# Beyond Traditional Measures of Teacher Quality:

# Incorporating Cultural Competence to Measure Classroom Community

By

Grant Van Eaton

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Learning, Teaching, and Diversity

December 16, 2017

Nashville, Tennessee

Approved:

Douglas B. Clark, Ph.D., Chair

Ilana Horn, Ph.D.

Barbara Stengel, Ph.D.

Sun-Joo Cho, Ph.D.

This dissertation is dedicated to all of the great teachers in my life, especially:

Darlene Rankin, Evelyn Saucier, Joy Killough, Chris Magnuson, and Mark Roche

ACKNOWLEDGEMENTS

Words cannot express the deep gratitude I hold for each and every mentor, colleague, teacher, student, and friend that has supported me to reach this milestone. Without your love, truth, and hugs I never would have made it this far.

First, none of this would have been possible without my mentor and champion, Doug Clark. What a journey we have completed together! Through all of the twists and turns you have been there for me, guided me, advocated for me, and given me the freedom to explore and find my own path, while also steering me back to safe ground when I went too far afield. Thank you for the dedication, support, and perseverance you have shown over these years.

Lani Horn, you never stopped pushing and you never would let me settle for mediocrity. I have learned so much about the craft of teaching, how students think, and how to look at the world of teaching with an objective, careful, critical, and respectful eye. I hope to take some of your wit, fierce advocacy for the profession of teaching, and attention to detail with me.

Barb Stengel, you have been a rock on stormy days. Not only did you widen my perspective on the world and myself, you did so with loving care and grace. I am a better academic for having the chance to learn with you, and a better person for having the chance to spend time with you.

Sun-Joo Cho, you opened my eyes to a new way of looking at the work we do as researchers and gave me the knowledge and tools to bring my ideas to life. In all that you do, you have shown that you care about the act of teaching just as much as you care about the act of researching. Thank you for all of the time and interest you have shown as I explored the world of measurement and the support as I sought to integrate that knowledge into my own craft.

In addition to a great committee, I've been blessed with an equally great group of family that picked me up on the dark days and were the first to celebrate on the best days. Mom, Dad, Ross, Grammy, and Nana, you all pretended to be interested in the beginning (when I highly doubt it was all that interesting), never stopped asking questions until one day it finally became interesting, and even more importantly – never stopped asking how the writing was going. And if it weren't for the unfailing, day-to-day support of Mark, Mary, Mia, Jonathan, and Adam there might not have been any writing to ask about! You all are my extended family and deserve as much of the credit as any for going on this journey with me.

Finally, I want to acknowledge all of the students I have had the honor of teaching, especially those at Paul Public Charter School. Whether we worked through the details of cellular respiration together, went on a journey through the Hunger Games, began to learn the fundamentals of teaching at summer institute, or enacted ambitious science teaching while balancing grad school and a full-time teaching career, I am a better person for having known and learned alongside each and every one of you.

Thank you all for being a part of my journey.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

# CHAPTER I

Over the last two decades, calls for greater public school accountability have dominated the political landscape. Federal intervention in the form of the No Child Left Behind Act (2001) and the Race to the Top competitive grant program, funded by the 2009 passage of the American Recovery and Reinvestment Act, increased accountability and oversight for the nation's K-12 public schools. In the wake of these federal programs, significant funds and attention were allocated to measuring teacher effectiveness, primarily through an increased focus on students' scores from state-administered standardized tests. In addition to pressuring states to include students' standardized test outcomes as a component of teachers' evaluation, these policy interventions pressed local districts and schools to redefine their teacher observation protocols and measures. In many districts, these observations took on greater significance in teacher evaluations and were primarily used as evaluative rather than formative tools.

Concurrently, with the adoption of the Common Core State Standards for Math and English Language Arts, researchers and districts took a renewed interest in deeply understanding the qualities of effective teachers in the context of the new, more rigorous standards. Across many disciplines, researchers partnered with districts and schools to better understand quality teaching in this new context, understand the broader social forces in schools and districts affecting the implementation of the new standards and teaching paradigms, and support teachers as they grappled with developing deeper content and pedagogical expertise. These collaborations led to the development of multiple frameworks and measures of quality teaching, some of which aligned with districts' accountability goals and others of which were solely focused on understanding and developing teachers' pedagogical expertise.

Looking back on the last two decades of measurement development, this dissertation seeks to better understand the current landscape of widely-adopted measures that focus on teacher quality, propose avenues for expanding the dominant construct of teacher quality as researchers and school districts design new measures, and assess the reliability and validity of a new measure of classroom community that includes aspects of cultural competence.

The remainder of Chapter I will provide a brief overview of Item Response Theory (IRT) models, the dominant statistical tool for creating and evaluating measures, in order to frame key measurement terminology and methodology for the remainder of the paper. Chapter II will then look across widely-adopted measures of teacher quality used both by schools and districts, as well as by the research community, to explore their implementation context, connections to the broader research literature, and placement of teacher and student actions. A second aim of Chapter II is to assess the dominant construct of teacher quality and its inclusion of cultural competence, with an eye toward gaps in the current construct.

Chapter III presents a new measure – the Vision for Student Learning – that seeks to shift classroom observation rubrics away from the traditional, evaluative measurement of teacher quality and toward formative measurement of classroom community. Furthermore, Chapter III benchmarks the inclusion of cultural competence in the Vision for Student Learning against the theoretical constructs described in Chapter II. After establishing the Vision for Student Learning as a measurement instrument that includes aspects of cultural competence, Chapter IV uses IRT to assess the validity and reliability of the measurement instrument. Chapter V concludes with key findings and recommendations.

## THE PROJECT OF MEASUREMENT

Researchers have classically defined measurement as the assignment of a number to an object or event according to a general set of rules (Stevens, 1946). In the context of this dissertation, the general aim of measurement is to assign a number that represents teachers' development toward leading a classroom that fosters a rigorous and equitable learning environment. In our daily lives, we measure a variety of directly observable things on a regular basis – the temperature outside, the distance from home to work, the speed at which we travel in the car while driving to work. Many characteristics that researchers attempt to measure in education, however, are not directly observable. Constructs like "teacher quality," "student achievement," and "school quality" do not lend themselves to direct measurement with tools like thermometers, yard sticks, and odometers. Attributes that are not directly observable are referred to as *latent constructs*. In order to measure latent constructs, researchers rely on collections of directly observable characteristics – such as the amount of time teachers wait between asking a question and providing the answer to students, or the number of students who raise their hand to answer a question – that they theorize serve as proxies for the latent construct they are attempting to measure. The constellation of observable characteristics that inform a broader understanding of latent constructs like "teacher quality," however, is under constant debate. This dissertation seeks to contribute to that debate through its analysis of current measures of teacher development, highlighting the theories of teacher development that different measures operationalize by virtue of the observable characteristics included in each measure.

One of the primary tools used by researchers interested in measuring and evaluating latent constructs is an *IRT model*, first conceptualized by Georg Rasch (1960). IRT models are used to explain the relationship between outcomes and a construct (such as teacher quality). One

way to evaluate the scale using an IRT model is through the Four Building Blocks approach developed by the Berkeley Evaluation and Assessment Research (BEAR) Center (Wilson, 2004). The four building blocks include *construct development, item design, defining the outcome space,* and *building a measurement model*.

In the Four Building Blocks approach, researchers must first *define the construct* which they seek to measure. This construct is a single latent characteristic that the measurement instrument is designed to measure, such as mathematical ability or quality of science teaching. In order to define the construct, researchers develop a construct map to serve as a visual representation of the construct. In a traditional IRT model, a construct must be unidimensional – that is, it can only represent a single latent variable. The construct is defined along a continuum, similar to a number line, with two variables ordered along the continuum: the first variable is the placement of individuals along the continuum from low to high levels of ability with regard to the construct (such a low or high levels of teaching quality), and the second variable is the items that make up the instrument, ordered by the level of the construct each item is most efficient at measuring. Appendix A provides an example of a construct map for quality of science classroom discourse.

The second building block is *item design*. After defining the construct to be measured and creating a construct map, researchers must design items that are theorized to relate to the construct map. Items can take various forms: multiple choice questions, Likert scales, or indicators that make up rows in a teacher observation rubric. A person's responses to items allow the IRT model to infer a relationship between the person and the construct, ultimately resulting in a ranking of persons along the construct map in relation to the items associated with the construct. In an IRT model, causality is inferred between outcomes (such as a rubric score)

and the construct – we assume that the construct causes individuals' responses to items. In the context of this analysis, items are different rubric rows from a classroom observation instrument. Each rubric row contains an observable action or trait that is theorized to have a relationship to the latent construct of teacher quality. If a person has higher teaching ability, then their observable actions while teaching will cause them to score higher on a rubric row item.

The third building block is to establish an *outcome space*. In defining an outcome space, researchers make decisions with regard to categorizing observations and then score them as indicators of the construct. In our analysis, the outcome space is an observation rubric in which observable teacher and student actions are translated into ordered levels to determine a score for that item/rubric row.

Finally, researchers must define the *measurement model*. The measurement model does the work of translating scored responses to the items (such as rubric scores) to locations on the construct map. IRT models specify not only how a person's responses to an item are related to the overall construct, but also how the item's own properties relate to varying levels of ability with regard to the construct. In other words, IRT models not only show the relationship between a teacher's observation scores and their teaching ability, but also how the properties of each indicator on the rubric relate to varying levels of teaching ability. These properties include an item's location (how difficult it is to receive a higher rating on that item) and an item's discrimination (how well the item differentiates between persons along the continuum of the latent construct).

A key component of IRT models is the ability to assess the validity and reliability of a measurement instrument. A reliable measure consistently measures a construct with the same level of confidence each time the instrument is used. IRT takes a nuanced approach to

establishing reliability through the use of test and item information, which can vary across the range of test scores. In other words, IRT models recognize that the precision of a measurement instrument is not uniform across all of the possible scores, with the scores possibly becoming less reliable at the very low and very high ends. Test and item information functions visualize the reliability of the instrument across the range of possible scores, rather than having a single reliability coefficient that speaks to the entire instrument. When looking at a test information function (Figure 1), we can see that the measure is most reliable (provides the most information) where the function peaks. In an ideal world, a measure would provide equal information across the entire range of possible scores (the x-axis). Figure 1 is a typical test information function, however, in that most measures are most reliable around the median score and become less reliable at the tails.

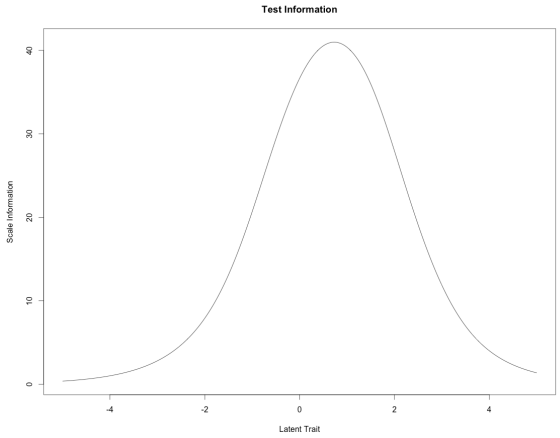



*Figure 1. Test Information Function*

IRT models also offer more sophisticated analyses of an instrument's validity, a test of whether a measurement instrument's items accurately correspond with the underlying latent construct when enacted in the real word. Validity, construed as how well an item fits or is able to contribute to placing a person along the continuum of the latent construct, can be tested at the

individual item level in IRT models, allowing researchers and developers to test an instrument using empirical data and then revise the instrument based on the individual fit of each item.

Each of the widely-adopted measures analyzed in this paper were designed and validated within the IRT modeling framework, with the aim of placing teachers on a continuum of teacher quality as a result of scores from items on a classroom observation rubric or questionnaire. While each of the measures analyzed in the subsequent sections seek to measure some aspect of teacher quality, each chooses to operationalize the latent construct of teacher quality through a different set of items on a rubric or survey. A basic understanding of IRT models will allow us to interrogate how the design characteristics from which measures of teacher quality either converge or diverge have implications for how teacher quality is defined and measured in current systems, as well as open the door for opportunities to expand the dominant theory of what determines quality teaching and what additional items would be necessary in order to measure an evolving construct of teacher quality.

## CHAPTER II

As politicians enacted No Child Left Behind and states began to draft the Common Core State Standards, academics and teachers concurrently grappled with the intersections between issues of race, class, accountability, and quality instruction in public schools. These lines of inquiry were not novel, but garnered renewed interest and urgency as No Child Left Behind highlighted achievement gaps and increased school systems' accountability toward the learning of diverse student populations.

Many researchers have sought to illuminate the interplay between race, class, and students' learning (Lee & Wong, 2004; Rist, 1970; Tate, 1997). Foundational studies sought to make visible the ways in which systems of power and entrenched social structures resulted in patterns of inequitable access to education based on students' race and class (Delpit, 1988, 1995; Oakes, 1983). These studies often look at how teachers reified social constructions of race and class in their classrooms. Feistritzer (2011) documented that public school teachers in the United States were overwhelmingly white, female, and from middle-class backgrounds. In contrast, KewalRamani, et al (2007) found that students in public schools were generally balanced with regard to gender and increasingly non-white. In this context, researchers observed biases where divergence existed in race and social class between teachers and students. Compared with teachers who share the same background as their students, white teachers from middle class backgrounds were found to discriminate against students who came from lower income backgrounds (Alexander et al., 1987). Furthermore, Ready and Wright (2010) found that literacy teachers overestimated the abilities of their white students and underestimated the abilities of Latino students. Similarly, African American students' classroom grades were found to be lower

than expected based on what their 10th grade state exams would predict with White students' classroom grades found to be higher than state exam scores would predict (Wildhagen, 2012).

Throughout the remainder of this dissertation, *culture* will be a chief focus. The term culture can take on many different meanings both within and outside of educational contexts. Three modifiers of culture are prevalent: *classroom* culture, culture *for learning*, and *students'* culture. Classroom culture refers to the habits, dispositions, norms, routines, student-student interactions, and teacher-student interactions that occur within the classroom setting. Culture for learning is a restrictive subset of classroom culture; in this context, culture for learning addresses classroom culture insofar as it promotes a classroom environment that supports students to engage with academic content. Finally, students' culture refers to the cultural patterns of students outside of school and while at home, and frequently correspond with students' race, language, and class. Classroom culture is typically defined by the norms of the white middle class, frequently resulting in a cultural mismatch for students who have different racial, ethnic, linguistic, and class backgrounds.

*Building Teachers' Cultural Competence.* Given the ways in which the intersections of race, class, teaching, and learning play out in the classroom, several researchers have proposed frameworks for supporting students and teachers through culturally responsive teaching and pedagogy (Banks, 1991, 1992; Gay, 2000; Ladson-Billings, 1994, 1995; Lee, 2007).

One of the most prominent examples is Gloria Ladson-Billings' framework for culturally relevant pedagogy (1994, 1995); Ladson-Billings' framework is not only the most prominent, but also one of the most widespread and adopted frameworks across the country, taught throughout teacher preparation programs and used as the focus of multi-day professional development

workshops. It is also the core framework provided to teachers by schools and districts to improve learning for diverse student populations. Culturally relevant pedagogy seeks to disrupt systems of inequity in schooling through the inclusion of students' cultural norms in classroom instruction, while also affirming students' diverse cultural backgrounds (Ladson-Billings, 1995; Gay, 2000). Ladson-Billings and Gay argue that teachers' cultural knowledge plays a significant role in designing and enacting culturally-responsive lessons, shifting deficit-based views of race and culture toward an asset-based, inclusive paradigm. In Ladson-Billings' framework, there are three components of culturally-responsive pedagogy:

1. Academic Achievement: teachers hold high academic expectations that simultaneously meet students where they are.
2. Cultural Competence: students understand their cultural background and learn about the cultural background of others.
3. Socio-political Competence: students develop a liberatory view of education that empowers them to challenge the status quo and current social order.

In addition to Ladson-Billings' culturally-responsive pedagogy, other models have sought to elevate the role that knowledge and incorporation of students' culture plays in fostering equitable and rigorous education. Another prominent model is Carol Lee's (2007) Cultural Modeling Framework. In her framework, Lee endeavors to make explicit connections between rigorous subject-matter learning and students' cultural knowledge, further augmented by the knowledge students bring to the classroom from their everyday interactions with the world. Lee argues that for many students, culture is a resource for learning, a position corroborated through research showing that students bring a wide array of cultural and linguistic resources to learning (Adler, 2001; Ball et al., 2003; Cobb & Hodge, 2002). Lee also insists that in order to make connections

between subject-matter content and students' culture visible for students, teachers must have deep knowledge of their content domain. In order to make content relevant to students' own lived experiences, teachers must be able to build connections between students' culture and the content to be learned. Only through intimate knowledge of both their students' backgrounds and their content domain can teachers make these rigorous connections.

Banks' Multicultural Education Framework (1991, 1992) is another widespread model for culturally competent instruction that aims to create an equitable learning environment for students from diverse backgrounds and identities. The Multicultural Education Framework consists of five dimensions: content integration, knowledge construction, prejudice reduction, equity pedagogy, and an empowering school culture and social structure. Similar to Ladson-Billings and Lee, Banks' framework focuses on the extent to which students' diverse cultures are incorporated into the educational setting. Each of the five dimensions makes explicit connections between culture and praxis. The domain of content integration is presented as the extent to which teachers use non-dominant and diverse cultures to illustrate key disciplinary concepts and theories. The knowledge construction domain explicitly calls out the role that race, ethnicity, and class influence the construction of knowledge. Banks' conceives equity pedagogy as the set of skills teachers need in order to promote the academic achievement of students from diverse backgrounds. Prejudice reduction and an empowering school culture also seek to make schools places where children can not only attain a quality education but also learn within a school environment that promotes and affirms their own cultures and the values of others.

Given that cultural knowledge plays an important role in designing and enacting equitable teaching (Banks, 2005; Gay, 2000; Irvine, 2003; Ladson-Billings, 1994; Nieto, 2000), and that a disjunct exists between the racial and cultural backgrounds of the teaching force and

students in public schools, research suggests that teacher education programs and in-service professional development should foster the development of teachers' cultural competence. The next section will analyze seven measures of teacher quality that currently enjoy widespread adoption, have been psychometrically validated, and have been used in multiple research studies as a measure of teacher quality. The subsequent analysis will provide a lens into the dominant construct of teacher quality employed by districts, schools, and researchers, during which measures of teacher quality will also be assessed with regard to their inclusion of students' culture as a feature of quality instruction.

## THE CONSTRUCT OF TEACHER QUALITY AS EVIDENCED IN WIDELY-ADOPTED MEASURES OF TEACHER QUALITY

Across the nation, there are a number of measures of teacher quality that are widely used by districts, schools, and academic researchers. How "teacher quality" is operationalized in each measure, however, differs with regard to the goals that influenced the development of each measurement tool. In some contexts, teacher quality is defined primarily as pedagogical ability; in other tools, teacher quality takes on broader scope, including not only teachers' pedagogical skills but also classroom management and culture. The designers of different measurement tools also operate within different contexts. Many widely-adopted measures were designed by researchers specifically looking to measure teachers' development of pedagogical skills aligned with the new Common Core and Next Generation standards. Others were designed by districts to measure the effectiveness of their teacher workforce vis-à-vis the accountability provisions of No Child Left Behind.

This diversity of purpose and aim with regard to measurement tool design muddles the teacher quality landscape and is a direct reflection of the lack of a shared understanding in the United States with regard to the purpose and goals of public education. This section attempts to clarify select aspects of measures that currently enjoy widespread adoption, paying particular attention to different measurement instruments' implementation context; whether a measure can be used for formative purposes; a measure's connection to the broader research literature; and placement of student and teacher actions within a measure's indicators. Measures are also assessed insofar as they include cultural competence as part of their construct of teacher quality.

A variety of measures designed by both researchers and practitioners will be analyzed with regard to these features, spanning the breadth of content-specific tools, such as the Mathematical Quality of Instruction (MQI) measure, as well as broader, content-neutral measures like The Danielson Framework. The majority of measures used in this analysis were included in the Gates Foundation's Measures of Effective Teaching (MET) study, supplemented with additional measures widely used in practice by districts and schools. Measures used in the MET study were chosen as exemplars given the significant research and design work that preceded the use of the measures by the MET study, the high rate of adoption of these measures nationwide, and the diversity of measures. These measures are not meant to be exhaustive in scope, but rather representative of the diversity of measures that are most widely used in current practice by researchers and practitioners. It is also important to note that the present analysis focuses solely on measures used to assess the quality of teachers who have full-time classroom duties and responsibilities; measures designed to gauge the learning progress of pre-service teachers are not within the scope of this analysis. A full list of measures included in this analysis can be found in Table 1.

