacher-level variables used for validation purposes include years of teaching experience and rating categories from classroom observations.

2.2.2 Selecting Instructional Demand Indicators

In order to construct a SIDI using IRT calibration, student-level variables related to instructional demand are redefined as discrete test items for compatibility with a hybrid IRT model. Continuous variables are collapsed into either two or four levels to create a combination of graded-response model (GRM) and two-parameter logistic (2PL) items. All items are coded such that the highest values correspond to the highest levels of demand. For instance, students who reached the advanced proficiency benchmark in the previous year are likely to pose lower levels of demand on their educators and are assigned the lowest value, 0, for the graded-response item constructed from this variable. Students who did not reach any of the performance benchmarks in the previous year are likely to pose higher levels of demand; these students are assigned the highest value, 3. The value definitions for all instructional demand indicators are shown in Table 2-2.

The complete list of indicators is used to estimate an initial SIDI, referred to as the “full” index, before undergoing an item reduction process for alternate versions of the SIDI. The full set of instructional demand indicators is reduced to a “compact” set based purely on empirical relationships between the indicators through an iterative analysis of internal consistency using Chronbach’s coefficient alpha (Cronbach, 1951). In each iteration of the item analysis, indicators with item-test correlations below 0.1 were identified and excluded from future iterations. Once there were no remaining items below 0.1, items with item-test correlations below 0.25 were excluded iteratively until no more remained. A third “restricted” indicator set was selected by repeating the same item reduction process, except that several demographic variables that could be controversial for accountability purposes (ethnicity, gender, free or reduced-price lunch eligibility, parent education level, and home language) were excluded from consideration prior to beginning the item analysis.

Table 2-2. Definitions of discrete levels for instructional demand indicators

Polytomous (GRM) items	**0**	**1**	**2**	**3**
State performance level: math State performance level: ELA	Advanced proficiency	Proficiency	Basic proficiency	Below basic proficiency
Course grade for achievement Course grade for effort	[3.5, 4.0]	[3.0, 3.5)	[2.0, 3.0)	[0.0, 2.0)
Highest degree held by parent	Advanced degree	College degree	High school diploma	None
Number of suspensions Number of days suspended	0	1	2	3+
Number of days absent	0-4	5-10	11-17	18+
Number of days enrolled	180	175-179	155-174	<155
Dichotomous (2PL) items			**0**	**1**
Gender			Female	Male
Eligible for gifted program(s) Native English speaker English is home language Consistent achievement/effort grades throughout year			Yes	No
Under-represented minority (URM) Student with disability (SWD) Eligible for free or reduced-price lunch (FRL) Limited English proficiency (LEP—current or previous)			No	Yes

The three sets of indicators (full, compact, and restricted) are outlined in Table 2-3. The full set, which consists of 73 items, has high internal consistency with *alpha*=0.92. Through the item reduction process, the full set reduces to 33 items in the compact item set and 31 items in the restricted set, both with *alpha*=0.94. While these are the exact sets of indicators used to estimate each IRT-calibrated SIDI, some of the indicators that are considered optimal for IRT

Table 2-3. Instructional demand indicators in the full, compact, and/or restricted SIDI.

Category	GRM items	F	C	R	2PL items	F	C	R
Math performance	Math achievement	X	X	X	Inconsistent math grades	X		
	Math effort*	X	X	X	Inconsistent math effort	X		
	State math level	X	X	X				
ELA performance	Reading achievement	X	X	X	Inconsistent reading grades	X		
	Writing achievement	X	X	X	Inconsistent writing grades	X		
	Speaking achievement	X	X	X	Inconsistent speaking grades	X		
	Listening achievement	X	X	X	Inconsistent listening grades	X		
	Reading effort	X	X	X	Inconsistent reading effort	X		
	Writing effort*	X	X	X	Inconsistent writing effort	X		
	Speaking effort	X	X	X	Inconsistent speaking effort	X		
	Listening effort	X	X	X	Inconsistent listening effort	X		
	State ELA level	X	X	X				
Performance (other subjects)	Science achievement	X	X	X	Inconsistent science grades	X		
	History achievement	X	X	X	Inconsistent history grades	X		
	Health achievement	X	X	X	Inconsistent health grades	X		
	Art achievement	X	X	X	Inconsistent art grades	X		
	PE achievement	X	X	X	Inconsistent PE grades	X		
	Science effort*	X	X	X	Inconsistent science effort	X		
	History effort*	X	X	X	Inconsistent history effort	X		
	Health effort	X	X	X	Inconsistent health effort	X		
	Art effort	X	X	X	Inconsistent art effort	X		
	PE effort	X	X	X	Inconsistent PE effort	X		
Learner needs & support structure	Days absent	X		X	URM status	X	X	
	Days enrolled	X			FRL eligibility	X	X	
	Times suspended	X			Male	X		
	Days suspended	X			Non-native English speaker	X		
	Parent education level	X	X		Home language not English	X		
	Number of disabilities	X	X	X	ELL	X	X	X
					Proficient, non-native speaker	X		
					Previously ELL	X		
					Gifted (any type)*	X	X	X
					Gifted (high intellectual ability)	X	X	X
					Gifted (high achievement)	X	X	X
					Gifted (multiple types)	X		
					SPED: any type*	X	X	X
					SPED: adapted PE	X		
					SPED: autistic	X		
					SPED: behavioral intervention	X		
					SPED: speech impairment	X		
					SPED: occupational therapy*	X		
					SPED: referred for counseling	X		
					SPED: resource specialist	X	X	X
					SPED: specific learning disability	X		
					SPED: any physical disability	X		
					SPED: rare disability	X		
					SPED: any related to learner needs	X	X	X
					SPED: any behavioral disability*	X		

*Item was excluded from regression-based SIDI due to collinearity. F: full, C: compact, R: restricted calibration are inappropriate for the regression analysis method due to collinearity with other indicators. These items, which are marked with an asterisk (*) in Table 2-3, are excluded from the regression-based SIDIs. After excluding collinear variables, there are no pairs of regression model covariates with correlations above 0.8 and no regression model with a variance inflation factor (VIF) above 4.0.

2.2.3 Constructing Indices of Student Instructional Demand

For the IRT-calibration method, the Graded Response Model (Samejima, 2016) shown in Equations 2-1 and 2-2 is used to estimate student levels of instructional demand with each item set (full, compact, and restricted). A single slope parameter is estimated for each item, and a unique difficulty parameter is estimated for each response option beyond the zero category. For the polytomous graded-response items defined in Table 2-2, this yields four total parameters (one discrimination parameter and three location parameters). For the dichotomous items, this reduces to a two-parameter logistic (2PL) model with one discrimination parameter and one location parameter.

$$P^{+}_{j,q}(d_i) = \frac{\exp\,[a_j(d_i - b_{j,q})]}{1 + \exp\,[a_j(d_i - b_{j,q})]} \qquad (2\text{-}1)$$

$$P_{j,q}(d_i) = P^{+}_{j,q}(d_i) - P^{+}_{j,q+1}(d_i) \qquad (2\text{-}2)$$

where d_i is the latent level of instructional demand for student *i*,
$P_{j,q}$ is the probability that a student is in category q of indicator *j*,
$P^{+}_{j,q}$ is the probability that a student is in category q or higher of indicator *j*,
a_j is the discrimination parameter for indicator *j*,
and $b_{j,q}$ is the location parameter for category q of indicator *j*

For each IRT-calibrated SIDI, a comparison SIDI is estimated using the regression analysis method with the same set of indicators (except for the collinear items that were excluded, as indicated in Table 2-3). The regression-based SIDI is computed by estimating a linear regression model with scores on the 2015-2016 state standardized assessment as an outcome variable and 2014-2015 instructional demand indicators as covariates (as shown in Equation 2-3). The instructional demand estimate is equivalent to the reverse of the predicted outcome for a student, as shown in Equation 2-4. Separate models are used for state math and ELA assessments; this means that there are twice as many regression-based SIDIs estimated as there are IRT-calibrated SIDIs, as the IRT calibrated SIDI is not subject-specific. Continuous indicators that were collapsed into discrete levels for the IRT-calibrated SIDI are used in their original forms when estimating the regression-based SIDI.

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 X_1 + \cdots + \beta_k X_k + \varepsilon \qquad (2\text{-}3)$$

$$d = -\hat{y}_t \qquad (2\text{-}4)$$

where y_{it} is the standardized achievement score at time t,
y_{t-1} is the standardized achievement score from the previous year,
X_1-X_k is a set of instructional demand indicators,
d is the instructional demand level for a student

2.3 Results

2.3.1 Comparison of SIDIs Across Indicator Sets and Estimation Methods

Test Information Functions for each of the IRT-calibrated SIDIs are shown in Figure 2-1. Even after drastically reducing the number of indicators for the compact and restricted SIDI, test information levels are impacted very little. The information functions for the compact and restricted SIDIs are nearly identical to one another, which suggests that the impact of excluding demographic indicators is negligible. The shapes of these curves are nearly identical across the two cohorts of students. Table 2-4 provides descriptive statistics for each of the SIDIs. The distributions of instructional demand are consistent across item sets and grade levels. The correlations between student-level demand estimates and future performance on standardized math and ELA assessments are also listed in Table 2-4: all are strong and negative, and correlations are stronger for regression demand indices than for IRT indices.

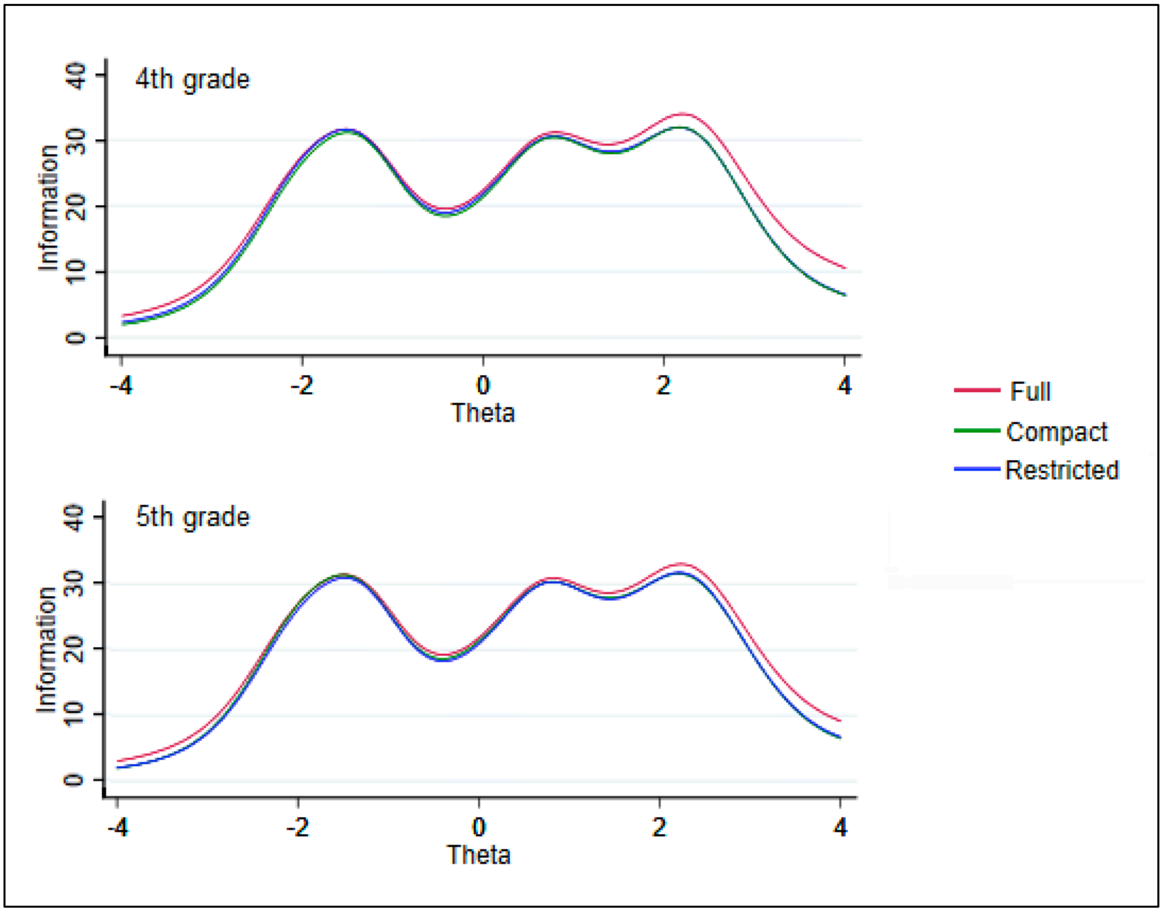


Figure 2-1. Test Information Functions for IRT-calibrated SIDIs

Table 2-4. Distributions of SIDIs and correlations with future state assessment scores

SIDI conditions			Descriptive statistics				Future score correlation	
Method	Grade	Indicators	Mean	SD	Min	Max	Math	ELA
IRT	4th	Full	0.00	0.98	-3.76	3.98	-0.67	-0.70
		Compact	0.00	0.98	-3.43	4.05	-0.68	-0.70
		Restricted	0.00	0.98	-3.34	4.00	-0.67	-0.69
	5th	Full	0.00	0.98	-3.63	4.01	-0.68	-0.71
		Compact	0.00	0.98	-3.30	3.92	-0.69	-0.71
		Restricted	0.00	0.98	-3.21	4.03	-0.68	-0.71
Regression (math)	4th	Full	-0.07	0.87	-2.81	2.56	-0.87	-0.81
		Compact	-0.07	0.87	-2.66	2.50	-0.87	-0.81
		Restricted	-0.07	0.86	-2.69	2.43	-0.87	-0.81
	5th	Full	-0.08	0.88	-2.84	2.60	-0.88	-0.83
		Compact	-0.08	0.88	-2.68	2.57	-0.88	-0.83
		Restricted	-0.08	0.88	-2.69	2.70	-0.88	-0.83
Regression (ELA)	4th	Full	-0.07	0.85	-2.58	2.55	-0.82	-0.86
		Compact	-0.07	0.85	-2.58	2.45	-0.82	-0.86
		Restricted	-0.07	0.85	-2.54	2.53	-0.82	-0.86
	5th	Full	-0.07	0.87	-2.58	2.63	-0.83	-0.87
		Compact	-0.07	0.87	-2.51	2.69	-0.84	-0.87
		Restricted	-0.07	0.86	-2.53	2.68	-0.83	-0.87

Table 2-5 provides correlations among estimates of instructional demand for the same students based on SIDIs estimated with different methods and sets of instructional demand indicators. There are strong, positive, linear relationships between estimates for all pairs of SIDIs. Estimates of instructional demand are nearly identical across pairs of SIDIs with the same estimation method, regardless of which set of instructional demand indicators is used to construct them. In comparison, the relationships between SIDIs constructed using opposite estimation methods are somewhat weaker. However, correlations among these estimates are still rather strong, ranging from 0.77 to 0.81.

Table 2-5. Correlations between estimates from different instructional demand indices

Method	Indicators		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
IRT	Full	(1)		0.99	0.99	0.78	0.78	0.78	0.81	0.81	0.81
	Compact	(2)	0.99		0.99	0.78	0.78	0.78	0.81	0.81	0.81
	Restricted	(3)	0.99	0.99		0.77	0.77	0.78	0.80	0.81	0.81
Regression (math)	Full	(4)	0.77	0.77	0.77		0.99	0.99	0.95	0.95	0.95
	Compact	(5)	0.77	0.78	0.77	0.99		0.99	0.95	0.95	0.95
	Restricted	(6)	0.77	0.78	0.77	0.99	0.99		0.95	0.95	0.95
Regression (ELA)	Full	(7)	0.81	0.81	0.80	0.94	0.95	0.94		0.99	0.99
	Compact	(8)	0.81	0.81	0.80	0.94	0.95	0.94	0.99		0.99
	Restricted	(9)	0.81	0.81	0.80	0.94	0.94	0.95	0.99	0.99	

**Below diagonal: 4th grade, above diagonal: 5th grade*

2.3.2 Comparisons of SRT Models with Different SIDIs

Table 2-6 provides the Akaike Information Criterion (AIC) values for different SRT models. The models differ in which target performance level is used as an outcome variable (basic proficiency, proficiency, or advanced proficiency) and which set of instructional demand indicators is used to compute the SIDI (full, compact, or restricted). AIC comparisons are only meaningful across models for the same grade level, as these estimates were derived from identical data. Across all conditions, the AIC values for models with the most difficult performance target (advanced proficiency) indicate the best fit. SRT models with regression-based SIDIs have AIC values that indicate better fit than the corresponding models with IRT-calibrated SIDIs. There are no major differences in model fit across sets of instructional demand indicators when all other conditions are the same.

Table 2-6. Akaike Information Criterion (AIC) for SRT models

		IRT				**Regression**			
		Math		ELA		Math		ELA	
Target	**SIDI**	4th	5th	4th	5th	4th	5th	4th	5th
Low	Full	-12703	-12865	-13107	-12044	-9088	-8766	-9852	-9026
	Compact	-12649	-12773	-13050	-12005	-9133	-8807	-9891	-9120
	Restricted	-12683	-12820	-13103	-12051	-9203	-8866	-9974	-9173
Middle	Full	-12410	-10270	-12541	-12222	-8317	-6585	-9315	-8949
	Compact	-12308	-10170	-12454	-12135	-8377	-6432	-9333	-9018
	Restricted	-12367	-10232	-12518	-12201	-8436	-6700	-9404	-9085
High	Full	-7105	-6576	-8786	-7709	-4595	-4217	-6295	-5544
	Compact	-7036	-6509	-8708	-7639	-4636	-4252	-6299	-5565
	Restricted	-7084	-6552	-8768	-7697	-4714	-4260	-6365	-5613

Table 2-7 lists ranges of correlations for capacity estimates among pairs of SRT models with specified properties. Correlations of capacity estimates are above 0.90 for all pairs of SRT models with the same performance target and estimation method but different sets of indicators. All of these pairs of models have correlations above 0.99 except when one model uses a regression-based SIDI with the restricted set of indicators (correlations range from 0.91 to 0.94 for these conditions). Correlations are lower when the pair of models uses different performance targets and the same estimation method; they are lower for nonconsecutive targets (one low and one high) and regression-based SIDIs, and about the same regardless of whether the same demand indicators are used. Pairs of models with different SIDI estimation methods show similar patterns across targets and sets of indicators, but correlations are lower overall. Because estimates of capacity are so consistent across sets of indicators, the remaining comparisons focus only on SRT models and SIDIs constructed from the restricted indicator set.

Table 2-7. Range of correlations among capacity estimates for pairs of SRT models.

Model 1 SIDI	Model 2 SIDI	Target	Indicators	Range of correlations
IRT	IRT	Same	Different	>0.99
		Different	Same	0.59 to 0.86
		Different	Different	0.58 to 0.86
Regression	Regression	Same	Different	0.91 to 0.99
		Different	Same	0.40 to 0.73
		Different	Different	0.40 to 0.70
IRT	Regression	Same	Different	0.52 to 0.66
		Different	Same	0.25 to 0.53
		Different	Different	0.25 to 0.53

Scatterplots of teacher capacity rankings[2] across pairs of SRT models across estimation methods and performance targets are provided in Figure 2-2. There are high concentrations of teachers close to the 45-degree line, indicating that their capacity rankings are consistent across the pair of models. However, there are still large quantities of teachers with inconsistent rankings across every pair of SRT models. Patterns of clustering indicate that capacity estimates across somewhat broad ranges in one model are concentrated within limited ranges in other models. This appears to be primarily driven by special education teachers. When these teachers are excluded, as in Figure 2-3, there are far fewer of these clustering patterns. The differences in between Figures 2-2 and 2-3 are most prominent among pairs of models with at least one regression-based SIDI. Regardless, there are still large proportions of teachers ranked inconsistently across models when only general education teachers are included.

[2] Capacity estimates were converted to percentile rankings for these scatterplots in order to improve visibility of patterns in concentrated areas of the capacity distribution.

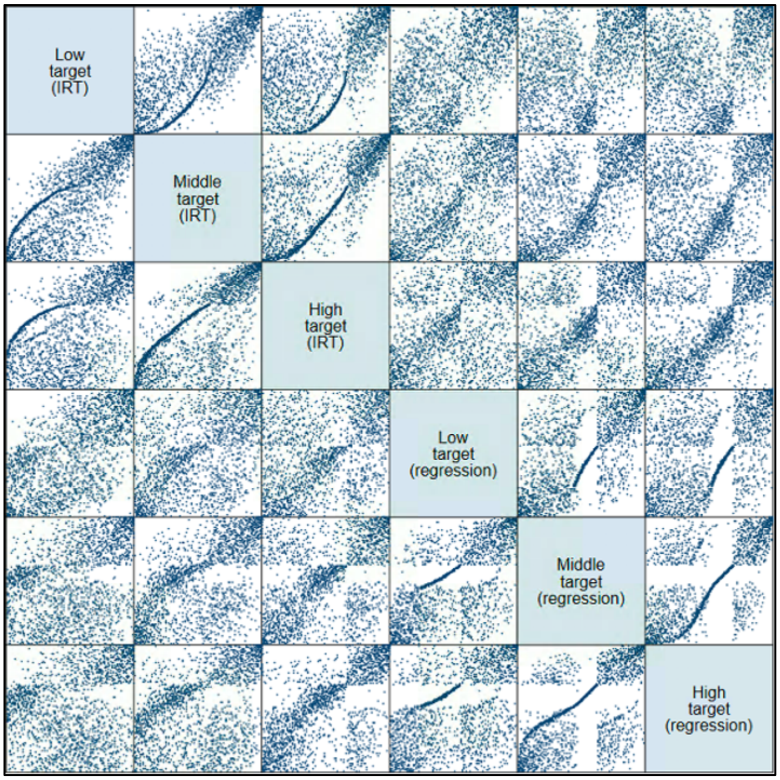


Figure 2-2. *Scatterplot of capacity rankings by SRT model specifications.*

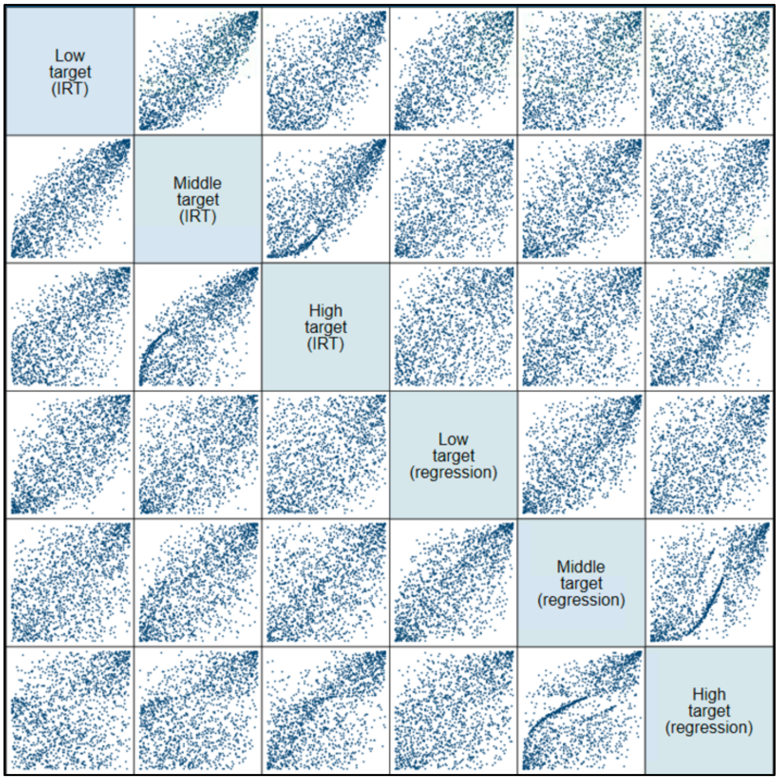


Figure 2-3. *Scatterplot of capacity rankings (special education excluded)*

2.3.3 Comparisons to Other Measures of Educator Quality

Figure 2-4 provides marginal density distributions of capacity percentile rankings across teachers rated in each overall performance category based on observations of their classroom teaching. Marginal density distributions are the most distinct across observation rating categories for the capacity rankings from SRT models with the lowest performance target. Teachers in the lowest observation rating category are generally ranked within a smaller range of the capacity distribution by models with IRT-calibrated SIDIs than by models with regression-based SIDIs. These teachers are also more likely to be ranked in the top half of the capacity distribution by SRT models with regression-based SIDIs than those with IRT-calibrated SIDIs across all performance targets and in both subject areas.

Figure 2-5 provides marginal density distributions of capacity percentile rankings by teaching experience level. Across both subjects and in all but the highest performance target, the distributions of capacity rankings are more distinct across experience categories for SRT models with IRT-calibrated SIDIs than for models with regression-based SIDIs. However, there is no distinction between marginal distributions of capacity rankings by teaching experience level for any SRT models that use the highest performance target as the outcome variable, regardless of the subject area and SIDI estimation method.