*Table 1. Current Measures of Teacher Development for Analysis*

| Measure | Primary Publication | Primary Users | Overview (as stated in the primary publication) |
|---|---|---|---|
| The Danielson Framework for Teaching | Danielson, C. (2011). | Districts and individual schools | The Framework for Teaching is a research-based set of components of instruction, aligned to the INTASC standards, and grounded in a constructivist view of learning and teaching. |
| Classroom Assessment Scoring System (CLASS) | La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). | Districts and individual schools | The Classroom Assessment Scoring System (CLASS) was developed to identify observable teacher-student interactions, to determine which interactions are effective in driving better developmental and academic student outcomes, and to support teachers as they improve their teaching practices. |
| Mathematical Quality of Instruction (MQI) | Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). | Academic researchers | The Mathematical Quality of Instruction (MQI) instrument is designed to provide scores for teachers on important dimensions of classroom mathematics instruction. The MQI is based on the perspective that the mathematical work that occurs in classrooms is distinct from classroom climate, pedagogical style, or the deployment of generic instructional strategies. |
| Instructional Quality Assessment (IQA) | Junker, B. W., Weisberg, Y., Matsumura, L. C., Crosson, A., Wolf, M., Levison, A., & Resnick, L. (2005). | Academic researchers | The IQA was conceptualized around a specific set of guidelines for instructional practice which integrates strong pedagogical knowledge with deeply rigorous subject matter knowledge called the *Principles of Learning.* The four principles included in the IQA include academic rigor, clear expectations, self-management of learning, and accountable talk. |
| Protocol for Language Arts Teaching Observations (PLATO) | Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). | Academic researchers | The Protocol for Language Arts Teaching Observations (PLATO) is a classroom observation protocol designed to capture features of English/Language Arts (ELA) instruction. It was originally developed for a study of the relationship between teachers' classroom practices and their impact on student achievement. It is currently being used as a professional development tool to support teachers' use of rigorous, research-based teaching practices. |

| | | | |
|---|---|---|---|
| Quality of Science Teaching (QST) | Schultz, S. E., & Pecheone, R. L. (2014). | Academic researchers | The Quality Science Teaching (QST) instrument is an evidence-based observation instrument representing the Next Generation Science Standards (NGSS). The original QST measures six domains of science practices; QST- MET, a condensed version of the original instrument, requires raters to gather evidence on biology teachers' instruction from videotaped lessons and use an analytical rubric to rate qualities of effective science teaching practices that promote student learning. |
| Tripod Student Surveys | Kane, T. J., & Staiger, D. O. (2012). | Districts and individual schools; academic researchers | The Tripod Project uses surveys to understand student perspectives on teaching practices, classroom learning conditions, and student engagement. Students complete Tripod surveys at the classroom level to report on key dimensions of their experiences. The 7C's of Effective Teaching measured by Tripod include care, confer, captivate, clarify, consolidate, challenge, and control. |

*Implementation Context.* Widely-adopted measures of teacher quality are primarily used by two groups of users: researchers in academic settings or administrators and practitioners in school and district settings. As a result of different users' context, the ease of use for a particular measure can vary. The measures included for this analysis broadly fall into three groups: measures intended for use during in-person classroom observations, measures intended for use with video recordings of teachers' instruction, and measures used to gauge students' perception of their teacher's quality. As a result of their format and intended user, the ease of use of each measure varies considerably.

The first broad group, *measures used during in-person classroom observation*, includes The Danielson Framework for Teaching, the Classroom Assessment Scoring System (CLASS), and Instructional Quality Assessment (IQA) measures. The ease of use of these measures is primarily dependent on the complexity of each measures' structure and the depth of training observers must go through in order to arrive at valid and reliable scores using each measure.

The Danielson Framework requires the observer to rate teacher actions across four domains, with each domain having five to six indicators to score during the observation. This amounts to 22 different indicators for observers to score on a scale of 1-4, with each level having precise descriptors to achieve a score for that level. These indicators are content-neutral; while the indicators do address teachers' pedagogical skill and content knowledge, indicators specific to individual disciplines are not present in the Framework, nor are any indicators related to cultural competence. To use the entire Framework takes considerable amount of time for both novice and more expert observers, resulting in many districts hiring personnel with the sole job of observing classrooms and scoring them using the Framework. The Danielson Group also

offers an extensive list of workshops and professional coaching experiences to assist observers in gaining proficiency with the observation tool.

The CLASS measure has a similar format, with three domains consisting of three to five indicators each. Three versions of the CLASS measure exist for the K-12 environment: a K-3 measure, an upper elementary measure, and a secondary measure. The K-3 measure has a total of 10 indicators, with the upper elementary and secondary measures having 12 indicators each. While the CLASS tool has fewer indicators that observers must pay attention to, the addition of different versions for different age groups requires additional specialization on the part of the observer. Like The Danielson Framework, the CLASS measures are content-neutral and do not contain specific rubrics for each content area, but rather general indicators for pedagogical skill and content knowledge; no indicators for cultural competence are present. Similar to The Danielson Group, Teachstone (the firm which markets the CLASS measure) offers extensive professional development opportunities to learn how to best use the CLASS measure and support teachers' growth across the various domains included in the measure.

Finally, the Instructional Quality Assessment (IQA) is a measure designed by researchers at the Learning Research & Development Center (LRDC) at the University of Pittsburgh. Consistent with the structure of The Danielson Framework and CLASS measures, the IQA measure is organized into three domains that include approximately 20 indicators within those domains. The IQA is not content-neural, however, and has developed pilot measures to assess reading comprehension and mathematics instruction at the lower and upper elementary levels. Content-specific pedagogies are deeply integrated into the IQA measures, requiring indicators focused on academic rigor to differ significantly for each content area. The implementation of the IQA is more complex than standard implementations of The Danielson Framework and

CLASS. The IQA is administered in three parts: 1) teachers fill out a pre-visit questionnaire and compile a portfolio of assignments that will be given to students during the lesson to be observed, 2) during the classroom observation, raters look for specific evidence aligned with the IQA rubrics and conduct short student interviews, as well as briefly interview the teacher after the lesson, and 3) assignment portfolios are assessed using IQA rubrics relating to academic rigor and students' self-management of learning. This three-step administration is highly thorough and rigorous, but also requires a considerable time investment and content expertise on the part of the observer. An extensive ecosystem of professional development resources also exists for the IQA, consisting of rubrics, observation protocols, and scoring guidelines. The LRDC, however, does not provide the same level of support through workshops and professional development as The Danielson Group and Teachstone.

In summary, measures focused on scoring in-person classroom observations are vast tools with one to two dozen indicators that are scored during a classroom observation. Each of the three measures are considered complex enough to justify extensive professional development opportunities to train district and school officials to use the tools, with large manuals of descriptors for each scoring level of each indicator. As a result of these complexities, considerable district and school resources are used to train administrators and teachers in the use of the tools. The content-neural nature of The Danielson Framework and CLASS provides both advantages and disadvantages for ease of use: observation scores can be compared across content area and a single observer could theoretically observe teachers teaching a variety of content areas using the same rubric. Since The Danielson Framework and CLASS do not provide specific guidance for what constitutes strong pedagogy and content knowledge in each discipline, however, observers must already possess this expertise and have a shared definition of what

excellence in this content area looks like across the observer cohort in order for scores to be meaningful for teacher development and reliable across content areas. The IQA addresses this challenge by developing rubrics specific to each content area with regard to pedagogy and content knowledge. These additions, however, increase the complexity of the tool and the expertise-demands on observers.

A second group of measures are *measures used to analyze video of classroom teaching*. These measures are largely used by academic researchers who analyze video observations, collected by research assistants or filmed by teachers themselves, to assess the quality of teaching demonstrated in the video recording. The three content-specific measures used in the MET study (MQI, PLATO, and QST) all fall within this category.

Measures based on the analysis of video have unique structural strengths and challenges. Similar to the measures focused on in-person observations, these tools are complex in structure and require considerable training for observers to achieve proficiency at reliably scoring classrooms. They do, however, offer observers the affordance of being able to review the same episode of teaching multiple times if they are unsure how to score a classroom on a particular indicator, which is impossible during a live, in-person observation.

Rubrics reliant on video to observe classrooms also pose unique demands on infrastructure. Whereas in-person observations merely require observers to find their way to classrooms and have a centralized database for recording scores, video observations require the presence of recording technology at each school site, sufficient video recorders to accommodate all teachers to record lessons within certain time periods, network connections and data storage to transfer videos so that observers can access them, as well as an interface to observe, score, and leave feedback on videos submitted for observation. In addition to the technical infrastructure,

these systems require significant investment in training teachers and/or support staff to capture salient aspects of instruction on the video, as well as submit the videos for scoring and access feedback. While these infrastructure demands are feasible for large-scale, grant-funded university research teams, these demands would be prohibitive at-scale for many districts and schools. If teachers are required to record and submit their own videos, this is a significant increase in time and training required to be observed as opposed to observations conducted in-person in teachers' classrooms.

A final group of measures consists of *measures used to gauge students' perception of teachers' abilities*. These measures are unique in that they use students' perceptions of their teacher as a proxy for teacher quality. In this analysis, the Tripod Student Survey developed at Harvard serves as an exemplar. While many districts and researchers have designed their own student surveys to fit their local context, many student surveys use some or all of the Tripod Student Survey questions as their core.

Administering the surveys requires significantly fewer resources than the administration of classroom observations, both for teachers and administrators. Typically, the primary infrastructure demands consist of distributing surveys to teachers, having students fill out the survey during class, and then collecting the surveys to be sent for analysis. Most student surveys have multiple forms for different age groups and are able to be scored automatically, requiring little additional technical knowledge on behalf of teachers and administrators. While survey authors must work diligently in advance to ensure that questions are understandable by students and do not have multiple interpretations, this work happens on the forefront and eliminates the need for multiple observers to norm around their interpretation of different rubric items on a classroom observation rubric. Most student survey analysis protocols, however, do require basic

data analysis skills of teachers and administrators as they interpret the results for their local context and individual classrooms. Student survey execution demands are significantly lower than classroom observations, however, making them easier to implement at scale than classroom observations.

*Formative Nature.* In addition to implementation context and ease-of-use, widely-adopted measures of teacher quality differ with regard to their utility as a formative tool for teacher development. Depending on their implementation, measures can span the continuum from formative to evaluative. Formative measures are primarily used to help teachers and administrators learn and improve. When used formatively, measures of teacher development provide concrete feedback to teachers on the current quality of their practice, as well as clear guidance as to what actions to take in order to improve practice. In contrast, evaluative measures are primarily used to determine whether a teacher is 'effective' or 'not effective' and often to make high-stakes decisions with regard to teacher retention. When used for evaluation, measures of teacher development usually have a cutoff score that determines whether or not a teacher is effective. Evaluative measures are rarely used to help teachers learn and improve, often providing little to no feedback to teachers other than their score, but rather are used to determine which teachers should be retained by a school system or let go.

In practice, each of the seven measures could be used as either formative or evaluative tools depending on their implementation; the CLASS measure even goes as far to explicitly state a goal of "support[ing] teachers as they improve their teaching practices." The Danielson Group takes a slightly different approaching, highlighting that The Danielson Framework not only has

utility as a formative tool to promote teacher growth and development, but also as an evaluative tool:

> The Framework may be used for many purposes, but its full value is realized as the foundation for professional conversations among practitioners as they seek to enhance their skill in the complex task of teaching. The Framework may be used as the foundation of a school or district's mentoring, coaching, professional development, and teacher evaluation processes, thus linking all those activities together and helping teachers become more thoughtful practitioners.

The language used by measurement tools to label rubric scoring categories also has an impact on whether the tool is perceived as formative or evaluative in nature. The most evaluative language used by measures included in this analysis explicitly refers to teacher effectiveness. The Danielson Framework, for example, has four scoring categories for each indicator within a domain: Ineffective, Developing, Effective, and Highly Effective. These scoring categories explicitly use the evaluative language of effectiveness, lending itself to be enacted in a system that prioritizes teacher evaluation over growth.

The next group of measures use numerical labels for their scoring categories. While the number of categories can shift from measure to measure, with the CLASS, IQA, and QST measures having four categories and the PLATO measure having seven, these scoring categories still imply a ranked order. One could argue, however, that measures with more categories present a finer gradient of teacher practice, and therefore have greater potential for being viewed as formative, as opposed to simply placing teachers in quartiles. By defining more points along the teacher quality continuum, measures with greater number of categories not only provide a more nuanced progression for teachers to follow in their growth, but also contribute to the measure's ability to discern change in quality over time as teachers continue to use the measure to chart their growth.

A final group eschews effectiveness and numerical labels for proficiency labels. The MQI only has three scoring groups for each indicator, labeling them Low, Medium, and High. While these scoring groups could easily be viewed from an evaluative lens, they can also flip into the formative paradigm by focusing on the level of skill development within each category and the framing of moving from Low to High skill over time. The Tripod Student Survey, while focused on measuring student perception of teacher quality and not teacher quality through direct observation, also favors proficiency labels, using a five-point Likert scale of Totally True, Mostly True, Somewhat, Mostly Untrue, and Totally Untrue in its survey instrument. While proficiency labels are often translated into numerical rankings for the purpose of quantitative analysis, they provide an alternative frame with which to view the developmental continuum that does not explicitly use effectiveness language or numerical ranking systems to convey scoring categories on a rubric.

Each of the seven measures analyzed has the potential to be used as either a formative or evaluative tool. Certainly, a district or school could simply use any of these measures once for evaluative purposes. The design of each measure, however, positions each to be used repetitively over time to inform the continued development of a teacher's practice over time. While the language each measure uses to define scoring categories and the number of scoring categories contribute to whether the tool lends itself to provide new feedback to teachers when used repeatedly over time, each possesses the potential to push teachers to reflect on their practice and develop over multiple observations.

*Connections to the Broader Research Literature*. In seeking to understand the current landscape of teacher quality measures, it is also important to consider how these measures are used by

academic researchers to develop new theories of teacher development and support teacher growth. To determine how measures were taken up by the research community peer-reviewed studies linked from measures' main website, as well as from academic research database searches for studies utilizing each measure, were analyzed. These searches yielded 33 studies across the seven measures, which were then categorized by research goal and outcome of interest, yielding four analytic categories: *design and implementation studies, psychometric validity and reliability studies, correlational studies with other measures,* and *studies that used the measures to answer a novel research question*.

The largest analytic category, encompassing 12 of the 33 studies (see Table 2), focused on the *design and implementation* of a particular measure. Each of the studies had a least one peer-reviewed study focused on its design and implementation, with the Danielson Framework for Teaching and CLASS having the most, at three each. These studies largely focused on the design rationale for a particular measure, provided copies of associated frameworks, rubrics, and coaching tools, as well as evidence from pilot studies as to a given measure's effectiveness.

*Table 2. Design and Implementation Studies for Each Measure*

| | |
|---|---|
| Danielson Framework | Danielson, 2008, 2011, 2013 |
| CLASS | Pianta et al, 2008 |
| IQA | Matsumura, 2006; Boston, 2012; Junker, 2005 |
| MQI | Hill et al, 2008; Learning Mathematics for Teaching Project, 2011 |
| QST | Schultz & Pecheone, 2014 |
| PLATO | Grossman et al, 2013 |
| Tripod Student Surveys | Ferguson, 2008 |

A closely related companion category is studies assessing the *psychometric validity and reliability* of a given measure, in which nine studies fit, with two overlaps from the design and implementation category. Five of the seven measures had peer-reviewed validity and reliability analyses: Classroom Assessment Scoring System (Downer, 2010; Hamre et al, 2013; LaPro, 2004; Pakarinen et al, 2010), Mathematical Quality of Instruction (Hill, 2012), Quality of

Science Teaching (Schultz & Pecheone, 2014), Protocol for Language Arts Teaching Observations (Cohen & Grossman, 2016; Lazarev et al, 2013), and Tripod Student Surveys (Polikoff, 2014).

The next largest category, incorporating six studies, focuses on *establishing correlations between measures of teacher development and other measures*. These studies largely fit into two-subcategories: those establishing correlations with teacher value-added measures (Balch, 2012; Grossman, 2014; Raudenbush & Jean, 2014) and those establishing correlations with student achievement, primarily through the use of standardized exam scores (Milanowski, 2004; Kimball et al, 2009; Allen, 2013). One study also examined correlations between higher scores on the Danielson Framework and increased Tripod Student Survey scores (Ferguson & Danielson, 2014). None of these studies sought to establish correlations with measures of cultural competence or critical consciousness. Four of the measures in this analysis had correlational studies associated with them: the Danielson Framework, CLASS, PLATO, and Tripod Student Surveys.

Finally, five studies sought to use measures of teacher development to *answer novel research questions*. These studies used measures of teacher development in a variety of ways to support emerging research. Associated research domains included validation of teacher learning from professional development experiences (Cohen et al, 2016), efficacy of traditionally- versus emergency-credentialed teachers (Nougaret, 2005), applicability of the CLASS measure in diverse settings (Downer et al, 2012), development of new measures (Benjamin, 2002), and validating conceptual models of teacher performance and intellectual development (Song, 2006).

*Cultural Competence and the Location of Teacher versus Student Actions*. The final analytic category, and the most salient from the perspective of this analysis, focuses on where students are found in each measure and the degree to which cultural competence is included as a construct measured by the instrument. For this analysis, each indicator from the seven measures was categorized based on whether its focus was on a teacher action or student action. Subsequently, indicators with student actions were coded based on the type of student action the measure sought to capture through its observation rubric. There are 91 total indicators across the seven measures, 73 of which explicitly focused on teacher actions. The remaining 18 indicators focused on student actions, resulting in 80.2% of indicators focused on teachers and 19.8% focused on students. The range of indicators focused on students is wide, spanning 0% to 60% of each individual measure, with the average across measures falling at 18%. Two measures (CLASS and QST) had zero indicators explicitly focused on student actions, with only one measure (IQA) having a majority of indicators focused on student actions, comprising 60% of the measure's indicators.

At face value, these descriptive statistics are not necessarily surprising. The measurement instruments are teacher observation rubrics, primarily focused on measuring teacher quality through constructs focused on instructional quality and classroom management. It is clear that the current construct of teacher quality heavily favors observable teacher actions to serve as proxies for the construct. Interestingly, however, when the gaze of the instruments turns to students there is no clear consensus across measures as to which student actions contribute to teacher quality. Six distinct categories of student action-focused indicators emerged from the indicator analysis. Of the emergent categories of student action-focused indicators, one category had indicators from three measures, three categories had indicators from two measures, and the

remaining two categories were only present in one measure. The emergent categories are: *content/task-based indicators; whole-class discourse; questioning, explanations, and classroom culture;* and *student self-monitoring*. Zero student-focused indicators focused on students' developing their own cultural competence or critical consciousness.

The category with the broadest consensus across measures, while still not garnering coverage from the majority, was student engagement with authentic, content-specific tasks (IQA, MQI, and PLATO). That this category would have the broadest support is not surprising – four of the seven measures included in this analysis are content-focused measures that specifically look for evidence that students are engaged in rich, content-centered instruction. While they make explicit connections between content and pedagogy, these indicators do not take the extra step of linking content, pedagogy, and incorporation of students' cultural backgrounds and identities. Additionally, while the remaining four measures do speak to task design at the teacher level, they do not ask observers to measure students' engagement with quality tasks.

Three categories of student action-focused indicators were present in at least two measures: student participation in whole-class discussion (IQA and MQI), explanation-rich student discourse (IQA and MQI), and students' participation in class discussion through asking questions (The Danielson Framework and MQI). Discourse-focused indicators centered on student participation and use of evidence and did not assess for the use of discourse practices to promote equitable engagement in instruction nor as a means to incorporate students' unique cultural and linguistic identities in instruction.

Finally, the remaining distinct categories were students' contributions to classroom culture and students' self-monitoring and assessment through peer and teacher discourse (both in The Danielson Framework). It is worth noting that The Danielson Framework is the only

measure that specifically calls out students' role in fostering a classroom culture based on respect and a culture of learning. While many of the content-focused measures contain teacher action-indicators focused on classroom culture, none specifically focused on students' contributions to classroom culture. It is also of note that indicators focused on classroom culture specifically focused on promoting a 'culture of learning' that allows for students to be curious and learn from failure. Classroom culture indicators did not, however, incorporate respect for students' diverse cultural and linguistic backgrounds, nor the creation of a space to learn about and respect the cultural identities of others.

That these measures lack any mention of students' backgrounds and cultures, nor push teachers to examine and develop their own cultural competence, is shocking. Existing, widely-adopted measures of teacher quality fail to address Ladson-Billings' categories of cultural competence and socio-political consciousness, ignore the inclusion of students' culture as a resource for learning at the center of Lee's Cultural Modeling Framework, and are silent on the push for equity pedagogies and classroom cultures that reduce prejudice as featured in Banks' Multicultural Framework. As already discussed, teachers' cultural competence impacts the ways in which teachers perceive students, as well as student outcomes in their classrooms. Inclusion of cultural competence in measurement systems is not impossible, however, as demonstrated by the medical education and social science literature discussed in the next section.

*Measures of Cultural Competence.* During the same time period that the seven widely-adopted measures of teacher quality were being developed, tested, and implemented, numerous measures of cultural competence were also in development – almost exclusively by medical and social services professionals. In their review of measures of cultural competence used by healthcare

professionals, Kumas-Tan et al. (2007) looked across 54 measurement instruments for cultural competence and identified the 10 most widely used. They found that culture is conceptualized primarily as a matter of ethnicity and race, and that culture is conceived as possessed by the Other. Measures in their review conceptualized the construct of cultural competence with the assumption that cultural *incompetence* lies in a practitioner's lack of familiarity with the Other. Therefore, the process of becoming culturally competent involves (predominantly white) practitioners becoming comfortable with others that do not share their backgrounds and identities.

The authors conclude that existing measures of cultural competence focus on increasing practitioners' knowledge of the Other, largely ignoring the power relations of social inequality in their construct of cultural competence; current constructs of cultural competence in the medical field were also found to assume that growth in an individual's knowledge and self-confidence is sufficient to disrupt systems of power and create environments that are less assaultive to individuals from non-dominant backgrounds. Moving forward, the authors recommend developing measures that assess cultural humility and the actual utilization of cultural competence in a patient-care setting in order to push practitioners to develop the requisite skills to engage patients in a culturally competent manner.

While 54 such measures exist for healthcare professionals, no such quantitative measures exist for education professionals. The closest measures within the field of education present themselves as measures of critical consciousness. Coined by Freire (1968), critical consciousness is the ability to perceive and take action against oppressive social, political, and economic elements of society. Critical consciousness is situated within Freire's liberatory lens toward education (echoed by Ladson-Billings as sociopolitical consciousness) and is a more

expansive construct than cultural competence. Moving beyond simple knowledge of students' culture and identities, critical consciousness pushes educators toward a stance of actively harnessing knowledge of students' cultures and backgrounds to empower students to disrupt social systems of power and inequity. Freire's stance is stronger than Kumas-Tan's critique of current measures of cultural competence in the health fields, pushing beyond the need for practitioners to merely understand systems of inequity and instead challenging educators to actively disrupt systems of inequity through instruction.

Three measures currently exist to measure critical consciousness, all developed recently. Diemer, et al (2014) developed the Critical Consciousness Scale (CCS) for adults to measure three domains of critical consciousness: an individual's ability to perceive inequality, their socio-political participation, and their attitudes toward egalitarianism. The CCS was later modified for use with adolescents as the Measure of Adolescent Critical Consciousness (MACC; McWhirter & McWhirter, 2016). The Critical Consciousness Inventory (Thomas et al., 2014) assess an individual's progression along the four stages of developing critical consciousness: pre-critical, beginning critical, critical, and post-critical.

While not possessing the same published record as a psychometrically validated measure, one additional tool deserves special note. The Culturally Responsive Instruction Observation Protocol (CRIOP; Powell et al., 2016) is a compilation of 24 indicators of culturally-responsive practice across seven domains: classroom relationships, family collaboration, assessment, curriculum, instruction/pedagogy, discourse, and socio-political consciousness. In addition to the observation protocol, CRIOP also includes parent and teacher interview protocols to help triangulate evidence from classroom observations.

CRIOP is inclusive of both culturally competent and critically conscious elements. When describing curriculum and instructional practices, CRIOP assesses for "learning experiences [that] use the knowledge and experience of students and their families in order to facilitate students' cultural knowledge and the academic knowledge of the school," as well as "instruction that is contextualized in students' lives and experiences." With a critical consciousness lens, CRIOP also looks for evidence that "students question the status quo and take action on real world problems;" furthermore, CRIOP turns educators' gaze toward deconstructing "instructional materials and popular texts… to uncover their ideological assumptions and biases."

The tool is currently in the pilot stage with limited adoption and use in 27 elementary classrooms. Preliminary findings are promising, however, with teachers who used the tool showing increasing levels of culturally responsive instruction from the fall to the spring, as well as increasing student achievement scores in reading and mathematics among high implementers versus low implementers.

*Expanding Widely-Adopted Measures to Include Cultural Competence.* Carter and Darling-Hammond (2016), in their review of teaching for diverse learners, argue for three broad practices for the effective teaching of diverse learners: using culture as a resource for learning, explicit teaching of skills and critical thinking, and cooperative learning as a path to achievement. In looking at the landscape of widely-adopted measures of teacher quality, existing measurement tools have emphases on both the explicit teaching of skills and critical thinking, as well as cooperative learning. Across the seven measures analyzed in the first section, 78% of indicators

focus on the teaching of skills and critical thinking, and 15% of indicators address cooperative learning; 0% of indicators focus on cultural competence or critical consciousness.

While an extensive research literature exists supporting the role that students' culture plays in creating rigorous and equitable classroom instruction, students' culture is not mentioned once in the seven measures analyzed. Research has established that teachers' knowledge and skill has a direct impact on the quality of instruction delivered, which in turn impacts student learning (Darling-Hammond, 1999; Ingersoll, 2002; Whitehurst, 2002). Furthermore, students respond to the inclusion of their cultural knowledge and linguistic patterns in instruction (Alim, 2004; Andrade-Duncant & Morrell, 2008; Emdin, 2010; Fisher, 2007; Kinloch, 2005). Teachers' default patterns of instruction, however, draw from a plethora of unexamined biases and assumptions about students based on teachers' own experiences and habits (Ball et al., 2003). In order to be culturally responsive educators, teachers must develop a multicultural and culturally responsive skill set (Sleeter, 2001), providing a lens to recognize the cultural resources that students bring to the classroom (Moll et al., 1992). In doing so, teachers will be better positioned to link their content domain with students' cultures.

Given the current state of widely-adopted tools for measuring teacher quality and the literature base in support of cultural competence as an essential tool for teachers of diverse learners, this analysis suggests that measurement designers should expand the dominant construct of teacher quality to include domains that focus on teachers' cultural competence and enactment of culturally responsive pedagogies. The CRIOP measure is one such model of a tool whose domains could serve as a reference point for the revision of existing tools. While CRIOP has not been studied at scale, revising existing measures that already enjoy widespread adoption could rapidly increase our knowledge of how to create valid and reliable measures of teacher

quality that incorporate cultural competence, as well as provide a large sample of observational data to serve as a comparison group from which to better gauge change in teaching practice as a result of implementing measures with expanded constructs of teacher quality.

## CONCLUSION

Chapter II assessed the current state of measures of teacher quality in light of the research literature on cultural competence, seeking to better understand tools that have found widespread use and adoption, as well as interrogate the current construct of 'teacher quality' as reified by existing measurement tools. Through this analysis, the current landscape of teacher quality measures was found to be diverse and complex. Depending on the designer, measures could be complex systems that spanned numerous domains of teaching competence, requiring significant training for districts to adopt at scale; conversely, measures could also be researcher-focused systems that primarily coded videos of teaching to determine the quality of a teachers' content pedagogy. Each measure was found to have formative properties that would allow them to be used by teachers and their mentors for coaching and the improvement of practice, while also possessing the possibly to be enacted in an evaluative manner to determine teacher effectiveness.

The research literature is equally diverse in its uptake of measurement tools. Studies describing measures, as well as validating their psychometric properties, dominate the research literature. Pushing an evaluative paradigm, another branch of the literature seeks to establish correlations between scores on measures of teacher quality and measures of student and teacher achievement such as standardized test scores and state-calculated teacher value-added measures. A final branch of the literature seeks to correlate measures of teacher quality with other constructs of interest, such as intellectual development.

Finally, students' presence in current measures of teacher develop was found to be limited, with only 20% of indicators focused on students' actions. These student-focused indicators were largely in relation to students' participation in whole-class discussions and rigorous content-based tasks; students' culture and backgrounds are not mentioned in any of these measures.

As a result of this analysis, one could conclude that the current construct of "teacher quality" is predominantly focused on teachers' actions to deliver instruction; manage a classroom; and build a cooperative, discourse-focused classroom community focused on students' enactment of rich tasks. A robust research literature, however, has established the importance of teachers' cultural competence for successfully teaching diverse learners who share different backgrounds than their own; an early-stage measure of teacher quality that is inclusive of students' culture exists, but does not yet enjoy widespread adoption by the practitioner nor research communities. Therefore, a gap exists in that widely-adopted measures of teacher quality do not take into account the cultural resources that students bring to the classroom, nor do they highlight teacher actions that incorporate students' cultural resources into the design and enactment of content-focused instruction. Without widespread adoption and psychometric validity, the dominant paradigm of teacher quality as assessed by schools, districts, and researchers remains one that is not inclusive of students' cultures.

Moving forward, this analysis suggests that those who design and research measures of teacher quality that are implemented by districts and schools should 1) expand their construct of teacher quality to include cultural competence and 2) collect data with expanded measures at scale to assess their psychometric validity and reliability. Whether through adding additional domains to existing frameworks for teaching quality, or infusing cultural competence throughout

existing indicators focused on instructional design and teacher knowledge domains, the inclusion of students' diverse backgrounds and cultures in tools that are widely used is necessary to fill the existing gap. As these new indicators are designed, researchers should also ensure that the psychometric properties of these expanded measures are properly established, giving districts and other academics confidence that the tools they are using to measure teacher quality are rigorous and sound.

# CHAPTER III

The conclusion of Chapter II called for the inclusion of cultural competence in measures of teacher quality that could be adopted at-scale, along with their psychometric validation. Chapters III and IV address this gap by assessing the psychometric properties of a measure of classroom community that is designed to specifically include cultural competence, and has the potential to be used at-scale. To this end, these chapters examine the construct validity of the Vision for Student Learning (see Appendix B), a new measure of classroom community which includes cultural competence as part of its construct and has the potential to be adopted at scale in 53 cities across the United States. An alternative-route teacher preparation program designed this measure in order to address a cultural competence-gap in their own measurement tools, and seeks to refine and strengthen this measure by better understanding its psychometric properties. The author of this study consulted with the regional design team to facilitate sessions that aimed to sharpen the focus of the measure and ensure that the phrasing of items was appropriate for the purposes of measurement. He did not make any final design decisions, however, nor did he play a significant role in shaping the theoretical foundations of the measure. If the measure is found to accurately capture elements of cultural competence through its design and implementation, the organization will consider adopting the measure across its entire network.

In order to establish construct validity, Chapters III and IV will explore four research questions. Chapter III will address the study methodology and first research question, while Chapter IV will address the second, third, and fourth research questions:

1. What are the features of the Vision for Student Learning (VfSL)?
2. What does the VfSL tell us about teachers' practices?
3. What evidence do we have that the VfSL is a valid instrument?
4. What are the psychometric properties of the VfSL?

## METHODOLOGY

*Setting, Participants, and Data Collection.* The data for this study were collected by a regional office of a national, alternative-route teacher preparation program located in a major metropolitan area in the mid-South United States. This region of the program places in two large, urban school districts, as well as two large charter networks. Across these four placement settings, a total of 358 novice teachers were trained and placed by the alternative-route certification program during academic year 2016-2017. 163 of these teachers are teaching for their second year, and 195 are in their first year of teaching. 264 teachers teach in a traditional public school setting, and 94 teach in charter schools. 161 of these teachers (45%) identify as people of color.

Due to capacity constraints on data collection, a sample of classrooms from the region was used for the present study. Sample and population demographic statistics can be found in Table 3; sample demographic statistics are within 5% of the population's demographic statistics.

*Table 3. Sample and Population Demographic Statistics*

| | Sample (n = 208) | Population (n=358) | Percent of Sample | Percent of Population | Difference |
|---|---|---|---|---|---|
| First-year Teachers | 111 | 195 | 53% | 54% | -1% |
| Second-year Teachers | 97 | 163 | 47% | 46% | 1% |
| District 1 | 118 | 205 | 57% | 57% | 0% |
| District 2 | 43 | 59 | 21% | 16% | 5% |
| Charter | 47 | 94 | 22% | 26% | -4% |
| Black | 28 | 63 | 13% | 18% | -5% |
| Latinx | 45 | 69 | 22% | 19% | 3% |
| Asian-American/ Pacific Islander | 19 | 29 | 9% | 8% | 1% |

Data were collected using the Vision for Student Learning classroom observation rubric, designed by the regional program team of the alternative-route certification program (see Appendix B for the complete rubric). The regional program team spent academic year 2015-2016 redesigning the observation rubric. In the spring of 2016, the coaching staff began the process of norming their use of the observation rubric. Time was carved out during weekly program meetings to explore each section of the rubric. The coaching team then had the chance to rate video observations of classrooms using the tool, norming on what each indicator looked like in classroom practice. Classroom coaches then had the opportunity to use the observation rubric during live classroom observations with their managers, further norming on how each would rate each classroom observation.

During the 2016-2017 school year, coaches used the observation rubric at minimum once a quarter while observing teacher classrooms. At the end of each quarter, instructional coaches discussed teacher ratings with their manager and their coaching "pod," consisting of 3-4 other coaches. In its inaugural year, the observation rubric ratings were primarily used to help coaches and program leaders make decisions with regard to where to focus teacher professional development and coaching resources. Some coaches used the observation ratings in coaching

conversations with teachers, but there was no expectation that coaches use the ratings during coaching conversations during the school year. Program leaders are currently working with managers and coaches to devise a protocol to integrate rubric ratings into coaching conversations during the school year.

The data analyzed in Chapters III and IV are from the final round of classroom observations, which took place in May of 2017. Each coach spent a minimum of 45 minutes in each classroom to conduct the in-person observation, scoring classrooms on each indicator in the VfSL. Using classroom observation data from May for the validation of the VfSL provides a snapshot of classrooms at their most advanced point during teachers' first and second year teaching – critical points in the development of teachers during the program. By using these data, we are able to test where teachers fall along the continuum of development for each indicator at the end of their first and second years of teaching. This not only provides a snapshot into where teachers' skills lie at the end of their first and second year, but also a glimpse at teachers' growth trajectory between their first and second years.

*Measurement Instrument History and Overview*. All teachers were rated by their instructional coach using the Vision for Student Learning (VfSL; Appendix B). The VfSL was explicitly designed to incorporate elements of culturally responsive teaching and cultural competence into a classroom observation rubric that can provide teachers with feedback on how to improve their practice, as well as data for program leaders to use as they make strategic decisions on where to focus teacher professional development.

It is important to note that the VfSL is not simply the adaptation of an existing teacher evaluation rubric to include an added layer or domain of cultural competence, but rather an

entirely new and reimagined formative tool that has cultural competence infused throughout each item and dimension. The crux of this transformation lies in the move from items that are written from the teacher-action perspective, to items written from the vantage point of student actions. Traditional teacher observation rubrics focus on just that – what the teacher is doing. To rate teachers on such rubrics, observers must pay attention to specific teacher actions such as how long teachers wait to provide students an answer after asking a question or the clarity with which teachers deliver instructions to students. As such, a classroom observer's gaze is primarily focused on the teacher, not the students.

The VfSL flips this paradigm by requiring raters to observe *student* actions, rather than teacher actions. As such, the VfSL does not squarely fit into the category of teacher observation rubrics that measure "teacher quality" as currently defined; nor is the VfSL a measure of "student quality". Instead, it uses the proxy of student actions to measure teacher quality, pushing teachers and observers to take stock of the actions in which all of the classroom actors are engaged, both students and teachers. In this framework, a teacher can't simply perform a list of actions to get checkmarks on a rubric; a teacher's actions must promote a classroom culture and level of engagement with content that results in students participating a certain way. As such, the VfSL is better construed as a measure of classroom community. Based on the actions of students in a classroom, teachers are able to reflect on their own practice to see how their actions are and are not contributing to the learning community. The VfSL provides aligned teacher actions for each dimension, supporting teachers to think through changes they could make to their own practice to improve on a certain dimension of practice.

The VfSL evolved from three tools previously used by the certification program. The initial rubric used by the program was a six-section rubric that promoted instruction that was

teacher-driven and focused on an "I do, you do, we do" pedagogical paradigm; instruction in this mode prepared students to memorize and regurgitate information, elevating pedagogy that prepared students to perform well on high-stakes assessments. The first iteration of the rubric did not focus on the culture of learning within a classroom and was primarily focused on classroom management, planning, and instructional execution. The next evolution of the coaching rubric used within the region was minor, expanding the rubric to include aspects of classroom culture. In this iteration, two new domains were added to the rubric: focused on 1) the "culture of learning" established in the classroom to 2) support students' "rigorous engagement with content". Both of these tools rated teachers' actions within the classroom and did not focus on observing students' actions in the classroom.

With the adoption of the Common Core State Standards for Mathematics and English Language Arts (CCSS), as well as the Next Generation Science Standards (NGSS), a pedagogical shift occurred nationally from pedagogies focused on acquisition to pedagogies focused on participation (Sfard, 1998). As a result of this paradigmatic shift, the measures used by the region to gauge teacher quality also needed to shift. In response, the region began to use a rubric designed by another well-known alternative-route teacher preparation program that was more aligned to the pedagogics of the CCSS and NGSS. This rubric, however, was expansive. With over 40 indicators across four domains, the design of the tool was unwieldly for regular classroom observation use. In addition, the tool focused exclusively on students' academic success, excluding aspects of cultural competence or non-cognitive student outcomes. The tool did, however, elevate pedagogies consistent with the shift from acquisition-focused to participation-focused learning, and shifted the gaze of the observer from teacher actions to student actions. As a result, coaches had a new emphasis when observing classrooms on what

students were doing during the lesson, rather than what actions teachers were taking during instruction.

Given the challenges inherent to implementing such a large rubric, however, the program team decided to create a new measure, the VfSL, which is streamlined and easier to use in classrooms, while also including aspects of cultural competence and culturally responsive teaching practices. The VfSL is grouped into three dimensions:

*Table 4. Dimensions of the Vision for Student Learning*

| | |
|---|---|
| *Safe, Brave, and Equitable Classrooms* | Classroom culture; cultural competence; and peer relationships |
| *Rigorous and Culturally Relevant Learning* | Academic expectations; methods of culturally responsive instruction; connections between content, discourse, and culture |
| *Perseverance to Goals* | Content mastery and access to path expanding opportunities |

Each dimension of the VfSL has 4-5 items that are rated based on observed student actions. In addition to the four to five items in each dimension, there are also aligned teacher actions associated with the items in each dimension. These banks of teacher actions were included by the rubric designers and program managers to serve as a starting point for coaches to reference when coaching teachers on different competencies assessed by items within each dimension.

It is worth underscoring that the VfSL is not meant to be an evaluative tool; rather, *the VfSL is formative in nature*. As a measure of classroom community, the VfSL is designed to help teachers improve their practice; secondarily, the tool aims to provide useful information to coaches, professional development designers, and district/program officials in order to help them better support their teachers. The VfSL is not designed to be used for teacher evaluations, nor to make high-stakes decisions around teacher retention. Many measures of teacher development, The Danielson Framework being a prime example, began their tenures as formative tools that

were later taken up by districts as evaluative measures. While one aim of this dissertation is to determine whether the VfSL is a valid and reliable measure of classroom community, the goal in doing so is not to establish the VfSL as appropriate for use in an evaluative setting.

## ANALYSIS OF THE VfSL INSTRUMENT

This section will address the first research question: *What are the features of the VfSL?* In response to Chapter II's analysis of current measures of teacher quality, the design of the VfSL attempts to explicitly include culturally competent practices and markers of culturally responsive pedagogy. This thrust is felt throughout many of the items, which are assessed at the student level (i.e., to what degree students enact these actions in classrooms). Using the cultural competence theoretical frameworks of Ladson-Billings, Lee, and Banks described in Chapter II, this section analyzes each dimension of the VfSL to assess its inclusion of aspects of cultural competence.