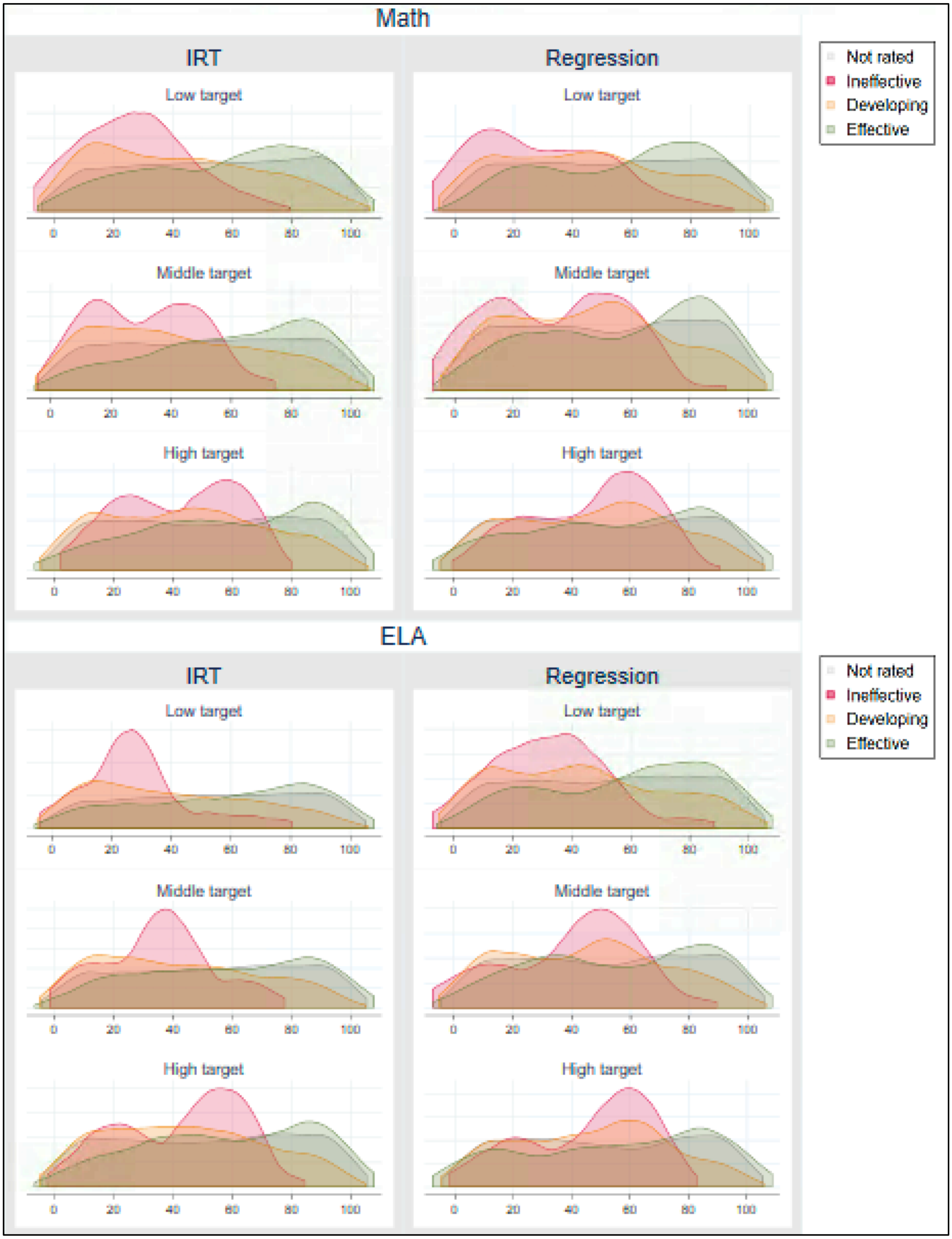


Figure 2-4. *Marginal density distributions by observation rating.*

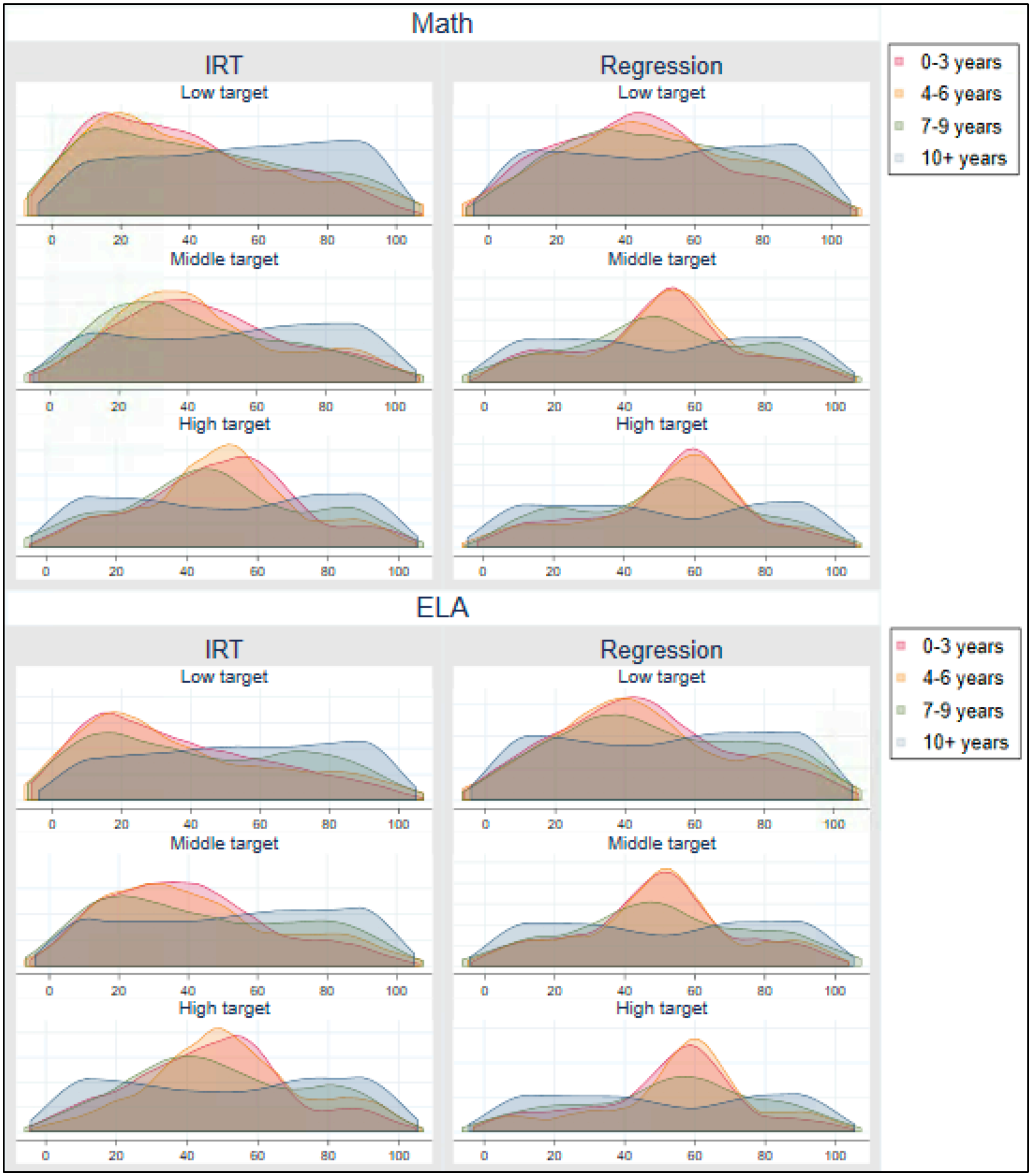


Figure 2-5. *Marginal density distributions by experience level.*

2.4 Discussion

2.4.1 Validity of the SIDI

The types of instructional demand indicators retained after the reduction processes reflect similar balances of academic and behavioral characteristics to the full set. Considering the minimal impact of removing approximately half of the indicators on the test information function and distribution of estimated instructional demand, it can be concluded that the excluded indicators were not making substantial contributions to estimates of instructional demand. These indicators tend to have either low discriminating power or difficulty parameters far outside the observed range of student instructional demand levels. The impact of removing demographic indicators is also negligible. This suggests that other, less controversial indicators may adequately account for the variation in performance associated with differences between demographic groups. Differences in instructional demand can be attributed directly to the types of indicators that were retained in the restricted set. However, there is no strong theoretic basis for attributing differences in instructional demand to demographic group membership. For this reason, the restricted set of indicators is likely the most appropriate for use in an accountability framework.

Similarly, regression-based SIDIs change very little depending on the set of instructional demand indicators used. However, the regression-based SIDIs are slightly more sensitive to these differences than the IRT-calibrated SIDIs. Estimates of instructional demand for the same students from different SIDIs, in general, are affected very little by the set of indicators used or the subject area but are somewhat sensitive to the estimation method. Both methods produce SIDIs that are negatively correlated with future assessment performance, suggesting associations between higher instructional demand levels and lower academic performance. A significant relationship between these quantities is desirable, as this provides evidence of convergent validity. However, some discordance between the two quantities is also desirable in order to demonstrate evidence of divergent validity. The magnitude of correlations for the IRT-calibrated SIDIs is only moderately strong, implying that these SIDIs capture a different underlying construct than the assessment scores. The correlations for regression-based SIDIs are notably stronger; this makes sense considering that this estimation method frames instructional demand estimates in reference to these same assessment scores. However, this does warrant consideration about whether the regression analysis method is capturing the appropriate construct. If instructional demand is a substantively different construct than expected achievement, the regression analysis method may not be capturing this.

2.4.2 Differences Across SRT Specifications

The comparatively better fit for SRT models that use the highest performance target may be misleading, as other findings suggest these measures are the least informative about educator performance. This improvement in model fit could be a result of low variation in the outcome variable, as there are very few students who reach the high performance targets. These results also suggest that the types of students who reach these targets can be predicted from demand indicators with greater accuracy than for the other performance targets. This implies that this outcome is related more to characteristics of the students and less to characteristics of their teachers, relative to the other performance targets. Although this translates to an improvement in model fit, if the outcome is unlikely to be affected by an educator it is probably not an appropriate indicator of educator performance. Differences in fit across SRT models with IRT-calibrated SIDIs and regression-based SIDIs may also be misleading. Regression-based SIDIs are estimated as predicted assessment performance. By definition, these instructional demand estimates account for the greatest possible amount of variation in assessment outcomes. This translates to “better” model fit by construction but does not necessarily imply that these models produce better estimates of educator capacity.

Similar to the estimates of instructional demand, estimates of capacity are quite consistent across different sets of instructional demand indicators, but somewhat less consistent for models with regression-based SIDIs than those with IRT-calibrated SIDIs. The choice of a performance target has a much larger impact, though once again, the impact is more pronounced for models with regression-based SIDIs than those with IRT-calibrated SIDIs. These relationships are likely driven in part by special education teachers. These teachers tend to fall within a restricted range of the capacity scale using regression-based SIDIs but are more dispersed across the capacity distribution using IRT-calibrated SIDIs. This suggests that IRT-calibrated SIDIs might be better able to discriminate between instructional demand levels of special education students and capacity levels of special education teachers. One possible explanation for this is that indicators of eligibility in some specific special education programs were included in IRT-calibrated SIDIs but excluded from regression-based SIDIs due to collinearity. Although these indicators are highly correlated with one another, they may be important to discriminating between the relative demand levels of special education students and the capacities of their teachers.

The marginal distributions of capacity rankings across classroom observation rating categories and teaching experience levels offer insight about the concurrent validity of capacity estimates from SRT models with different specifications. Across all conditions, the highest performance target does not discriminate between these groups of teachers in intuitive ways. The high performance target is unreasonably difficult for the vast majority of students in the district, and as a result, it is rather uninformative about educator quality in most cases. The two lower targets do separate these groups to a greater degree, especially for ELA. Models with IRT-calibrated SIDIs are much more likely to rank "ineffective" teachers in lower parts of the distribution than those with regression-based SIDIs, and much less likely to assign low rankings to "effective" teachers. Across all specifications, models with regression-based SIDIs do not show meaningful distinctions between the distributions of rankings for teachers at different experience levels, while IRT-calibrated SIDI are able to do so in both subject areas for all targets except the highest.

2.5 Conclusions

These results demonstrate that IRT calibration is a feasible method for constructing the student instructional demand index for an SRT analysis. It is important to recognize that the indicators of instructional demand in this study were selected in a manner that is optimal for IRT-calibration. This demonstrates that, with proper selection methods, IRT calibration is a useful tool for demand scale construction, and that these SIDIs outperform regression-based SIDIs in several ways. While results from prior studies favored the regression analysis method, this is likely a result of the type of indicators used to construct the SIDI. This study demonstrates that selecting indicators in a different way can favor a different method. This does not necessarily mean that IRT calibration is a more effective method of scale construction for SRT, but rather that the selection of indicators should be appropriate for the scale construction method of choice.

The IRT calibration method offers several advantages over regression analysis. Its independence from current year achievement allows for instructional demand to be estimated far earlier than with a regression-based SIDI, allowing for this information to influence planning processes for an upcoming school year. The potential ability to equate scales across grade levels and years using parameters of instructional demand indicators also could allow for the measurement of absolute changes (or growth) in teaching capacity over time, rather than simply changes in the relative position of a teacher within the capacity distribution. Considering the benefits IRT calibration may offer to schools and evaluation systems, it may also be worthwhile to invest in the collection of higher-quality or more appropriate instructional demand indicators before implementing an SRT model in a practical setting, rather than relying on a limited but readily-available set of indicators.

The study also highlights a few areas of concern that warrant further examination. Capacity estimates may be less reliable for special education teachers. This is likely related to a mismatch between the instructional demands of their students and the performance targets against which they are assessed. Similarly, when the student performance target in an SRT model is unreasonably high, estimates of capacity are less consistent with other measures of educator quality, even for general education teachers. In order for these tools to be informative and useful in practice, the performance targets considered in SRT models must be both challenging and realistic. Setting appropriate benchmarks for students and teachers will be critical in order for SRT to contribute valuable feedback to an education system.

CHAPTER 3. CONTRIBUTIONS OF SRT TO SUMMATIVE EVALUATION

3.1 Introduction

Within a teacher accountability framework, value-added models (VAMs) and similar statistical growth models typically operate as summative assessments of teaching performance. Estimates of effectiveness are compared against a standard or benchmark to identify high or low performing educators and administer rewards or sanctions (Harris, 2011). This process is conceptually similar to end-of-year student achievement testing to determine a final grade or performance level for the year. Student Response Theory (SRT) bridges the student assessment and educator assessment contexts, using item response theory (IRT) methodology to define a statistical growth model for educators (Reckase & Martineau, 2013). Previous research in this area focuses primarily on estimating and comparing teacher effects across SRT and other statistical growth models like value-added and growth percentile models (Ham, 2014; Martineau, 2016). In this study, the main comparison of interest is between the IRT and SRT frameworks. Procedures commonly used in IRT analyses are studied within the SRT context, with particular regard to the meaning and quality of information they provide and implications of differences across the two frameworks.

3.1.1 Item and Student Response Functions

An item-response function (IRF) is a mathematical equation that relates the probability of a correct response to properties of the test item and the examinee (Lord, 2012). In typical applications of IRT, various quantities derived from the IRF are useful for item selection, outcome prediction, standard setting, and comparing test forms. In SRT, a student response function (SRF) plays a comparable role to the IRF, by modeling the relationship between performance outcomes and characteristics of students and teachers. However, the IRF and SRF are not perfectly analogous to each other, and the ways in which they differ have implications for the interpretation and use of these functions. Corresponding components of the IRF and SRF are shown side-by-side in Table 3-1. The outcome variable and latent variable are both directly analogous across the two functions. The binary outcome variable, *X,* takes on a value of 1 in IRT if a student gives a correct response to a particular test item. In SRT, this variable takes on a value of 1 for a teacher if a pre-defined performance target is achieved by a particular student in their class. The latent variable, θ, represents performance of a student in IRT and of an educator in SRT. The location parameter, d, describes the difficulty of a test item in IRT and the level of instructional demand posed by a student in SRT. While the two location parameters are conceptually similar, they differ in how and when they are estimated. The item difficulty parameter is typically estimated simultaneously with the latent variable and slope parameter. The instructional demand parameter is estimated separately and entered into the SRT model as a constant (Reckase & Martineau, 2014).

The most prominent difference between the IRF and SRF is the nature of the slope parameter, a. In IRT, the slope is estimated at the item level. This parameter describes the extent to which a test item discriminates between examinees with latent ability levels just above and just below the item difficulty level. In order to estimate a perfectly-analogous student slope parameter in an SRT model, students would need to work with many teachers simultaneously and the same achievement target would need to be measured separately with each of these teachers. This is not feasible in the SRT context, as it is incompatible with the way that students and teachers are placed with one another and with the way classroom instruction takes place. For this reason, the SRT slope parameter is estimated at the teacher level. This parameter describes the consistency of expected outcomes across students posing different levels of instructional demand to a teacher. Because of this difference, the SRF is a function of one first-level variable (instructional demand) and two second-level variables (educator capacity and educator consistency), while the IRF is a function of two first-level variables (item difficulty and item discrimination) and one second-level variable (examinee ability).

Table 3-1. Comparison of item and student response functions.

	IRT	**SRT**
Response function	$P(X_i) = \dfrac{exp\left[a_i(\theta_j - d_i)\right]}{1 + exp\left[a_i(\theta_j - d_i)\right]}$	$P(X_s) = \dfrac{exp[a_e(\theta_e - d_s)]}{1 + exp[a_e(\theta_e - d_s)]}$
Dependent variable	Correct response to item (X_i)	Successful student outcome (X_s)
1st level unit	Item (i)	Student (s)
2nd level unit	Examinee (j)	Educator (e)
Location parameter	Item difficulty (d_i)	Student instructional demand (d_s)
Slope parameter	Item discrimination (a_i)	Educator consistency (a_e)
Latent variable	Examinee ability (θ_j)	Educator capacity (θ_e)

3.1.2 Item Characteristic Curves and their SRT Counterparts

In IRT, each individual item has an item characteristic curve (ICC) that is defined by substituting the difficulty and discrimination parameters for the item into the IRF. The ICC illustrates the probability of a correct response to a particular item for examinees at each location along the latent scale (Lord, 2012). The direct analog to the ICC in SRT is a student characteristic curve (SCC) which illustrates the probability that a student with a given level of instructional demand will reach a performance target with teachers of varying capacity levels.

However, there is only one student-level parameter in the SRF and two teacher-level parameters, so substituting the one student-level parameter would leave two unknown quantities in the SCC (capacity and consistency). In order to hold constant two variables that describe the same unit of analysis, Martineau (2016) defined an educator characteristic curve (ECC) that illustrates the relationship between the probability of a successful student outcome for a particular teacher and the level of instructional demand posed by a student.

By defining the characteristic curve in this way, the expected performance of similar students across different teachers can be compared beyond inferences about their relative capacities. If the SRT slope parameter varies among teachers, and equivalently, if the IRT slope parameter varies among items, the characteristic curves for different items or educators will intersect with one another and the relative order of their heights will vary along the latent scale. In SRT, this variability in relative rankings could be particularly useful for evaluating different aspects of teaching performance. This type of analysis may offer more insight into the strengths or weaknesses of a particular teacher in terms of their expected performance with different types of students, compared to a single summative measure.

3.1.3 Item and Test Information Functions and their SRT Counterparts

In IRT, a mathematical function of the slope parameter and the ICC called an item information function (IIF) indicates the amount of Fisher information a particular item provides about examinees at each location along the latent scale. The test information function (TIF), computed as the sum of the IIF accros every item on a test, is inversely related to the standard error of an estimate at a particular location on the latent scale (Lord, 2012). Because of its relationship to estimation precision, the TIF can be used to inform standard setting processes, item selection algorithms, and the development or identification of equivalent test forms (Samejima, 1977a). The direct analog to the IIF in SRT would be a student information function (SIF). The sum of SIFs across all students in the same class would then constitute a class information function (CIF), analogous to the TIF.

In variable-length adaptive testing, items are continually administered until a pre-determined stopping rule, which typically corresponds to a minimum acceptable level of precision, is achieved (Dodd, Koch, & De Ayala, 1993). For instance, if the stopping rule is a reliability coefficient of 0.9, items are administered until the height of the TIF at the estimated ability level of the examinee corresponds to a reliability of 0.9. For a latent trait with a standard deviation of one, test reliability reaches 0.9 when the standard error and test information level are approximately 3.16 and 10, respectively. Within the SRT framework, similar benchmarks could be useful for determining the range of latent teacher capacities for which a particular group of students provides reliable estimates.

The TIF is also used to determine whether different forms of a test can be considered equivalent. Two tests consisting of different items that measure the same construct can be considered or "weak parallel" forms if their TIFs are the same (Samejima, 1977b). This is an important step in ensuring fairness for students taking different forms of the test. Similar techniques can be used to address concerns about fairness for educators evaluated based on different groups of students. Comparisons of classroom-level information functions across teachers may be informative about the demand of a class as a whole, quality of information provided by an estimate, and which groups of teachers are most appropriate to compare with one another.

3.1.4 Differences Between Models and Contexts

A few key differences between IRT and SRT complicate the extensions of these procedures from one context to the other. The primary methodological difference is that the two response functions differ in the number of variables at each level of analysis. The IRF is a function of two first-level variables and one second-level variable, while the SRF is a function of one first-level variable and two second-level variables. This means that, although substituting the first-level (item) parameters into the IRF defines an equation with a single unknown quantity, substituting first-level (student) parameters into the SRF results in an equation with two unknown quantities. Several contextual differences arise from differences in what factors are considered when administering items to an examinee versus assigning students to a teacher. Adjustments may be necessary in order for some equations and procedures developed for IRT to be useful when applied within the SRT framework. This requires careful consideration of the implications of the differences between models and contexts and of the adjustments made to address them.

This study explores possible uses of the SRF and its derivations within a summative teacher evaluation framework. Specifically, the SRF is used to define cut-scores and rating categories with intuitive meanings to facilitate result comprehension. SRT-specific information functions are analyzed to determine the extent of differences across classrooms and implications of these differences for comparing teachers to one another. A cluster analysis of information functions is then used to identify groups of educators whose sets of students constitute reasonably-equivalent tests of teaching performance. Finally, sample reporting materials are presented to illustrate ways this information may be communicated to and interpreted by stakeholders.

3.2 Data and Methodology

3.2.1 Estimates of Demand, Capacity, and Consistency

The study draws on data from an anonymous large urban school district in a major U.S. city. The analytic sample consists of 5th grade students and teachers in the 2015-2016 school year. Estimates of instructional demand, educator capacity, and educator consistency were constructed according to the process described in Chapter 2 and Table 3-2 provides descriptive statistics for each of these quantities. The student instructional demand index (SIDI) was estimated from characteristics of these same students in previous school year (as 4th graders in 2014-2015) in the restricted set of demand indicators using the IRT calibration method. Educator capacity and consistency parameters were estimated for three different SRT models, each with the same SIDI but for three different student performance targets. The three targets correspond to sequential state-determined performance level categories for a standardized mathematics assessment. The labels "basic proficiency", "proficiency", and "advanced proficiency," are used interchangeably with "low," "middle," and "high" to describe the three targets.

Table 3-2. Descriptive statistics for SRT model parameters

	Mean	**SD**	**Min**	**Max**
Student instructional demand	0.00	0.98	-3.21	4.03
Educator capacity: low target	-0.13	2.18	-5.42	6.23
Educator capacity: middle target	-4.57	2.23	-9.85	3.81
Educator capacity: high target	-8.41	2.03	-12.60	1.72
Educator consistency: low target	1.89	0.34	0.17	2.79
Educator consistency: middle target	2.02	0.29	0.25	2.93
Educator consistency: high target	2.09	0.17	1.20	2.67

3.2.2 Defining Characteristic Curves and Information Functions for SRT

Several mathematical functions, comparable to different characteristic curves and information functions in IRT, are defined from the SRT model parameters. Each of these functions is listed in Table 3-3. Martineau (2016) introduces an educator characteristic curve (ECC) that addresses the different numbers of parameters at each level in the item and student response functions. However, a student-level characteristic curve is a closer analog to the ICC and may be of interest for different purposes. Two types of student-level functions are defined, both of which express the probability of achieving a performance target as a function of only one variable, educator capacity, by substituting both a student-level variable (instructional demand) and a teacher-level variable (educator consistency) as constants.

Table 3-3. Mathematical functions of SRT model parameters

$$P_t(x_s \geq x_t) = \frac{exp(a_{et}(\theta_{et} - d_s)}{1 + exp(a_{et}(\theta_{et} - d_s)}$$

$$I_t(\theta \mid d_s, a_t) = a_{et}{}^2 P_t[1 - P_t]$$

Variables:
P: probability
x: achievement level
a: slope
θ: capacity
d: instructional demand
I: information

Subscripts:
s: student
e: educator
t: target
•*:* mean

Name	Function	Level	IRT counterpart
ECC – educator characteristic curve	$P_t(d \mid \theta_e, a_e)$	teacher	ICC
SCC – student characteristic curve	$P_t(\theta \mid d_s, a_\bullet)$	student	ICC
E-SCC – educator-student characteristic curve	$P_t(\theta \mid d_s, a_e)$	student	ICC
SIF—student information function	$I_t(\theta \mid d_s, a_\bullet)$	student	IIF
E-SIF—educator-student information function	$I_t(\theta \mid d_s, a_e)$	student	IIF
CIF—class information function	$\sum_{s(e)} I_t(\theta \mid d_s, a_\bullet)$	teacher	TIF
E-CIF—educator-class information function	$\sum_{s(e)} I_t(\theta \mid d_s, a_e)$	teacher	TIF

The first of these functions, referred to simply as the student characteristic curve (SCC), substitutes the average slope parameter across all teachers as a constant. This function expresses the probability that a particular student will reach a performance goal with a teacher of average consistency, as a function of the capacity of the teacher. The SCC may be useful for judgements about groups of students independently from properties of the teachers to whom they are assigned. For other purposes, such as computing Fisher Information standard errors of teacher capacity estimates, the properties of a specific student-teacher pairing are appropriate. The second student-level characteristic function is informed by properties of both the teacher and student. The educator-student characteristic curve (E-SCC) describes the probability associated with a particular student and a teacher at a particular educator consistency level as a function of educator capacity.

Constructing analogous information functions for SRT also requires careful consideration of the meaning and purpose of the function and implications of the different nature of the slope parameters. If the purpose is to identify equivalent groups of students, then the function should be defined solely by student-level parameters. The student information function (SIF), which is informed by student instructional demand and the mean slope across educators, serves this purpose. The SIF is summed across all students in the same class, yielding a class information function (CIF), that describes the total amount of information that a particular group of students contributes to estimates of capacity for teachers with average slope parameters at each location along the capacity scale. However, if an information function is to be used to judge the precision of a capacity estimate or to determine a cut-score, then it is more appropriate to incorporate properties of both the student and the teacher. The information function defined from the E-SCC is referred to as the educator-student information function (E-SIF). Summed across all students in a class, the resulting educator-class information function (E-CIF) is informative about the information a particular group of students contributes to an estimate of capacity for a particular teacher and can be used as a basis for deriving standard errors of these estimates.

3.2.3 Ranking and Categorizing Teachers

Teachers are assigned percentile rankings according to several different measures. The primary measure is the capacity estimate from the SRT model, which is equivalent to the location on the instructional demand scale where the height of the ECC is exactly 0.5. Teachers are then ranked according to several alternate measures of effectiveness derived from the SRF and its derivations. The first alternate measures are the locations of the instructional demand scale that correspond to ECC heights of 0.25 and 0.75. These measures, referred to as P25 and P75, respectively, indicate the instructional demand level of a student expected to have a probability of reaching a performance target of 0.25 or 0.75 given the capacity and consistency parameters for a particular teacher. The equations used to construct these measures and the steps used to derive them are shown in the Appendix.

Next, teachers are ranked according to the heights of their ECCs at fixed locations along the instructional demand scale that correspond to meaningful distinctions between types of students. These fixed values are selected based on the location parameters for different instructional demand indicators from the IRT calibration process described in Chapter 2. For instance, the location parameter for a student having 5 or more absences is 0.04, indicating that the probability that a student with instructional demand of 0.04 was absent for 5 or more days is 0.5. Students with instructional demand levels above 0.04 have probabilities above 0.5 and students with instructional demand levels below 0.04 have probabilities below 0.5. Three key values are selected, describing low-demand, average-demand, and high-demand students based on location parameters for characteristics associated with these types of students. An instructional demand level of -1.5 was selected as the low-demand value, as students below this level have probabilities above 0.5 of receiving top marks from their teachers for their effort in all subject areas. An instructional demand level of 0 was selected as the average-demand value, as students below this instructional demand level have probabilities of above 0.5 of reaching basic proficiency. An instructional demand level of 1.5 was selected for the high-demand value, as students above this demand level have high probabilities of receiving failing grades in their courses. These measures are referred to as D1, D2, and D3, corresponding to the low-demand, average-demand, and high-demand values, respectively. D1 ranks teachers according to the probability that a low-demand student will reach a particular performance target in their class, while D2 and D3 rank teachers based on the probabilities for average-demand and high-demand students.

These three fixed probabilities and three fixed demand levels each define a unique effectiveness measure derived from the same SRF. These measures, outlined in Table 3-4, are each computed for the three different performance targets, yielding 18 total measures according to which teachers are ranked. Teachers are then assigned categorical effectiveness levels in several ways. For each of the fixed-probability measures, teachers are placed into one of four categories that are separated using the three fixed demand levels as cut-points. Similarly, for each of the fixed-demand measures, teachers are placed into one of four categories that are separated using the three fixed probabilities as cut-points. Teachers are assigned to groups according to each of the six categorization schemes (also outlined in Table 3-4) for each of the

Table 3-4. Definitions of effectiveness measures and levels derived from the SRF.

Effectiveness measures descriptions and equations			**Effectiveness levels** classification criteria and descriptions			
Fixed-probability measures			**L1** $d < -1.5$	**L2** $-1.5 \leq d < 0$	**L3** $0 \leq d < 1.5$	**L4** $d \geq 1.5$
P25	Demand when probability=.25	$\theta_j - \ln\frac{1}{3}/a_j$	Low-demand student very likely to miss target	Low-demand student somewhat likely to hit target	Average-demand student somewhat likely to hit target	High-demand student somewhat likely to hit target
P50	Demand when probability=.50	θ_j	Low-demand student likely to miss target	Low-demand student likely to hit target	Average-demand student likely to hit target	High-demand student likely to hit target
P75	Demand when probability=.75	$\theta_j - \ln 3/a_j$	Low-demand student somewhat likely to miss target	Low-demand student very likely to hit target	Average-demand student very likely to hit target	High-demand student very likely to hit target
Fixed-demand measures			**L1** $p < 0.25$	**L2** $0.25 \leq p < 0.5$	**L3** $0.5 \leq p < 0.75$	**L4** $p \geq 0.75$
D1	Probability if demand=-1.5	$\frac{\mathrm{e}^{a_j(\theta_j+1.5)}}{1+\mathrm{e}^{a_j(\theta_j+1.5)}}$	Unlikely for low-demand student to hit target	Somewhat likely for low-demand student to hit target	Likely for low-demand student to hit target	Very likely for low-demand student to hit target
D2	Probability if demand=0	$\frac{\mathrm{e}^{a_j\theta_j}}{1+\mathrm{e}^{a_j\theta_j}}$	Unlikely for average-demand student to hit target	Somewhat likely for average-demand student to hit target	Likely for average-demand student to hit target	Very likely for average-demand student to hit target
D3	Probability if demand=1.5	$\frac{\mathrm{e}^{a_j(\theta_j-1.5)}}{1+\mathrm{e}^{a_j(\theta_j-1.5)}}$	Unlikely for high-demand student to hit target	Somewhat likely for high-demand student to hit target	Likely for high-demand student to hit target	Very likely for high-demand student to hit target

three performance targets, yielding 18 total categorical assignments for each teacher in the sample.

Groups of teachers with equivalent or comparable groups of students were identified by conducting a cluster analysis of the heights of teachers' CIFs at periodic locations along the instructional demand scale. CIF heights are computed at all integer values between and including -4 and 4. Similar to the manner in which test forms with matching TIFs are classified as parallel test forms, sets of classrooms with similar CIFs are considered reasonably-equivalent comparison groups (or parallel classrooms). Using a hierarchical cluster analysis with Ward's linkage, teachers are classified into comparison groups of teachers whose CIFs are most similar to theirs, indicating that the students they teach comprise equivalent levels of instructional demand.

Statistical and practical properties of clustering solutions are both considered. The Duda-Hart index is one of the statistical properties considered. This index gives the ratio of the within-cluster sum of squared errors after partitioning the data to the sum of squared errors before partitioning the data (Duda & Hart, 1973). The corresponding pseudo T-squared statistic is a measure of cluster similarity taking into account the number of observations (Halpin, 2016). The preferred clustering solutions should correspond to increases in the Duda-Hart index and decreases in pseudo T-squared, relative to the solution with one fewer cluster. Practical considerations include the number and size of clusters in a solution. In order for teachers to be compared meaningfully within clusters, the number of teachers per cluster must be sufficiently large. However, the number of groups must also be adequately high to distinguish between types of classrooms in a meaningful way.

3.3 Results

3.3.1 Distributions of Characteristic Curves and Information Functions

All students in the sample have their own unique SCCs and E-SCCs determined by their estimated instructional demand and either the mean consistency estimate across teachers for a particular performance target or the actual consistency estimate of their classroom teacher for the target. The distributions of these characteristic curves across the sample of students is shown in Figure 3-1, where the 5th, 25th, 50th, 75th, and 95th percentiles of SCC and E-SCC heights are shown at each location along the educator capacity scale. When teacher capacity is at either extreme end of the distribution, there is little variation in student probabilities. The distribution of student probabilities is fairly constant across performance targets, even though the targets vary in difficulty. This is because the instructional demand scale is fixed across targets. Differences in target difficulty, rather, are reflected in downward shifts of the capacity distribution for the higher targets.

The SCC is a function of student instructional demand and the mean slope across the sample of teachers. Instructional demand is constant across performance targets, so the only source of differences in the distribution of SCCs across performance targets is differences in the distributions of teacher slopes for the different performance targets. Although there are slight differences in these mean slopes (as shown previously in Table 3-2), these do not translate to visually-perceptible differences in the corresponding SCC distributions. E-SCC is a function of student instructional demand (which is constant across performance targets) and the slope for a particular student's teacher. Similar to the distributions of the SCC, any differences in the distributions of E-SCCs across performance targets must be attributed to differences in teacher

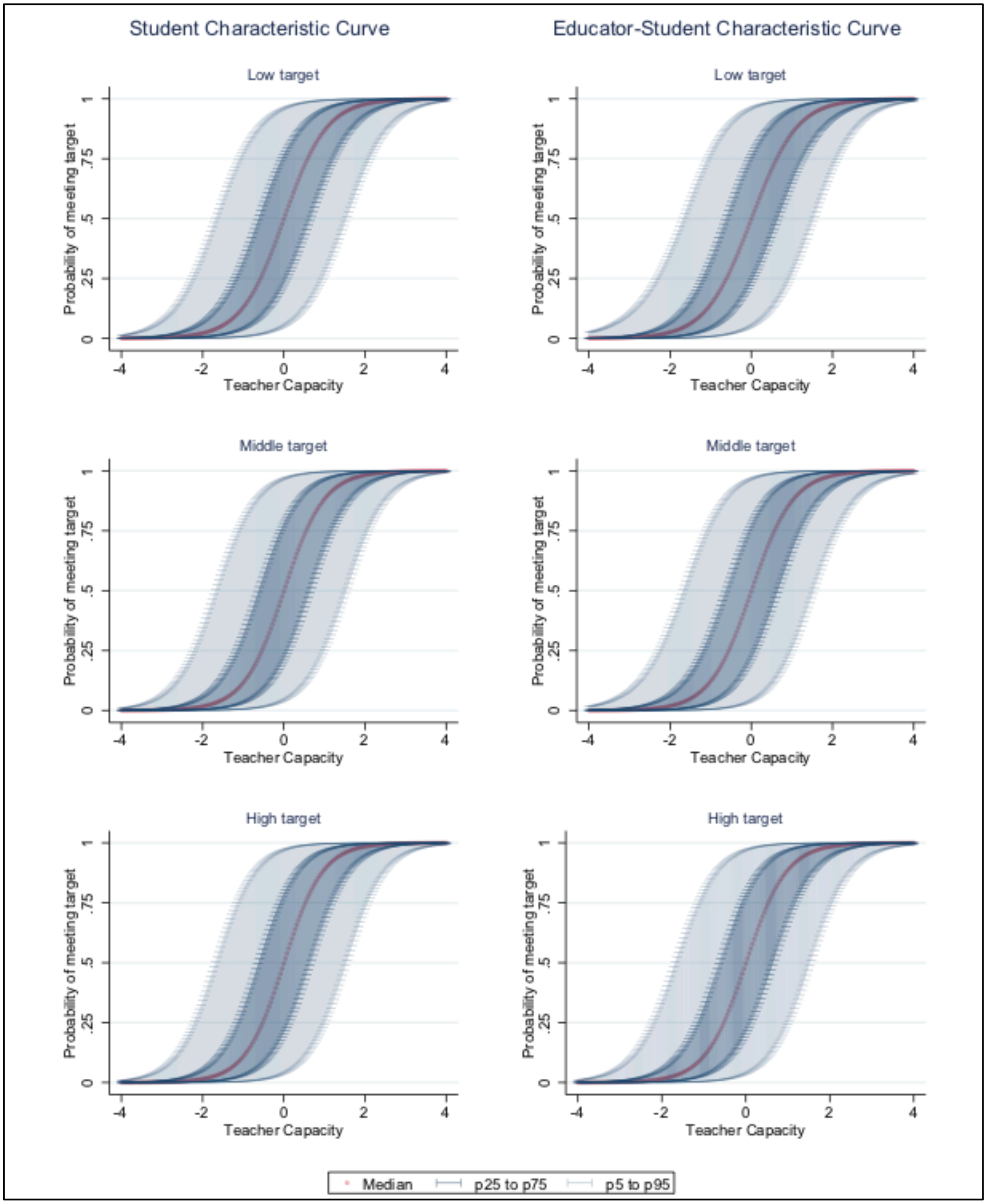


Figure 3-1. Distributions of student and educator-student characteristic curves.

slopes across targets. As was the case with the SCCs, no such differences in E-SCCs are evident from the figure.

The distributions of the SIF and E-SIF across all students are shown in Figure 3-2. The SIF distributions are similar in shape across the three performance targets. However, the peak heights of the 95th curves vary, reaching just below 1, approximately 1, and just above 1 for the low, middle, and high targets, respectively. This is a result of slightly larger slopes, on average, for the higher targets. While the shapes of the 5th, 25th, 50th, and 75th percentiles resemble bell curves, the 95th percentile has a flatter distribution, as the maximum information for any given student is limited by the fixed slope parameter. Excluding the 95th percentile, the distributions of the E-SIF resemble bell curves as well and are similar across targets. While the 95th percentile for the low target is shaped similarly to these, the 95th percentile for the middle and high targets show a comparatively lower level of information for higher-demand students, resulting in an asymmetric curve.

The distributions of the ECC, CIF, and E-CIF across all teachers are shown in Figure 3-3. The educator characteristic curve (ECC) differs most drastically across targets. While there is a high level of variation in probabilities across the entire range of observed student demand values for the low target, the ECC heights for the middle and high targets have very little variation. For the middle target, the median teacher has a probability close to zero for almost the entire range of demand values. While the 75th and 95th percentile have probabilities above zero for a small range of values on the demand scale, probabilities are close to zero for all higher-demand students. For the highest target, even teachers at the 95th percentile have probabilities near zero across most of the instructional demand scale, and no teachers below the 75th percentile have a probability visually distinguishable from zero anywhere in the range of observed demand values. The class CIF and E-CIF distributions are similar in shape across targets. The 95th percentile of the E-CIF peaks higher than the 95th percentile of the CIF, as the teacher slope is not fixed at its mean.

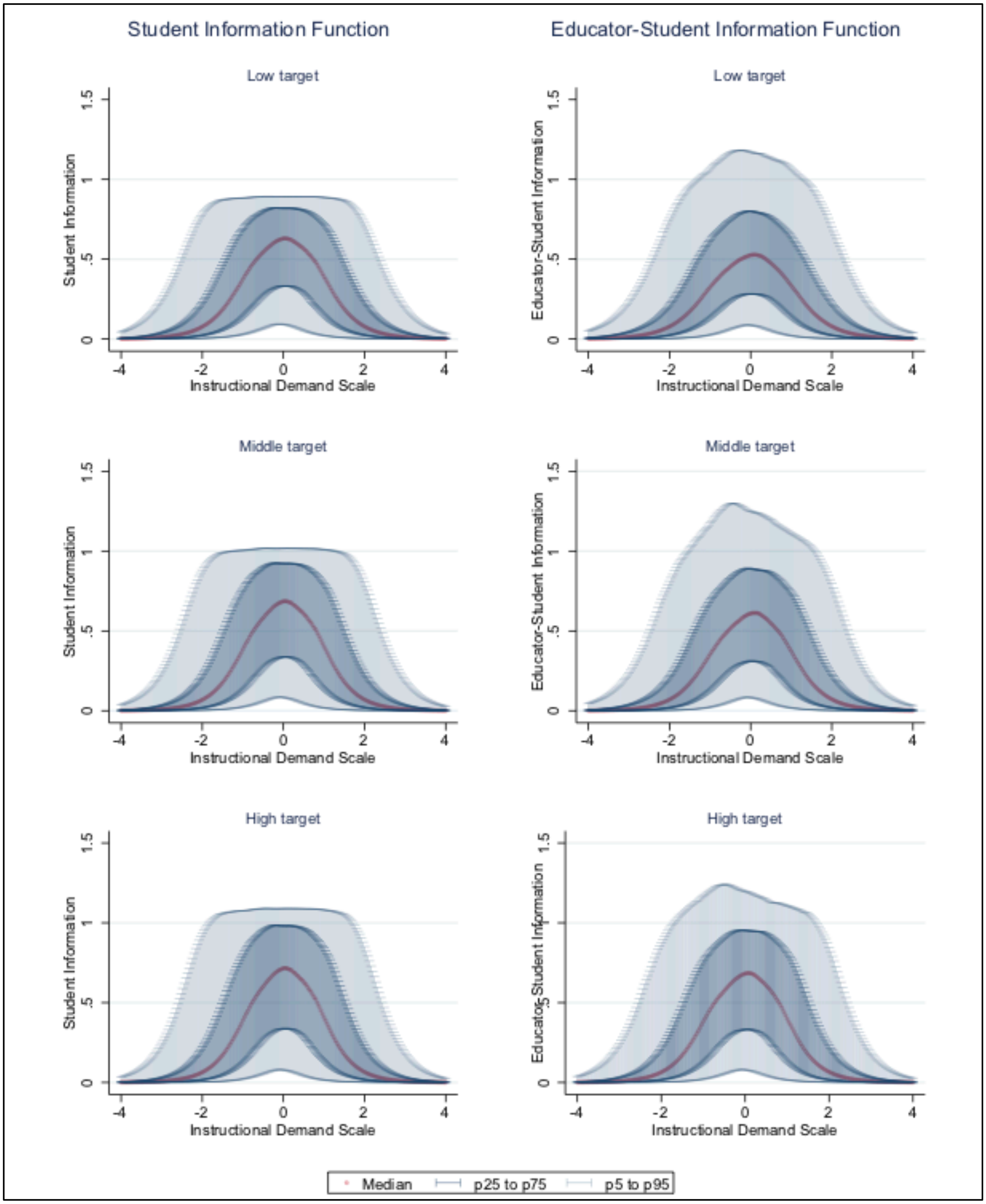
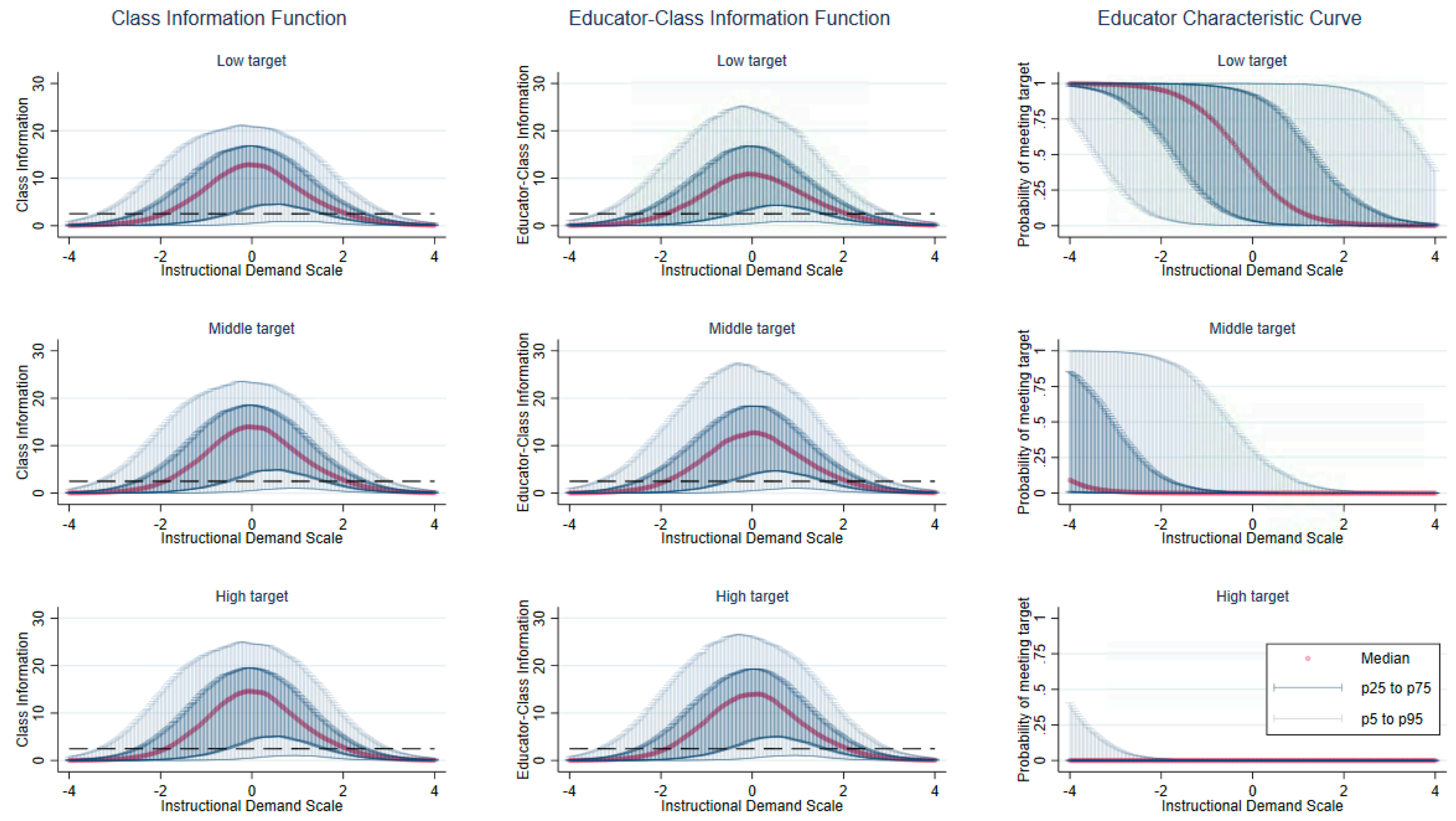


Figure 3-2. Distributions of student and educator-student information functions



Note: dashed line shows minimum information level (2.5) required for a reliability of 0.9

Figure 3-3. *Distributions of class and educator-class information functions and educator characteristic curves*

The dashed lines on the CIF and E-CIF plots highlight the minimum level of information (approximately 2.5) required for a reliability coefficient of 0.9 when the distribution of latent capacity has a standard deviation of about 2.0. Across the three performance targets, the median E-CIF is above this threshold for demand levels between approximately -2 and 2. Capacity estimates are within this range for 61% of teachers in the sample for the low target, 14% of teachers for the middle target, and fewer than 1% of teachers for the high target. At the 75th percentile, the range of capacities above this threshold extends from approximately -2.5 to 2.5. Capacity estimates are within this range for 73%, 19% and 1% of teachers for the low, middle, and high targets, respectively. At the 25th percentile, the range is only from about -0.5 to 1.5. Only 31% of teachers have capacity estimates within this range for the low target, 4% for the middle target, and fewer than 1% for the high target.

3.3.2 Comparing Alternate Measures of Effectiveness

Table 3-5 provides Spearman rank correlations among the different measures of effectiveness derived from the SRF. Rankings based on the same performance target but different probabilities or demand levels are very similar with correlations above 0.90. These relationships are strongest for the same type of measure (i.e. both measures are based on a fixed probability or both are based on a fixed demand level) with correlations above 0.97. Correlations for measures with the same target but different types of rankings (one fixed demand and one fixed probability) are strongest for the low target, followed by the middle, then the highest target. Correlations among pairs of measures based on different targets are weakest when the targets are farther apart (i.e. low target and high target) and highest when both measures are based on fixed probabilities. Correlations of measures for different targets are similar in strength when both are based on fixed demand levels and when one is based on a fixed probability and the other on a fixed demand level.

Table 3-5. Spearman rank correlations of teacher effectiveness measures derived from the SRF

		Low target						Middle target						High target					
		P50	P25	P75	D1	D2	D3	P50	P25	P75	D1	D2	D3	P50	P25	P75	D1	D2	D3
Low target	P50	1																	
	P25	.99	1																
	P75	.99	.98	1															
	D1	.99	.98	.99	1														
	D2	.99	.98	.99	.97	1													
	D3	.99	.99	.97	.99	.99	1												
Middle target	P50	.78	.78	.77	.77	.76	.75	1											
	P25	.78	.78	.77	.76	.76	.75	.99	1										
	P75	.79	.79	.78	.77	.77	.76	.99	.99	1									
	D1	.74	.75	.73	.71	.73	.74	.98	.99	.98	1								
	D2	.72	.73	.70	.71	.71	.72	.97	.97	.96	.99	1							
	D3	.69	.70	.68	.69	.74	.66	.95	.96	.94	.99	.99	1						
High target	P50	.55	.54	.55	.54	.53	.51	.71	.71	.72	.67	.64	.62	1					
	P25	.54	.53	.55	.54	.53	.51	.71	.70	.72	.67	.65	.62	.99	1				
	P75	.55	.54	.56	.54	.53	.51	.71	.71	.72	.67	.64	.62	.99	.99	1			
	D1	.50	.49	.50	.48	.45	.48	.68	.68	.69	.67	.65	.65	.95	.96	.95	1		
	D2	.48	.48	.48	.48	.47	.46	.67	.67	.67	.64	.65	.63	.93	.94	.93	.99	1	
	D3	.46	.46	.46	.47	.45	.45	.66	.66	.66	.66	.64	.63	.91	.92	.90	.99	.99	1

3.3.3 Parallel Classroom Analysis

Solutions from the hierarchical cluster analysis are shown in Table 3-6. The solutions associated with an increase in the Duda-Hart index and a decrease in pseudo T-squared relative to the solution with one fewer cluster were considered. Solutions with fewer than 3 clusters or fewer than 100 teachers per cluster were excluded from consideration, as these solutions do not adequately distinguish between types of classrooms or maintain a sufficient number of teachers per group for within-cluster comparison purposes. Only the 3, 6, and 8-cluster solutions meet all of the statistical and practical criteria discussed. Ultimately, the 6-cluster solution was selected because its pseudo T-squared is the smallest, indicating dissimilarity of clusters, taking into account the number of observations (Halpin, 2016). The dendrogram for this solution is shown in Figure 3-4.

Table 3-6. Hierarchical cluster analysis solutions

Number of clusters	Duda-Hart index	Pseudo T-squared
2	0.618	860.67
3	0.707	424.73
4	0.627	430.65
5	0.460	738.81
6	0.626	219.51
7	0.495	303.49
8	0.681	239.89
9	0.468	237.9
10	0.572	118.21
11	0.557	175.8
12	0.565	185.9
13	0.595	197.55
14	0.578	150.94
15	0.580	158.77

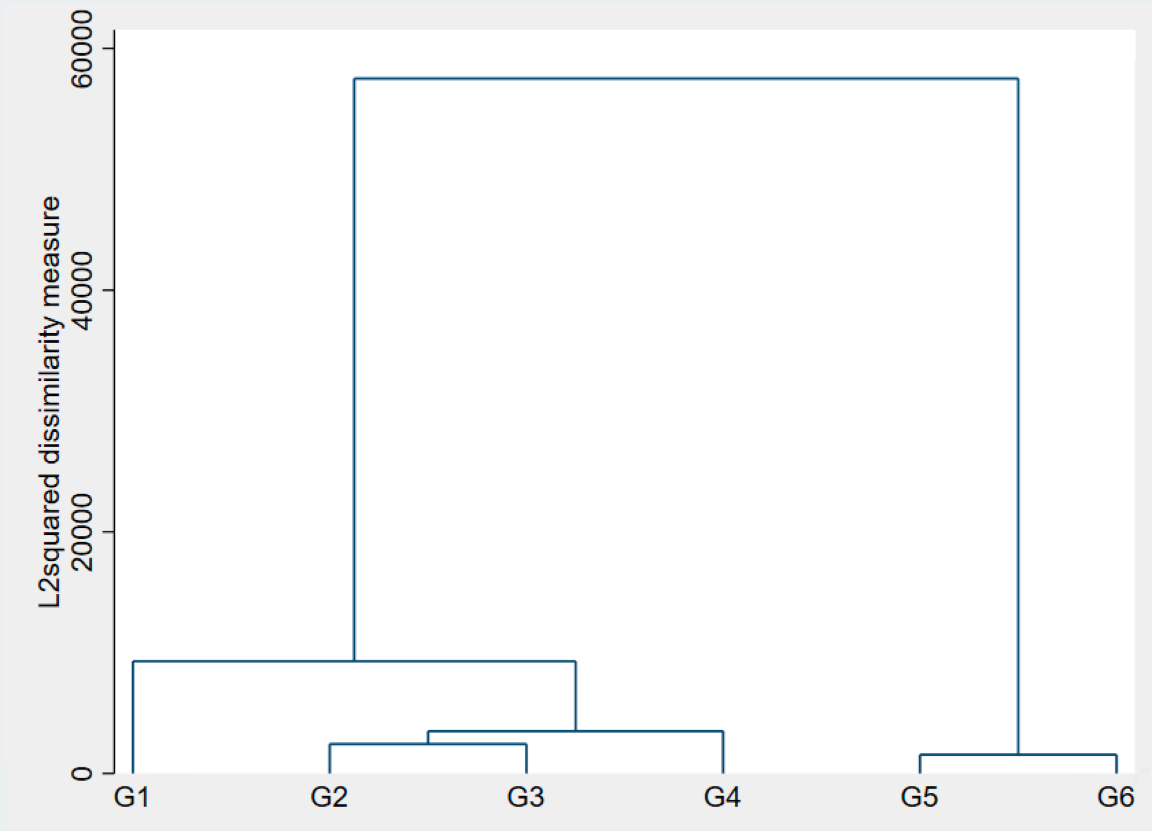


Figure 3-4. *Dendrogram for CIF cluster analysis (6-cluster solution).*

Table 3-7 provides characteristics of teachers and students in each cluster of classrooms. Clusters 1 through 4 have similar levels of variation in instructional demand within classrooms. Clusters 5 and 6 have somewhat less variation, as these classes are comprised mainly of special education students who tend to have high demand levels. Estimates of capacity for teachers with different classroom types do vary to some extent. Teachers in Cluster 1, on average, have the highest estimates of capacity for all three performance targets. Teachers in Cluster 6, who primarily teach special education students and have unusually low student counts, have the lowest average capacity for the low target, but perform similarly to the remaining clusters for the middle and high targets. Slope parameters, on average, differ very little between clusters.

Table 3-7. Characteristics of students, teachers, and classes by cluster

Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**All**
Percent of teachers	18.2%	25.4%	10.4%	14.8%	12.0%	19.1%	100%
Proportion SPED	0.0	0.0	0.0	0.0	0.6	0.9	0.2
	(0.0)	(0.0)	(0.0)	(0.0)	(0.5)	(0.3)	(0.4)
Number of students	29.6	28.1	23.6	21.5	9.0	2.9	19.8
	(3.0)	(2.9)	(3.5)	(3.9)	(2.3)	(1.5)	(10.7)
Mean student demand	-0.8	0.2	0.7	-0.2	0.7	1.1	0.2
	(0.4)	(0.3)	(0.3)	(0.3)	(0.8)	(0.7)	(0.8)
SD student demand	0.7	0.7	0.7	0.8	0.5	0.5	0.7
	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)	(0.4)	(0.3)
Capacity (low target)	1.5	0.1	-0.6	0.2	-1.0	-1.5	-0.1
	(2.0)	(2.1)	(1.9)	(1.9)	(2.4)	(1.5)	(2.2)
Capacity (middle target)	-3.0	-4.7	-5.5	-4.6	-4.9	-5.3	-4.6
	(2.4)	(2.4)	(1.9)	(2.2)	(2.0)	(1.2)	(2.2)
Capacity (high target)	-7.0	-8.6	-9.2	-8.7	-8.5	-8.8	-8.4
	(2.6)	(2.2)	(1.4)	(1.9)	(1.8)	(0.8)	(2.0)
Slope (low target)	1.9	1.9	1.9	1.8	2.0	2.0	1.9
	(0.4)	(0.4)	(0.4)	(0.4)	(0.3)	(0.2)	(0.3)
Slope (middle target)	2.0	2.0	2.0	2.0	2.1	2.1	2.0
	(0.4)	(0.3)	(0.2)	(0.4)	(0.3)	(0.1)	(0.3)
Slope (high target)	2.1	2.1	2.1	2.1	2.1	2.1	2.1
	(0.3)	(0.2)	(0.1)	(0.2)	(0.1)	(0.0)	(0.2)

The left side of Figure 3-5 shows the proportions of teachers at each percentile of the capacity scale represented by each classroom cluster. While there are some differences in cluster composition across capacity levels for all three targets, these differences become more prominent as performance targets become more difficult. Teachers in Cluster 6, in particular, are unlikely to fall outside of a restricted range of capacity scores, especially for the middle and high target. Cluster 1, which consists of large, low-demand classes, has the most consistent distribution of rankings across targets of all the clusters. These teachers tend to rank towards the top of the capacity distribution for all three targets. The right side of Figure 3-5 provides scatterplots of within-cluster percentile rankings and overall percentile rankings of capacity parameters for each performance target. Within-cluster rankings are most similar to overall rankings for teachers in

Clusters 2 and 4, although even the rankings for these groups become less similar as performance targets become more difficult.