At its core, the Safe, Brave, and Equitable Classrooms dimension aims to promote a classroom culture that affirms students' identities and allows them to take risks and become more confident as they engage with subject-matter content. Table 5 maps items from this dimension onto the three cultural competence frameworks.

*Table 5. Cultural Competence Markers in the Safe, Brave, and Equitable Classrooms Dimension*

| | Openness | Expectations | Routines | Engagement | Inclusion |
|---|---|---|---|---|---|
| GLB: Academic Achievement | | | | X | |
| GLB: Cultural Competence | X | | | | X |
| GLB: Socio-political Consciousness | | X | X | | X |
| Lee: Culture as a Resource for Learning | X | | | | X |
| Lee: Connections between Content and Culture | | | | | |
| Lee: Deep Knowledge of Content Domain | | | | | |
| Banks: Content Integration | | | | | |
| Banks: Knowledge Construction | | | | | |
| Banks: Prejudice Reduction | X | | | | X |
| Banks: Equity Pedagogy | X | | | X | X |
| Banks: Empowering School Culture | X | | | | X |

Items in this dimension map across all three domains of Ladson-Billings' framework, with the greatest overlap in the domains of cultural competence and socio-political consciousness. There is also concentrated representation of concepts from Banks' Multicultural Framework, especially around prejudice reduction, equity pedagogies, and fostering an empowering school culture. Representation from these strands of Banks and Ladson-Billings' theoretical works is unsurprising, given the focus of this VfSL dimension on fostering a classroom culture that incorporates students' cultures into classroom learning and fosters an environment in which students learn from and support one another as people who bring multiple identities to the classroom.

The Rigorous and Culturally Relevant Learning dimension of the VfSL also has rich theoretical connections to the literature on cultural competence and learning (Table 6). The dimension's explicit focus on academic achievement and connections between content and students' cultures are the cornerstone of this dimension's operationalization of culturally responsive pedagogy. The Discourse and Connections items have the broadest theoretical coverage due to their focus on equitable access to instruction through classroom discourse practices (Cornelius & Herrenkohl, 2004; Moschkovich, 2002; O'Connor & Michaels, 1993; Windschitl et al, 2012) and attention to meaningful connections between what students are learning inside school and what they experience both within and outside school.

*Table 6. Cultural Competence Markers in the Rigorous and Culturally Relevant Learning Dimension*

| | Opportunity | Discourse | Evidence | Connections |
|---|---|---|---|---|
| GLB: Academic Achievement | X | X | X | X |
| GLB: Cultural Competence | | X | | X |
| GLB: Socio-political Consciousness | | | | X |
| Lee: Culture as a Resource for Learning | | X | | X |
| Lee: Connections between Content and Culture | | | | X |
| Lee: Deep Knowledge of Content Domain | X | X | X | X |
| Banks: Content Integration | X | X | X | X |
| Banks: Knowledge Construction | X | X | X | X |
| Banks: Prejudice Reduction | | | | |
| Banks: Equity Pedagogy | | X | X | X |
| Banks: Empowering School Culture | | | | |

The Perseverance to Goals dimension also exemplifies aspects of cultural competence as described in the theoretical literature (Table 7). Items in this dimension are strongly aligned with Ladson-Billing's three pillars of culturally responsive pedagogy and support Banks' emphasis on equitable instruction and building an empowering school culture.

*Table 7. Cultural Competence Markers in the Perseverance to Goals Dimension*

| | Yearly Progress | Daily Mastery | Perseverance | Resources |
|---|---|---|---|---|
| GLB: Academic Achievement | X | X | X | X |
| GLB: Cultural Competence | X | | X | X |
| GLB: Socio-political Consciousness | X | | X | X |
| Lee: Culture as a Resource for Learning | | | | X |
| Lee: Connections between Content and Culture | | | | X |
| Lee: Deep Knowledge of Content Domain | | | | |
| Banks: Content Integration | X | X | | |
| Banks: Knowledge Construction | | | | |
| Banks: Prejudice Reduction | X | | | X |
| Banks: Equity Pedagogy | X | | X | X |
| Banks: Empowering School Culture | | | X | X |

Across the VfSL, items in all three dimensions operationalize theoretical aspects of culturally responsive pedagogy and multicultural education (Table 8). The VfSL shows strong alignment with Ladson-Billings' culturally-responsive pedagogy framework, which is unsurprising as designers of the VfSL relied heavily on this framework during its creation. VfSL items also cover each major strand of Lee and Banks' frameworks, although at lesser frequencies than with the strands of Ladson-Billings' framework. The VfSL appears weakest with regard to

explicitly promoting connections between academic content and students' cultures. Only two items of the VfSL – Connections and Resources – reference this theme.

*Table 8. Number of VfSL Items that Address Cultural Competence Markers*

| | Number of VfSL Items |
|---|---|
| GLB: Academic Achievement | 9 |
| GLB: Cultural Competence | 7 |
| GLB: Socio-political Consciousness | 8 |
| Lee: Culture as a Resource for Learning | 5 |
| Lee: Connections between Content and Culture | 2 |
| Lee: Deep Knowledge of Content Domain | 4 |
| Banks: Content Integration | 6 |
| Banks: Knowledge Construction | 4 |
| Banks: Prejudice Reduction | 4 |
| Banks: Equity Pedagogy | 9 |
| Banks: Empowering School Culture | 4 |

## CONCLUSION

Chapter III described the design history of the VfSL, the characteristics of the dataset collected using the VfSL, and addressed the first research question: What are the features of the Vision for Student Learning? As a result of these analyses, the VfSL qualifies as a measurement tool that includes aspects of cultural competence. Developed by the regional office of an alternative-route teacher preparation program, the VfSL operationalizes constructs of culturally responsive pedagogy and multicultural education across each of its three dimensions.

Chapter IV will report findings from analyses of the psychometric validity and reliability of the VfSL. Based on the descriptive statistics reported in this chapter, the VfSL dataset is a representative subsample of the teacher population supported by this regional office and appropriate for analysis.

# CHAPTER IV

In response to the call in Chapter II for a validated measurement tool that includes aspects of cultural competence and critical consciousness, Chapter III established the Vision for Student Learning (VfSL) as a potential measure that includes aspects of cultural competence and critical consciousness. Chapter IV addresses the validity and reliability of the VfSL by answering the remaining three research questions:

2. What does the VfSL tell us about teachers' practices?
3. What evidence do we have that the VfSL is a valid instrument?
4. What are the psychometric properties of the VfSL?

## TEACHERS' PRACTICES AS OBSERVED USING THE VfSL

The descriptive statistics for each item of the VfSL for this validation study are presented in Table 9, grouped by dimension. This validation study analyzes these data in terms of outcomes for the teachers' classrooms in order to determine how the VfSL captures different characteristics of teachers' classrooms. Analyses of trends within these descriptive statistics, in conjunction with analysis of the psychometric properties of the items, will also yield insights into the utility of each item for measuring the construct of classroom community in classrooms similar to those of the sample. In other words, while the purpose of this validation study is not to determine the efficacy of the teachers themselves or their training per se, we analyze their data to explore the sensitivity and affordances of VfSL itself.

Teachers in the sample exhibited the highest means across the instrument for the Safe, Brave, and Equitable Classroom dimension, with an average score of approximately 4 across the dimension. Items in the Rigorous and Culturally Relevant Learning dimension were the most

challenging, with an average score of approximately 3.3. Items in the Perseverance to Goals dimension had a slightly higher average score of approximately 3.5.

Based on these descriptive statistics, the students observed in this sample were most successful at independently following behavioral expectations, assuming responsibility for classroom routines, and remaining engaged in learning during the lesson – pointing to teachers' strengths at setting expectations, teaching and releasing control of classroom routines, and promoting engagement during lessons. Students had the most difficulty engaging in peer-to-peer discourse during lessons and making connections between course content and their social, cultural, political, and historical position in the world. These lower scores point to an increased need to develop the skill of teachers in this sample at promoting classroom discourse and connections between course content and students' lived realities. These descriptive statistics point to the ability of the VfSL to provide useful information with regard to strengths and areas for growth across each of its three dimensions, providing administrators and teachers a high-level snapshot of the current state of teaching in their school or district. They also provide insight into areas for principals and teacher leaders to target coaching and teacher development.

*Table 9. Vision for Student Learning Item Descriptive Statistics*

| Item | Observations | Mean | SD | Min | Max |
|---|---|---|---|---|---|
| Openness | 179 | 3.63 | 1.06 | 1 | 5 |
| Expectations | 182 | 4.40 | 0.82 | 1 | 5 |
| Routines | 181 | 4.34 | 0.85 | 2 | 5 |
| Engagement | 196 | 4.20 | 0.90 | 1 | 5 |
| Inclusion | 177 | 3.50 | 1.22 | 0 | 5 |
| Opportunity | 182 | 4.07 | 1.05 | 1 | 5 |
| Evidence | 194 | 3.44 | 1.13 | 0 | 5 |
| Discourse | 178 | 2.82 | 1.25 | 0 | 5 |
| Connections | 173 | 2.75 | 1.38 | 0 | 5 |
| Yearly Progress | 186 | 3.35 | 1.38 | 0 | 5 |
| Daily Mastery | 162 | 3.80 | 0.85 | 2 | 5 |
| Perseverance | 195 | 3.53 | 1.14 | 1 | 5 |
| Resources | 189 | 3.46 | 1.31 | 0 | 5 |

*Distribution of Scores for All Teachers.* The distribution of scores for each item within each dimension are worth noting. For the Safe, Brave, and Equitable Classroom dimension (Figure 2), the vast majority of first and second year teachers ended the year with a score of 4 or 5. Teachers only received a score of 0 on the Inclusion item, and less than 15% of teachers received a score of 1 or 2 on any item within the dimension. These distributions suggest that the sample of teachers was largely proficient at establishing a strong classroom culture at the end of their first and second years. Used with a different sample, however, the VfSL might show increased variation throughout this dimension, allowing principals and coaches to better target the professional development needs of teachers with regard to creating a strong classroom culture.

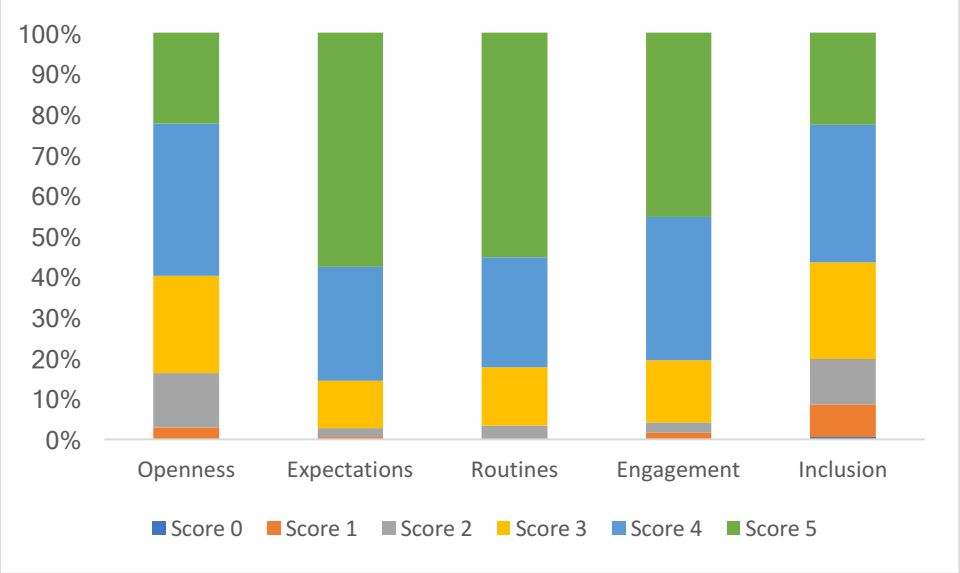



*Figure 2. Distribution of Scores on the Safe, Brave, and Equitable Classroom Items*

For items in the Rigorous and Culturally Relevant Learning dimension, score distributions were more varied (see Figure 3). While a majority of teachers received a score of 4 or 5 on the Opportunity and Evidence items, less than 30% of teachers received scores of 4 or 5 on the Discourse and Connections items, with the majority receiving a score of 2 or 3. Very few

teachers received a score of 0, however, and scores of 1 were also infrequent. There is a clear pattern, however, of top scores decreasing and lower scores increasing as the items progress in the dimension – it appears more difficult to receive a higher score on the Discourse and Connections items than the Opportunity and Evidence items. These findings point to the sensitivity of the VfSL to measure different aspects of students' engagement with content, positioning it as a useful measurement tool to learn more about teachers' pedagogical practices and align support accordingly.

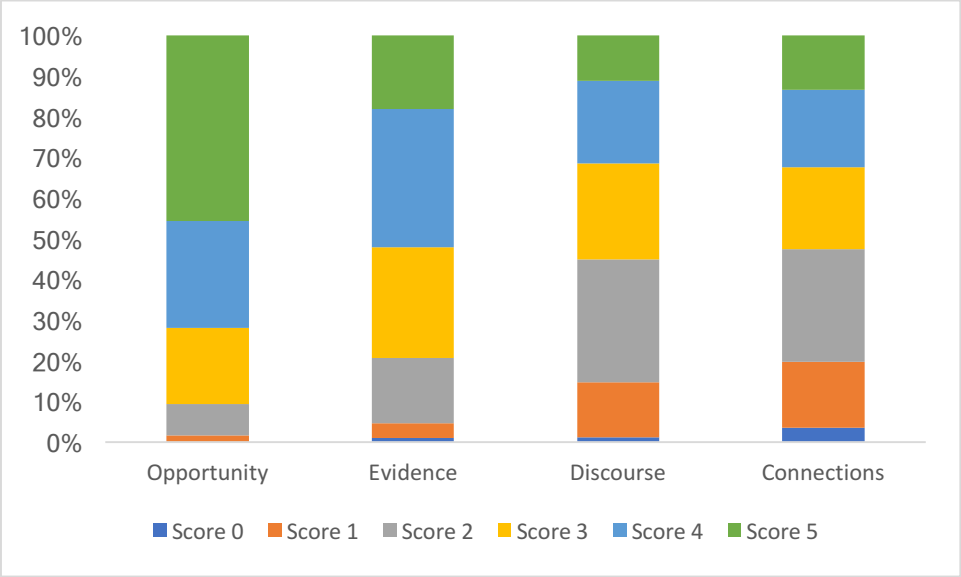



*Figure 3. Distribution of Scores on the Rigorous and Culturally Relevant Learning Items*

For the Perseverance to Goals dimension, items had markedly different score distributions despite their overall means trending higher than the Rigorous and Culturally Relevant Learning dimension (see Figure 4). More teachers received a score of 4, driving up the mean, and fewer teachers received a score of 2. Scores of 0 remained rare, with fewer scores of 1, as well. This suggests that either the VfSL does not function as well to measure variation within this dimension of classroom community, or that this sample of teachers simply has less

variation in ability within this dimension. While the IRT model will take into account these differences in score distribution when determining the degree to which each item contributes to measuring the construct of classroom community, these distribution statistics are worth noting insofar as they point to the particular skillset of this sample of teachers as rated by their instructional coaches.

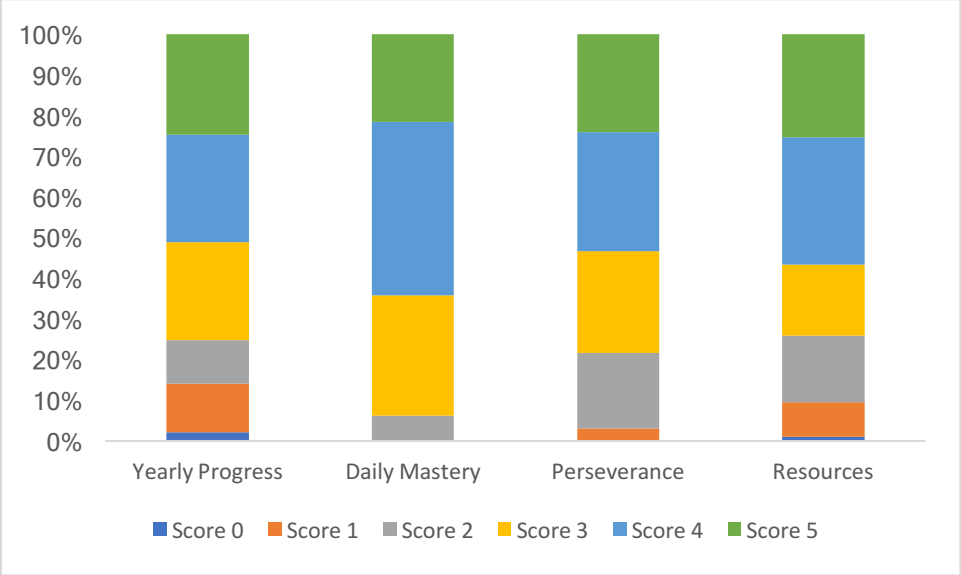



*Figure 4. Distribution of Scores on the Perseverance to Goals Items*

It is also worth investigating whether any trends in score distribution emerge based on demographic characteristics. Appendix C presents graphs showing the distribution of scores on each VfSL item by three demographic cuts: first- and second-year teacher status, gender, and race.

*Distribution of Scores by First- and Second-Year Teacher Status.* As one would expect, classrooms led by second-year teachers generally score higher on VfSL items than first-year teachers. On the Openness, Routines, Inclusion, Opportunity, Evidence, Discourse, Yearly

Progress, Daily Mastery, Perseverance, and Resources items, second year teachers have a higher percentage of scores of 4 and 5. On the Expectations, Engagement, and Connections items first- and second-year teachers have a similar percentage of scores of 4 and 5, with second-year teachers having a greater incidence of 5's than 4's. The general trend of second-year teachers having higher scores than first-year teacher holds across each of the three dimensions of the VfSL, as well. This pattern suggests that the VfSL successfully differentiates among novice teachers, given patterns of higher ratings for second-year teachers over first-year teachers.

*Distribution of Scores by Gender.* Classrooms led by women generally score higher when looking at VfSL score distributions by gender. For the Openness, Expectations, Engagement, Opportunity, Evidence, Discourse, Yearly Progress, Daily Mastery, Perseverance, and Resources items, women earn scores of 4 or 5 more frequently than men. For the Inclusion item, men earn scores of 4 or 5 more frequently than women. And for the Routines and Connections items, men and women earn scores of 4 and 5 at relatively equal frequencies. These differences could point to the ability of the VfSL to be sensitive to differences along its various dimensions. Programs who are potentially looking to implement the VfSL in their context should be careful, however, to ensure that these differences are not due to bias on the part of the raters or the items themselves.

*Distribution of Scores by Race.* When looking at the distribution of scores on the VfSL by the race of the classroom teacher, no clear trend emerges. On the Openness, Opportunity, and Discourse items, white teachers receive scores of 4 or 5 more frequently than teachers of color. On the Engagement and Resources items, teachers of color received scores of 4 or 5 more

frequently than white teachers. On the Expectations, Inclusion, Evidence, Connections, and Perseverance items, teachers of color and white teachers receive scores of 4 or 5 at roughly equivalent frequencies. On the Routines item, teachers of color and white teachers received approximately the same number of scores of 4 and 5, although teachers of color had more scores of 5 than 4, compared to white teachers. On the Yearly Progress and Daily Mastery items, white teachers and teachers of color received approximately the same number of scores of 4 and 5, although white teachers had more scores of 5 than 4, compared to teachers of color. Once again, these differences suggest that the VfSL is sensitive to differences along its various dimensions. Programs should remain cautious, however, and ensure that these differences are not the result of bias on the part of observers or the items themselves.

## VALIDITY OF THE VfSL

The Vision for Student Learning (VfSL) was designed to measure the construct of classroom community through three dimensions: Safe, Brave, and Equitable Classrooms; Rigorous and Culturally Relevant Learning; and Perseverance to Goals. In order to examine the construct validity of the VfSL, an exploratory factor analysis (EFA) is necessary to determine whether the items in the VfSL actually function together as a dimension thereby indicating that the items as a group measure the intended construct.