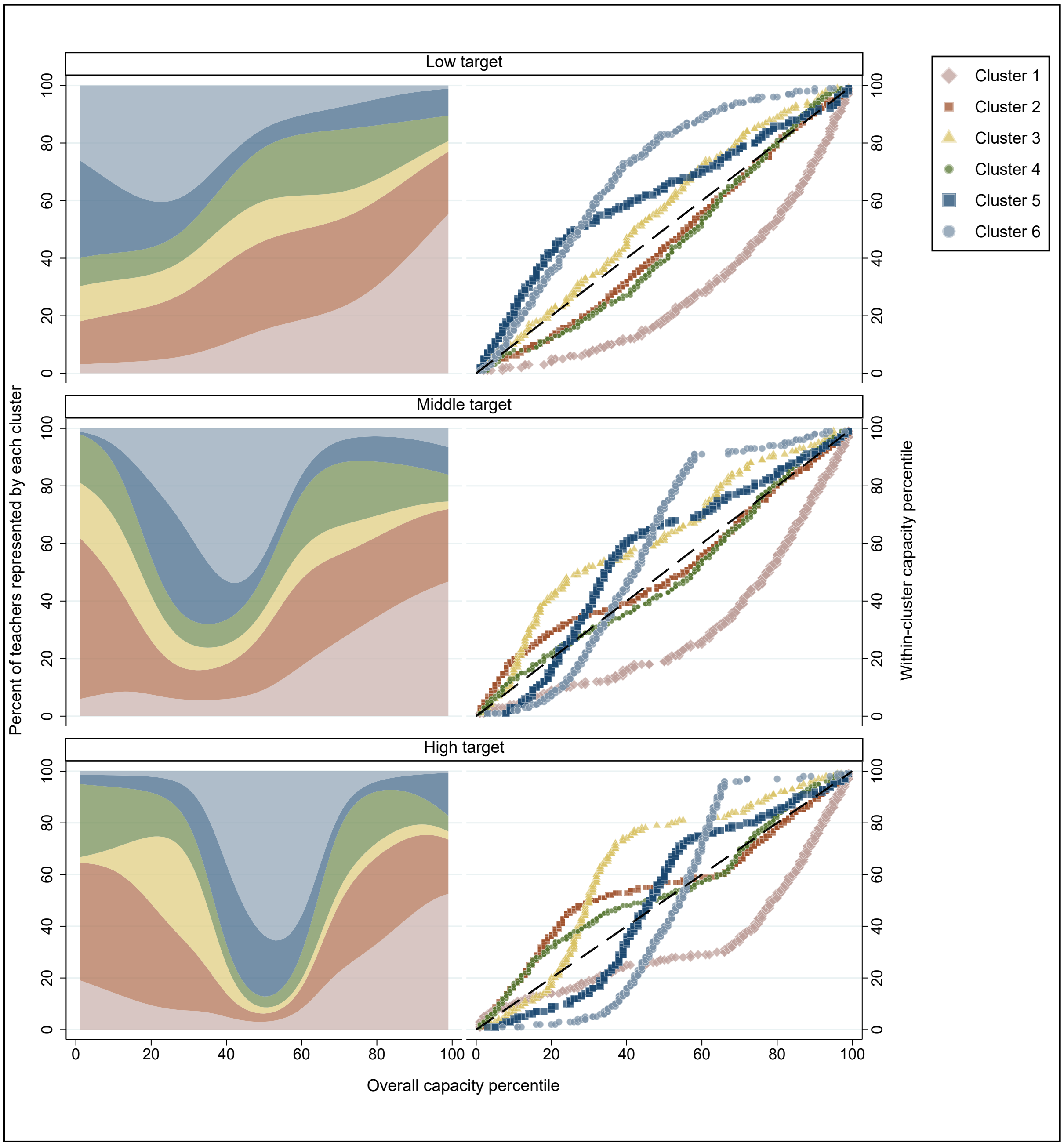


Figure 3-5. *Relationships between cluster membership and capacity rankings*

3.3.4 Comparing Alternate Rating Categories

Table 3-8 shows the percentages of teachers assigned each performance rating level based for the classification schemes defined in Table 3-4, along with the breakdown of these percentages by cluster. For the high target, nearly every teacher is placed in the lowest category across all measures and clusters. The same is true for the middle target, although to a lesser extent. The overwhelming majority of teachers are placed in the bottom category in these cases, with teachers of Cluster 1 classes appearing in the higher categories at somewhat higher rates. For the low target measures, more substantial proportions of teachers receive ratings in each of the four categories. Teachers are split most evenly among categories for the measures based on fixed probabilities (P25, P50, P75), while the measures based on fixed demand levels (D1, D2, D3) tend to have much higher frequencies in either the top of bottom category. The dominant category for these fixed demand measures varies by cluster.

Because of the low level of variation in ratings within and between classification schemes for the middle and high targets, comparisons of ratings for the same individual teachers across measures focus only on the low target. Table 3-9 provides frequencies for all 14 observed combinations of ratings across these measures. About 37% of teachers receive the same rating on all six measures. 20% of teachers are in bottom group on all measures while the other 17% are placed in the top group on all measures. No teachers receive uniform ratings in either of the middle categories. All teachers placed in the bottom category for P25 or for D1 are also in the bottom category on the other five measures; more than half of these teachers are from Clusters 5 and 6. All teachers placed in the top category for P75 or for D3 are also in the top category on the five other measures; nearly half of these teachers are from Cluster 1, and a quarter are from Cluster 2.

Table 3-8. Percent of teachers in each performance category (L1 to L4)

		P25				P50				P75			
	Cluster	**L1**	**L2**	**L3**	**L4**	**L1**	**L2**	**L3**	**L4**	**L1**	**L2**	**L3**	**L4**
Low target	1	4	8	25	63	6	17	25	51	12	21	27	41
	2	13	22	29	35	20	29	26	25	31	30	22	17
	3	22	29	29	20	33	30	24	13	44	32	16	9
	4	12	1	33	35	19	27	30	25	27	33	24	16
	5	44	17	14	24	51	16	16	16	58	17	12	13
	6	36	40	18	6	53	32	11	4	70	22	6	2
	all	21	23	25	32	29	26	22	24	39	26	18	17
Middle target	1	64	21	10	5	76	14	7	4	83	11	5	2
	2	84	10	3	2	89	8	3	1	92	5	2	1
	3	92	5	2	0	95	5	0	0	97	3	0	0
	4	86	9	4	1	92	6	2	0	96	3	1	0
	5	88	7	4	1	91	5	3	1	94	4	1	0
	6	96	4	0	0	98	2	0	0	99	1	0	0
	all	84	10	4	2	89	7	3	1	93	5	2	1
High target	1	97	2	1	0	98	1	1	0	99	1	1	0
	2	99	1	0	0	99	1	0	0	100	0	0	0
	3	100	0	0	0	100	0	0	0	100	0	0	0
	4	99	0	0	0	100	0	0	0	100	0	0	0
	5	99	1	0	0	100	0	0	0	100	0	0	0
	6	100	0	0	0	100	0	0	0	100	0	0	0
	all	99	1	0	0	99	0	0	0	100	0	0	0

		D1				D2				D3			
	Cluster	**L1**	**L2**	**L3**	**L4**	**L1**	**L2**	**L3**	**L4**	**L1**	**L2**	**L3**	**L4**
Low target	1	4	2	6	88	12	11	9	67	37	12	11	41
	2	13	6	12	69	35	13	13	39	65	10	9	17
	3	22	11	10	56	51	13	12	24	80	7	4	9
	4	12	6	8	73	32	14	15	40	65	11	9	16
	5	44	7	7	42	61	6	8	25	76	8	4	13
	6	36	17	17	30	77	9	6	8	94	2	1	2
	all	21	8	10	61	43	1	11	35	68	8	7	17
Middle target	1	64	11	7	17	85	5	4	6	95	1	2	2
	2	84	4	4	8	95	1	2	2	98	1	1	1
	3	92	2	2	3	98	2	0	0	100	0	0	0
	4	86	6	4	4	95	2	1	1	99	0	0	0
	5	88	3	3	6	95	1	2	2	99	1	0	0
	6	96	2	1	1	100	0	0	0	100	0	0	0
	all	84	5	4	7	94	2	2	2	98	1	1	1
High target	1	97	1	1	1	99	0	1	1	100	0	0	0
	2	99	0	0	0	100	0	0	0	100	0	0	0
	3	100	0	0	0	100	0	0	0	100	0	0	0
	4	99	0	0	0	100	0	0	0	100	0	0	0
	5	99	0	0	0	100	0	0	0	100	0	0	0
	6	100	0	0	0	100	0	0	0	100	0	0	0
	all	99	0	0	0	100	0	0	0	100	0	0	0

Table 3-9. Teacher classification patterns and their frequencies for low-target measures.

Tree diagram of rating categories across measures						Frequency of pattern	Breakdown by cluster					
P25	P50	P75	D1	D2	D3		1	2	3	4	5	6
L1	L1	L1	L1	L1	L1	20.5%	0.8%	3.4%	2.3%	1.8%	5.3%	7.0%
L2	L1	L1	L2	L1	L1	8.1%	0.3%	1.6%	1.2%	0.9%	0.9%	3.2%
	L2	L1	L3	L1	L1	9.3%	0.8%	2.6%	0.9%	1.1%	0.7%	3.1%
		L2	L4	L1	L1	5.1%	0.2%	1.4%	0.9%	0.8%	0.4%	1.3%
L3	L2	L1	L3	L2	L1	0.7%	0.2%	0.3%	0.1%	0.1%	0.0%	0.0%
		L2	L4	L2	L1	10.3%	1.9%	2.9%	1.2%	2.0%	0.7%	1.7%
	L3	L2	L4	L3	L1	9.7%	1.4%	3.2%	1.2%	2.0%	0.7%	1.2%
		L3	L4	L4	L1	4.3%	1.1%	1.1%	0.5%	0.8%	0.3%	0.4%
L4	L2	L1	L3	L2	L2	0.2%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%
	L3	L1	L3	L3	L2	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
		L2	L4	L3	L2	0.8%	0.2%	0.2%	0.0%	0.1%	0.2%	0.0%
		L3	L4	L4	L2	7.2%	1.8%	2.2%	0.6%	1.4%	0.7%	0.4%
	L4	L3	L4	L4	L3	6.6%	1.9%	2.2%	0.4%	1.3%	0.4%	0.2%
		L4	L4	L4	L4	16.9%	7.5%	4.2%	0.9%	2.3%	1.5%	0.4%

For the remaining 63% of teachers, ratings vary across the six categorization schemes. About 15% of teachers receive ratings in two different categories; in all of these cases, the two categories are consecutive (either the two bottom categories or the two top categories). 27% are rated in three categories: 23% are rated in three consecutive categories. For the other 4%, five of the six ratings are in the top two categories and the other is in the bottom category. This pattern is observed most frequently among Cluster 1 teachers. 20% are rated in each of the four categories across the six classification schemes. These inconsistent classification patterns are observed most frequently among Cluster 2 teachers. Because each rating is a function of the teacher's SRT parameters, descriptive statistics for the capacity and slope estimates of teachers with each rating pattern are given in Table 3-10 for reference. Uniform classification patterns are associated with very high or very low capacity estimates. Inconsistent classification patterns are associated with smaller slope estimates and capacity estimates that fall between the demand levels used as cut-points.

Table 3-10. Descriptive statistics for low-target SRT parameters by rating pattern.

Rating category						**Estimated capacity**				**Estimated slope**			
P25	P50	P75	D1	D2	D3	Mean	SD	Min	Max	Mean	SD	Min	Max
L1	L1	L1	L1	L1	L1	-3.02	0.75	-5.42	-1.92	2.02	0.23	1.18	2.79
L2	L1	L1	L2	L1	L1	-1.78	0.18	-2.30	-1.50	1.90	0.30	1.02	2.54
	L2	L1	L3	L1	L1	-1.20	0.18	-1.50	-0.79	1.84	0.34	1.04	2.62
		L2	L4	L1	L1	-0.76	0.12	-1.08	0.49	2.07	0.24	1.52	2.74
L3	L2	L1	L3	L2	L1	-0.70	0.16	-0.96	-0.41	1.08	0.13	0.81	1.32
		L2	L4	L2	L1	-0.30	0.19	-0.69	0.00	1.79	0.34	0.91	2.58
	L3	L2	L4	L3	L1	0.27	0.17	0.01	0.70	1.79	0.36	0.90	2.57
		L3	L4	L4	L1	0.75	0.13	0.50	1.09	2.07	0.26	1.58	2.72
L4	L2	L1	L3	L2	L2	-0.25	0.14	-0.43	-0.13	0.45	0.19	0.22	0.64
	L3	L1	L3	L3	L2	0.20	0.17	0.08	0.32	0.38	0.29	0.17	0.58
		L2	L4	L3	L2	0.62	0.28	0.05	0.93	0.95	0.20	0.71	1.36
		L3	L4	L4	L2	1.19	0.17	0.81	1.50	1.79	0.35	0.93	2.62
	L4	L3	L4	L4	L3	1.79	0.17	1.51	2.34	1.84	0.35	0.94	2.79
		L4	L4	L4	L4	3.30	0.90	1.98	6.23	1.96	0.28	1.17	2.67

3.4 Discussion

These results repeatedly demonstrate that estimates for and conclusions about teachers vary depending on the student performance target selected. Different peak information levels across targets suggest that the predictive relationship between instructional demand and target attainment is stronger for the higher targets; this is consistent with findings from the model fit analysis in Chapter 2. However, the educator characteristic curves (ECCs) shown in Figure 3-3 affirm that very few students are expected to reach the high target even with exceptional teachers. The same is true of the middle performance target, although to a lesser extent. The distribution of ECCs shows little variation in the probability of attaining the middle target for students at most instructional demand levels. Because attainment of these higher targets is both rare and strongly related to observable characteristics of students, they may not be appropriate benchmarks against which to evaluate teachers. Similarly, categorization schemes for the middle and high target measures place the overwhelming majority of teachers into the lowest performance level. These classifications could be useful for identifying a very small number of higher-performing teachers but would be less helpful for discriminating among other types of teachers. The categorization schemes for the low target measures do a better job of discriminating among the full sample of teachers.

Four of the categorization schemes for the low target identify either the top or bottom group of teachers in perfect agreement with the remaining 5 measures. The P25 and D1 measures are strong choices for confidently identifying a low-performing group of teachers, as this group has the lowest probabilities of success with the least-demanding students. Similarly, P75 and D3 are strong choices for identifying a high-performing group of teachers, as those identified will have the highest probability of success with the most demanding students. However, the vast majority of teachers are placed in different categories across measures. This stresses that, aside from those who perform extremely well or extremely poorly, a combination of measures may offer a more nuanced description of educator performance than any single measure on its own.

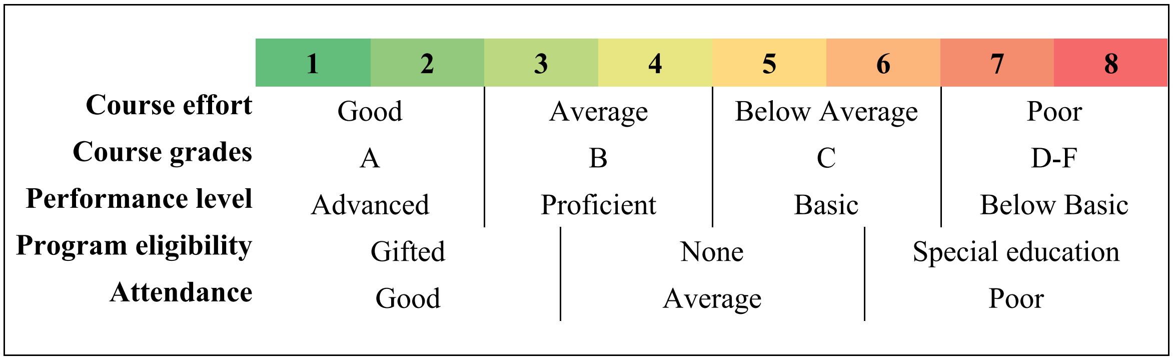
The clusters of parallel classrooms highlight differences not only in the groups of students assigned to different teachers, but also in the properties of capacity estimates for these teachers. For instance, Cluster 6 teachers work with very small groups of students who are typically in special education programs. Considering the low student counts and high demand levels, using this information to assess the effectiveness of these teachers is equivalent to administering a test that is very short and very difficult. Figure 3-5 highlights that, although these teachers' estimated capacities for the low target span the entire scale, they are disproportionally concentrated within much smaller ranges of the capacity scale for the higher targets. This is likely more indicative of an assessment that does not adequately discriminate between the capacities of these teachers than it is of a lack of variation in their true capacities. In contrast, Cluster 1 teachers, who have the most students and lowest demand levels, have the most consistent distribution of results across targets. This is equivalent to an assessment of effectiveness that is much longer and easier than the assessment for teachers in Cluster 6.

Relationships between classification patterns and cluster membership are evident in Table 3-10. A greater proportion of cluster 5 teachers receive uniform classifications than any other cluster, with 57% falling in either the top or bottom category on all 6 measures. Cluster 1 teachers follow behind this group with 45% receiving uniform ratings. While sorting and assignment patterns could play a role in these differences, the main attributes that define each cluster (student counts and the distribution of instructional demand) also influence the precision with which capacity is estimated. Cluster 1 teachers have the most students and lowest demand levels. The "longer" test of effectiveness allows for greater precision, which could possibly result in more consistent ratings. The "easier" test could result in a ceiling effect. This would explain the lack of variation in ratings across measures that have inherently different meanings, and a higher target measure may be more appropriate for evaluating this particular group of teachers. Cluster 5 teachers, on the other hand, are likely to receive uniform ratings (typically in the lowest category) despite having fewer students and higher demand levels. While this cannot be a result of higher precision due to test length, it is possible that Cluster 5 teachers are experiencing a floor effect. If this is the case, this implies that even the lowest target is inappropriately difficult for most students in Cluster 5 classrooms.

In general, the different measures vary more in their meanings and interpretations than they do in quality. Thus, rather than pointing towards a recommended measure or subset of measures for reporting purposes, these results affirm that SRT can offer a degree of flexibility for districts or administrators to select measures that align with their objectives and are relevant for a particular audience. The criterion-referenced nature of SRT measures connects these estimates to concepts that are already familiar to many stakeholders. The section that follows uses sample reporting materials to illustrate a few possible ways this information can be framed, presented, and interpreted.

3.4.1 Contextualizing the Instructional Demand Scale

While the student instructional demand index (SIDI) estimated in Chapter 2 is a standardized measure, it has been rescaled for simplicity and user-friendliness for reporting purposes. The rescaled SIDI was computed by adding 4.5 to the SIDI, so that all observed demand levels take on positive values. The scale bar in Figure 3-6 divides the scale into one-unit intervals that are each labeled with a consecutive integer that corresponds to the midpoint of the interval (i.e. the interval labeled “1” spans from 0.5 to 1.5 on the rescaled SIDI). Typical characteristics of students in each interval are indicated below the scale bar, providing a framework to associate student characteristics with numbers or colors on the scale. For example, students within the fourth interval (whose rescaled demand estimates are between 3.5 and 4.5) tend to have average marks for effort, earn B’s in their courses, reach proficiency, not be eligible for gifted or special education programs, and have average attendance.



<table>
<tr><th></th><th colspan="2">1</th><th colspan="2">2</th><th colspan="2">3</th><th colspan="2">4</th><th colspan="2">5</th><th colspan="2">6</th><th colspan="2">7</th><th colspan="2">8</th></tr>
<tr><td><b>Course effort</b></td><td colspan="4">Good</td><td colspan="4">Average</td><td colspan="4">Below Average</td><td colspan="4">Poor</td></tr>
<tr><td><b>Course grades</b></td><td colspan="4">A</td><td colspan="4">B</td><td colspan="4">C</td><td colspan="4">D-F</td></tr>
<tr><td><b>Performance level</b></td><td colspan="4">Advanced</td><td colspan="4">Proficient</td><td colspan="4">Basic</td><td colspan="4">Below Basic</td></tr>
<tr><td><b>Program eligibility</b></td><td colspan="5">Gifted</td><td colspan="6">None</td><td colspan="5">Special education</td></tr>
<tr><td><b>Attendance</b></td><td colspan="5">Good</td><td colspan="6">Average</td><td colspan="5">Poor</td></tr>
</table>

__Figure 3-6.__ Mapping of instructional demand scale to typical student characteristics.

The same scale bar appears alongside several other sample reporting materials to connect the information presented with types of students and instructional challenges. Figure 3-7, for instance, illustrates the distributions of instructional demand across the entire district and three randomly-selected sample schools in reference to the scale bar. Sections of the distribution are shaded according to percentage of students within the corresponding instructional demand interval. The distributions shown in Figure 3-7 indicate that each of the sample schools has lower-demand students, on average, than the district as a whole. This provides a context for interpreting a report.

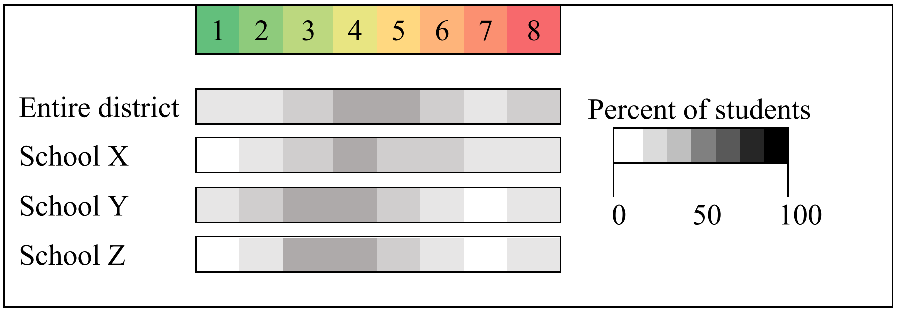
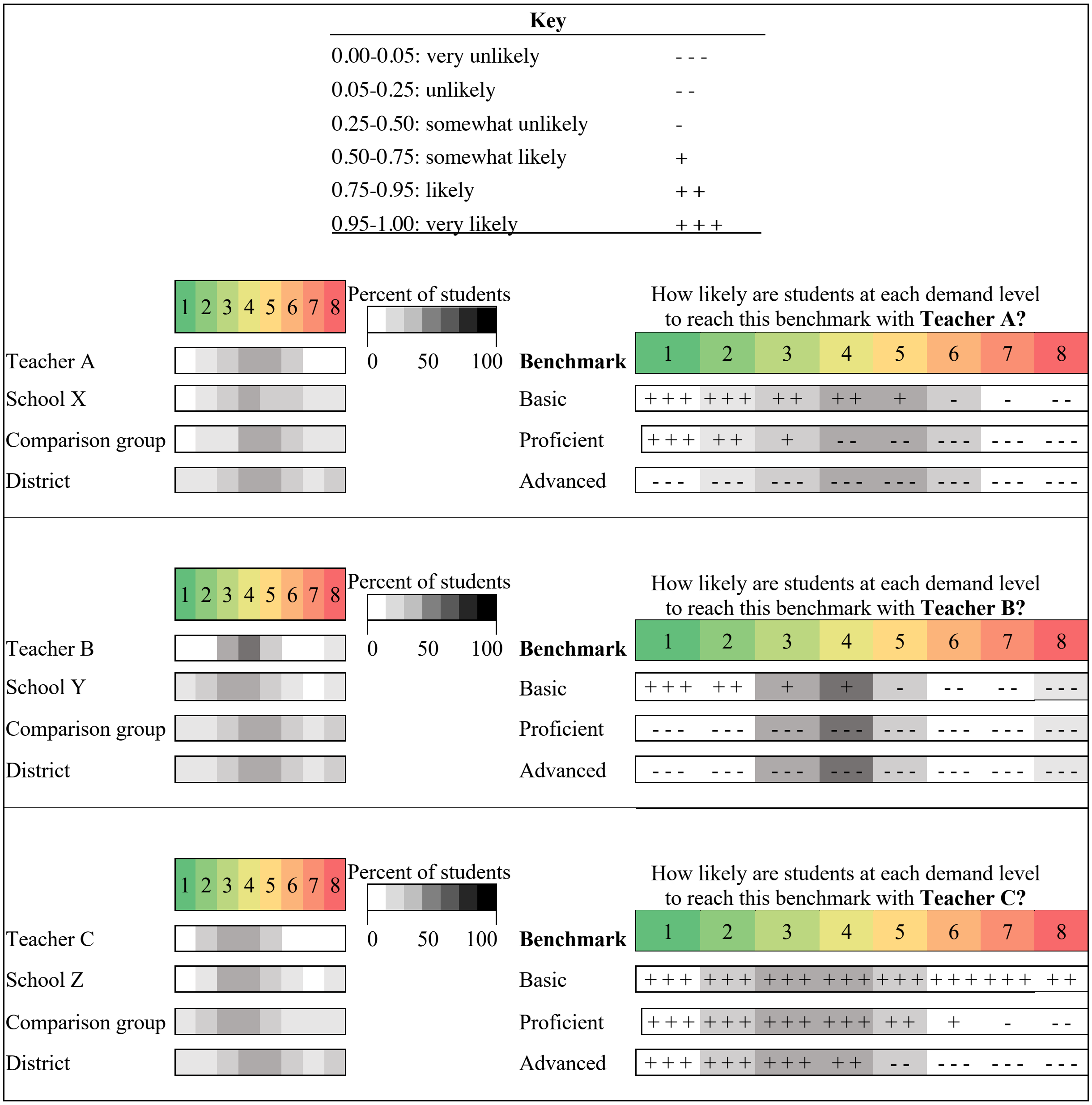


Figure 3-7. Distribution of instructional demand across district and sample schools.

3.4.2 Simplifying the ECC

Figure 3-8 combines this information with results for three sample teachers (one from each sample school) from the SRT analysis. On the left-hand side, the distribution of instructional demand for a particular teacher is shown alongside the distributions for the school, all students in the same classroom cluster (the term "comparison group" is used in place of cluster because it is more familiar to many audiences), and all students in the district. The right-hand side condenses the educator characteristic curve (ECC) into a simpler format, while also tying this information to the color-coded scale bar. Symbols are used to indicate how likely different types of students are to reach each performance benchmark with the teacher, according to the height of the teacher's ECC at the midpoints of each interval on the demand scale. For instance, students with demand levels greater than or equal to 6 are "very unlikely" to reach the high target with Teacher C; however, the distribution of demand for Teacher C reveals that there are no students at this level in Teacher C's class. Nearly all of Teacher C's actual students are "likely" or "very likely" to reach all of the performance targets.

The report does not explicitly show the capacity or slope estimate for any teacher, but this information is still communicated in the report. The location where probabilities change from "unlikely" to "likely" categories is the location of the capacity estimate and the abruptness of



Key	
0.00-0.05: very unlikely	- - -
0.05-0.25: unlikely	- -
0.25-0.50: somewhat unlikely	-
0.50-0.75: somewhat likely	+
0.75-0.95: likely	+ +
0.95-1.00: very likely	+ + +

How likely are students at each demand level to reach this benchmark with **Teacher A?**

Benchmark	1	2	3	4	5	6	7	8
Basic	+ + +	+ + +	+ +	+ +	+	-	-	- -
Proficient	+ + +	+ +	+	- -	- -	- - -	- - -	- - -
Advanced	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -

How likely are students at each demand level to reach this benchmark with **Teacher B?**

Benchmark	1	2	3	4	5	6	7	8
Basic	+ + +	+ +	+	+	-	- -	- -	- - -
Proficient	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -
Advanced	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -

How likely are students at each demand level to reach this benchmark with **Teacher C?**

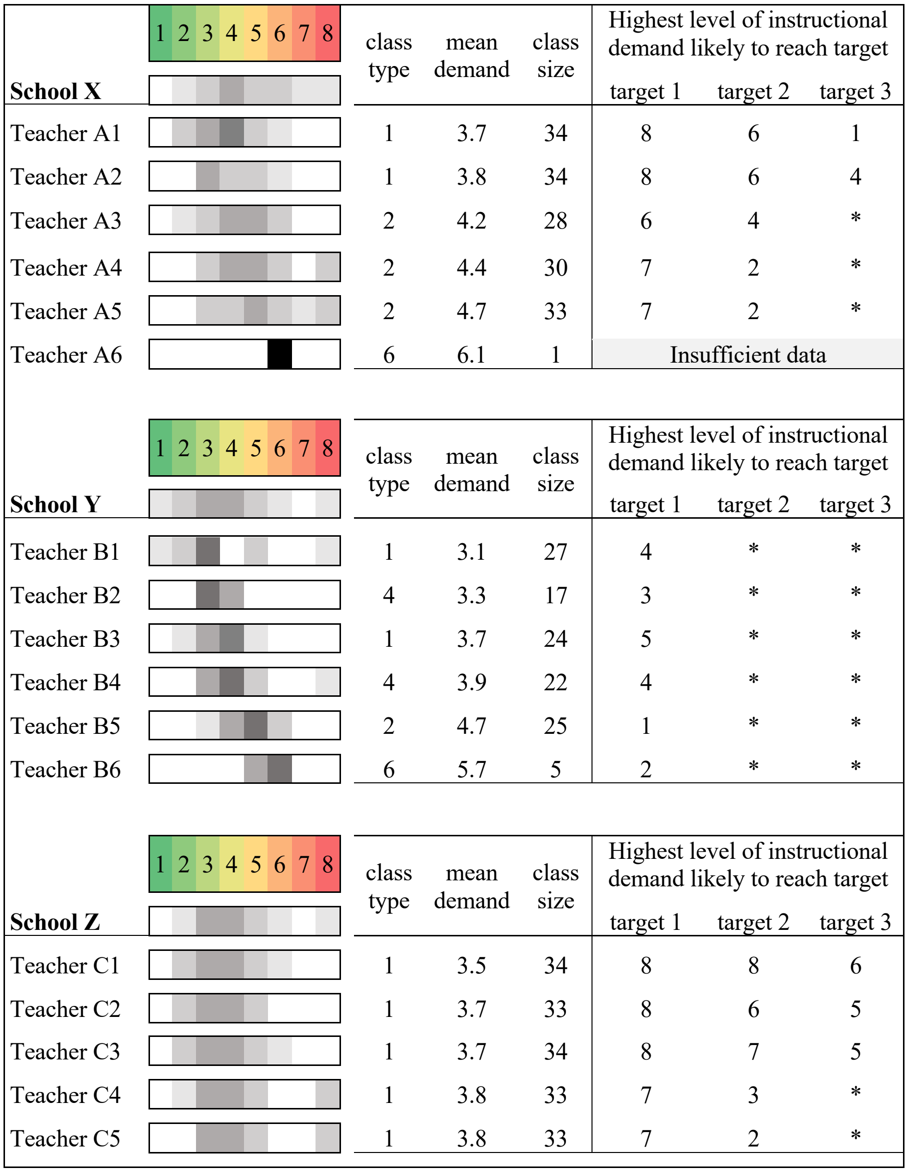
Benchmark	1	2	3	4	5	6	7	8
Basic	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +	+ +
Proficient	+ + +	+ + +	+ + +	+ + +	+ +	+	-	- -
Advanced	+ + +	+ + +	+ + +	+ +	- -	- - -	- - -	- - -

Figure 3-8. *Sample reports for one teacher from each sample school.*

transitions between probability categories communicates the magnitude of the slope. For instance, Teacher A's capacity for the low target is between 5 and 6, as the probability category changes from a single plus symbol ("somewhat likely") in the 5th interval to a single minus symbol (“somewhat unlikely”) in the 6th interval. Teacher C’s probability for the high target changes more abruptly than this: there is a change from two plus symbols (“likely”) in the 4th interval to two minus symbols (“unlikely”) in the 5th interval, indicating that the capacity parameter is between 4 and 5, and that the slope is larger than Teacher A’s low-target slope. Because the parameter values themselves may not be inherently meaningful to many stakeholders, a simplified form of the ECC can communicate the relevant concepts in terms of probabilities, performance benchmarks, and types of students.

3.4.3 Describing Teacher Performance in Terms of Demand

Figure 3-9 summarizes the performance of all teachers of the same grade level in each of the sample schools. The scale bars on the left compare the distributions of demand across these teachers and are supplemented by information about the mean demand, size, and type (or cluster) of each class. Differences in the distribution of demand for teachers in the same school suggest that assignment practices vary across these three locations. For instance, the overall distributions of demand in School Y and School Z are quite similar, but School Y places students with similar demand levels in the same classrooms while School Z distributes instructional demand equitably across teachers with each classroom distribution mirroring the schoolwide distribution. In School X, classrooms vary in size, mean demand, and variation in demand. The right-hand side provides information about the capacity estimates for each teacher. The number reported is largest integer value on the demand scale for which a teacher has a probability above 0.5 for a particular target. This frames the estimate for a teacher in terms of the types of students (with respect to instructional demand) that are likely to reach a particular benchmark in their class.   Many capacity estimates, particularly those for the highest target, are outside the range of instructional demand values that are observed for students in the district. In these cases, an asterisk (*) is reported in place of a demand level, as there is no realistic level of instructional demand that corresponds to a probability of student success above 0.5. For example, no teacher in School Y has a capacity estimate within the range of observed demand values for the middle or high target.



School X	class type	mean demand	class size	Highest level of instructional demand likely to reach target		
				target 1	target 2	target 3
Teacher A1	1	3.7	34	8	6	1
Teacher A2	1	3.8	34	8	6	4
Teacher A3	2	4.2	28	6	4	*
Teacher A4	2	4.4	30	7	2	*
Teacher A5	2	4.7	33	7	2	*
Teacher A6	6	6.1	1	Insufficient data		

School Y	class type	mean demand	class size	Highest level of instructional demand likely to reach target		
				target 1	target 2	target 3
Teacher B1	1	3.1	27	4	*	*
Teacher B2	4	3.3	17	3	*	*
Teacher B3	1	3.7	24	5	*	*
Teacher B4	4	3.9	22	4	*	*
Teacher B5	2	4.7	25	1	*	*
Teacher B6	6	5.7	5	2	*	*

School Z	class type	mean demand	class size	Highest level of instructional demand likely to reach target		
				target 1	target 2	target 3
Teacher C1	1	3.5	34	8	8	6
Teacher C2	1	3.7	33	8	6	5
Teacher C3	1	3.7	34	8	7	5
Teacher C4	1	3.8	33	7	3	*
Teacher C5	1	3.8	33	7	2	*

Figure 3-9. *Sample reports for all teachers in the sample schools.*

3.4.4 Referencing Estimates to Populations of Teachers and Students

While norm-referenced measures that describe a teacher's relative position in the distribution are commonly featured in evaluation systems, SRT scores can also be presented in reference to distributions of students. Each estimate of educator capacity corresponds to a location on the student instructional demand scale. The percentage of students with demand levels below a teacher's capacity estimate indicates how many students would have probabilities of reaching a performance target with this teacher that are above 0.5. By reporting this percentage, teachers' capacities are framed in terms of how many students would, theoretically, be more likely than not to reach a particular performance target if placed in their class. Table 3-12 shows two versions of this metric: the first is the percentage of all students in the district with probabilities above 0.5 for the teacher, and the second is the percentage of all students in a particular school with probabilities above 0.5 for the teacher. While the first percentage is useful for comparisons of teachers across the entire district relative to the student population, the second is informative about how well a teacher meets the needs of students in their own school.

The table also includes district and school norms as comparison points, further contextualizing the performance of the teacher relative to other teachers district-wide and in the same school. For example, only 33% of students in the district have probabilities above 0.5 of reaching the lowest performance target with Teacher B, meaning that 33% of students in the district have instructional demand estimates that are greater than the capacity estimate for this teacher. In comparison, 45% of students in the district have probabilities above 0.5 of reaching this same target with a teacher whose capacity is equal to the district average. However, 60% of the students in Teacher B's school have probabilities above 0.5 with Teacher B, compared to only 32% of these students with a teacher whose capacity is equal to the school average.

Although the capacity of Teacher B, in reference to the districtwide distributions of capacity and demand, may suggest poor performance, Teacher B is expected to perform well with the majority of students in the school where he/she teachers, and expected to perform significantly better than the average teacher in that school.

Table 3-11. Sample teacher-level reports: capacity relative to student demand distribution.

	Percentage of students more likely than not to reach target with Teacher A					
	All students in district			All students in Teacher A's school		
	Teacher A	Average teacher in district	Average teacher in School X	Teacher A	Average teacher in district	Average teacher in School X
Basic proficiency	92%	45%	95%	97%	76%	98%
Proficiency	19%	4%	35%	33%	17%	38%
Advanced proficiency	0%	0%	6%	0%	0%	9%

	Percentage of students more likely than not to reach target with Teacher B					
	All students in district			All students in Teacher B's school		
	Teacher B	Average teacher in district	Average teacher in School Y	Teacher B	Average teacher in district	Average teacher in School Y
Basic proficiency	33%	45%	18%	60%	76%	32%
Proficiency	0%	4%	0%	0%	17%	0%
Advanced proficiency	0%	0%	0%	0%	0%	0%

	Percentage of students more likely than not to reach target with Teacher C					
	All students in district			All students in Teacher C's school		
	Teacher C	Average teacher in district	Average teacher in School Z	Teacher C	Average teacher in district	Average teacher in School Z
Basic proficiency	98%	45%	98%	99%	76%	99%
Proficiency	97%	4%	50%	99%	17%	51%
Advanced proficiency	49%	0%	35%	83%	0%	46%

Both Figure 3-9 and Table 3-12 present information about the capacity estimate for a teacher, which is equivalent to the location on the demand scale corresponding to a probability of 0.5 (referred to as P50 in Sections 3.2 and 3.3). However, equivalent versions of these reports could be developed for any measure derived from the ECC. The SRT framework affords flexibility to administrators or policymakers to determine what information is most relevant to their priorities and their context. Relevant measures may include, but are not necessarily limited to, probabilities associated with each teacher at a fixed demand level that corresponds to specific student attributes, demand levels associated with a probability above an agreed-upon threshold, or percentages of students expected to achieve a particular outcome with each teacher.

3.5 Conclusions

Item-response functions, item and test characteristic curves, and item and test information functions serve a variety of purposes in testing contexts. Analogous forms of these functions also offer meaningful contributions to the context of educator evaluation with Student Response Theory. Some of these contributions apply concepts or procedures from IRT directly to the SRT framework; for instance, matching information functions to identify parallel test forms. While fundamental differences between how tests are constructed and how students are assigned to classrooms require different procedures to be used, the underlying concept of identifying an equivalent assessment is the same. Comparisons of SRT information functions highlighted some concerns about differences in the properties of classrooms across students, and the implications of these differences for estimating the capacity of a teacher. Differences in the number of students assigned to different teachers equate to varying lengths and precision levels of the assessments for different teachers. Differences in the distribution of demand across classrooms equate to varying difficulty levels of assessments for different teachers. While varying lengths and difficulty levels are common in some areas of IRT, particularly in adaptive tests, equal precision and fairness are valid concerns if comparisons are to be made across the entire population of teachers. However, by identifying groups of teachers with reasonably-equivalent information functions, many of these concerns can be alleviated.

Other applications are unique, arising from key differences and incompatibilities among IRT and SRT. In particular, the teacher-level slope results is the equivalent of an IRT model with only one item-level parameter and two examinee-level parameters. This key difference raises concerns about whether some of the technology that has been developed and studied extensively in IRT contexts is truly compatible with the SRT framework. However, this difference also introduces new possibilities in an SRT analysis that do not have a direct parallel in IRT. The additional teacher-level parameter that relates properties of educators to the instructional demand levels of their students is especially useful for deriving alternate performance measures from the same SRT models. These alternate measures allow for flexibility to choose those that are most useful and meaningful for the specific purposes of a district, state, or program.

The different measures derived from the SRF for the same performance target are quite similar, but measures are much less consistent across targets. Some of this can be explained by the difficulty level of the middle and high targets; very few students reach either of these performance benchmarks regardless of their instructional demand levels and the capacities of their teachers. As a result, these targets are often not informative about an educator's performance. The measures derived from the low target are more appropriate for the vast majority of students and teachers. While the different low-target measures are highly correlated and rank teachers consistently, there are slight differences in how they are defined and interpreted. When combined, these different measures offer a more nuanced summary of educator performance than is possible with any single measure.

The criterion-referenced nature of SRT also contributes to the flexibility in report design; results can be framed in a variety of ways that are connected directly to stakeholders' knowledge of students and teaching. While the sample reports illustrate some of the ways this information may be presented in practice, these are only preliminary suggestions. Additional work is needed, perhaps with focus groups comprised of different stakeholders, to determine the best ways to develop reports that empower stakeholders to understand the results and their implications and take appropriate actions in response that improve future outcomes for students and educators. With proper development, SRT could be a useful for presenting objective information about teacher performance with direct connections to the varying levels of instructional demand posed to teachers, properties of a specific population of students, and concrete performance standards for both teachers and students.

CHAPTER 4. CONTRIBUTIONS OF SRT TO FORMATIVE EVALUATION

4.1 Introduction

4.1.1 Summative and Formative Evaluations of Teachers

Summative and formative assessments are both important components in evaluations of student learning. Summative assessments typically occur at the end of a unit or course to evaluate whether a particular goal has been met, while formative assessments occur on an ongoing basis to evaluate student needs and adjust future instruction accordingly. In evaluations of teaching performance, the role of value-added models (VAMs) and similar statistical growth models has been typically been as a summative assessment of teaching performance, where ratings are assigned at the conclusion of a school year and associated with consequences for teachers. In a formative teacher evaluation framework, schools or districts use evaluative information about teaching performance to guide decisions about assignment, professional development, and resource allocation. This occurs through an ongoing process in order to best support the needs of teachers and evaluate whether these decisions have the desired outcomes in future evaluations.

Student Response Theory (SRT) applies the underlying methodology of item response theory (IRT) to an assessment of educator performance, offering an alternative to VAMs and other statistical growth models. In IRT, the probability of an examinee responding correctly to a test item is predicted as a function of the latent ability of the examinee and properties of the test item (Lord, 2012). SRT frames students as test items that assess the performance of their educators. The probability that a student will reach a pre-defined performance target with a particular teacher (equivalent to the probability of a correct item response in IRT) is predicted as a function of the latent capacity of the teacher and properties of the student (Reckase & Martineau, 2014). The 2-parameter logistic (2PL) models in IRT and SRT define similar location parameters. In IRT, this parameter indicates the difficulty level of an item, while in SRT, it indicates the level of instructional demand posed by a student to an educator. Instructional demand is a construct that describes the difficulty level associated with helping a particular student reach a performance target.

SRT may be better equipped to play a formative role in the teacher evaluation process than commonly-used statistical growth models like VAMs. Due to the criterion-referenced nature of SRT scales, the information produced by these models relates directly to characteristics of students and teachers and performance standards for students and teachers. This forms a clearer link between the content of a report and instructional or administrative practices. As a result, the information produced by an SRT model could potentially provide educators and administrators with actionable feedback about their practices. With norm-referenced VAMs and similar models, changes in teachers' ratings over time typically only denote changes in their relative ranking within a distribution of teachers. IRT equating procedures may allow for a consistent longitudinal scale to be established for evaluating teachers within the SRT framework. This type of scale introduces the possibility of measuring absolute growth over time, both for individual teachers and for the distribution of teachers as a whole. The combination of the consistent longitudinal scale and the probabilistic nature of the model also provides a means for administrators to predict outcomes of different types of students with different types of teachers, consider these predictions in decisions about assignment and resource allocation in an upcoming school year, and monitor changes over time in meaningful ways.

4.1.2 Differential Item and Student Functioning

In IRT, differential item functioning (DIF) occurs when the probability of a correct response is different for examinees with the same latent ability who differ on a characteristic that is irrelevant to the construct being measured (Holland & Wainer, 2012). The purpose of a DIF analysis is to improve the quality of a test by revising or removing items that exhibit bias. The analog to DIF in SRT is differential student functioning (DSF). While a DIF analysis explores whether items function differently for certain types of examinees, a DSF analysis investigates whether students perform differently than expected in certain types of classrooms. Tests of DSF would investigate whether students placed in classrooms or schools or with teachers with certain characteristics tend to perform differently than predicted given their levels of instructional demand.

However, a DSF analysis would likely serve different purposes and warrant different actions than a typical DIF analysis due to fundamental differences between the IRT and SRT contexts. Revising or removing students when DSF is detected is neither a reasonable nor a desirable response. Rather, DSF detection would ideally prompt changes to instructional or administrative practices to address the underlying problem that resulted in DSF. The primary purpose of a DSF analysis would be to provide diagnostic information about subgroups to target for intervention. In an IRT framework, this is equivalent to responding to DIF by training subgroups of examinees to respond to an offending item differently rather than removing or revising the item.

DSF analysis could also provide diagnostic information about the student instructional demand index (SIDI) and SRT model with respect to underlying model assumptions. For instance, misspecification of the SIDI is a threat to the unidimensionality assumption (Ham, 2014). If a DSF test indicates that certain characteristics are associated with different model performance, this implies that the SIDI does not appropriately account for those characteristics. Similarly, DSF for an aggregate classroom-level characteristic would raise concerns about the conditional independence assumption. Ham (2014) finds evidence of conditional dependence in an SRT analysis and discusses both misspecification of the SIDI and peer effects as possible explanations. A DSF analysis may offer additional insight into the source, magnitude, and practical significance of conditional dependence among students in the same class. While this particular purpose does not play a direct role in formative evaluation, it is a necessary step in establishing whether the necessary conditions for equating are met. Many of the proposed uses of SRT in formative evaluation, such as measuring growth of educators and predicting future performance of students with different potential teachers, hinge on the ability to equate SRT scales over time.

4.1.3 Establishing a Consistent Scale

Test equating procedures in IRT allow for scales to be linked across different test forms by placing the parameters from each test form onto a common scale. With common-item equating, a subset of items that appear on two different test forms operate as anchors that define this scale (Ryan & Brockmann, 2009). Some items may appear on both forms but relate differently to the underlying construct for the groups of examinees taking each form or across testing occasions due to differences or changes in curriculum, item exposure, or context. For this reason, the relationships between parameters of common items are first analyzed before selecting a set of anchors. Through this process, items that exhibit evidence of parameter drift are identified and excluded as anchors (Dorans et al., 2010).

Because norm-referenced statistical growth measures do not have fixed scales across years, multi-year comparisons can only focus on changes in the relative standing of teachers, rather than on the absolute growth of teachers. The same process can be extended to SRT instructional demand indices from different years or grade levels in order to assess longitudinal growth and compare educators of different grade levels using a consistent scale. An analysis of SIDI indicator parameters from different years and grade levels can identify the extent of variation in the relationships between indicators and instructional demand over time and across contexts. Then, the indicators that are consistent across these conditions can be used as a basis for building a common scale. This allows for the effectiveness of individual educators to be analyzed over multiple years and for rates of educator growth to be calculated. Relationships between classroom or school-level factors and growth rates of early career teachers may shed light on how administrators can support them more effectively.

4.1.4 Optimizing the Assignment Process

When IRT item parameters are already known, this information is sometimes used to select the best item(s) for a particular purpose. Sometimes this purpose is to construct a new form of a test such that the information function (TIF) matches that of an existing form as closely as possible (Samejima, 1977a). In an adaptive testing context, the purpose is to select the most informative item for a particular examinee, given their responses to previous items (Cella et al., 2007). These decisions are typically subject to certain constraints, such as item exposure and content balance. These processes parallel the assignment of students to teachers in an SRT framework. Although the objectives and constraints are likely to differ tremendously across these two contexts, similar optimization procedures may still be applicable and useful in different ways within the SRT framework.

Probabilistic outcomes from an SRT model for students with all potential teachers in an upcoming year could be useful for administrators in determining which students and teachers to place with one another for an upcoming year. Optimal assignments can be configured countless ways, affording flexibility to administrators to determine how different factors are prioritized. For instance, assignments can be made in a way that maximizes the number of students expected to meet a proficiency standard in a particular subject, optimally matches student demand levels with teacher capacities, or distributes instructional demand across teachers as equitably as possible, subject to constraints like class size restrictions, groups of students who must be placed either together or separately, or avoidance of tracking practices where students of similar ability levels are placed together.

Comparisons between actual assignments and different types of optimized assignments may also serve as feedback for administrators about inefficiencies or implications of their existing assignment practices. Prior studies find that schools vary widely in the relative degrees of influence from principals, teachers, and parents in assignment decisions (Monk, 1987; Paufler & Amrein-Beardsley, 2014) and that teachers with more prominent positions in formal leadership and informal advice networks tend to be assigned higher-achieving students (Kim, Frank, & Spillane, 2018). Comparing the assignments determined through traditional processes to assignments that optimize a particular objective function, implications of existing practices on expected outcomes may surface.

4.1.5 Significance

While the use of statistical growth models in teacher evaluation is widely criticized, the need for objective performance measures is far less disputed than the manner in which these measures are used. A model that offers evaluative information in a formative context, helping administrators to better support the needs of teachers and students, may garner significantly more support within the field than one that simply focuses on making summative judgements and determining high-stakes personnel decisions (Goe et al., 2017). This study explores possible contributions of SRT within a formative evaluation framework. Procedures extended from traditional IRT applications offer potential tools for diagnosing problems in an educational system, monitoring changes over time, and making informed decisions to improve future outcomes for both students and teachers.

4.2 Data and Methodology

4.2.1 Data and Measures

The study draws on data from an anonymous large school district in a major U.S. city for 5th grade teachers and their students during the 2015-2016 and 2016-2017 school years. SRT measures of educator capacity and consistency for both math and ELA were estimated according to the procedures outlined in Chapter 2. Estimates for the low performance target are the primary focus in this study, as findings from Chapter 3 suggest that these measures are most informative about the population of students and teachers in this district. The sample of students is restricted to those with data available from the previous school year, as this information is used to compute the student instructional demand index (SIDI). The SIDI is estimated using the IRT calibration method and restricted set of instructional demand indicators according to the procedures outlined in Chapter 2. In order to compare estimates for the same teachers with the same students across subject areas, 5th grade students with different teachers for math than for ELA (approximately 3% all students each year) and teachers with none of the same students for both subjects (approximately 3% of all teachers each year) are excluded from the sample. For analyses of teacher growth, the sample is further restricted to teachers that appear in the data as 5th grade teachers in both school years (69% of all teachers) and students placed with these teachers (74% of all students).

Table 4-1 describes the full samples of students and teachers, the restricted sample used in cross-sectional analyses, and the further restricted sample used for longitudinal analyses. The rate of special education eligibility is slightly higher for the first cohort than the second. This corresponds to a slightly lower ratio of students to teachers, as special education and inclusion classes tend to be smaller than general education classes. For both cohorts, the mean of the instructional demand distribution is approximately zero for the full sample and decreases as more restrictions are imposed. This indicates that the students excluded from the sample tend to have higher instructional demand levels than those that are retained. Special education students, who generally have high instructional demand levels, likely contribute to this pattern. For instance, students with specific learning disabilities may be more likely to work with someone other than their primary classroom teacher in either math or ELA. Because the special education populations vary in size across the two cohorts, special education teachers may be more likely to move between grade levels between years, and therefore less likely to meet the criteria for inclusion in the growth analyses. Potential implications of these differences on the quality of information about teachers of high-demand and special education students are considered throughout the study.

Table 4-1. Full and restricted samples of teachers and students.

	Student Cohort 1			**Student Cohort 2**		
Sample restrictions	*5^{th} grade 2015-2016*			*5^{th} grade 2016-2017*		
Students with same teacher for math and ELA		X	X		X	X
Teachers who taught 5^{th} grade both years			X			X
Percent of all students included	100%	97%	74%	100%	97%	73%
Percent of all teachers included	100%	97%	66%	100%	97%	71%
Student-teacher ratio	19.5	19.6	21.8	20.7	20.9	21.4
Percent eligible for special education services	12%	12%	11%	11%	11%	11%
Distribution of instructional demand	0.00 (0.98)	-0.03 (0.98)	-0.09 (0.98)	0.00 (0.98)	-0.04 (0.97)	-0.08 (0.97)

4.2.2 Differential Student Functioning

Tests of DSF are conducted using the Mantel-Haenszel (MH) DIF index (Mantel & Haenszel, 1959; Holland & Thayer, 1988). The MH DIF statistic is computed from a set of 2x2 contingency tables, with a separate table for each of *j* latent ability levels, containing the counts of correct and incorrect responses to a particular item *i* for members of the focal and reference groups. Applying these same procedures for a DSF analysis, there is a separate 2x2 table for each of *t* latent capacity levels that contains counts of students at an instructional demand level *d* who did and did not meet a particular performance target in the focal and reference groups (the SRT adaptation of the Mantel-Haenszel DIF table is shown in Table 4-2). In a DIF analysis, the values in these contingency tables represent responses to the same item across different examinees. In the DSF context, however, each student is only observed with one teacher. In this case, the counts in each cell of the contingency table instead reflect groups of students with approximately equal levels of instructional demand.

Table 4-2 Mantel-Haenszel DSF 2x2 contingency table

	Student reached target	
	Yes	**No**
Reference group	A_{st}	B_{st}
Focal group	C_{st}	D_{st}

The distribution of student instructional demand estimates, shown in Figure 4-1, is slightly bimodal. Each peak represents a commonly observed instructional demand level among the sample of students, one slightly above and one slightly below the mean. Students with instructional demand levels within 0.15 of the lower peak are grouped together for one set of DSF analyses, and students within 0.15 of the upper peak are grouped together for another set. These groups of students provide insight about performance of educators with slightly below-average and slightly above-average demand levels, across several key characteristics of teachers and classrooms. For the purpose of the DSF tests, each of these groups of students operates like a single test item administered to many examinees, rather than a set of similar items that are each administered to one examinee. These groups were selected because their high frequencies allow for more teachers to be included in the DSF tests. 76% of all teachers in the sample have at least one student in at least one of the two demand intervals and 55% have at least one student in each of the two demand intervals.