As demonstrated by Schweig (2014), the assumption of cross-level invariance when conducting factor analyses can distort proper analysis of the factor structure, leading researchers and policymakers to form improper inferences. The current dataset includes classroom-level data taken from 69 different schools, with clusters of one to 17 classrooms in each school and an average cluster size of three classrooms. To investigate variability due to schools clustering, the

intraclass correlations (ICCs) of each item were tested. To do so, maximum likelihood estimates of the within-group correlation matrix and between-group level correlation matrix were obtained using Mplus version 8 (Muthén & Muthén, 2017). ICCs range from around 0.13 to around 0.5 for the VfSL (see Table 10), indicating that individual teacher scores within school clusters share a nontrivial amount of similarity. ICCs of this size provide sufficient evidence that a multilevel exploratory factor analysis (MEFA) is warranted.

*Table 10. VfSL Intraclass Correlations*

| Item | ICC |
|---|---|
| Openness | 0.357 |
| Expectations | 0.186 |
| Routines | 0.156 |
| Engagement | 0.266 |
| Inclusion | 0.409 |
| Opportunity | 0.137 |
| Evidence | 0.288 |
| Discourse | 0.478 |
| Connections | 0.408 |
| Yearly Progress | 0.466 |
| Daily Mastery | 0.133 |
| Perseverance | 0.424 |
| Resources | 0.501 |

MEFA was conducted on the matrix, with oblique (geomin) rotation used the leave the factors free to correlate. Retaining factors with an Eigenvalue greater than 1, MEFA indicated that there were two within-school factors and one to two between-school factors present in the data. At the within-schools level, the two factors were statistically significant and moderately correlated (0.61). At the between-schools level, both a 1-factor and 2-factor structure were statistically significant. The two factors in the 2-factor model were highly correlated (0.942), however; as a result, only one factor was retained at the between-schools level.

The final factor structure is outlined in Table 11. Model-fit indices for this factor structure indicate good model fit. RMSEA values of less than 0.05 are considered good; in this

sample, the RMSEA value is 0.036. CFI and TLI values greater than 0.9 also indicate good fit; in this sample, the CFI value is 0.981 and the TLI value is 0.975.

*Table 11. Rotated Factor Loadings: Multilevel EFA*

| | Within schools | | Between schools |
|---|---|---|---|
| | Factor | | Factor |
| Item | 1 | 2 | 1 |
| Openness | **0.581*** | 0.295* | 0.897* |
| Expectations | **0.935*** | 0.003 | 0.800* |
| Routines | **0.902*** | 0.004 | 0.895* |
| Engagement | **0.998*** | -0.147 | 0.936* |
| Inclusion | **0.682*** | 0.182 | 0.900* |
| Opportunity | **0.390*** | **0.403*** | 0.895* |
| Evidence | 0.001 | **0.756*** | 0.955* |
| Discourse | 0.192 | **0.681*** | 0.724* |
| Connections | -0.308* | **0.784*** | 1.014* |
| Yearly Progress | 0.218 | **0.588*** | 0.942* |
| Daily Mastery | **0.488*** | **0.477*** | 0.908* |
| Perseverance | **0.456*** | **0.371*** | 0.939* |
| Resources | **0.387*** | **0.343*** | 0.970* |
| Cross-loadings | 4 | 4 | 0 |

* indicates that the loading is significant at the 5% level

The within-school (classroom-level) factor analysis shows two factors – one that could be broadly conceptualized as classroom culture, and a second focused on rigorous instruction. The classroom culture factor is dominated by items from the Safe, Brave and Equitable Classrooms dimension of the VfSL. Furthermore, the rigorous instruction factor is dominated by items from the Rigorous and Culturally Relevant Learning dimension of the VfSL. This provides strong evidence that the current configuration of the two dimensions is empirically valid. This factor structure also suggests that individual teachers vary in their ability to construct a positive classroom culture and to execute rigorous instruction. Some teachers are adept at developing strong classroom cultures, but less adept at designing and executing rigorous instruction, and vice versa. Four item cross-load across the two factors: Opportunity, Daily Mastery, Perseverance, and Resources. The Opportunity item comes from the Rigorous and Culturally

Relevant Learning dimension; Daily Mastery, Perseverance, and Resources all come from the Progress to Goals dimension.

The between-school (school-level) structure of the VfSL differs from the within-school structure. Only one factor is present at the between-school level, which could be conceptualized as a classroom community factor. This suggests that at the school level, classroom community is a holistic construct that includes elements of both classroom culture and rigorous instruction. At the classroom level, however, classroom culture and rigorous instruction are distinct skills.

## PSYCHOMETRIC PROPERTIES OF THE VfSL

*Model Fit*. Since VfSL data are categorical in nature, as well as hierarchical and multi-factor in structure, a multilevel, multidimensional graded response model (MMGRM) was fit in Mplus version 8 to assess the instrument's psychometric properties. The model was fit with two within- and one between-level factors, using the loading patterns from the MEFA. Additionally, at least five observations had to be present in each category to ensure proper model fit; in order to achieve this target, the category for score 0 was collapsed into the category for score 1 for the following items: Expectations, Routines, Engagement, Opportunity, and Daily Mastery. Furthermore, the Perseverance and Resources items would not converge. Given their problematic cross-loading in the MEFA, both items were dropped from the final MMGRM.

Item discriminations for the VfSL, which are similar to factor loadings in a factor analysis, show that items are highly loaded on the within- and between-schools factors with the exception of the Connections item (see Table 12).

*Table 12. Vision for Student Learning Item Discriminations*

| Item | Within schools (2 factors) | Between schools (1 factor) |
|---|---|---|
| Openness | **1.937** (2.56) | 2.333 (2.94) |
| Expectations | **4.425** (6.14) | 2.440 (4.36) |
| Routines | **3.498** (2.72) | 2.726 (3.98) |
| Engagement | **4.731** (6.93) | 2.378 (4.37 |
| Inclusion | **2.023** (1.38) | 1.707 (2.01) |
| Opportunity | 2.182 (2.49) | 1.211 (1.46) |
| Evidence | 1.637 (1.67) | 1.965 (1.64) |
| Discourse | 1.684 (1.47) | 2.714 (1.57) |
| Connections | 0.791 (1.29) | 2.199 (1.42) |
| Yearly Progress | 1.863 (1.91) | 2.289 (1.43) |
| Daily Mastery | 4.175 (11.60) | 3.197 (5.43) |

**Bold** indicates items associated with the first factor.
Standard Errors are reported in parentheses.

A graded response model produces item thresholds, which represent item location information for each item on the sum of the two dimensions, weighted by item discriminations. Threshold 1 represents the transition point from a score of 1 (Few) on the VfSL, to scores of 2-5 (Some, About Half, Most, Almost All); Threshold 2 reflects the transition point from scores of 1-2 to scores of 3-5; Threshold 3 reflects the transition point from scores of 1-3 to scores of 4-5; and Threshold 4 reflects the transition point from scores of 1-4 to a score of 5. For all items, thresholds were in order (see Table 13). Item thresholds cover a wide range of the measurement continuum; there is significantly more coverage on the lower end of the continuum, however, with only the upper two thresholds representing positive (above average) ability on the continuum.

*Table 13. Vision for Student Learning Thresholds, Between-Schools Level*

| Item | T1 (SE) | T2 (SE) | T3 (SE) | T4 (SE) |
|---|---|---|---|---|
| Openness | -5.838 (5.038) | -3.101 (3.686) | -0.978 (2.239) | 2.306 (2.928) |
| Expectations | -9.469 (12.002) | -4.932 (3.325) | -0.928 (2.150) | |
| Routines | -7.698 (6.262) | -3.782 (3.161) | -0.603 (2.582) | |
| Engagement | -8.998 (11.902) | -4.499 (6.586) | 0.440 (3.152) | |
| Inclusion | -4.176 (3.099) | -2.680 (1.896) | -0.736 (1.858) | 2.100 (1.772) |
| Opportunity | -3.694 (2.082) | -1.543 (1.085) | 0.431 (1.582) | |
| Evidence | -4.849 (4.576) | -2.556 (1.808) | -0.520 (1.433) | 2.258 (2.106) |
| Discourse | -3.692 (2.669) | -1.035 (2.004) | 0.863 (1.514) | 3.364 (2.218) |
| Connections | -2.418 (1.791) | -0.647 (1.547) | 0.628 (1.424) | 2.481 (1.674) |
| Yearly Progress | -3.693 (1.739) | -2.553 (1.333) | -0.587 (1.439) | 1.709 (1.732) |
| Daily Mastery | -7.620 (17.925) | -2.006 (4.588) | 3.541 (8.117) | |

*Instrument and Item Information.* In order to determine reliability – the precision of an instrument – IRT models utilize "information". Information is typically graphed as a function, indicating that an instrument's precision can change at different points within the range of possible scores. In general, one expects an instrument to have more measurement error for scores at its extremes; as a result, total (whole-instrument) and item information functions typically look bell-shaped.

At the school level, the VfSL is most reliable in the low and average score ranges (see Figure 5). Information drops steadily at the upper end of possible scores. This is potentially due to two possible reasons: 1) the items on the VfSL are best suited to only measure the lower range of the construct of classroom community and/or 2) there were not enough responses in the upper score range to accurately model the VfSL's information function.

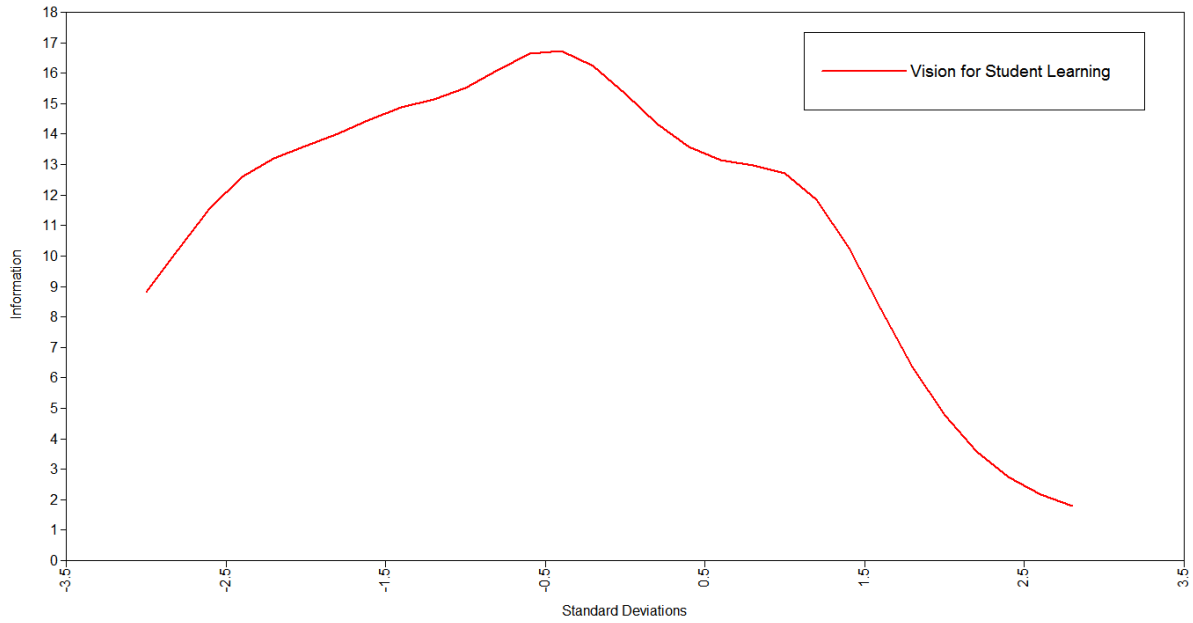



*Figure 5. VfSL Total Information Function for the Within-Schools Construct*

To investigate the first possible reason – that individual items on the VfSL only contributed to measuring the lower ranges of the construct of classroom community – individual item information curves were analyzed (see Figures 6 and 7). Items in the Safe, Brave, and Equitable Classrooms dimension provide the most information at the lower and average ranges of possible VfSL scores. These items provide less information at the higher end of the scale. The average score for classrooms on these items, however, was 4. Given the relatively high scores classrooms received on these items, it is likely that items from this dimension are best suited for measuring the lower to middle ranges of the classroom community construct.

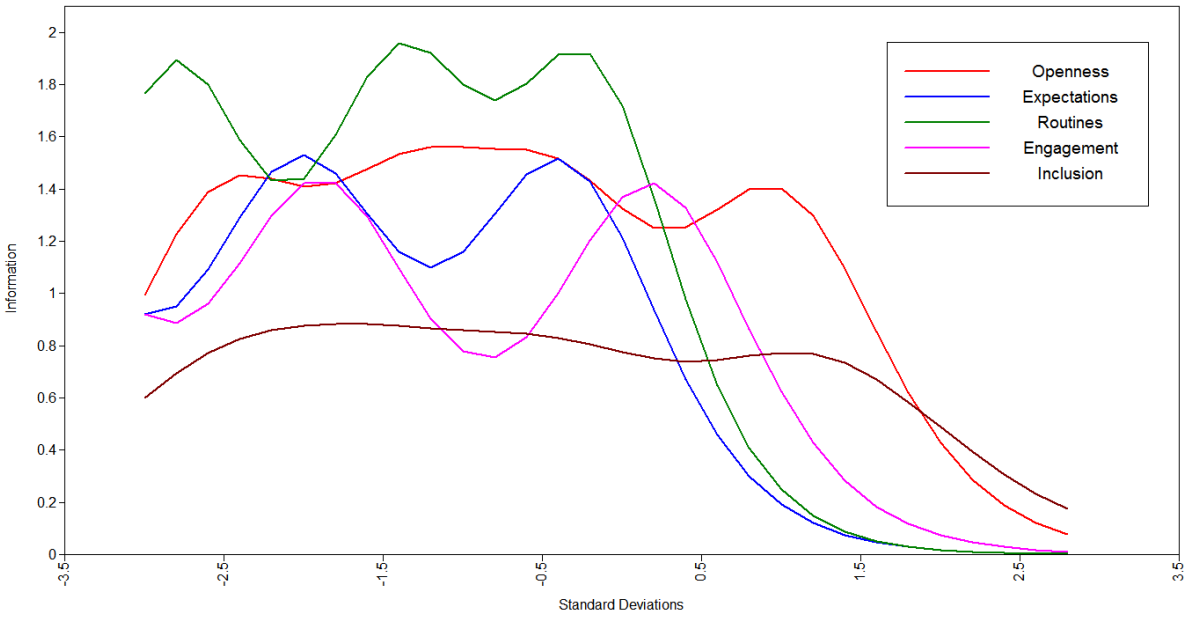



*Figure* 6. *Safe, Brave, and Equitable Classrooms Dimension Item Information Functions for the Within-Schools Construct*

Items in the Rigorous and Culturally Relevant Learning dimension provide mixed levels of information. The Evidence item provide the most information across the widest range of possible scores, with the Connections, Yearly Progress, and Discourse items also providing consistently strong information across the range of possible scores. The Opportunity indicator provides relatively low levels of information across the range of possible scores. The Daily Mastery indicator provides inconsistent information and appears to function poorly as an item.

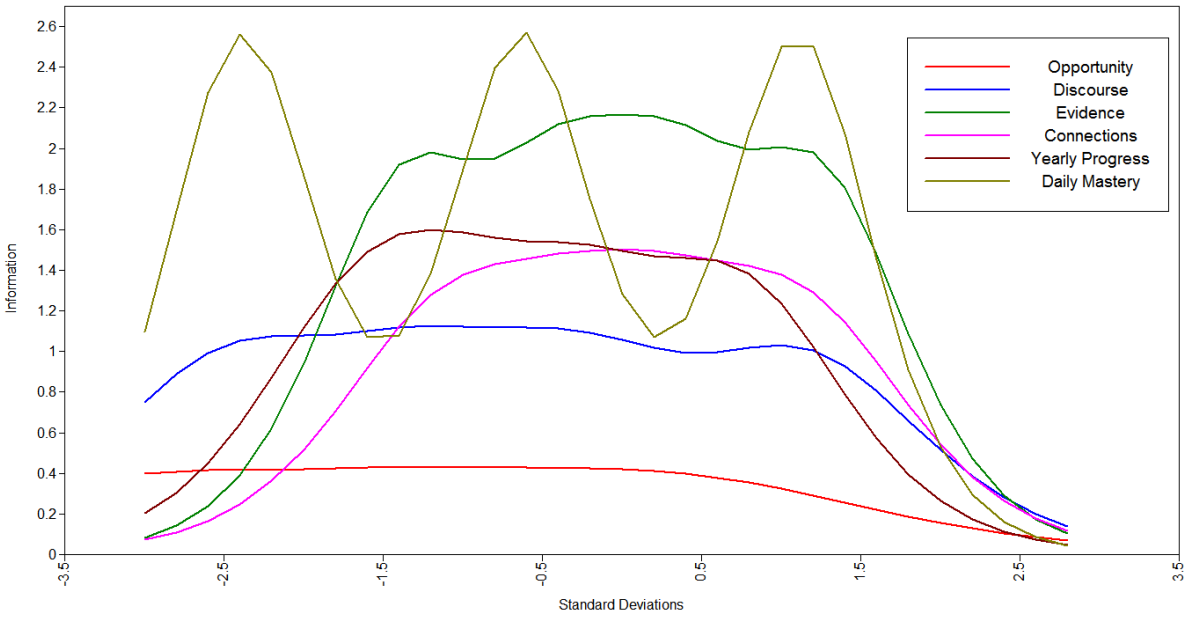



*Figure 7. Rigorous Learning Dimension Item Information Functions for the Within-Schools Construct*

## CONCLUSION

Chapter IV set out to answer three research questions: What does the VfSL tell us about teachers' practices?; What evidence do we have that the VfSL is a valid instrument?; and What are the psychometric properties of the VfSL?

With regard to the first research question, the data collected using the VfSL suggest that students of teachers in the validation sample are engaged, capable of independently enacting established classroom routines, and meeting their teachers' behavioral expectations at the end of teachers' first and second years. The VfSL is able to point to the fact that students in this dataset struggle, however, to engage in peer-to-peer discourse and to make connections between what they are learning and its social, cultural, political, and historical context and implications. Looking across the distribution of scores within items, teachers are by and large proficient at establishing Safe, Brave, and Equitable classrooms. Within the Rigorous and Culturally

Relevant Learning dimension, a majority of teachers were proficient at giving students an opportunity to express their learning and use evidence to support their claims. Teachers were less proficient, however, at supporting students to engage in discourse with one another around content, as well as make connections between what they were learning in class and the world around them. A majority of teachers were proficient at the competencies captured in the Perseverance to Goals dimension. This variation provides evidence that the VfSL is capable of measuring aspects of classroom management, academic rigor, and cultural competence, and that items are measuring different dimensions of classroom community.

Disaggregating score distributions on the VfSL by first or second year teaching, gender, and race also provided more information about classrooms in the sample. In general, classrooms of second-year teachers scored higher on the VfSL than classrooms of first-year teachers. Classrooms led by female teachers generally received higher scores on the VfSL than classrooms led by male teachers. With regard to race, teachers of color and white teachers had similar score distributions on about half of VfSL items, and classrooms of white teachers had scores higher than classrooms of teachers of colors on about half of items; classrooms of teachers of color only had higher VfSL scores than white teachers on two items. This variation also points to the sensitivity of the VfSL, while also reminding those who would implement the tool to ensure raters are properly trained and not exhibiting bias in their rating of classrooms.

With regard to the validity of the VfSL, findings are generally positive. The VfSL was found to have a different factor structure at the between- and within-schools levels. At the within-schools level, two factors were retained that roughly mapped on to the Safe, Brave, and Equitable Classrooms dimension, as well as the Rigorous and Culturally Responsive Learning dimension. The items in the Perseverance to Goals dimension demonstrated problematic cross-

loadings, suggesting that the items might be poor fits for the VfSL construct. At the between-schools level, one factor was retained. These findings suggest that the theoretical structure of the first two dimensions of the VfSL is empirically valid. Practically, these findings indicate that the construct of classroom community at the school level is holistic and includes elements of both classroom culture and rigorous instruction. At the classroom level, however, the skills of building classroom culture and enacting rigorous instruction are distinct.

# CHAPTER V

This dissertation began by exploring the current state of measures of teacher quality, finding a significant gap in the inclusion of aspects of cultural competence in current valid and reliable measures. Based on this finding, Chapter III explored the Vision for Student Learning (VfSL) as a possible measure that includes aspects of cultural competence. Chapter IV then examined the validity and reliability of the VfSL through a multilevel exploratory factor analysis and multilevel, multidimensional graded response model. This chapter presents key findings and recommendations based on the data presented in Chapters II, III, and IV. This analysis guided by the four original research questions:

1. What are the features of the Vision for Student Learning (VfSL)?
2. What does the VfSL tell us about teachers' practices?
3. What evidence do we have that the VfSL is a valid instrument?
4. What are the psychometric properties of the VfSL?

## KEY FINDINGS AND RECOMMENDATIONS

First, *cultural competence is a key domain of expertise required of high-quality teachers that is noticeably lacking in current, widely-adopted measures of teacher quality*. While existing measures of teacher quality are varied and complex in nature, those that enjoy widespread adoption are strong when assessing classroom management and delivery of academic instruction, but noticeably lacking in their inclusion of aspects of cultural competence. While new measures that include cultural competence are emergent, they do not yet enjoy adoption at scale, nor have they undergone analysis for psychometric validity and reliability.

Second, *the Vision for Student Learning (VfSL) is a measure of classroom community that includes aspects of cultural competence and has potential to be adopted at scale* (RQ1). The VfSL makes a shift away from evaluative measures of teacher quality that lack aspects of cultural competence; instead, the VfSL positions itself as a measure of *classroom community*, encompassing aspects of cultural competence. Mapped against Ladson-Billings' culturally responsive pedagogy framework, as well as Lee's Cultural Molding Framework and Banks' Multicultural Education Framework, items in the VfSL include aspects of each major domain in the three theoretical frameworks. The VfSL does not put as strong of an emphasis on connecting classroom content to students' culture, however, which should be made more explicit across the measurement instrument in future iterations.

Third, *the VfSL provides useful information about teachers' classrooms using student-level descriptors* (RQ2). Items on the VfSL, which are written from the perspective of student actions, not teacher actions, are useful to measure the construct of classroom community. Given the promising findings from EFA and IRT models, the VfSL shows that items focused on classroom actions taken by students are able to effectively measure the construct of classroom community. In fact, this shift toward student-level descriptors is one of the key reasons the VfSL positions itself as a formative measure of classroom community as opposed to an evaluative measure of teacher quality. By using student-level descriptors, the VfSL does not directly measure the act of teaching, nor the person of the teacher – it is no longer a measure of teacher nor teach*ing* quality. Instead, by using student actions as a proxy for teacher actions, the VfSL encompass the acts of the entire classroom community in its measurement construct.

Because the VfSL is a measure of the entire classroom community and not simply the teacher, however, it is especially important to emphasize the intent of this observation rubric to

be used as a formative, not evaluative, measure. By expanding the observational frame beyond the actions of the teacher to encompass the entire classroom, it is no longer measuring any act nor ability inherent to the teacher alone.

Instead, the VfSL on the acts of the entire classroom community, which could be confounded by any number of variables depending on the time of day, week, or year in which the classroom is observed. Whether the classroom is observed at the beginning of the day or at the end of the day, whether it is an honors or advanced class, the number of students in a particular classroom with individualized education plans (IEPs) – all of these factors could affect the rating of a particular classroom on any given day at any given time. As a result, this measure finds its strength in providing formative feedback for a teacher on the observed state of his or her classroom on that particular day, at that particular time, with a particular group of students. Given the number of factors that could affect a classroom's rating on any given day, this measure should never be used for evaluative purposes.

Fourth, *the VfSL is sensitive with regard to years of teaching and demographic variables* (RQ2). The data demonstrate sensitivity across demographic groups, and the potential ability of the VfSL to point program leaders toward differentiated professional development opportunities, especially among novice teachers. The VfSL as currently written is best positioned to measure novice teaching ability, providing the most information at the lower range of possible scores. If revised to include items that increase the ability of the VfSL to measure the higher ranges of possible scores, it could potentially serve as a useful rubric to observe more experienced teachers. As written, however, the tool is best positioned for formative use in novice teacher classrooms.

It will also be important in future research to determine whether sensitivity across demographic groups is actually a function of differences amongst the teachers or a function of potential rater bias. While one study found that teachers' use of rubrics when assessing student performance evaluations decreases bias by constraining teachers' possible responses (Jonsson & Svingby, 2007), another study of responses from a rubric-based evaluation of university professors reported that the gender of the professor influences students' responses (Laube et al, 2007). Assessments of elementary students' reading abilities have also been found to show racial and gender bias (Kranzler & Miller, 1999). While IRT analyses of item bias have been conducted for a variety of psychological and cognitive scales (Cole et al, 2000; Hunter & Schmidt, 2000; Jane et al, 2007; Teresi et al, 1995; Uebelacker et al, 2009), analyses of gender and racial bias using ratings from widely-adopted measures of teacher development have not been conducted.

Fifth, *the structure of the Safe, Brave, and Equitable Classrooms dimension, as well as the Rigorous and Culturally Relevant Learning dimension are empirically valid* (RQ3). The factor structure of the VfSL differs at the between-schools and within-schools level, with two factors present at the within-schools level and one factor present at the between-schools level. Based on this factor structure, policy makers and program leadership should think of the construct of classroom community as a holistic construct at the school level; at the classroom level, however, classroom culture and rigorous instruction present as two distinct skills. As a result, coaching and professional development should focus on developing both of these skills to improve overall classroom quality.

Sixth, *the VfSL should be revised, removing the Opportunity, Daily Mastery, Perseverance, and Resources items* (RQ3 and RQ4). All four indicators exhibited problematic

cross-loadings in the factor analysis. Furthermore, both the Perseverance and Resource items would not converge when fitting the item response model, indicating poor item fit. While the Opportunity item would converge, it provided low levels of information to the total information function. The Daily Mastery indicator also had a problematic information function, providing inconsistent levels of information across the range of possible outcomes. As a result, all four items should be removed from the VfSL. Since the Yearly Progress item would then be the only remaining item in the Perseverance to Goals dimension, this analysis supports moving it to the Rigorous and Culturally Relevant Learning dimension; the Yearly Progress item loaded onto the same factor as the other items in this dimension.

It is important to note, however, what is lost by the removal of these items. Table 9 outlines the changes in the number of VfSL items that address cultural competence markers from the theoretical literature. While the marker with the greatest change was Gloria Ladson-Billings' academic achievement, the VfSL still has fairly strong representation across her three domains of culturally relevant pedagogy. Instead, the tools appear to be most weakened with regard to promoting connections between classroom content and student culture, as well as promoting an empowering school culture – markers that were already sparsely represented.

Removal of the Perseverance to Goals dimension does not mean that the items within – which sought to measure students' daily mastery of content, students' perseverance in the face of challenges, and their ability to seek out additional resources to support their learning – should not be measured or are not important aspects of culturally competent instruction. Instead, this finding simply means that measuring these aspects of culturally competent instruction through an observation rubric is not the best method. This is not entirely surprising; the chances of observing a majority of students persevering through a task or seeking out additional resources

over the course of a single classroom observation is slim. Utilizing other methods to capture such actions, such as student mindset inventories, student journals, long-term course portfolios and projects, or qualitative interviews with teachers and students, are potential avenues to capture information about these aspects of instruction that are perhaps better suited to the task than a classroom observation rubric.

*Table 9. Change in Number of VfSL Items that Address Cultural Competence Markers*

| | Prior Number of VfSL Items | Current Number of VfSL Items | Difference |
|---|---|---|---|
| GLB: Academic Achievement | 9 | 5 | -4 |
| GLB: Cultural Competence | 7 | 5 | -2 |
| GLB: Socio-political Consciousness | 8 | 6 | -2 |
| Lee: Culture as a Resource for Learning | 5 | 4 | -1 |
| Lee: Connections between Content and Culture | 2 | 1 | -1 |
| Lee: Deep Knowledge of Content Domain | 4 | 3 | -1 |
| Banks: Content Integration | 6 | 4 | -2 |
| Banks: Knowledge Construction | 4 | 3 | -1 |
| Banks: Prejudice Reduction | 4 | 3 | -1 |
| Banks: Equity Pedagogy | 9 | 7 | -2 |
| Banks: Empowering School Culture | 4 | 2 | -2 |

While these changes don't negate the VfSL as a measure of classroom community that includes aspects of cultural competence, they do provide design opportunities for future iterations of the instrument. As items are revised or added to increase the ability of the instrument to measure the full range of the construct, these two aspects of cultural competence should be focus areas for item design. Other areas with lower representation – such as deep knowledge of the content domain, knowledge construction, and prejudice reduction – also serve as design guideposts that could increase the ability of the VfSL to measure classroom community at the upper range of the construct.

Seventh, *the VfSL is a psychometrically valid measure of classroom community* (RQ4). Thresholds for the VfSL items that would converge were all in order and broadly covered the continuum of possible outcomes. Items provided the most information at the lower ranges of possible scores; classrooms, however, generally received high scores on the VfSL. This discrepancy indicates that the VfSL appears to best measure classrooms in the lower to average ranges of possible scores. As a result, the VfSL may either be revised to include items that better measure the upper ranges of the construct of classroom community, or be constrained in use to measure the growth and development of novice teachers who are establishing the foundations of classroom culture and rigorous instruction.

Psychometric validity does not ensure ecological validity, however. As mentioned previously, many issues could affect the ratings that teachers receive during an observation, such as the time of day during which the observation took place, social and political issues playing out in the community outside of school, and student social-emotional factors. While some of these factors can be controlled or standardized through protocols like observing all teachers at a fixed time of day, or minimized with a large sample size, many of them cannot. One way in which the VfSL accounts for this variation is by grouping score categories by ranges rather than absolute targets (e.g., having the top score of 5 represent observing 80-100% of student actions aligned to the item, rather than a fixed, rigid requirement of 100%). These actions alone, however, might be insufficient to ensure ecological validity. As the features and validity of the VfSL are continually investigated and refined, special attention should also be paid to the ecological validity of the data collected for psychometric analysis.

# CONCLUSION

This study explores the necessity, validity, and feasibility of including cultural competence as a dimension of the construct of quality teaching and classroom community. Through an analysis of existing measures, cultural competence was found to be missing across widely-adopted measures of teacher quality; in contrast, the literature on cultural competence in education, as well studies documenting the importance and validity of including cultural competence in measures of quality medical education, point to the importance of expanding the dominant construct of teacher quality to include aspects of cultural competence. Statistical and psychometric analyses of the Vision for Student Learning establish the reliability and validity of a measure which includes cultural competence as part of its construct for classroom community, providing initial evidence that inclusion of cultural competence in observation rubrics is both possible and desirable when used strictly to provide formative feedback.

Moving forward, these analyses suggest an evolution not only in the instruments that are used to measure classroom community but also in way in which such instruments are implemented. While most classroom observation rubrics are written from the vantage point of teacher actions, writing items from the vantage point of students is also a valid means to capture the quality of teaching in a classroom. This perspective shift also provides the benefit of shifting the gaze of the observer toward how students are interacting and experiencing their classroom, a shift in perspective that exemplifies the push to develop teachers' ability to teach in a culturally competent manner. As a result of this shift, however, it is inappropriate to use such measures for evaluative purposes. Constructing a measure that bases observer ratings on the actions of the entire classroom community positions the measure to provide strong formative feedback for teachers to better understand dynamics at play in their classrooms and make adjustments to their

own practice in order to better meet the needs of students. Such measures, however, are not focused solely on the actions or abilities of the teacher. Given the number of variables that could influence classroom ratings on any particular day (e.g., time of day, number of advanced students in a particular class, number of students with IEPs in a particular class), observation scores should never be used in a high-stakes, evaluative manner.

Measures that include cultural competence possess the ability to help teachers grow in their craft, rather than simply be evaluated as effective or not effective. As the importance of developing teachers' cultural competence is brought into increasing focus, having measurement tools that help teachers gauge their progress and make adjustments is critical. Such tools also equip principals and coaches to provide differentiated professional development to their teachers in order to facilitate growth. Moving forward, evaluative and high-stakes measures should give way to formative measures that aid teachers in their development while supporting the realization of classroom environments in which all students are able to thrive.

Researchers and administrators should remain cautious, however, to ensure that any observed difference in teaching quality across lines of race and gender are not the function of biased items within the observation instrument or biases of the raters themselves. As districts and schools adopt measures of classroom community that include aspects of cultural competence at scale, instrument designers and researchers must partner with districts and schools to conduct IRT analyses of rubric ratings focused on detecting item or rater bias. The VfSL shows preliminary evidence of sensitivity toward differing levels of novice teacher skill. As adoption increases, however, additional analyses should be conducted to ensure that VfSL items do not exhibit racial or gender bias.

## REFERENCES


Adler, J. (2001). Teaching mathematics in multilingual classrooms. Boston: Kluwer academic publishers.

Alexander, K. L., Entwisle, D. R., & Thompson, M. S. (1987). School performance, status relations, and the structure of sentiment: Bringing the teacher back in. *American Sociological Review,* 52(5),665-682.

Alim, H. S. (2004). *You know my steez: An ethnographic and sociolinguistic study of styleshifting in a Black American speech coml1lunity.* Durham, NC: Duke University Press.

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B., & Pianta, R. (2013). Observations of effective teacher-student interactions in secondary school classrooms: Predicting student achievement with the classroom assessment scoring system-secondary. *School Psychology Review*, *42*(1), 76.

Andrade-Duncan, J., & Morrell, E. (2008). *The art of critical pedagogy.* New York: Peter Lang.

Balch, R. T. (2012). *The validation of a student survey on teacher practice.* (Doctoral dissertation, Vanderbilt University).

Ball, D. L., & Bass, H. (2003). Toward a practice-based theory of mathematical knowledge for teaching. In B. Davis & E. Simmt (Eds.), Proceedings of the 2002 annual meeting of the Canadian Mathematics Education Study Group. Edmonton, AB: CMESG/GCEDM.

Banks, J. A. (1991). The dimensionsof multiculturaleducation. *Multicultural Leader*, 4, 5-6.

Banks, J. A. (1992). Multiculturaleducation: Approaches, developments, and dimensions. In J. Lynch, C. Modgil, & S. Modgil (Eds.), *Cultural diversity and the schools, Vol. 1,*

*Education for cultural diversity: Convergence and divergence* (pp. 83-94). London: The Falmer Press.

Banks, C. (2005). *Improving multicultural education: LessollS from the intergroup education movement.* New York: Teachers Col- lege Press.

Benjamin, W. J. (2002). Development and Validation of Student Teaching Performance Assessment Based on Danielson's Framework for Teaching.

Boston, M. (2012). Assessing instructional quality in mathematics. *The Elementary School Journal*, *113*(1), 76-104.

Carter, P. & Darling-Hammond, L. (2016). "Teaching Diverse Learners." *Handbook on Research and Teaching*, edited by Courtney Bell and Drew Gitomer, Washington, D.C.: *American Educational Research Association.*

Cobb, P., & Hodge, L. (2002). A relational perspective on issues of cultural diversity and equity as they play out in the mathematics classroom. Mathematical Thinking & Learning 4, no, 2 & 3, 249-284.

Cohen, J., & Grossman, P. (2016). Respecting complexity in measures of teaching: Keeping students and schools in focus. *Teaching and Teacher Education*, *55*, 308-317.

Cohen, J., Schuldt, L. C., Brown, L., & Grossman, P. (2016). Leveraging Observation Tools for Instructional Improvement: Exploring Variability in Uptake of Ambitious Instructional Practices. *Teachers College Record*, *118*(11), n11.

Cole, S. R., Kawachi, I., Maller, S. J., & Berkman, L. F. (2000). Test of item-response bias in the CES-D scale: experience from the New Haven EPESE study. *Journal of clinical epidemiology*, *53*(3), 285-289.

Cornelius, L. L., & Herrenkohl, L. R. (2004). Power in the classroom: How the classroom environment shapes students' relationships with each other and with concepts. *Cognition and Instruction*, *22*(4), 467-498.

Danielson, C. (2011). *Enhancing professional practice: A framework for teaching*. ASCD.

Danielson, C. (2011). Evaluations that help teachers learn. *Educational Leadership*, *68*(4), 35-39.

Danielson, C. (2013). *The framework for teaching: Evaluation instrument*. Princeton, NJ: Danielson Group.

Darling-Hammond, L. (1999). Teacher quality and student achievement: A review of state policy evidence. Seattle, WA: Center for the Study of Teaching and Policy.

Delpit, L. D. (1988). The silenced dialogue: Power and pedagogy in educating other people's children. Harvard Educational Review, 58 (3), 280-297.

Delpit, L. (1995). Other people's children: Cultural conflict in the classroom. New York: New Press.

Downer, J. T., Booren, L. M., Lima, O. K., Luckner, A. E., & Pianta, R. C. (2010). The Individualized Classroom Assessment Scoring System (inCLASS): Preliminary reliability and validity of a system for observing preschoolers' competence in classroom interactions. *Early childhood research quarterly*, *25*(1), 1-16.

Downer, J. T., López, M. L., Grimm, K. J., Hamagami, A., Pianta, R. C., & Howes, C. (2012). Observations of teacher–child interactions in classrooms serving Latinos and dual language learners: Applicability of the Classroom Assessment Scoring System in diverse settings. *Early Childhood Research Quarterly*, *27*(1), 21-32.

Emdin, C. (2010). Affiliation and alienation: Hip-hop, rap, and urban science education. *Journal of Curriculum Studies, 42(1),* 1-25.

Fisher, M. (2007). *Writing in rhythm: Spoken word poetry in urban classrooms.* New York: Teachers College Press.

Ferguson, R. (2008). The tripod project framework. *The Tripod Project*.

Ferguson, R. F., & Danielson, C. (2014). How framework for teaching and tripod 7Cs evidence distinguish key components of effective teaching. *Designing teacher evaluation systems*, 98-143.

Feistritzer, C. (2011). *Profile of teachers in the U.S. 2011.* Washing- ton, DC: National Center for Education Information.

Freire, P. (1968). *Pedagogy of the oppressed*. New York: Seabury Press.

Gay, G. (2000). Culturally responsive teaching: Theory, research, and practice. New York: Teachers College Press.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The Test Matters The Relationship Between Classroom Observation Scores and Teacher Value Added on Multiple Types of Assessment. *Educational Researcher*, 0013189X14544542.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education. 119*(3), 445-470.

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., ... & Brackett, M. A. (2013). Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *The Elementary School Journal*, *113*(4), 461-487.

Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical Knowledge for Teaching and the Mathematical Quality of Instruction: An exploratory study. Cognition and Instruction, 26, 430-511.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56-64.

Hunter, J. E., & Schmidt, F. L. (2000). Racial and gender bias in ability and achievement tests: Resolving the apparent paradox. *Psychology, Public Policy, and Law*, *6*(1), 151.

Ingersoll, R. M. (2002). Out-of-field teaching, educational inequity and the organization of schools: An exploratory analysis. (Research Report No. R-02-1). Seattle, WA: Center for the Study of Teaching and Policy, University of Washington.

Irvine, J. (2003). *Educating teachers for diversity: Seeing with a cultural eye.* New York: Teachers College Press.

Jane, J. S., Oltmanns, T. F., South, S. C., & Turkheimer, E. (2007). Gender bias in diagnostic criteria for personality disorders: an item response theory analysis. *Journal of Abnormal Psychology*, *116*(1), 166.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review*, *2*(2), 130-144.

Junker, B. W., Weisberg, Y., Matsumura, L. C., Crosson, A., Wolf, M., Levison, A., & Resnick, L. (2005). *Overview of the instructional quality assessment*. Regents of the University of California.

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

KewalRamani, A., Gilbertson, L., Fox, M. A., & Provasnik, S. *(2007). Status and trends in the education of racial and ethnic minorities* (NCES 2007-039). Washington, DC: National Cen- ter for Education Statistics.

Kimball, S. M., White, B., Milanowski, A. T., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education*, *79*(4), 54-78.

Kinloch, V. (2005). Revisiting the promise of "students' right to their own language": Pedagogical strategies. *College Composition and Communication,* 57(1), 83-113.

Kranzler, J. H., Miller, M. D., & Jordan, L. (1999). An examination of racial/ethnic and gender bias on curriculum-based measurement of reading. *School Psychology Quarterly*, *14*(3), 327.

Kumas-Tan, Z., Beagan, B., Loppie, C., MacLeod, A., & Frank, B. (2007). Measures of cultural competence: examining hidden assumptions. *Academic Medicine*, *82*(6), 548-557.

Ladson-Billings, G. (1994). The Dreamkeepers: Successful teachers of African American children (1st ed.). San Francisco: Jossey-Bass Publishers.

Ladson-Billings, G. (1995). Making mathematics meaningful in multicultural contexts. In W. G. Secada, Fennema, Elizabeth, Byrd Adajian, Lisa (Ed.), New directions for equity in mathematics education (pp. 126-145): Cambridge University Press.