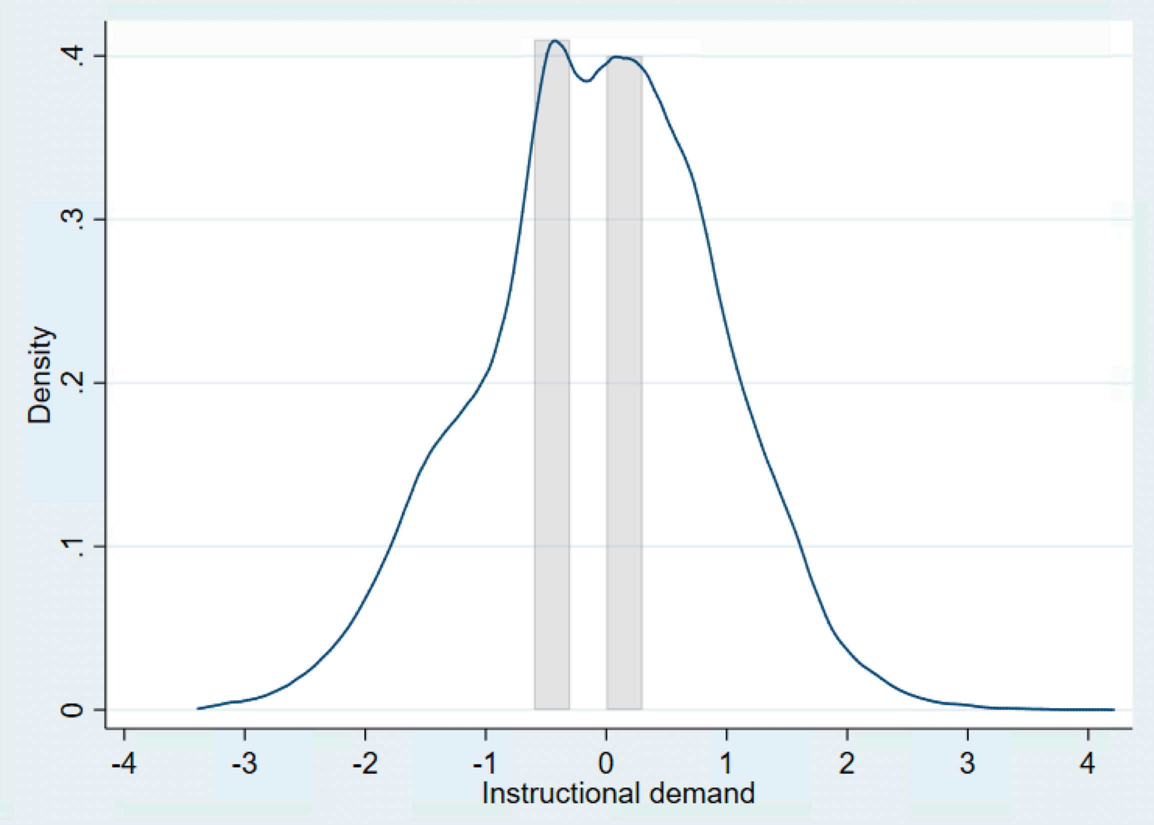


Figure 4-1. *Distribution of instructional demand with shaded areas in DSF intervals.*

In Chapter 3, the instructional demand and educator capacity scales were divided into eight one-unit intervals for reporting purposes. These same eight intervals (which each span approximately half of a standard deviation of the capacity distribution) are used as the latent capacity levels for the DSF analysis. The MH common odds ratio, shown in Equation 4-1, is computed using the counts in each of the contingency tables across the eight capacity levels. This ratio is then transformed into a "Mantel-Haenszel delta difference" (Dorans & Holland, 1993), which is typically abbreviated as *MH DIF* but shown in Equation 4-2 as *MH DSF* for consistency with SRT terminology.

$$\hat{\theta}_{MH} = \frac{\sum_t [A_{st}D_{st}/(A_{st} + B_{st} + C_{st} + D_{st})]}{\sum_t [B_{st}C_{st}/(A_{st} + B_{st} + C_{st} + D_{st})]} \tag{4-1}$$

$$MH\ DSF = -2.35 \ln\left(\hat{\theta}_{MH}\right) \tag{4-2}$$

The resulting DSF statistics are classified using the Educational Testing Service (ETS) guidelines for categorizing DIF. The ETS classification system assigns a letter (A, B, or C) indicating the level of DIF ("negligible DIF," "slight to moderate DIF," or "moderate to large DIF," respectively) and a symbol ("+" or "-") indicating the direction of DIF, based on the magnitude of MH DIF and statistical significance of the corresponding chi-square test (Zieky, 1993; Zwick, Thayer, & Lewis, 1999). For instance, "B+" indicates slight to moderate DIF in favor of the focal group, while "C-" indicates moderate to large DIF in favor of the reference group. Table 4-3 outlines these ETS classification criteria, with terminology from SRT in place of analogous IRT terms.

Table 4-3. DSF categories based on the ETS DIF classification system.

		MH chi-square test	
Direction	**Magnitude**	$p \leq 0.05$	$p > 0.05$
Positive DSF (favors focal group)	MH DSF ≥ 1.5	**C+**	**A**
	1.0 ≤ MH DSF < 1.5	**B+**	**A**
	MH DSF < 1.0	**A**	**A**
Negative DSF (favors reference group)	MH DSF > -1.0	**A**	**A**
	-1.5 < MH DSF ≤ -1.0	**B-**	**A**
	MH DSF ≤ -1.5	**C-**	**A**

4.2.3 Equating the Instructional Demand Index Across Years

First, test-level properties of the student instructional demand indices (SIDIs) from each year are compared to assess a few key assumptions of test form equating. In order for two test forms to be considered equated, or for scores from each form to be used interchangeably, both forms must measure the same construct with the same level of precision and have the same level of difficulty (Holland & Dorans, 2006). Because the SIDIs were constructed using IRT calibration, plots of their test information functions (TIFs) and conditional standard errors can be examined to compare the difficulty and precision levels from each year. Each cohort of students is then assigned an estimate of instructional demand based on the SIDI from the opposite cohort, so that estimates for the same individuals can be compared across indices. A high degree of consistency between these estimates is desired to support the assumption that both SIDIs measure the same construct.

Next, parameters and standard errors of individual instructional demand indicators are compared across the two SIDIs. Although both SIDIs were constructed using identical sets of instructional demand indicators, relationships between some indicators and the latent construct could differ between the two years. Relationships between the location parameter estimates, standard errors of location parameter estimates, slope parameter estimates, and standard errors of slope parameter estimates, for the same indicators across the two years are examined so that a set of consistent indicators can be identified. These indicators are referred to as “anchors,” which form the basis for equating the SIDIs using an anchor test design (Kolen & Brennan, 2014). Indicators are selected for the anchor test using the following steps: 1) estimates from the SIDI for the second cohort are regressed on estimates from the first cohort, 2) instructional demand indicators with standardized residuals greater than 3 or less than -3 are excluded from consideration, and 3) the same regression model is fit without these indicators. These steps are repeated until all remaining instructional demand indicators fall between -3 and 3, and these remaining indicators comprise the anchor test.

The anchor test is then assessed against criteria for test length, content balance, and parameter stability to establish its suitability for common-item equating. The anchor test must meet a minimum length requirement of about 20-25% of the total number of items in each full test form (Hambleton et al., 1991; Kolen & Brennan, 2014). The indicators included in the anchor test must also be representative of the balance of content on each form, such that the anchor test looks like a "mini version" of the two full test forms (Kolen & Brennan, 2014). Lastly, the parameters of instructional demand indicators must be sufficiently stable across the two forms, with correlations of at least 0.95 and the ratio of their standard deviations between 0.9 and 1.1 (Huynh & Meyer, 2010).

The SIDI for the second cohort of students is rescaled by applying a linear transformation that results in the location parameters of anchor indicators having the same mean and standard deviation as they do for the SIDI for the first cohort of students. The rescaled SIDI is then used to fit SRTs models in order to estimate educator capacity and consistency for the second school year on the same scale as the capacity and consistency estimates from the first school year. DSF tests, as described in Section 4.2.2, are conducted to compare results from the first and second school years as reference and focal groups, respectively. The purpose of these DSF tests is to determine whether there are differences in model performance across teachers with similar capacities estimated from models for different years. If DSF is nonnegligible, this could indicate nonequivalence of the two scales.

4.2.4 Identifying Growth of Teachers

Growth of an educator over the course of a particular year *y*, represented by the variable Δ_{yt}, is computed using Equation 4-3. The variables θ_{yt} and $\theta_{(y-1)t}$ represent the capacity estimates for teacher *t* in years *y*-1 and *y* (in this study, these are the 2015-2016 and 2016-2017 school years, respectively). The standard errors of the two capacity estimates are derived from the educator-class information functions (E-CIFs) for each teacher, as described in Chapter 3. The height of the E-CIF at the estimated capacity level is inversely related to the standard error of the capacity estimate, as shown in Equation 4-4. The height of the educator-student information function (E-SIF) is computed for each student *s* assigned to teacher *t* in year *y*, given the slope and capacity estimates for teacher *t* in year *y* (θ_{yt} and a_{yt}, respectively) and the instructional demand level of student *s* (represented by d_{ys}). These heights are summed across all students assigned to the same teacher that year, and the reciprocal of the square root of this sum is the standard error of the capacity estimate. The standard error of the growth estimate, s_{Δ_t}, is computed using Equation 4-5, where $s_{\theta_{(y-1)t}}$ and $s_{\theta_{yt}}$ are the standard errors of $\theta_{(y-1)t}$ and θ_{yt}, respectively.

$$\Delta_{yt} = \theta_{yt} - \theta_{(y-1)t} \tag{4-3}$$

$$s_{\theta_{yt}} = \sqrt{\sum\nolimits_{ys(t)} a_{yt}^2 \left[\frac{e^{a_{yt}(\theta_{yt}-d_{ys})}}{1+e^{a_{yt}(\theta_{yt}-d_{ys})}}\right]\left[1-\frac{e^{a_{yt}(\theta_{yt}-d_{ys})}}{1+e^{a_{yt}(\theta_{yt}-d_{ys})}}\right]}^{\,-1} \tag{4-4}$$

$$s_{\Delta_t} = \sqrt{\left(s_{\theta_{(y-1)t}}\right)^2 + \left(s_{\theta_{yt}}\right)^2} \tag{4-5}$$

Using these estimates of growth and their standard errors, three main questions are explored. First, under what conditions can growth of teachers be detected? In order for growth to be detectable, the standard error of each capacity estimate must be reasonably small. Because these standard errors are derived from the sums of information functions across all students in a class, they are likely sensitive to the number of students in a class and the distance between estimates of student instructional demand and estimates of educator capacity. The next question, "how much growth is detectable?" is addressed by determining the proportions of observed growth estimates that have standard errors below the threshold required to detect different amounts of growth. The third question, "what characteristics are associated with different types of growth?" is addressed for a subset of teachers with standard errors that are sufficiently small to detect moderate levels of growth. Relationships between observable teacher characteristics and significant negative, nonsignificant or zero, and significant positive changes in capacity are explored using a series of chi-square tests of association.

4.2.5 Assignment Optimization

Optimal assignments of students and teachers to one another are explored for rising 5th graders in one sample school. The school was selected based on the number of general education classrooms for the 5th grade, and relatively high variation in student instructional demand and teacher capacity estimates compared to other schools of similar sizes. There are 126 5th grade students, 6 general education 5th grade classrooms, and one teacher for each of these classrooms in the sample school. Table 4-4 shows the distribution of instructional demand for each of these classrooms, along with the average predicted probabilities of reaching the low and middle performance targets across students in each class, based on estimates of capacity and consistency of their assigned classroom teachers. Capacity estimates from the initial year are used because results from 2016-2017 would not be available at the time assignment decisions are made. The six classrooms vary somewhat in size and in mean instructional demand. On average, most students are very unlikely to reach the middle target for either math or ELA, although these probabilities are higher for students with Teacher 3. For the low performance targets, there is more variation in mean probabilities for math across classrooms, while the probabilities for ELA are very high for students with most teachers.

Table 4-4. Student demand and probability estimates by teacher.

Teacher/ Class	Student Count	Instructional Demand				Average Probability			
		Mean	SD	Min	Max	Math targets		ELA targets	
						Low	Middle	Low	Middle
1	22	-0.41	1.00	-1.44	2.25	0.57	0.00	0.93	0.03
2	20	0.91	0.75	-0.85	2.22	0.42	0.00	0.80	0.00
3	22	-0.70	1.17	-0.21	2.13	0.96	0.11	1.00	0.73
4	17	0.23	1.18	-1.43	2.15	0.65	0.00	0.53	0.01
5	22	0.45	0.98	-1.26	2.81	0.46	0.00	0.08	0.00
6	23	0.42	1.17	-1.60	2.62	0.70	0.00	0.99	0.20

Using an evolutionary optimization algorithm, different arrangements of these 126 students with these 6 teachers are generated in an effort to maximize the sum of probabilities of reaching performance targets across all students and both subjects. For most students, the probabilities considered in the objective function reflect the low targets for both math and ELA. However, if a student has a probability above 0.5 for the low target in a particular subject across all 6 potential teachers, the middle target probability is used in its place. Theoretically, if a student had a probability above 0.5 across all 6 teachers for the middle target, the high target would be considered instead. Because no students meet this description, the high targets are not considered at all.

Optimization procedures are repeated across four conditions, and each condition is repeated with 5 different random seeds, yielding 20 suggested assignments. These are compared to the actual assignments as well as randomly-generated assignments. The four conditions differ according to how initial values are set prior to beginning the optimization procedure (actual assignments or random assignments) and the mutation rate for the evolutionary algorithm (0.075 or 0.15). A higher mutation rate allows for more changes in assignments in each iteration of the optimization process, while a lower mutation rate places greater weight on preserving the initial assignments. In order to ensure that class sizes are reasonably equitable across teachers, a minimum and maximum number of students per teacher (17 and 25, respectively) are included as integer constraints.

4.3 Results

4.3.1 Differential Student Functioning Results

Results of DSF tests by teacher ethnicity, teacher gender, evaluation status, student count, mean instructional demand in a class, standard deviation of instructional demand in a class, classroom cluster (as defined in Chapter 3), and indicators for whether the majority of students in a class are nonwhite, eligible for special education services, classified as English Language Learners (ELL), or eligible for free or reduced-price lunch (FRL), are provided in Table 4-5. Each MH DSF statistic and corresponding chi-square significance test is assigned a category in accordance with the ETS DIF classification system. DSF is negligible across nearly every characteristic tested. There are a few exceptions to this: classes with mostly special education students compared to those with mostly general education students (for both modal demand levels and both subjects), high demand classes compared to average demand classes (lower mode and ELA only), Cluster 6 classes compared to Cluster 2 classes (lower mode for math and upper mode for ELA), and Cluster 1 classes compared to Cluster 2 classes (lower mode for ELA only). DSF is negative for each of these comparisons, indicating lower performance in the focal group than the reference group.

4.3.2 SIDI Equating Results

The test information functions (TIFs) and conditional standard errors for the SIDIs from each year (prior to equating) are shown in Figure 4-2. The corresponding curves from opposite years are nearly indistinguishable from each other upon visual inspection. Estimates of instructional demand for students based on the SIDI for the opposite cohort are also nearly identical to those for their own cohort, with correlations above 0.99. As Figure 4-3 illustrates, instructional demand estimates on the two SIDIs fall along the identity line for nearly all students. There are very few points that deviate far enough from this line for visual identification, and these deviations are extremely small.

Parameter estimates are also consistent across the two SIDIs. The only indicators flagged as outliers and excluded from the anchor set are those for the gifted programs and the total number of disabilities. Parameter estimates for the remaining indicators are correlated above 0.99 across the two years. The ratio of anchor indicator standard deviations for the 2015 and 2016 SIDIs is 1.02. Table 4-6 shows results from DSF tests using the two years as comparison groups

Table 4-5. Results from differential student functioning (DSF) analysis

		Lower modal student								Upper modal student							
		Math				**ELA**				**Math**				**ELA**			
Comparison groups																	
Focal	Reference	MH DSF	chi^2	sig.	ETS DSF	MH DSF	chi^2	sig.	ETS DSF	MH DSF	chi^2	sig.	ETS DSF	MH DSF	chi^2	sig.	ETS DSF
Nonwhite	White	0.22	1.32	0.17	A	0.22	0.89	0.14	A	0.01	0.00	0.00	A	-0.31	2.92	0.17	A
Male	Female	0.12	0.33	0.07	A	-0.13	0.25	0.06	A	-0.13	0.48	0.10	A	-0.13	0.46	0.09	A
Evaluation year	Not evaluated	-0.12	0.32	0.07	A	0.85	10.9	0.01	A	-0.04	0.03	0.01	A	0.25	1.90	0.18	A
Low student count	Average	-0.36	2.52	0.18	A	-0.54	4.19	0.13	A	-0.54	7.42	0.05	A	-0.32	2.65	0.18	A
High student count	Average	0.33	1.73	0.18	A	-0.70	5.01	0.10	A	-0.06	0.04	0.01	A	-0.01	0.00	0.00	A
High mean demand	Average	-0.64	2.91	0.17	A	-1.22	9.09	0.02	B-	0.03	0.01	0.00	A	0.01	0.00	0.00	A
Low mean demand	Average	-0.19	0.65	0.12	A	-0.22	0.60	0.11	A	0.48	3.24	0.16	A	0.26	0.89	0.14	A
High SD demand	Average	-0.35	2.44	0.18	A	-0.33	1.50	0.18	A	0.27	1.46	0.18	A	0.29	1.64	0.18	A
Low SD demand	Average	0.09	0.14	0.03	A	0.02	0.00	0.00	A	0.10	0.20	0.05	A	-0.57	9.20	0.02	A
Cluster[3] 1	Cluster 2	0.06	0.05	0.01	A	-1.00	12.8	0.01	B-	0.30	1.59	0.18	A	0.05	0.03	0.01	A
Cluster 3	Cluster 2	0.03	0.00	0.00	A	-0.62	1.77	0.18	A	-0.12	0.16	0.04	A	-0.07	0.05	0.01	A
Cluster 4	Cluster 2	-0.45	4.22	0.13	A	-0.22	0.61	0.11	A	-0.02	0.00	0.00	A	0.14	0.35	0.07	A
Cluster 5	Cluster 2	-1.57	3.50	0.15	A	-1.68	3.26	0.16	A	-0.69	1.92	0.18	A	-0.70	1.52	0.18	A
Cluster 6	Cluster 2	-7.81	11.4	0.01	C-	-4.66	6.85	0.06	A	-2.98	6.61	0.06	A	-3.28	9.82	0.02	C-
Mostly nonwhite	White	0.04	0.00	0.00	A	0.17	0.05	0.01	A	-0.39	0.83	0.14	A	-0.14	0.06	0.02	A
Mostly Special ed.	General ed.	-8.96	9.12	0.02	C-	-11.7	18.7	0.00	C-	-4.43	14.1	0.00	C-	-5.97	25.0	0.00	C-
Mostly ELL	Not ELL	0.03	0.02	0.01	A	0.35	2.34	0.18	A	-0.43	6.43	0.07	A	-0.16	0.76	0.13	A
Mostly FRL	Not FRL	-0.43	1.45	0.18	A	0.51	1.49	0.18	A	-0.69	4.53	0.12	A	-0.00	0.00	0.00	A

[3] These are the same clusters defined from class information functions (CIFs) in Chapter 3. Descriptions of each cluster are as follows: 1) high student count and low mean demand, 2) high student count and average mean demand, 3) average student count and high mean demand, 4) average student count and below-average mean demand, 5) low student count and high mean demand (inclusion classes with combination of special education and general education students), 6) very low student count and very high mean demand (dedicated special education classrooms).

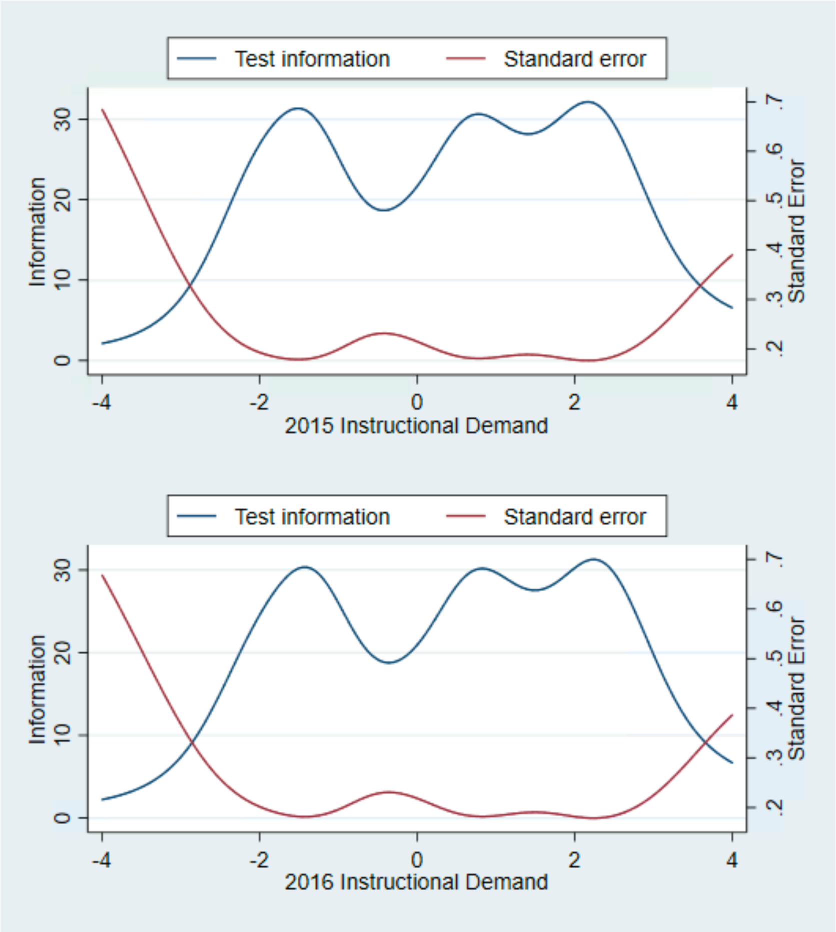


Figure 4-2. *Test information functions and conditional standard errors for each SIDI.*

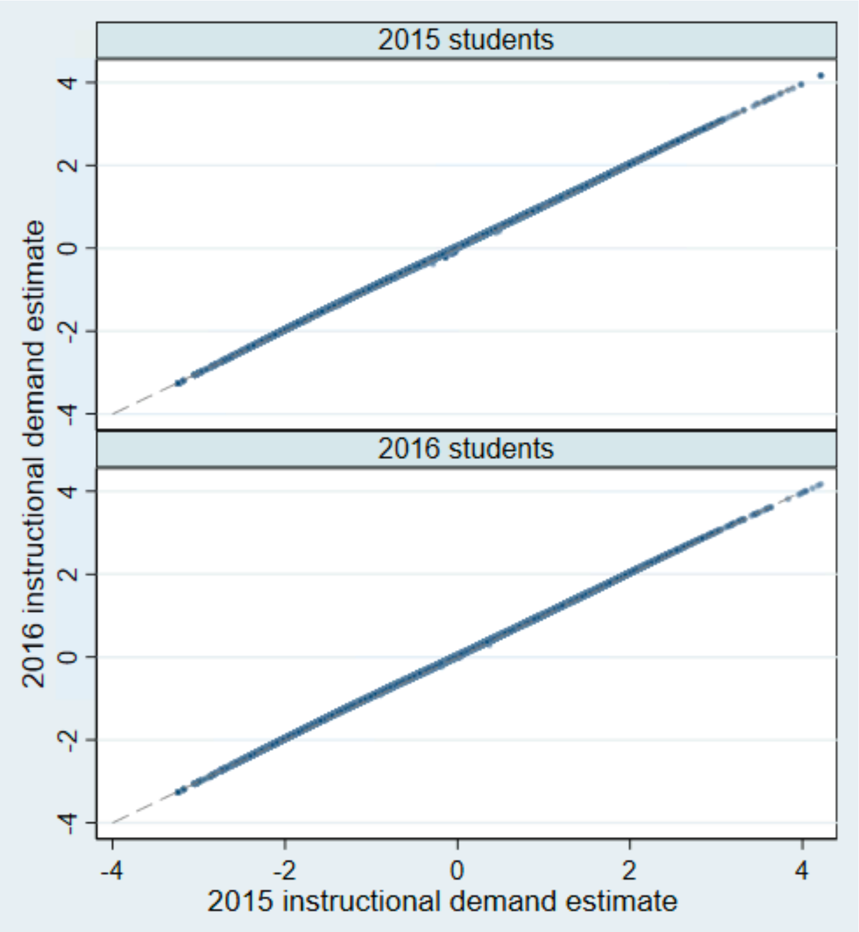


Figure 4-3. *Instructional demand estimates based on each SIDI.*

for SRT models using the low performance targets and equated SIDIs. DSF is negligible for both modal instructional demand levels across both subject areas.

Table 4-6. DSF* by year for equated SRT measures.

		MH DSF	chi^2	sig.	ETS DSF
Math	Lower mode	-0.02	7116	0.00	A
	Upper mode	0.01	884	0.00	A
ELA	Lower mode	0.04	12700	0.00	A
	Upper mode	0.20	1483	0.00	A

**The focal and reference groups are 2017 and 2016 for these tests, respectively..*

4.3.3 Teacher Growth Analysis Results

On average, teacher capacities decreased by about 0.07 in math and 0.02 in ELA between the two school years. These changes correspond to about 0.04 and 0.01 standard deviations of the capacity distribution[4], respectively. Figure 4-4 provides histograms of math and ELA capacity changes. The magnitude of change is within one standard deviation of the capacity distribution for approximately 70% of teachers in math and 65% of teachers in ELA. The standard errors of these estimated changes vary widely across teachers. For math, the mean standard error is 1.28 with a standard deviation of 1.40. For ELA, the mean is 1.62 with a standard deviation of 1.82. The distributions of standard errors, which are also provided in Figure 4-3, are highly skewed in the negative direction. The median standard errors of 0.72 for math and 0.82 for ELA are more reflective of the typical magnitudes observed in the sample, as the means are more sensitive to a

[4] The standard deviations of capacity measures are slightly larger across the entire of samples (approximately equal to 2) than for the analytic sample (about 1.80 and 1.74 for math and ELA, respectively). These values reflect the standard deviation of 2015-2016 capacities across the analytic sample of teachers in the growth analysis.

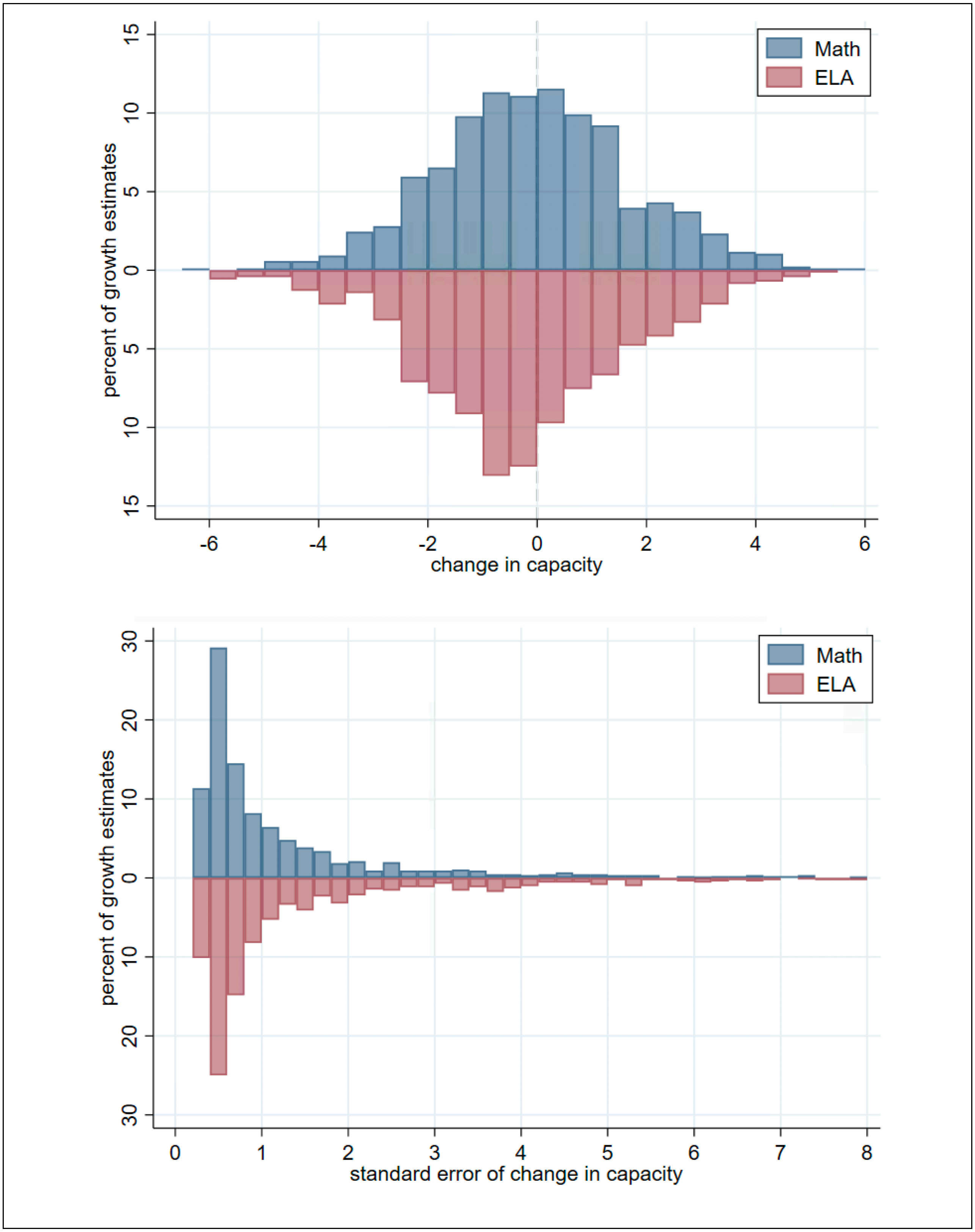


Figure 4-4. *Distributions of changes in capacity and their standard errors.*

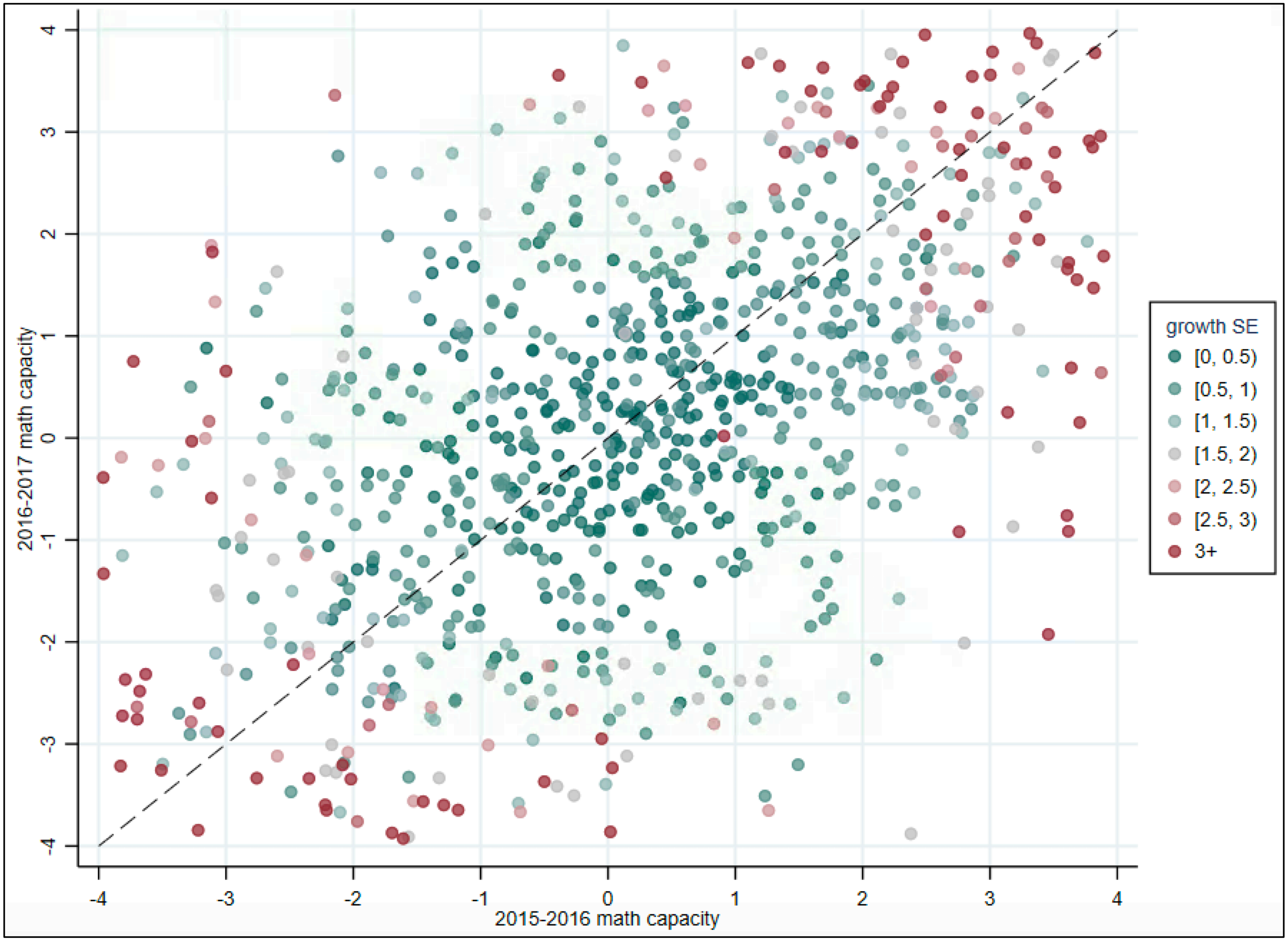


Figure 4-5. *Scatterplot of math capacity estimates colored according to the standard errors of corresponding math growth estimates.*

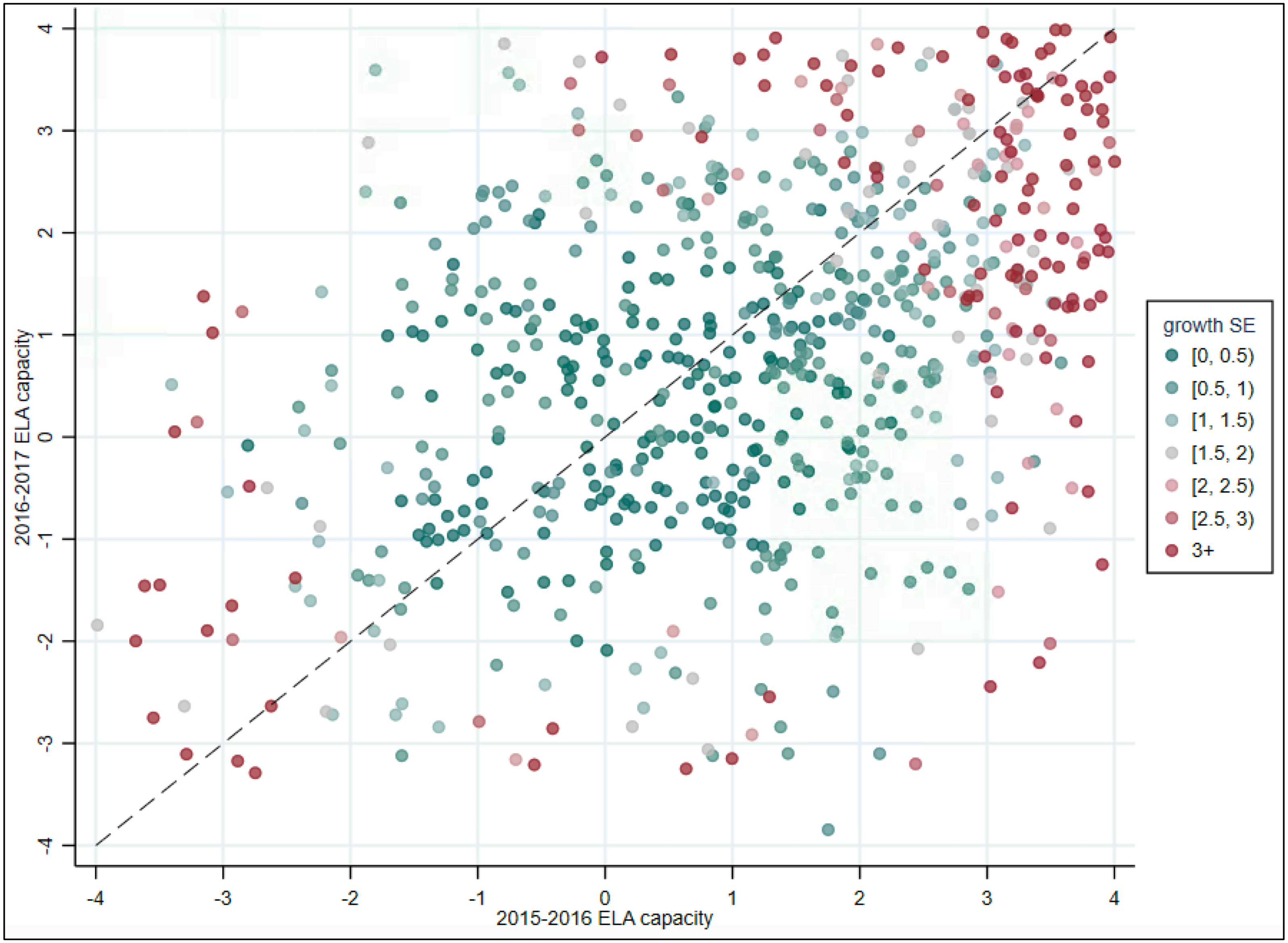


Figure 4-6. *Scatterplot of ELA capacity estimates colored according to the standard errors of corresponding ELA growth estimates.*

relatively small number of extremely large standard errors. Figures 4-5 and 4-6 provide scatterplots of 2015-2016 and 2016-2017 capacity estimates for the same subject with points colored according to the standard error of the growth estimate (or the change between the two capacity estimates). Standard errors are generally largest when the capacity estimate in one of the years is on one of the far ends of the distribution. However, there are teachers with capacities in these ranges with standard errors in the lower ranges.

Student counts and gaps between teacher capacity and class mean instructional demand were explored as possible drivers of differences in standard errors across teachers, as both of these quantities are directly related to the educator-class information function (E-CIF) from which standard errors are derived. Figure 4-7 provides the mean standard error among groups of teachers with a given student count in each of the two years. There is no obvious visual relationship between the student counts for each year and the mean standard error. In fact, some cells corresponding to very low student counts have small mean standard errors and some cells corresponding to very high student counts have very large mean standard errors.

Figures 4-8 and 4-9 provide scatterplots of the distance between a teacher's capacity estimate and the mean instructional demand estimate of students in the teacher's class for each year, where the color of each point corresponds to the standard error of the growth estimate for the teacher. There are clear visual patterns in the magnitude of standard errors based on these capacity-demand gaps; teachers with larger gaps in either direction have larger standard errors. There are a few exceptions to this, where a teacher without an extreme capacity-demand gap in either year has a standard error in the highest interval (appearing as a dark red point in a location populated mostly by bright green points). These correspond to teachers with unusually small consistency (or slope) estimates in at least one of the two years.

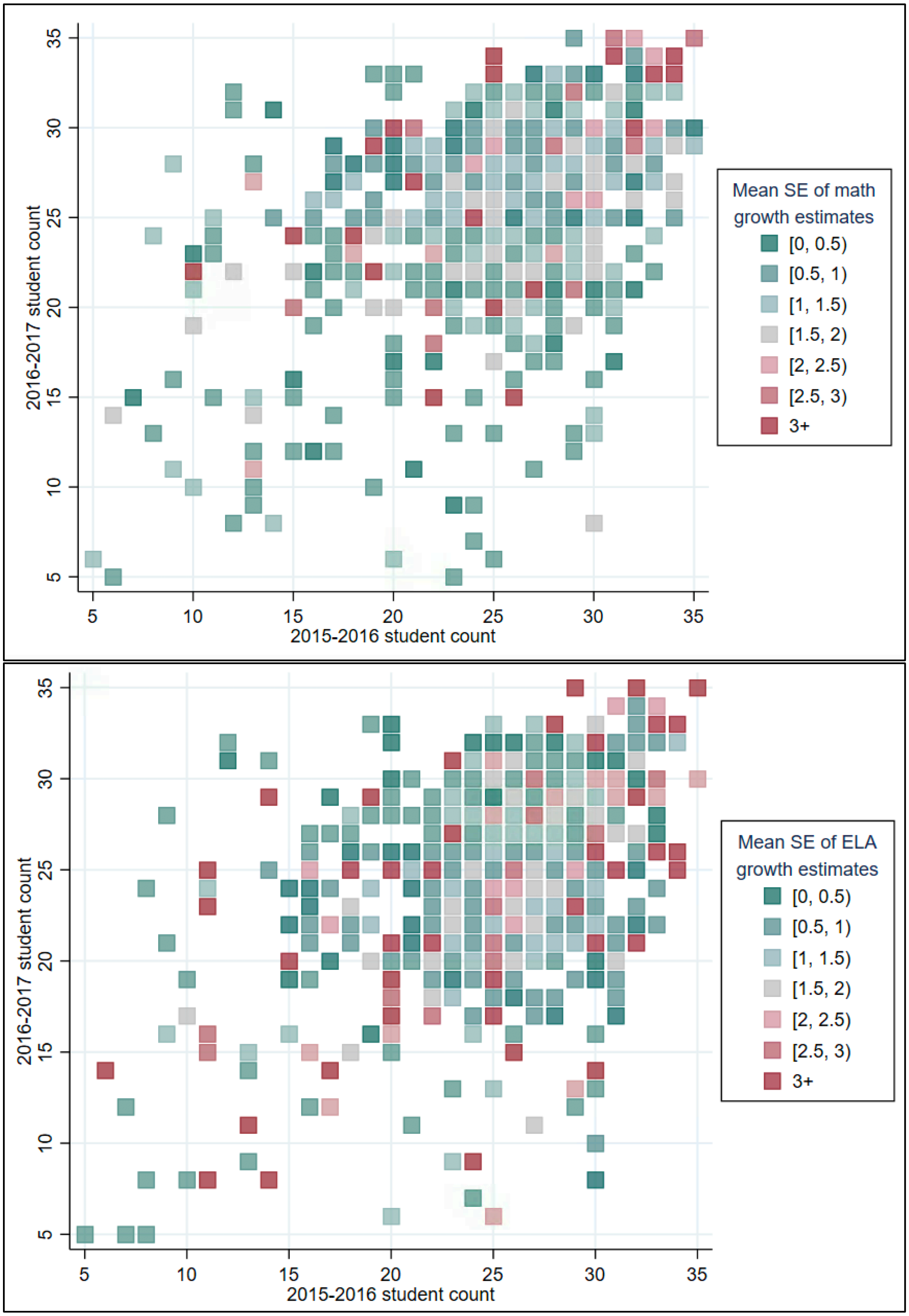


Figure 4-7. *Mean standard errors of growth estimates by student counts.*

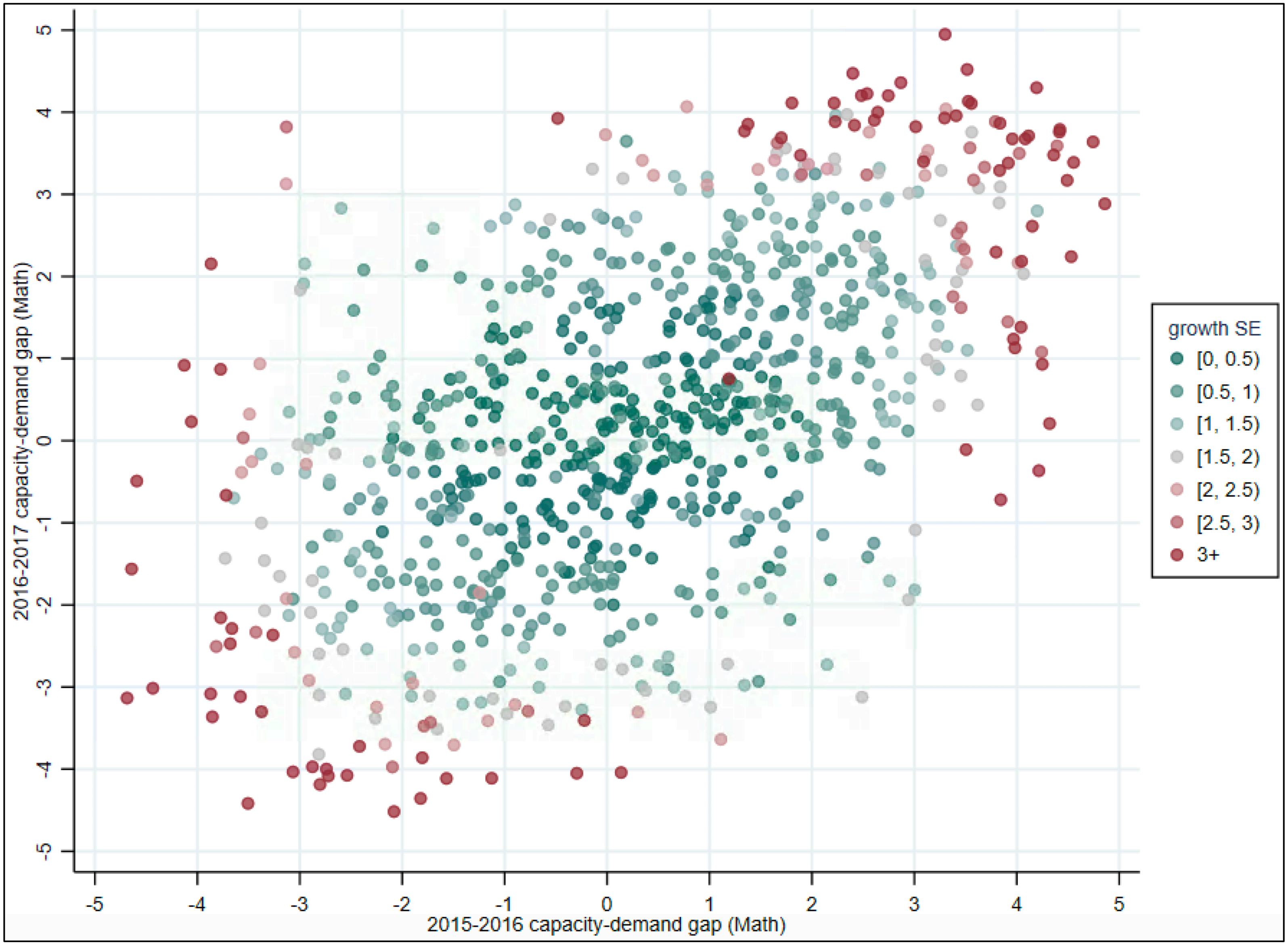


Figure 4-8. *Standard errors of growth estimates by gap between math capacity and class mean instructional demand.*

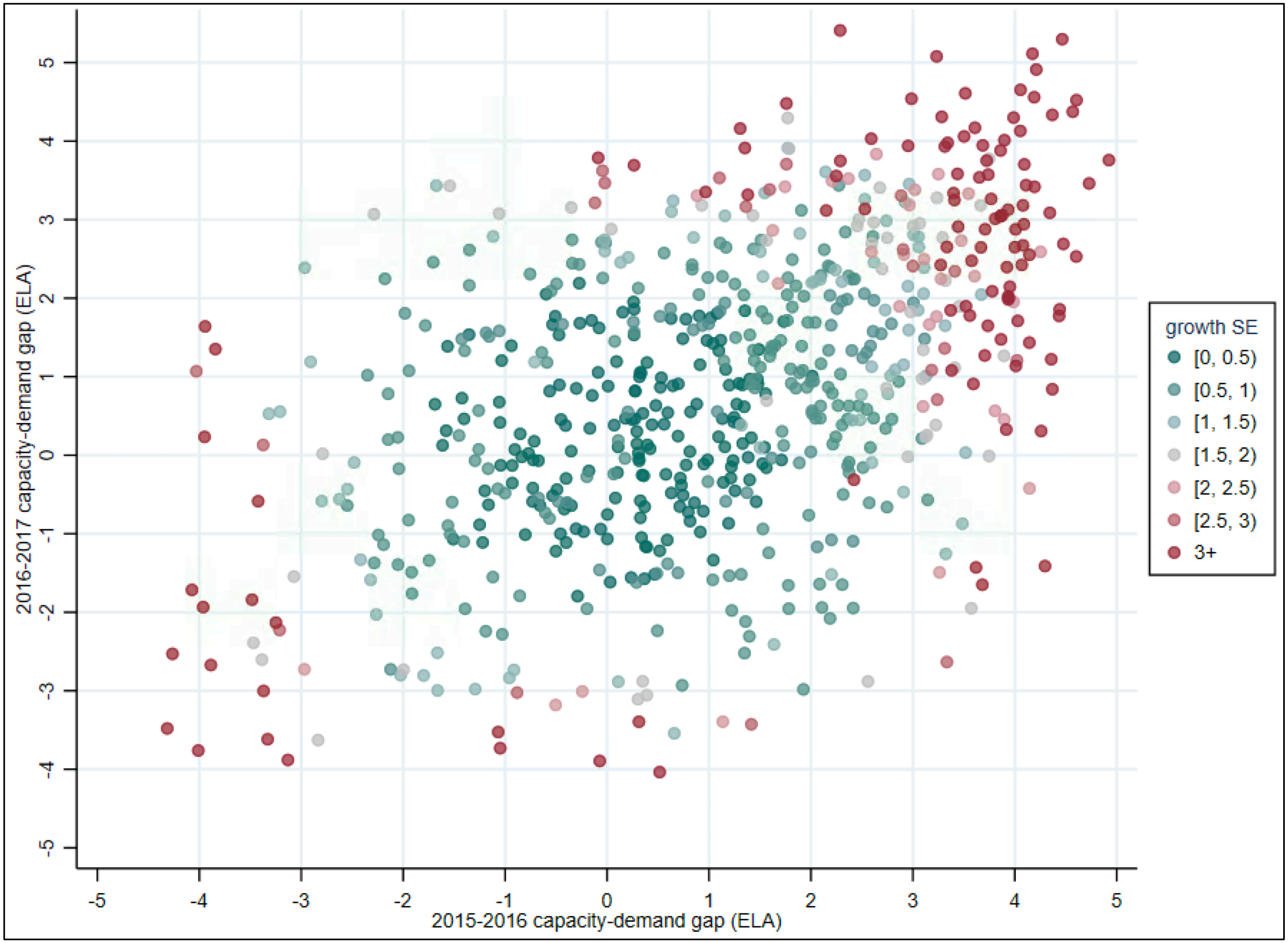


Figure 4-9. *Standard errors of growth estimates by gap between ELA capacity and class mean instructional demand.*

Table 4-7 provides details about the standard errors necessary for different magnitudes of growth to be considered statistically significant. The smallest magnitude of growth in math capacity with statistical significance is 0.64. This corresponds to about one third of a standard deviation of the capacity distribution. While about 72% of all math growth estimates are larger than this (including both positive and negative growth estimates), only 3% of math growth estimates have sufficiently small standard errors for this change to be statistically significant. Similarly, only 7% of ELA growth estimates have sufficiently small standard errors for the smallest observed significant growth (0.73 capacity units) to be statistically significant. More than half of all growth estimates have sufficient standard errors to detect significant changes in capacity of one full standard deviation unit, however, growth of this magnitude is not frequently observed, particularly among the subset of teachers with sufficient standard errors to detect it. About 31% of math growth estimates and 35% of ELA growth estimates are equal to or above this level, and fewer than half of these are statistically significant.

Table 4-7. Detectable and observed magnitudes of growth.

Math	Minimum detectable	0.5 SD	1 SD
Magnitude of growth (change in capacity)	0.64	0.90	1.80
Maximum SE for this magnitude of growth to be significant	0.33	0.46	0.92
Percent of teacher growth estimates above this magnitude	72%	60%	31%
Percent of teachers with sufficient SE to detect growth	3%	22%	60%
Percent of teachers with significant growth above this magnitude	1%	8%	13%
ELA	Minimum detectable	0.5 SD	1 SD
Magnitude of growth (change in capacity)	0.73	0.87	1.74
Maximum SE for this magnitude of growth to be significant	0.37	0.44	0.89
Percent of teacher growth estimates above this magnitude	67%	62%	35%
Percent of teachers with sufficient SE to detect growth	7%	16%	53%
Percent of teachers with significant growth above this magnitude	3%	7%	14%

For the subset of teachers with standard errors below the threshold for detecting one standard deviation of growth, percentages of teachers with significant negative, nonsignificant or zero, and significant positive changes in capacity are shown in Table 4-8. About half of all teachers have significant changes in capacity across the two years, but significant changes are more often negative than positive. Early career teachers are the only group more likely to have a positive change in math capacity than a negative change, and no groups of teachers are more likely to have positive changes than negative changes in ELA capacity. Rates of negative changes are particularly high among probationary teachers (in both subjects) and mid-career teachers (in ELA only). There are statistically significant associations between ELA growth and both teacher minority status and teaching experience. There are no significant associations found between math growth and observable teacher characteristics.

Table 4-8. Comparison of teachers by change type (if 1SD change is detectable)

	Math Growth				**ELA Growth**			
	Negative	None	Positive		Negative	None	Positive	
All teachers	25%	54%	20%		29%	49%	22%	
Employment status								
Probationary	42%	50%	8%	χ^2=2.30	40%	45%	15%	χ^2=1.41
Tenured	25%	55%	21%	*p*=0.32	28%	50%	22%	*p*=0.50
Master's degree								
Yes	24%	57%	19%	χ^2=0.88	26%	51%	23%	χ^2=0.89
No	26%	53%	21%	*p*=0.64	30%	49%	21%	*p*=0.64
Teacher Ethnicity								
Nonwhite	26%	53%	21%	χ^2=1.43	30%	46%	24%	χ^2=6.23
White	24%	58%	18%	*p*=0.49	25%	60%	15%	*p*=0.04
Teaching Experience								
1-5 years	20%	50%	30%	χ^2=2.27	26%	48%	26%	χ^2=12.84
6-9 years	33%	48%	18%	*p*=0.69	59%	33%	7%	*p*=0.01
10+ years	25%	55%	20%		26%	51%	23%	
Evaluation Year								
Yes	22%	59%	19%	χ^2=1.61	26%	54%	20%	χ^2=1.14
No	26%	53%	21%	*p*=0.45	30%	48%	22%	*p*=0.57

4.3.4 Optimal Assignment Results

The top panel of Table 4-9 shows properties of actual class assignments and a random set of assignments. When students are placed with their actual assigned teacher, the objective function (the sum of probabilities across 126 students and 2 subjects) is equal to 124.06. The within-student sums of probabilities (math probability plus ELA probability) have a standard deviation of 0.52. These same values were computed for randomly generated assignments. Across 100 replications of randomly placing students and teachers with one another, the mean of the objective function is 121.15 with a standard deviation of 4.73. This suggests that the expected outcomes for the actual assigned classes are within the range of typical outcomes when classes are assigned completely at random. The mean standard deviation of within-student sums across the 100 random assignment replications is 0.58 with a standard deviation of 0.02. This suggests that, although the sum of probabilities is similar for actual and random assignments, there is less variability in student probabilities.

Properties of the actual assignments, a set of random assignments with a similar objective function value to the actual assignments, and distributions of properties across optimization replications from each condition, are provided in Table 4-9. Although the objective functions have similar values for the actual and random assignments, students and demand levels are distributed more equitably across teachers with the actual assignments than with random assignments. Optimized classes have more equitable class size distributions than the random assignments, but they are slightly less equitable than the actual assignments. The standard deviation of mean demand across teachers is higher for all of the optimized assignments than for the actual assignments, and highest for the conditions with actual assignments as starting values. On average, optimization procedures lead to greater increases in the objective function when the actual assignments are used as starting values. Even after optimization, the value of the objective function for conditions with random starting values is similar in to that of the actual assignments. Optimization conditions with the higher mutation rate have more consistent results across different random seeds.

Table 4-9. Actual, random, and optimized class assignments.