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the prekindergarten year. *The Elementary School Journal*, 409-426.

Laube, H., Massoni, K., Sprague, J., & Ferber, A. L. (2007). The impact of gender on the evaluation of teaching: What we know and what we can do. *NWSA Journal*, *19*(3), 87-104.

Lazarev, V., Newman, D., & Grossman, P. (2013). Developing an Aggregate Metric of Teaching Practice for Use in Mediator Analysis. *Society for Research on Educational Effectiveness*.

Learning Mathematics for Teaching Project. (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, *14*, 25-47.

Lee, C.D. (2007). Culture, Literacy, and Learning: Taking Bloom in the Midst of the Whirlwind. New York: Teachers College Press

Lee, J., & Wong, K., (2004). The Impact of Accountability on Racial and Socioeconomic Equity: Considering both school resources and achievement outcomes. American Educational Research Journal 41(4), 797-832.

Matsumura, L. C., Slater, S. C., Junker, B., & Peterson, M. (2006). Measuring reading comprehension and mathematics instruction in urban middle schools: A pilot study of the Instructional Quality Assessment (CSE Technical Report 681).

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education*, *79*(4), 33-53.

Moll, L. C., Amanti, C., Neff, D., & Gonzalez, N. (1992). Funds of knowledge for teaching: Using a qualitative approach to connect homes and classrooms. *Theory Into Practice,* 31(2), 132-141.

Moschkovich, J. (2002). A situated and sociocultural perspective on bilingual mathematics learners. *Mathematical Thinking and Learning*, *4*(2-3), 189-212. doi:10.1207/s15327833mtl04023_5

Muthén, L. K., & Muthén, B. O. (1998-2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén.

Nieto, S. (2000). *Affirming diversity: The sociopolitical context of multicultural education* (3rd ed.). New York: Longman.

Nougaret, A. A., Scruggs, T. E., & Mastropieri, M. A. (2005). Does teacher education produce better special education teachers?. *Exceptional Children*, *71*(3), 217-229.

Oakes, J. (1983). Tracking and ability grouping in American schools: Some constitutional questions. Teachers College Record, 84(4), 801-819.

O'Connor, M. C., & Michaels, S. (1993). Aligning academic task and participation status through revoicing: Analysis of a classroom discourse strategy. *Anthropology & Education Quarterly*, *24*(4), 318-335. doi:10.2307/3195934

Pakarinen, E., Lerkkanen, M. K., Poikkeus, A. M., Kiuru, N., Siekkinen, M., Rasku-Puttonen, H., & Nurmi, J. E. (2010). A validation of the classroom assessment scoring system in Finnish kindergartens. *Early Education and Development*, *21*(1), 95-124.

Pianta, R. C., Karen, M., Paro, L., & Hamre, B. K. (2008). *Classroom assessment scoring system (CLASS) manual, pre-K*. Baltimore, MD: Paul H. Brookes Publishing Company.

Polikoff, M. S. (2014). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education*, *121*(2), 183-212.

Powell, R., Cantrell, S. C., Malo-Juvera, V., & Correll, P. (2016). Operationalizing Culturally Responsive Instruction: Preliminary Findings of CRIOP Research. *Teachers College Record*, *118*(1).

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests.(Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Raudenbush, S. W., & Jean, M. (2014). To what extent do student perceptions of classroom quality predict teacher value added. *Designing teacher evaluation systems*, 170-202.

Ready, D., & Wright, D. L. (2010). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal,* 48(2), 335- 360.

Rist, R. (1970). Student social class and teacher expectations: The self-fulling prophesy in ghetto education. Harvard Educational Review, 40(3), 411-451.

Schultz, S. E., & Pecheone, R. L. (2014). Assessing Quality Teaching in Science. *Designing Teacher Evaluation Systems*, 444-492.

Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. *Educational Evaluation and Policy Analysis*, *36*(3), 259-280.

Sleeter, C (2001). Preparing teachers for culturally diverse schools: Research and the overwhelming presence of Whiteness. *Journal of Teacher Education.* 52(2),94-106.

Song, K. H. (2006). A conceptual model of assessing teaching performance and intellectual development of teacher candidates: A pilot study in the US. *Teaching in Higher Education*, *11*(2), 175-190.

Stevens, S. S. (1946). *On the theory of scales of measurement*. Bobbs-Merrill, College Division.

Tate, W. F. (1997a). Equity, mathematics reform, and research. Journal for Research in Mathematics Education, 28, 650-651.

Teresi, J. A., Golden, R. R., Cross, P., Gurland, B., Kleinman, M., & Wilder, D. (1995). Item bias in cognitive screening measures: comparisons of elderly white, Afro-American, Hispanic and high and low education subgroups. *Journal of clinical epidemiology*, *48*(4), 473-483.

Uebelacker, L. A., Strong, D., Weinstock, L. M., & Miller, I. W. (2009). Use of item response theory to understand differential functioning of DSM-IV major depression symptoms by race, ethnicity and gender. *Psychological medicine*, *39*(4), 591-601.

Whitehurst, G. (2002). Scientifically-based research on teacher quality: Research on teacher education and professional development. Paper presented at the White House Conference on Preparing Tomorrow's Teachers.

Wildhagen, T. (2012). How teachers and schools contribute to racial differences in the realization of academic potential. *Teachers College Record,* 114(7),1-27.

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Routledge.

Windschitl, M., Thompson, J., Braaten, M., & Stroupe, D. (2012). Proposing a core set of instructional practices and tools for teachers of science. *Science Education*, *96*(5), 878-903. doi:10.1002/sce.21027

| Level | | | | Respondents | Responses to Items | Example |
|---|---|---|---|---|---|---|
| 6 | Thematically-centered | Community | Entrenched discourse community; **established classroom discourse norms** that are consistently met; engage in discourse modalities (verbal, written, or representational) with one or more peers in order to respond to puzzling questions **connected to overarching course themes** | Most advanced use of discourse in classrooms that connects to broader course goals | Teacher uses discourse strategies to engage students in critical thinking around an overarching course theme | Teacher regularly has students write evidenced-based explanations using models to explain real-world phenomena |
| 5 | Evidence-centered | | Established discourse community; students aware of and **regularly meet classroom discourse norms**; engage in discourse modalities (verbal, written, or representational) with one or more peers in order to craft **evidence-based explanations** | Advanced forms of discourse in classrooms focused on warranting claims | Teacher uses discourse strategies to prompt students to provide evidence-based explanations | Teacher regularly has students write evidence-based explanations using models to content-based, academic questions |
| 4 | Content-centered | | Emerging discourse community; **emerging discourse norms**; students engage in discourse modalities (verbal, written, or representational) with one or more peers around a **content topic** | Beginning stages of utilizing constructivist discourse strategies with content material | Teacher uses discourse strategies to engage students in a discussion around content | Teacher has students write constructed responses |
| 3 | Probing | Dyadic | Classroom discourse remains **teacher-driven**; third turn of talk expands beyond evaluation and feedback to **begin uncovering student thinking** with questions such as 'why?' | Moving toward reform discourse during instruction; beginning to make student thinking visible | Responses begin in the IR pattern and end in a third turn that prompts students to make their thinking visible | T: What route did you take to school today?<br>S: [response]<br>T: Why did you take that route? |
| 2 | Feedback | | **Initiation-Response-Feedback structure**; similar to level one with an expansion in the third turn to providing feedback around, rather than evaluation of, student responses | Traditional classroom discourse; moving beyond evaluation to provide feedback | Teacher-centered discourse conforming to typical IRF pattern | T: Why did you divide by 2?<br>S: [response]<br>T: Good thinking. Another way you could do that is [x]. |
| 1 | Evaluation | | **Initiation-Response-Evaluation structure**; very little writing or use of visual modes; infrequent peer-to-peer talk | Traditional classroom discourse; likely mirrors teachers' own classroom experience | Teacher-centered discourse conforming to typical IRE pattern | T: What is the capitol of TN?<br>S: Nashville<br>T: Correct |
| 0 | Lecture | | All teacher talk; **no student discourse** | Lecture-driven instructors | Teacher uses no discourse moves beyond lecturing | Teacher talks for the entire class |

# Safe, Brave, and Equitable Classrooms

*so that students are affirmed in their identity, can take risks, and grow confidence*

| Student Indicators | Scale | | | | | | |
|---|---|---|---|---|---|---|---|
| | *N/O* | *None* | *Few* | *Some* | *About Half* | *Most* | *Almost All* |
| | **N/O** | **0%** | **1-20%** | **21-40%** | **41-60%** | **61-80%** | **81-100%** |
| **Openness**: Students affirm and express curiosity about the lived experiences, ideas, and opinions of others; they exchange ideas and beliefs in an open-minded way that affirms the value or dignity of others. | | | | | | | |
| **Expectations**: Students independently follow behavioral expectations and directions, requiring little or no direction/narration. | | | | | | | |
| **Routines**: Students assume responsibility for routines and procedures, executing them independently and efficiently, requiring little or no direction/narration. | | | | | | | |
| **Engagement:** Students spend class time on learning with minimal to no missed opportunities. | | | | | | | |
| **Inclusion**: In group work, students include others, express empathy when people are excluded or mistreated because of their identities; they speak up and take action when they themselves or others experience bias. | | | | | | | |
| **Additional sources of evidence, which could consist of:** *Frequency & Type of Student Action; Number of Students Engaged in Action & Strategies Used; Description of a Teacher-Student or Student-Student Interaction; Teacher/Student Language* | | | | | | | |

## Teaching As Leadership Capacities for Safe, Brave, and Equitable Classrooms

| Self as Leader | Building Relationships | Culturally Responsive Pedagogy | Academic Expectations |
|---|---|---|---|
| Teachers understand their own cultural background and how their cultural background is similar or different to students' cultural backgrounds.<br><br>Teachers balance and manage their own emotional responses to student behavior.<br><br>Teachers depersonalize student behavior. | Teachers demonstrate awareness of the relationship dynamics between students and between teacher and students.<br><br>Teachers actively seek to learn about the cultural backgrounds and interests of their students<br><br>Teachers build relationships with students and share about themselves.<br><br>Teachers respond consistently to students as a way of building trust. | Teachers set and communicate age appropriate and culturally responsive behavior expectations.<br><br>Teachers give clear what to do directions and subsequent positive narration.<br><br>Teachers use age appropriate and culturally responsive redirection and/or consequences to respond to misbehavior. | Teachers connect how academic practices have an impact on classroom culture.<br><br>Teachers set and communicate a high expectation for what this class will accomplish together.<br><br>Teachers treat class time as a precious resource. |

# Rigorous and Culturally Relevant Learning

*so that students challenge their worldview and make progress to high academic expectations*

| Student Indicators | Scale | | | | | | |
|---|---|---|---|---|---|---|---|
| **Opportunity:** Students take advantage of opportunities to formally and informally express learning through writing and/or explanations. | *N/O* **N/O** | *None* **0%** | *Few* **1-20%** | *Some* **21-40%** | *About Half* **41-60%** | *Most* **61-80%** | *Almost All* **81-100%** |
| **Discourse**: Students ask and answer questions in peer-to-peer discourse related to content. | *N/O* **N/O** | *None* **0%** | *Few* **1-20%** | *Some* **21-40%** | *About Half* **41-60%** | *Most* **61-80%** | *Almost All* **81-100%** |
| **Evidence**: Students provide aligned oral or written evidence to support their thinking and/or can generate an informed opinion about a subject. | *N/O* **N/O** | *None* **0%** | *Few* **1-20%** | *Some* **21-40%** | *About Half* **41-60%** | *Most* **61-80%** | *Almost All* **81-100%** |
| **Connections**: In their work and responses, students are making sense of their social, cultural, political, and historical position in the world. | *N/O* **N/O** | *None* **0%** | *Few* **1-20%** | *Some* **21-40%** | *About Half* **41-60%** | *Most* **61-80%** | *Almost All* **81-100%** |
| **Additional sources of evidence, which could consist of:** *Frequency & Type of Student Action; Number of Students Engaged in Action & Strategies Used; Description of a Teacher-Student or Student-Student Interaction; Teacher/Student Language* | | | | | | | |

## Teaching As Leadership Capacities for Rigorous and Culturally Relevant Learning Experiences

| Self as Leader | Building Relationships | Culturally Responsive Pedagogy | Academic Expectations |
|---|---|---|---|
| Teachers recognize the limits of their own perspective about how to set-up and evaluate student learning and seek outside perspectives.<br><br>Teachers recognize how their actions in the classroom can both support and/or inhibit students' engagement in learning. | Teachers pose meaningful questions that drive discussion, challenge beliefs, and/or support students making connections.<br><br>Teachers use conversation starters, sentence stems, accountable talk routines, etc… to support students sharing and discussing. | Teachers design lessons and materials that affirm and respect students' lived experiences, ideas, and opinions.<br><br>Teachers create opportunities for students to examine social, cultural, political, and historical contexts through content.<br><br>Teachers generate rich tasks aligned to students' interests that affirm their funds of knowledge and meet or exceed the rigor of the state standards.<br><br>Teachers affirm and support the learning of all students by incorporating multiple learning modalities throughout lessons. | Teachers set and communicate expectations for exemplar student work and student discourse.<br><br>Teachers ask and follow up on purposeful checks for understandings.<br><br>Teachers explore and bring social, cultural, political and historical content connections to their classrooms. |

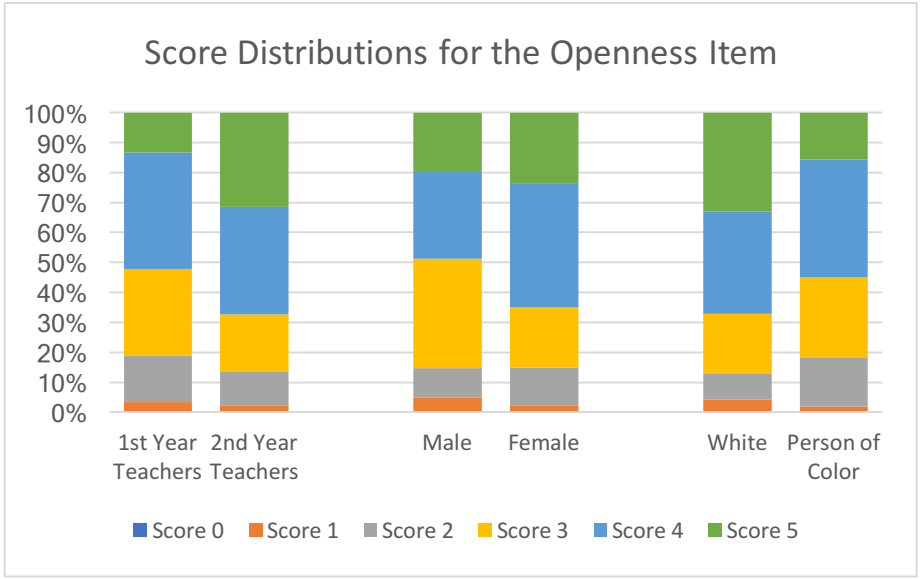
# Perseverance To Goals

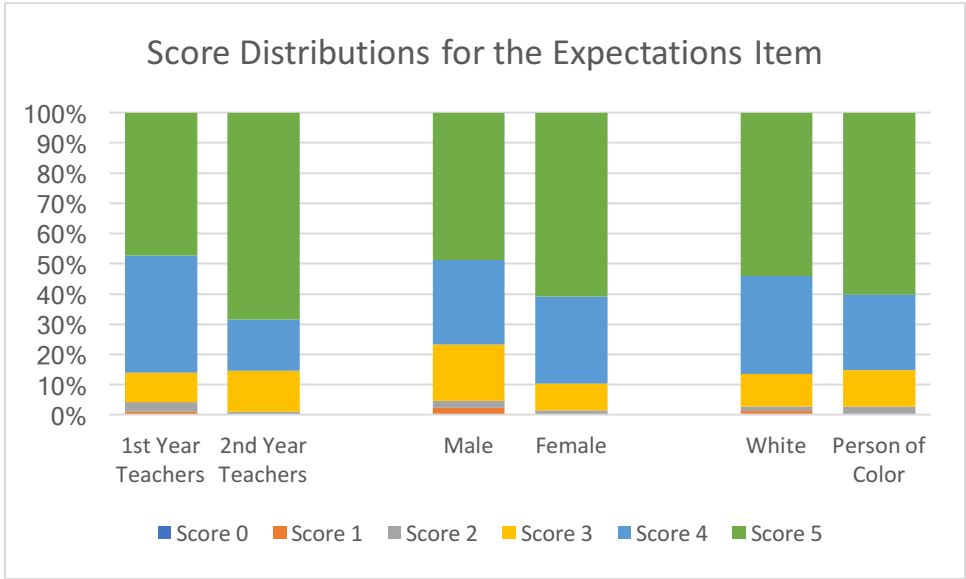
*so that students are best equipped to navigate their future academic and life trajectory*

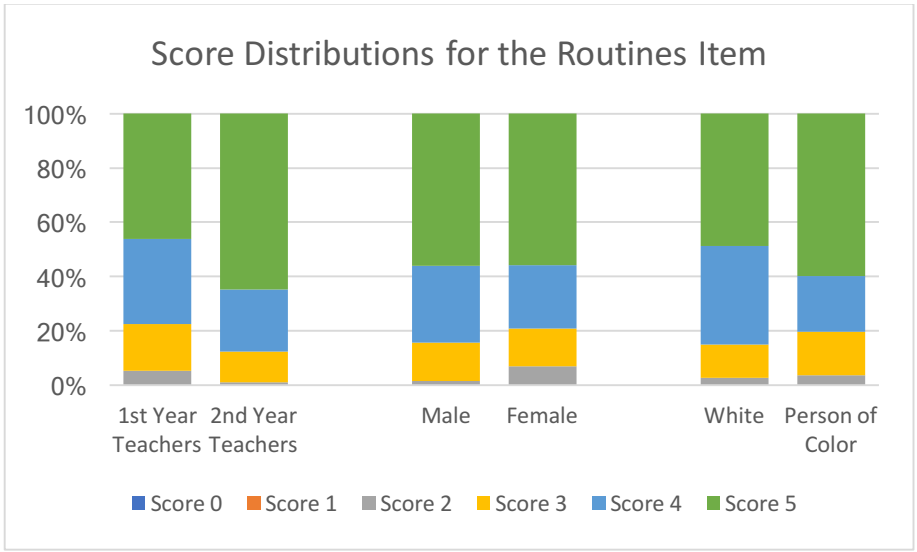
| Student Indicators | Scale | | | | | | |
|---|---|---|---|---|---|---|---|
| **Yearly Progress**: Students monitor and can name evidence to demonstrate progress toward their own academic goals and developing cultural competence and socio-political consciousness. | *N/O*<br>**N/O** | *None*<br>**0%** | *Few*<br>**1-20%** | *Some*<br>**21-40%** | *About Half*<br>**41-60%** | *Most*<br>**61-80%** | *Almost All*<br>**81-100%** |
| **Daily Mastery**: Student responses, work, and interactions demonstrate that they are on track to master the lesson objectives. | *N/O*<br>**N/O** | *None*<br>**0%** | *Few*<br>**1-20%** | *Some*<br>**21-40%** | *About Half*<br>**41-60%** | *Most*<br>**61-80%** | *Almost All*<br>**81-100%** |
| **Perseverance**: When faced with challenges, students reassess and determine a new course of action in order to meet their academic goals and/or continue to develop their cultural competence, and socio-political consciousness. | *N/O*<br>**N/O** | *None*<br>**0%** | *Few*<br>**1-20%** | *Some*<br>**21-40%** | *About Half*<br>**41-60%** | *Most*<br>**61-80%** | *Almost All*<br>**81-100%** |
| **Resources**: Students seek out and/or make connections to resources that support their success inside (e.g. peers, online tools, notes, etc.) and outside of school (e.g. tutoring, enrichment, mentors). | *N/O*<br>**N/O** | *None*<br>**0%** | *Few*<br>**1-20%** | *Some*<br>**21-40%** | *About Half*<br>**41-60%** | *Most*<br>**61-80%** | *Almost All*<br>**81-100%** |
| **Additional sources of evidence, which could consist of:** *Frequency & Type of Student Action; Number of Students Engaged in Action & Strategies Used; Description of a Teacher-Student or Student-Student Interaction; Teacher/Student Language* | | | | | | | |

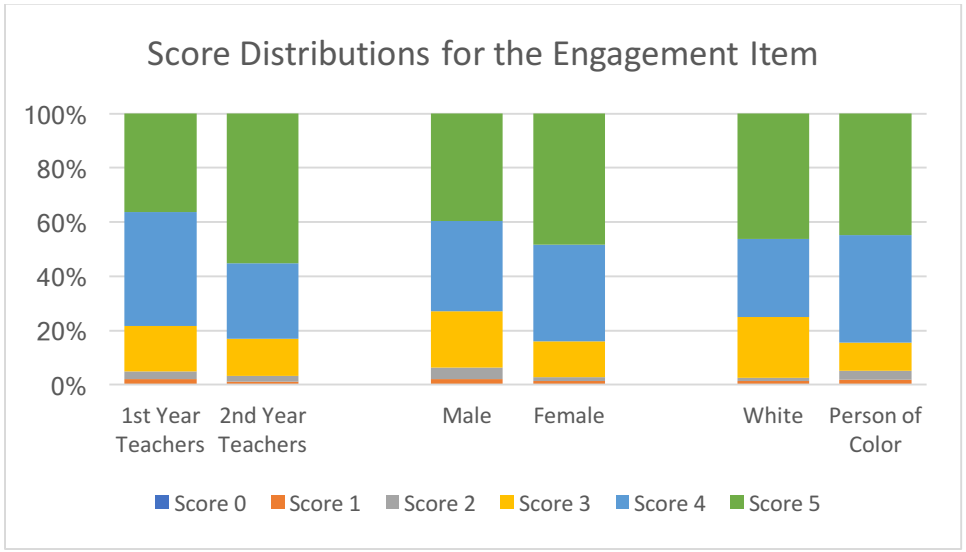
## Teaching As Leadership Capacities for Perseverance to Goals

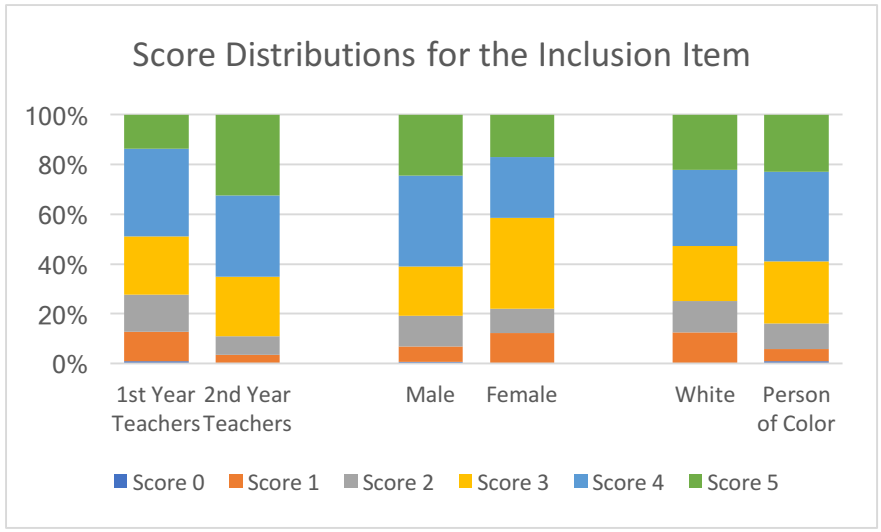
| Self as Leader | Building Relationships | Culturally Responsive Pedagogy | Academic Expectations |
|---|---|---|---|
| Teachers reflect on what changes they make in their practice to improve the situation for students in their class.<br><br>Teachers model and share what it looks like to have a growth mindset.<br><br>Teachers weigh positive and negative evidence equally when considering progress to goals. | Teachers set goals alongside students and their influencers.<br><br>Teachers regularly communicate and partner with students' influencers to best support students' progress.<br><br>Teachers invest students and their influencers in goals and in monitoring progress towards goals. | Teachers set goals that are culturally and academically responsive to students in their classroom.<br><br>Teachers implement systems for students and influencers to monitor and reflect on progress toward academic and personal goals.<br><br>Teachers celebrate successes with individual students and with groups of students.<br><br>Teachers connect students to resources that can best help them navigate the educational, employment, and life opportunities ahead. | Teachers guide goal setting with students by drawing on rigorous expectations and an acknowledgement of students' starting place academically.<br><br>Teachers regularly share academic feedback - both positive and negative - with students and influencers.<br><br>Teachers spend class time with students' making sense of growth and setting goals for the future. |

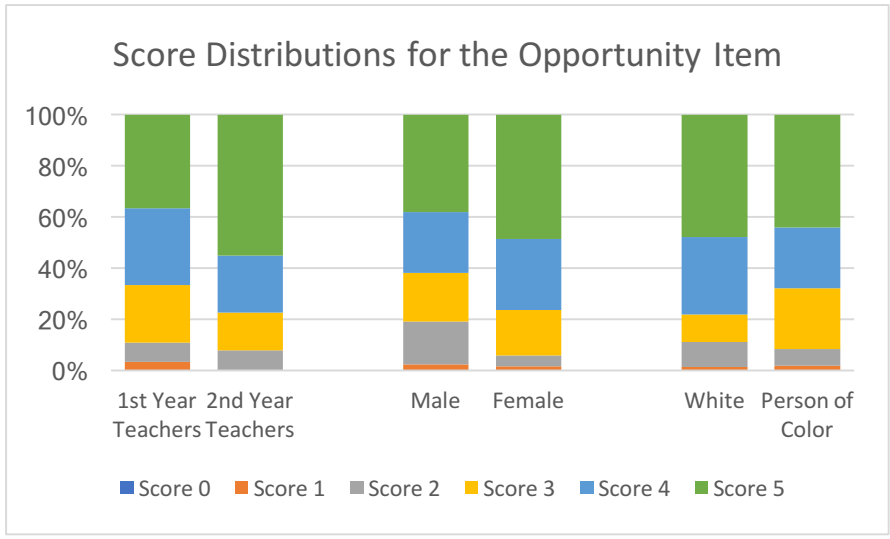

Score Distributions for the Openness Item
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1st Year Teachers
2nd Year Teachers
Male
Female
White
Person of Color
Score 0
Score 1
Score 2
Score 3
Score 4
Score 5

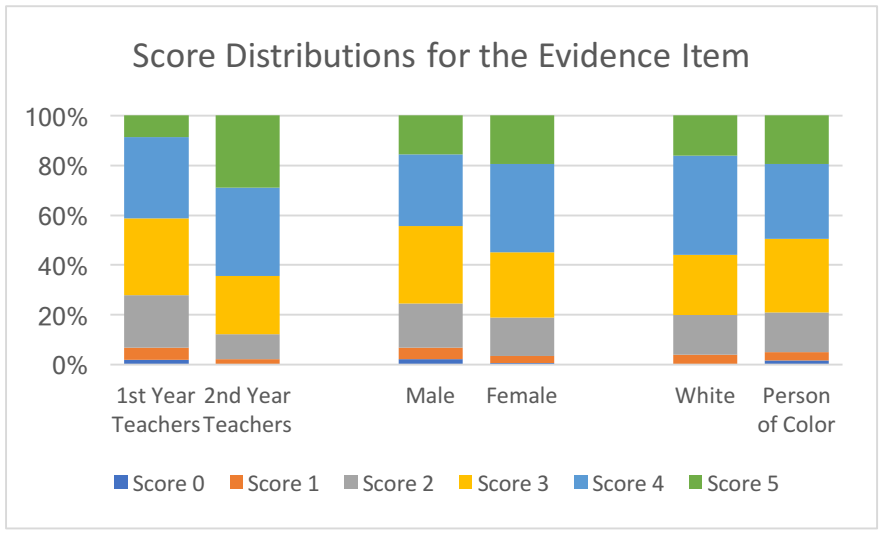

Score Distributions for the Expectations Item
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1st Year Teachers
2nd Year Teachers
Male
Female
White
Person of Color
Score 0
Score 1
Score 2
Score 3
Score 4
Score 5

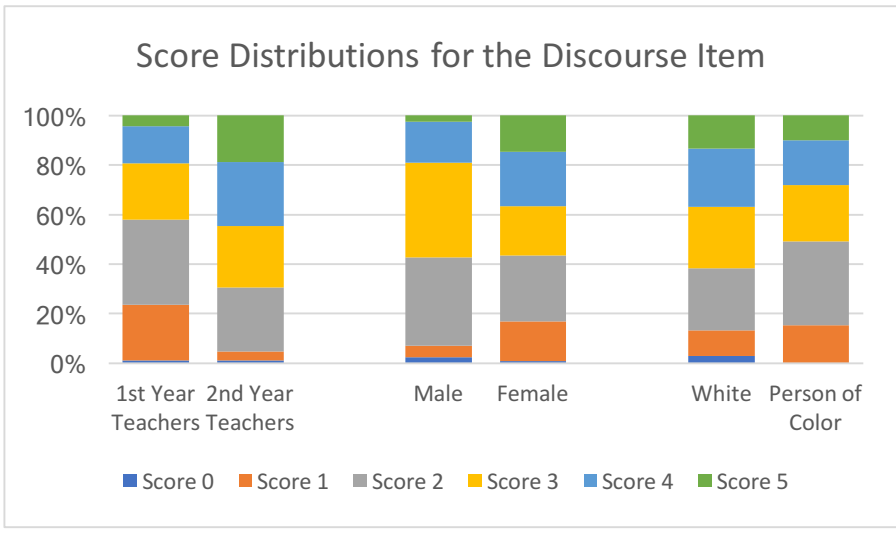

Score Distributions for the Routines Item
100%
80%
60%
40%
20%
0%
1st Year Teachers
2nd Year Teachers
Male
Female
White
Person of Color
Score 0
Score 1
Score 2
Score 3
Score 4
Score 5

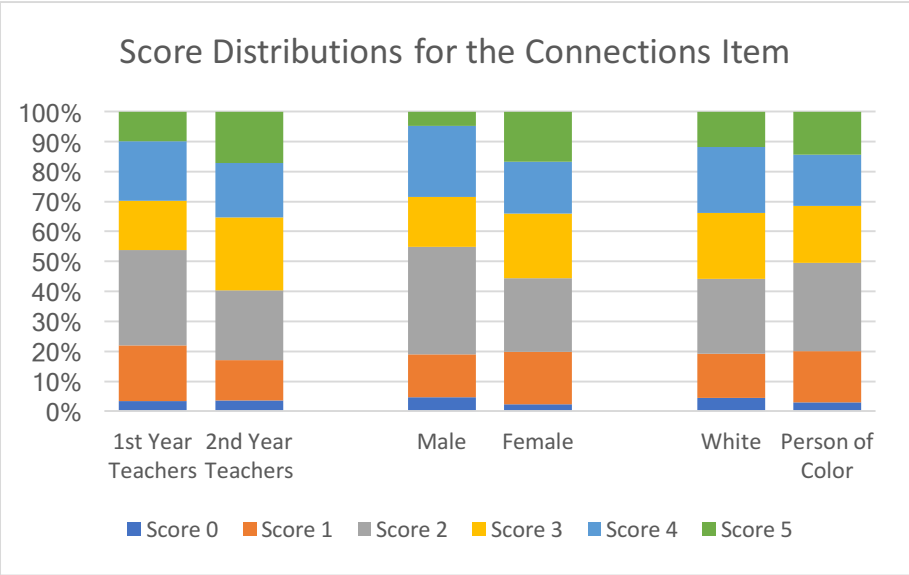

Score Distributions for the Engagement Item
100%
80%
60%
40%
20%
0%
1st Year Teachers
2nd Year Teachers
Male
Female
White
Person of Color
Score 0
Score 1
Score 2
Score 3
Score 4
Score 5

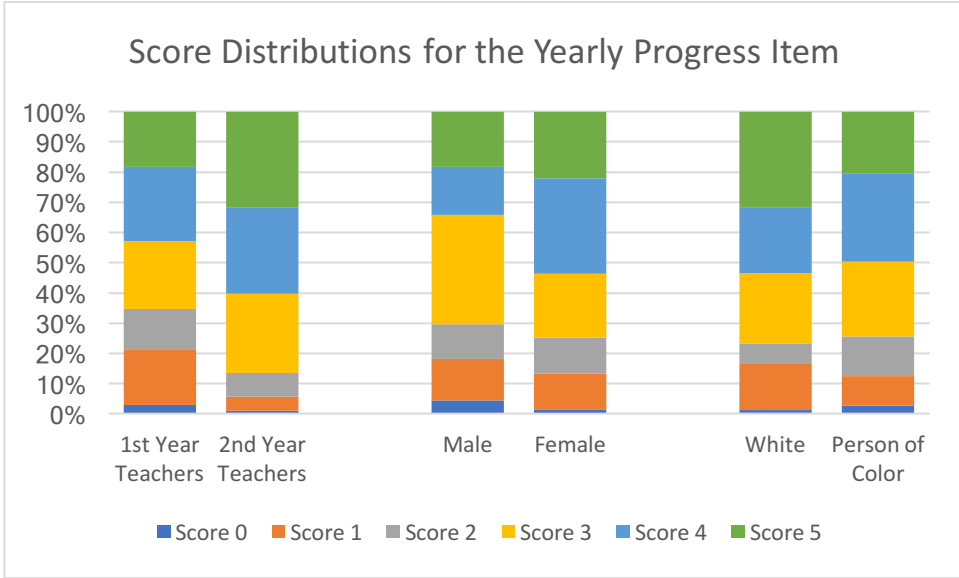

Score Distributions for the Inclusion Item
100%
80%
60%
40%
20%
0%
1st Year Teachers
2nd Year Teachers
Male
Female
White
Person of Color
Score 0
Score 1
Score 2
Score 3
Score 4
Score 5

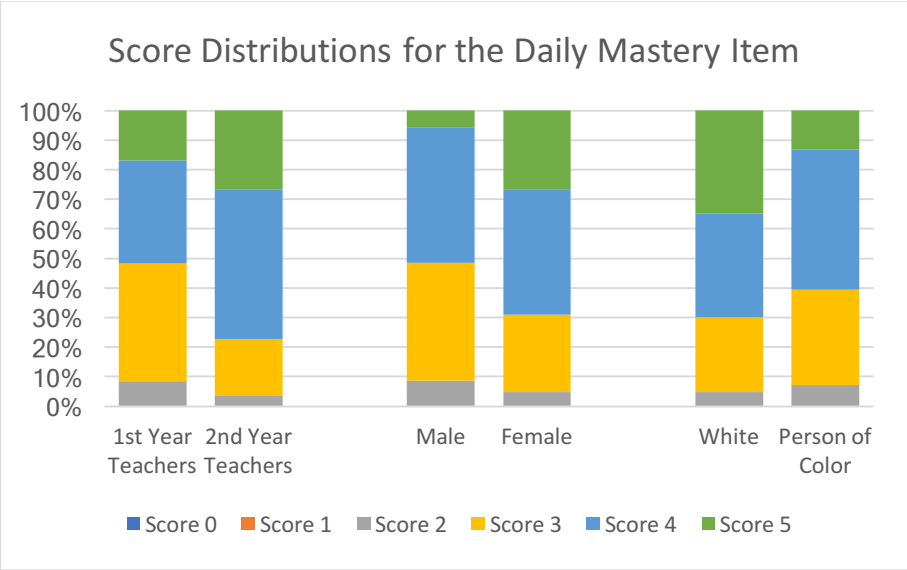

Score Distributions for the Opportunity Item
100%
80%
60%
40%
20%
0%
1st Year Teachers
2nd Year Teachers
Male
Female
White
Person of Color
Score 0
Score 1
Score 2
Score 3
Score 4
Score 5

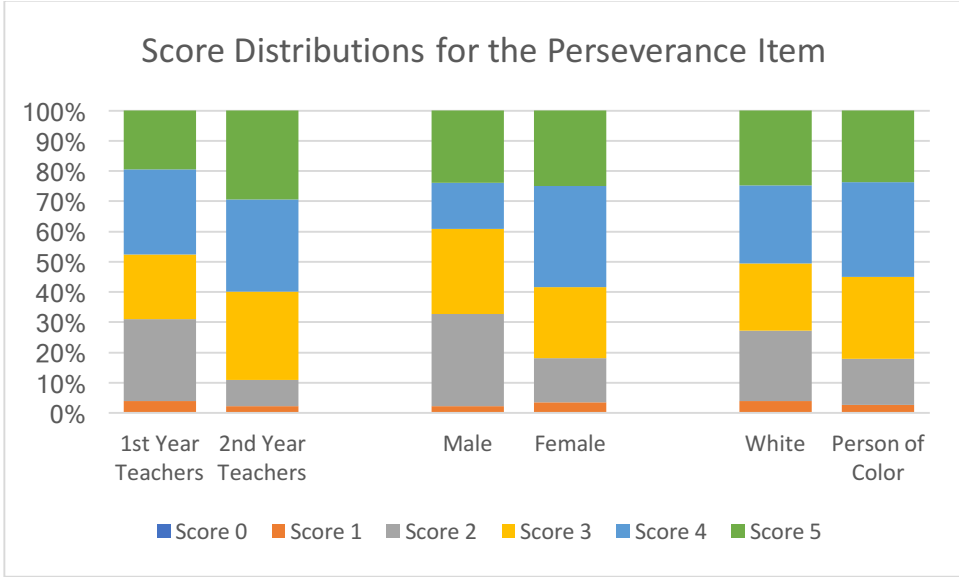

Score Distributions for the Evidence Item
100%
80%
60%
40%
20%
0%
1st Year Teachers
2nd Year Teachers
Male
Female
White
Person of Color
Score 0
Score 1
Score 2
Score 3
Score 4
Score 5

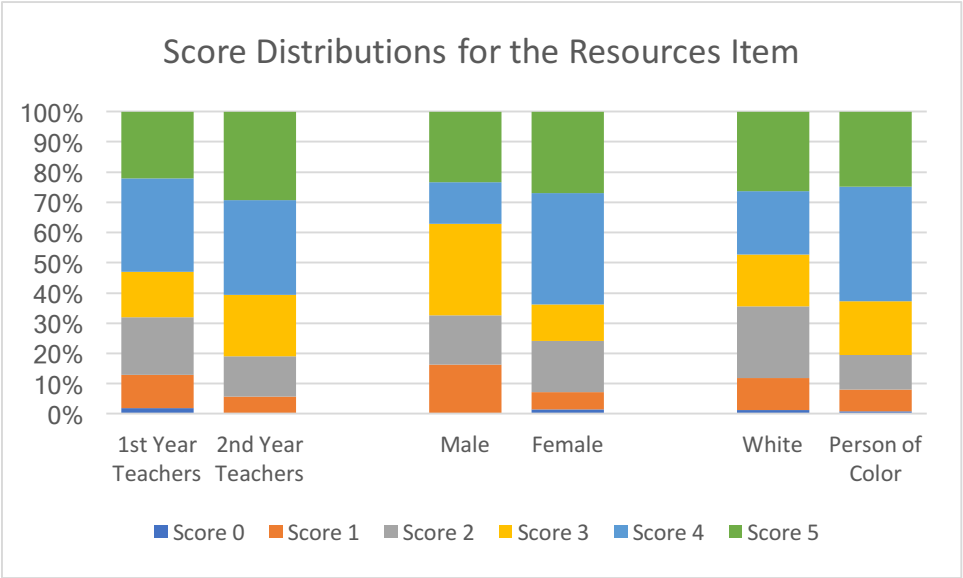

Score Distributions for the Discourse Item
100%
80%
60%
40%
20%
0%
1st Year Teachers
2nd Year Teachers
Male
Female
White
Person of Color
Score 0
Score 1
Score 2
Score 3
Score 4
Score 5


Score Distributions for the Connections Item
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1st Year Teachers
2nd Year Teachers
Male
Female
White
Person of Color
Score 0
Score 1
Score 2
Score 3
Score 4
Score 5


Score Distributions for the Yearly Progress Item
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1st Year Teachers
2nd Year Teachers
Male
Female
White
Person of Color
Score 0
Score 1
Score 2
Score 3
Score 4
Score 5


Score Distributions for the Daily Mastery Item
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1st Year Teachers
2nd Year Teachers
Male
Female
White
Person of Color
Score 0
Score 1
Score 2
Score 3
Score 4
Score 5


Score Distributions for the Perseverance Item
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1st Year Teachers
2nd Year Teachers
Male
Female
White
Person of Color
Score 0
Score 1
Score 2
Score 3
Score 4
Score 5


Score Distributions for the Resources Item
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1st Year Teachers
2nd Year Teachers
Male
Female
White
Person of Color
Score 0
Score 1
Score 2
Score 3
Score 4
Score 5