	Conditions		**Objective function**		**SD across teachers**	
	Initial	Mutation rate	Final	Change	Student count	Mean demand
Actual	----	----	124.06	----	2.19	0.15
Random	----	----	123.81	----	6.00	0.24
Optimal	Actual	0.075	136.49 (5.80)	12.44 (5.80)	2.92 (0.61)	0.38 (0.04)
Optimal	Actual	0.15	140.86 (1.09)	16.80 (0.98)	3.59 (0.41)	0.40 (0.10)
Optimal	Random	0.075	127.41 (4.94)	3.59 (4.94)	3.26 (0.47)	0.23 (0.07)
Optimal	Random	0.15	126.99 (2.33)	3.18 (2.33)	3.52 (0.67)	0.18 (0.02)

4.4 Discussion

These results offer two types of diagnostic feedback. First, they identify concerns about the performance and specifications of the SRT model. Second, they identify patterns in the performance of students and teachers. While the second type of feedback is most relevant to teachers, administrators, and other stakeholders, the first type provides evidence of validity and shortcomings in these measures that should be considered carefully when interpreting the more practical findings and when developing an appropriate SRT model for use in a formative evaluation system.

4.4.1 Differential Student Functioning

Most of the nonnegligible DSF results occur in similar groups of teachers that are identified in slightly different ways. There is likely a lot of overlap in the groups of teachers with special education classes, cluster 6 classes, and classes with high mean instructional demand levels, so it is not surprising that findings are similar across DSF tests for each of these focal groups. Teachers of high demand, special education, or cluster 6 classes tend to have fewer students in general. When viewing the classroom as a test in the SRT framework, these teachers are given shorter, more difficult tests, than teachers of an average class. The types of students in these classes are also more likely to have different teachers for math and ELA; these students were excluded from this study, potentially exacerbating the effects of low student counts for some of these teachers. Near-average students tend to perform worse in these classes than similar students with similar teachers in other types of classes. It is also possible that the instructional demand index does not adequately capture the added challenges associated with teaching many high-demand students simultaneously.

The only nonnegligible DSF test that was not for one of these same groups was for Cluster 1 relative to Cluster 2 for the lower mode in ELA. In this case, both the focal and reference groups have high student counts but differ in their mean demand levels. This result indicates that students with slightly below-average demand levels in large, low-demand classes are less likely to reach the ELA target than similar students with similar teachers in large, average-demand classes. This could indicate that teachers of Cluster 1 classes direct their instruction more towards low-demand students, while instruction in Cluster 2 classes is generally directed towards average-demand students. Similarly, tendencies to direct instruction towards higher-demand students in higher-demand classes could contribute to the significant DSF findings for special education, Cluster 6, and high-demand classes. If given this information as part of a formative evaluation process, an administrator may choose to plan professional development activities that focus on improving instructional practices to reach a broader range of students. An administrator may also consider this information when determining class assignments in the following year.

4.4.2 Equating the SIDI

The TIFs shown in Figure 4-1 provide strong evidence that, prior to equating, the SIDIs from each year can be considered equivalent forms. The high level of consistency for instructional demand estimates from each SIDI further supports that the forms can be used interchangeably, and neither form would be preferred over the other. The content balance of the anchor test is nearly identical to that of the complete set of indicators used to estimate each SIDI. The indicators for gifted students and the total number of disabilities do operate somewhat differently on the two forms, however. This may be related to differences between the groups of gifted students or students with multiple disabilities in each cohort. For instance, students in the first cohort with multiple disabilities may have qualitatively different disabilities than those in the second cohort. It is also possible that the eligibility criteria for different gifted programs or special education services changed slightly after the first year. Aside from these few exceptions, which were excluded from the set of anchors used to equate the scales, relationships between instructional demand indicators and the latent construct measured by the SIDI are quite stable across the two forms. The DSF tests comparing performance in the two years are all classified as negligible, affirming that the two scales are comparable.

4.4.3 Teacher Growth

Under certain conditions, there is relatively little power to detect significant changes in equated educator capacity estimates due to their unreasonably large standard errors. However, mismatch between student demand levels and teacher capacities poses a far greater threat to standard errors of growth estimates than low student counts. These findings emphasize that appropriate matching of the needs of students with the capabilities of teachers is critical in order to make inferences about the performance and growth of teachers in an SRT analysis. It may be more feasible to measure growth over longer periods of time, as changes occurring over multiple years may be larger relative to the standard errors of the initial and final capacity estimates.

Teachers at either extreme end of the capacity distribution are disproportionately likely to have large gaps between their capacity estimates and the mean instructional demand levels of their students, large standard errors, and large estimates of growth. Because of these relationships, the teachers with growth estimates above a given level tend not to be the same teachers with sufficiently small standard errors for this level of growth to be statistically significant. In future work, it will be important to address differences in SRT and IRT models and contexts, such as the level of the slope parameter and interactions among students in the same class, that are likely to contribute to these enlarged standard errors.

The frequency with which negative growth (or decreases in capacity from one year to the next) is observed raises concerns about possible changes in the difficulty of a performance target over time. If the cut-score for a performance level designated by the state is more difficult in the second year than the first, students will be less likely to reach the target than students of the same instructional demand levels in the previous year. Because the capacity scale is defined by the relationship between instructional demand and target attainment, this sort of change could result in a systematic shift of capacity estimates for all teachers regardless of whether their capabilities as educators have actually changed. In order to prevent incorrect inferences about declining capacities or underestimation of positive growth, comparability of performance targets should be established in future SRT studies involving longitudinal comparisons.

4.4.4 Optimal Assignment

Differences between the properties of actual assignments and randomly-generated assignments reveal that, although the expected sum of probabilities is similar, the actual assignments correspond to fewer students with probabilities very close to 0 or 1 than the random assignments. This suggests that the actual assignments do a better job of matching student instructional demand levels and teacher capacity levels. As a result, students are less likely to be placed with teachers with whom they have very low probabilities of reaching the performance target than if they were assigned at random. Despite beginning with approximately the same initial value for the objective function, the optimization process is more successful in improving successful outcomes when the actual assignments are used as initial values, compared to the random assignments. Although improvements to the objective function vary across random seeds, results are much more consistent across replications for conditions with the higher mutation rate. Results are most consistent for the condition with actual assignments as starting values and the higher mutation rate. Optimal assignments in this condition correspond to approximately 10 more students (of the 126 5th grade students in the school) expected to reach performance targets than with the actual assignments.

The standard deviation of class sizes tends to be larger for optimized assignments than for the actual assignments. The assignment procedures used in the sample school may prioritize equitable class sizes to a greater extent than simply imposing a minimum and maximum size. The class size restrictions may need to be revised in order to generate assignments that align with other priorities and concerns of the school. Similarly, the standard deviation of class mean instructional demand levels is lowest for the actual assignments. In order to maximize expected student outcomes, the optimal assignments focus more on matching of students and teachers, while assignment practices in the school may focus more on equitable distribution of instructional demand across teachers. If this is the primary objective of teachers and administrators in this particular school, then optimal assignments that minimize the standard deviation of instructional demand across teachers may be of greater interest. However, providing administrators with optimal assignments for different objectives, and comparisons of these with actual assignments, may be helpful in evaluating implications of their current assignment practices. For instance, while the sample school may currently focus more on equitable distribution than on expected student outcomes when determining assignments, an administrator may choose to emphasize expected student outcomes more in future assignment decisions after reviewing this type of report.

4.5 Conclusions

These findings affirm that the mean instructional demand level of a class, and its relationship to the capacity of a teacher or instructional demand of an individual student, impacts both the performance of students and the quality of information provided about a teacher.

Commonalities between special education, high mean demand, and cluster 6 classrooms are likely related more to discrepancies between the instructional demand levels of students in these classes and the capacities of most teachers than they are to low student counts. This is equivalent to assessing the performance of these groups of teachers using a significantly more difficult form of a test than for other teachers. When the latent capacity of a teacher is far below the demand level of their students, this may result in a floor effect. These patterns could also indicate a tendency of teachers to direct instruction towards the level of most students in the class. This would also explain differential performance of average students in very low-demand classes, compared to similar students with similar teachers in average classes.

The instructional demand indices for the two cohorts of students are successfully equated to a common scale, however, standard errors of teacher growth estimates are too large to make judgements about changes in most teachers' capacities between the two years. The magnitudes of these standard errors are most sensitive to gaps between capacity and mean demand and impacted little by student counts in comparison. With greater emphasis on matching student instructional demand levels with teacher capacity levels in assignment processes, this type of growth analysis may be more feasible. However, findings from the sample school and in previous studies suggest that some teachers and administrators prioritize equitable distribution of students and instructional demand across teachers, as opposed to matching of students with the most appropriate teachers.

Further development of SRT methods are necessary in order to adequately capture growth of educators when there are gaps between capacity and instructional demand as well as interactions among students in the same class and their contributions to the instructional demands posed to an educator. However, the negligible DSF findings, feasibility of longitudinal equating, and improvements in expected outcomes through assignment optimization are suggest that these procedures can contribute to formative teacher evaluation processes in novel and impactful ways.

CHAPTER 5. OVERALL CONCLUSION AND DISCUSSION

5.1 Summary of Findings

The main objective of the first paper is to demonstrate whether the student instructional demand index (SIDI) can be constructed in ways that are more beneficial to an evaluation system and less controversial politically. Although the IRT calibration method was less successful than regression analysis for estimating the SIDI in previous studies, this study is the first to use such an expansive set of instructional demand indicators. It is also the first to explicitly choose indicators based on their optimality for IRT calibration.

The results suggest that the IRT calibration method produces an index of instructional demand that is related to future achievement closely enough to demonstrate evidence of convergent validity but differs enough from future achievement to offer evidence of divergent validity. This suggests that the construct this SIDI captures is related to but different from future achievement. The regression analysis method, in comparison, creates a SIDI that is essentially an indicator of future performance. This is problematic within an accountability framework; if characteristics of a student before beginning the school year are so highly predictive of performance at the end of the school year, this implies that teachers have relatively little influence in these outcomes.

Estimates of teacher capacity from SRT models with IRT-calibrated SIDIs are also more consistent with other measures of teacher quality. The IRT-calibrated SIDI discriminates between teachers in different rating categories based on observations of their classroom teaching, as well as between different levels of teaching experience. However, when a difficult performance target is used as a standard for evaluating student and teacher performance, neither type of SIDI is able to discriminate among these groups of teachers.

Results from the second paper also emphasize that when student performance targets are unreasonably difficult, SRT models are not very informative about educator capacity. The lowest performance target for the state assessment taken by students in this sample is the only benchmark that provides reliable estimates of teacher performance across a wide range of the instructional demand scale. The middle performance target provides reliable information about some students and teachers in the district, but the highest performance target is generally uninformative across the entire student and teacher distributions.

Different measures derived from the low target are highly correlated with one another. This suggests that, although it might not be particularly useful to report multiple measures if the information they provide is redundant, administrators should feel comfortable choosing a measure derived from the SRF based on the relevance of its interpretation to the objectives and priorities of the educational system conducting the evaluation, as long as the underlying SRT model uses a performance target that is informative about the populations of students and teachers in that system.

Differences in class information functions (CIFs) raise some alarm about whether comparisons across all teachers are fair or appropriate. However, commonalities among the CIFs across large subsets of teachers, similar to parallel forms of a test, provide a promising solution to this problem. CIF matching allows for meaningful within-cluster comparisons, however, differences between within-cluster rankings and overall rankings suggest that overall rankings should be interpreted with caution.

Results from the third paper indicate that, under most conditions, model performance is consistent across characteristics of teachers and classes. Most exceptions to this are for classes with very high-demand students, including special education classes. The only other exception is for slightly below-average demand students placed in large, low-demand classes compared to large, average-demand classes. While low student counts in the high-demand and special education classes offer one possible explanation for these differences in model performance, the presence of a similar pattern across large classes with different demand levels could not be explained in the same way. Another explanation is that the differences in model performance relate to mismatch between the instructional demand levels of individual students and the instructional demand levels of their classmates. It is not clear whether performance of these students is more likely affected by the demand levels of their classmates, or whether the instruction administered in these types of classrooms is targeted towards the mean demand level of the class. If the latter is true, students far above or below that mean demand level might not benefit as much as they would in an average-demand classroom where instruction is designed for a different type of student.

Parameter estimates for instructional demand indicators are rather stable across the two years, lending well for common item equating of the SIDI across the two cohorts of 5th grade students. The only instructional demand indicators that were flagged for inconsistency were for gifted programs and students with multiple disabilities. These indicators represent opposite ends of the instructional demand distribution, where estimates tend to be less precise in general. These differences could also reflect changes in the criteria for program eligibility between the two years or differences in the relatively small groups of students from each cohort that fall within these categories. Standard errors of teacher growth estimates are rather large for some teachers, resulting in little power to detect changes in capacity in these cases. The primary factor driving these large standard errors is mismatch between the capacity of a teacher and the mean demand level of students in their class. This discrepancy makes a far greater impact than the number of students in a class. Small slope estimates also correspond to large standard errors, however, this affects far fewer teachers than capacity-demand mismatch.

Although inappropriate matching of students and teachers is a limiting factor for analyzing teacher growth, assignment optimization procedures provide guidance to schools about how to improve the ways students and teachers are assigned to one another. However, an administrator would only be interested in this information if student-teacher matching aligns with their objectives and priorities in the assignment process. In the sample school selected for the optimization analysis, the actual assignments appeared to prioritize the equitable distribution of students and instructional demand across teachers. This type of practice would likely result in more similar CIFs across teachers in the same school, facilitating comparisons among the different teachers. However, it is not necessarily the best practice for improving student outcomes, accurately assessing the performance of a teacher, or monitoring and supporting teacher growth over time.

5.2 Implications

The importance of setting realistic performance standards for students and teachers is emphasized in several findings across the three papers. For this district, only one of the three performance targets is informative about large proportions of students and teachers. Without reliable information about performance based on the other two targets, the potential benefits of using multiple performance outcomes to generate different types of measures about teachers cannot be assessed. Different types of IRT models with polytomous or continuous response variables would likely capture more of the variation in student performance than the single dichotomous outcome that was informative in this study. However, these models are not as easily-interpretable as the 2PL and may be less accessible to stakeholders. Many of the other advantages of SRT over VAMs relate to IRT-specific procedures that have been studied extensively within the context of dichotomous response models. Continuous response models, in particular, are rarely used in research or in practice, and the types of IRT technology that may be most beneficial to the evaluation context has not been developed for these types of models. Interpretability and application of well-studied IRT technology are focal points of this study, so the most appropriate resolution for these same purposes would be to seek different dichotomous targets to use in place of the higher targets. Like the state-determined proficiency levels, these targets should correspond to concrete standards of performance. However, they must be both reasonable and challenging for a significant proportion of students in the district.

Another common theme throughout these findings is mismatch between the average instructional demand level of a class and either the capacity of a teacher or the instructional demand level of an individual student. Large discrepancies between capacity and demand result in large standard errors of capacity and growth estimates. Discrepancies between individual students' instructional demand levels and class mean demand correspond to differences in performance compared to similar students with similar teachers but different classmates. One explanation is that these differences arise from interactions between students in the class, or peer effects. Another explanation is that teachers alter their instructional practices to target the types of students that comprise the majority of a class. Students who are either above or below the demand levels of most students in their classes may not benefit as much as they would in a class where instruction is targeted towards students like them. The first explanation could potentially be addressed with further development of the SRT model and instructional demand index. The second explanation would be relevant feedback for administrators and could prompt them to train teachers on instructional practices that benefit a wider range of students or to actively avoid placing students in classes where most students are far above or far below their instructional demand levels.

Although several findings point to possible differences in model performance for special education students and special education teachers, as well as for teachers with very low student counts, most of these relationships are better explained by discrepancies between the high demand levels in these classes and the capacity estimates of these teachers. These types of classes are typically associated with low student counts, and are also more likely to have students missing from the data, however, the results suggest that the capacity-demand gap makes a far greater difference than the number of students placed with a teacher.

APPENDIX

Table A1. Location and slope parameters for instructional demand indicators [5]

Indicators	**location**	**slope**	**Indicators (cont'd)**	**location**	**slope**
Not gifted (high achievement)	-2.62	1.42	C for listening effort	0.78	2.71
B in writing	-2.23	2.95	C for history effort	0.90	3.02
Not gifted (intellectual ability)	-2.19	1.23	C for science effort	1.00	2.95
B in science	-2.12	2.63	C in health	1.09	2.52
B in history	-2.06	2.96	C for speaking effort	1.14	2.30
Proficient in math	-2.04	1.66	Limited English proficiency	1.15	1.07
B in math	-2.04	2.56	C for health effort	1.39	2.63
B in phys ed	-2.03	1.52	D-F in math	1.57	2.56
B in health	-2.02	2.52	D-F in writing	1.68	2.95
B in reading	-2.00	2.72	D-F in art	1.74	1.91
B in art	-1.81	1.91	D-F in reading	1.75	2.72
B in speaking	-1.77	2.31	Student with IEP	1.90	1.30
Not designated as gifted	-1.68	1.45	Learning disability	1.92	1.29
Proficient in ELA	-1.67	1.76	C for art effort	1.94	1.86
B in listening	-1.58	2.92	One disability	1.94	1.26
B for writing effort	-1.41	2.91	C in phys ed	1.97	1.52
B for PE effort	-1.39	1.49	C for PE effort	2.09	1.49
B for health effort	-1.37	2.63	D-F in history	2.11	2.96
B for science effort	-1.35	2.95	D-F for math effort	2.14	2.91
B for history effort	-1.34	3.02	D-F for writing effort	2.14	2.91
B for speaking effort	-1.29	2.30	D-F for reading effort	2.20	3.06
B for reading effort	-1.25	3.06	D-F in science	2.22	2.63
B for listening effort	-1.25	2.71	D-F in listening	2.32	2.92
B for math effort	-1.21	2.91	D-F for listening effort	2.42	2.71
B for art effort	-1.17	1.86	D-F for history effort	2.44	3.02
Basic math proficiency	-0.90	1.66	D-F for science effort	2.56	2.95
Basic ELA proficiency	-0.76	1.76	D-F in speaking	2.67	2.31
C in writing	-0.23	2.95	D-F in health	2.67	2.52
C in math	-0.12	2.56	Two disabilities	2.86	1.26
C in reading	-0.01	2.72	Resource specialist	2.92	1.07
Below basic ELA proficiency	0.04	1.76	D for health effort	2.93	2.63
Absent 5-10 days	0.04	0.36	D for speaking effort	2.98	2.30
Below basic math proficiency	0.41	1.66	At least 3 disabilities	3.63	1.26
C in history	0.48	2.96	D-F in art	3.91	1.91
C in science	0.49	2.63	Absent 11-17 days	3.92	0.36
C in listening	0.55	2.92	D-F for art effort	4.13	1.86
C for writing effort	0.55	2.91	D-F for PE effort	4.58	1.49
C for reading effort	0.64	3.06	D-F in phys ed	4.60	1.52
C for math effort	0.66	2.91	Absent 18 or more days	7.12	0.36
C in speaking	0.77	2.31			

[5] IRT-calibrated instructional demand index with restricted item set (cohort 1).

Table A2. Deriving equations for the P25 and P75 measures (Chapter 3)

$$P_t(d) = \frac{\exp\left(a_{jt}(\theta_{jt} - d)\right)}{1 + \exp\left(a_{jt}(\theta_{jt} - d)\right)} \tag{A1}$$

$$P_t(d)[1 + \exp\left(a_{jt}(\theta_{jt} - d)\right)] = \exp\left(a_{jt}(\theta_{jt} - d)\right) \tag{A2}$$

$$P_t(d) + P_t(d) * \exp\left(a_{jt}(\theta_{jt} - d)\right) = \exp\left(a_{jt}(\theta_{jt} - d)\right) \tag{A3}$$

$$P_t(d) = \exp\left(a_{jt}(\theta_{jt} - d)\right) - P_t(d) * \exp\left(a_{jt}(\theta_{jt} - d)\right) \tag{A4}$$

$$P_t(d) = \exp\left(a_{jt}(\theta_{jt} - d)\right) * (1 - P_t(d)) \tag{A5}$$

$$\frac{P_t(d)}{1 - P_t(d)} = \exp\left(a_{jt}(\theta_{jt} - d)\right) \tag{A6}$$

$$\ln\left(\frac{P_t(d)}{1 - P_t(d)}\right) = a_{jt}(\theta_{jt} - d) \tag{A7}$$

$$d + \ln\left(\frac{P_t(d)}{1 - P_t(d)}\right)/a_{jt} = \theta_{jt} \tag{A8}$$

$$d = \theta_{jt} - \ln\left(\frac{P_t(d)}{1 - P_t(d)}\right)/a_{jt} \tag{A9}$$

$$\mathrm{P25} = \theta_{jt} - \ln\left(\frac{0.25}{1 - 0.25}\right)/a_{jt} = \ \theta_{jt} - \ln\left(\frac{0.25}{0.75}\right)/a_{jt} = \theta_{jt} - \ln\left(\frac{1}{3}\right)/a_{jt} \tag{A10}$$

$$\mathrm{P75} = \theta_{jt} - \ln\left(\frac{0.75}{1 - 0.75}\right)/a_{jt} = \ \theta_{jt} - \ln\left(\frac{0.75}{0.25}\right)/a_{jt} = \theta_{jt} - \ln(3)/a_{jt} \tag{A11}$$

BIBLIOGRAPHY

BIBLIOGRAPHY

Baker, B. D., Farrie, D., & Sciarra, D. G. (2016). Mind the gap: 20 years of progress and retrenchment in school funding and achievement gaps. *ETS Research Report Series*, *2016*(1), 1-37.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of educational and behavioral statistics*, *29*(1), 37-65.

Ballou, D. (2005). Value-added assessment: Lessons from Tennessee. *Value added models in education: Theory and applications*, 272-297.

Betebenner, D. W. (2011). A technical overview of the student growth percentile methodology: student growth percentiles and percentile growth projections / trajectories. Paper presented at *The National Center for the Improvement of Educational Assessment*. New Hampshire.

Booher-Jennings, J. (2005) "Below the bubble: "Educational triage" and the Texas accountability system." *American educational research journal* 42(2), 231- 268.

Braun, H. I. (2005). Using Student Progress to Evaluate Teachers: A Primer on Value- Added Models. Policy Information Perspective. *Educational Testing Service*.

Castellano, K. E., & McCaffrey, D. F. (2017). The accuracy of aggregate student growth percentiles as indicators of educator performance. *Educational Measurement: Issues and Practice* 36(1), 14-27.

Cella, D., Gershon, R., Lai, J. S., & Choi, S. (2007). The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research*, *16*(1), 133-141.

Chatterjee, S., & Hadi, A. S. (2015). *Regression analysis by example*. John Wiley & Sons.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

De Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory.* (T. D. Little, Ed.). New York: The Guilford Press.

Dee, T. S., Jacob, B., & Schwartz, N. L. (2013). The effects of NCLB on school resources and practices. *Educational Evaluation and Policy Analysis*, *35*(2), 252-279.

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. *Educational and psychological measurement*, *53*(1), 61-77.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). Principles and Practices of Test Score Equating. *Educational Testing Service*, Princeton, NJ.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis* (Vol. 3). New York: Wiley.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2013). Selecting growth measures for school and teacher evaluations. *National Center for Analysis of Longitudinal Data in Education Research (CALDER). Working Paper*, *80*.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Goe, L., Wylie, E. C., Bosso, D., & Olson, D. (2017). State of the states' teacher evaluation and support systems: A perspective from exemplary teachers. *ETS Research Report Series*, *2017*(1), 1-27.

Grissom, J. A., Kalogrides, D., & Loeb, S. (2013). Strategic Staffing: Examining the Class Assignments of Teachers and Students in Tested and Untested Grades and Subjects. *American Education Finance and Policy Conference,* New Orleans, LA.

Guarino, C., Reckase, M., Stacy, B., & Wooldridge, J. (2014). A comparison of Student Growth Percentile and Value-Added models of teacher performance. *Statistics and Public Policy,* 2(1): 1-11.

Halpin, B. (2016). Cluster analysis stopping rules in Stata. Working Paper WP2016-01, Department of Sociology, University of Limerick. https://osf.io/rjqe3.

Ham, E. H. (2014). Comparison between educator performance function-based and educator production function-based teacher effect estimation. Unpublished doctoral dissertation *Michigan State University*, East Lansing, MI.

Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory*. Newberry Park, CA: Sage.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review*, 267-271.

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? An examination of the statistical properties and policy alternatives. *Education*, *4*(4), 319-350.

Harris, D. N. (2011). *Value-Added Measures in Education: What Every Educator Needs to Know*. Harvard Education Press. 8 Story Street First Floor, Cambridge, MA 02138.

Harris, D. N., & Sass, T. R. (2006). Value-added models and the measurement of teacher quality. *Unpublished manuscript*.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. *Educational measurement*, *4*, 187-220.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. *Test validity*, 129-145.

Holland, P. W., & Wainer, H. (2012). *Differential item functioning*. Routledge.

Horoi, I., & Ost, B. (2015). Disruptive peers and the estimation of teacher value added. *Economics of Education Review*, *49*, 180-192.

Huynh, H., & Meyer, P. (2010). Use of robust z in detecting unstable items in item response theory models. *Practical Assessment, Research & Evaluation*, *15*(2), 1-8.

Kim, C. M., Frank, K. A., & Spillane, J. P. (2018). Relationships among Teachers' Formal and Informal Positions and Their Incoming Student Composition. *Teachers College Record*, *120*(3), n3.

Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function*. National Center on Performance Incentives, Vanderbilt, Peabody College.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. Springer Science & Business Media.

Ladd, H. F., & Lauen, D. L. (2010). Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and Management*, *29*(3), 426-450.

Lee, L. (2011). What Did the Teachers Think? Teachers' Responses to the Use of Value- Added Modeling as a Tool for Evaluating Teacher Effectiveness. *Journal of Urban Learning, Teaching, and Research*, *7*, 97-103.

Lissitz, B., & Doran, H. (2009). Modeling Growth for Accountability and Program Evaluation: An Introduction for Wisconsin Educators. *Wisconsin Department of Public Instruction*.

Lord, F. M. (2012). *Applications of item response theory to practical testing problems*. Routledge.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, *22*(4), 719-748.

Martineau, J. A. (2016). Introduction to and Preliminary Evaluation of Student Response Theory. Unpublished manuscript, *Center for Assessment,* Dover, NH.

Monk, D.H. (1987). Assigning Elementary Pupils to Their Teachers. *The Elementary School Journal. 88*(2), 166–87.

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, *92*(2), 263-283.

National Council on Teacher Quality. (2017). *State Teacher Evaluation Policy Yearbook: National Summary.* Washington, DC.

Paufler N.A. & Amrein-Beardsley, A. (2014). The Random Assignment of Students Into Elementary Classrooms: Implications for Value-Added Analyses and Interpretations. *American Education Research Journal*. *51*(2), 328-362.

Reckase, M. D. & Martineau, J. A. (2014). The Evaluation of Teachers and Schools Using the Educator Response Function (ERF) In Lissitz, R. W., *Value Added Modeling and Growth Modeling with Particular Application to Teacher and School Effectiveness.*

Rothstein, J. (2008). Teacher quality in educational production: Tracking, decay, and student achievement (No. w14442). *National Bureau of Economic Research*.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education*, *4*(4), 537-571.

Ryan, J., & Brockmann, F. (2009). A Practitioner's Introduction to Equating with Primers on Classical Test Theory and Item Response Theory. *Council of Chief State School Officers*.

Samejima, F. (1977a). A use of the information function in tailored testing. *Applied psychological measurement*, *1*(2), 233-247.

Samejima, F. (1977b). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika*, *42*(2), 193-198.

Samejima, F. (2016). Graded response models. In *Handbook of Item Response Theory, Volume One* (pp. 123-136). Chapman and Hall/CRC.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, *8*(3), 299–311.

Thum, Y. M. (2003). Measuring progress toward a goal: Estimating teacher productivity using a multivariate multilevel model for value-added analysis. *Sociological Methods & Research*, *32*(2), 153-207.

U.S. Department of Education (2009). Growth Models: Non-Regulatory Guidance. (January 12). *http://www2.ed.gov/policy/gen/guid/significant-guidance.doc*.

Walsh, E., & Isenberg, E. (2014). How does value added compare to student growth percentiles? *Statistics and Public Policy,* 2(1), 1-13

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, *36*(1), 1-28.
83%