



University of Kentucky  
UKnowledge

---

Theses and Dissertations--Education Science

College of Education

---


2019

## VALIDATION OF A SCHOOL CLIMATE INSTRUMENT USING A RASCH RATING SCALE MODEL

Audrey Conway Roberts

University of Kentucky, [audreycroberts@uky.edu](mailto:audreycroberts@uky.edu)

Author ORCID Identifier:

 <https://orcid.org/0000-0002-1402-1887>

Digital Object Identifier: <https://doi.org/10.13023/etd.2019.088>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

---

### Recommended Citation

Roberts, Audrey Conway, "VALIDATION OF A SCHOOL CLIMATE INSTRUMENT USING A RASCH RATING SCALE MODEL" (2019). *Theses and Dissertations--Education Science*. 49.

[https://uknowledge.uky.edu/edsc\\_etds/49](https://uknowledge.uky.edu/edsc_etds/49)

This Doctoral Dissertation is brought to you for free and open access by the College of Education at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Education Science by an authorized administrator of UKnowledge. For more information, please contact [UKnowledge@lsv.uky.edu](mailto:UKnowledge@lsv.uky.edu).

## **STUDENT AGREEMENT:**

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

## **REVIEW, APPROVAL AND ACCEPTANCE**

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Audrey Conway Roberts, Student

Dr. R. Joseph Waddington, Major Professor

Dr. Margaret Bausch, Director of Graduate Studies

VALIDATION OF A SCHOOL CLIMATE INSTRUMENT USING A RASCH  
RATING SCALE MODEL

---

DISSERTATION

---

A dissertation submitted in partial fulfillment of the  
requirements for the degree of Doctor of Philosophy in the  
College of Education  
at the University of Kentucky

By

Audrey Conway Roberts

Lexington, Kentucky

Co- Directors: Dr. R. Joseph Waddington, Assistant Professor of Educational Policy  
Studies and Evaluation

and Dr. Kelly Bradley, Professor of Educational Policy Studies and Evaluation  
Lexington, Kentucky

2019

Copyright © Audrey Conway Roberts 2019  
[<https://orcid.org/0000-0002-1402-1887>]

## ABSTRACT OF DISSERTATION

### VALIDATION OF A SCHOOL CLIMATE INSTRUMENT USING A RASCH RATING SCALE MODEL

A new ESSA indicator of school quality and student success provides flexibility to broaden a states' definition of school and student success. Educational research has found school success is in part determined by a school's climate and should be considered in improvement/reform strategies (Cohen et al., 2009; Thapa et al., 2013). Yet, school climate research is often difficult and time consuming, and employs a variety of conflicting definitions and dimensions, instruments, and empirical approaches to determining school climate. Given these significant limitations with current measures, the purpose of this study was to validate an instrument measuring school climate based on the four most commonly accepted dimensions of school climate, using items adapted from a well-regarded and established theoretical framework to provide an effective measure for educators and researchers.

The sample selected for this study was a portion of teachers who indicated teaching 3<sup>rd</sup> or 8<sup>th</sup> grade as their primary teaching assignment (n=500) from the larger study sample (n=4974). A Rasch Rating Scale Model was used to evaluate unidimensionality, item fit and difficulty, reliability, and potential differential item functioning on a 23-item school climate survey. Results of the study showed the instrument was not unidimensional and was split into two subdimensions: student-centered and teacher/school support. All items were retained and displayed appropriate fit. Significant differential item functioning (DIF) was found between 3<sup>rd</sup> and 8<sup>th</sup> grade teachers on both subdimensions, further suggesting multidimensionality in the scale. Study findings suggest researchers should be mindful of any school climate instrument not validated at the item level for unidimensionality, and that an instrument may perform differently for teachers at different grade levels.

**KEYWORDS:** Rasch Rating Scale Model, School Climate, Differential Item Functioning, Survey Research, Education Policy

---

Audrey Conway Roberts

---

03/28/2019

---

Date

VALIDATION OF A SCHOOL CLIMATE INSTRUMENT USING A RASCH  
RATING SCALE MODEL

By  
Audrey Conway Roberts

---

Dr. R. Joseph Waddington  
Co-Director of Dissertation

---

Dr. Kelly Bradley  
Co-Director of Dissertation

---

Dr. Margaret Bausch  
Director of Graduate Studies

---

03/28/2019

Date

## ACKNOWLEDGMENTS

I would like to start this acknowledgement section by thanking my co-chairs, Dr. Kelly Bradley and Dr. R. Joseph Waddington. Dr. Bradley has been a source of advice and encouragement for many years. Even when I had my doubts, her guidance has helped to shape me into the researcher I am today. Dr. Waddington and I started at the same time in the Educational Policy Studies and Evaluation department, and because of that, I think we share a special bond. I have been fortunate enough to serve as his student, teaching assistant, and research assistant during my time as a doctoral student. My eternal thanks goes out to him for his patience, support, and commitment to my journey as a researcher these past years. Without the support and opportunities provided from both Dr. Bradley and Dr. Waddington, this dissertation would not have been possible.

I would also like to thank my other committee members, Dr. John Thelin and Dr. Wayne Lewis. I will miss the wonderful and thought-provoking conversations with Dr. Thelin and his seemingly endless knowledge of the history of higher education and my alma mater. Dr. Lewis, thank you for keeping me focused in the nature of educational policies and their implications in the classroom. Your flexibility to be available, even amidst your busy schedule does not go unnoticed. Special thanks to Dr. Amanda Potterton for serving as the outside examiner. I would be remiss if I didn't mention Dr. Shannon Sampson, who has long served as a research collaborator and person to bounce ideas with. Thank you for always taking the time to listen to my ideas, no matter how long-winded, and providing guidance along the way. I would also like to extend my thanks to Dr. Mark Berends, who, in collaboration with Dr. Waddington, lead the grant that made it possible for me to be funded as a doctoral student and gain invaluable experiences as a research team member.

I cannot imagine completing this journey without the support of my family and my friends. Mom and Dad, your unwavering love, support, and encouragement throughout my educational journey has made it possible for me to reach my ultimate educational goal. There is no way I could've done it without you two. To my UK friends: Abbey Love, Jie (Grace) Dai and Laura Carter-Stone, thank you so much for being shoulders to cry on, ears to vent frustrations to, and being available to just generally comment on the nuances of being a graduate student and a human being. To my newest friends, Blaire Gallagher and Daniel Leake, thank you for providing the best distractions from academia. My knowledge of sports, fitness, music, and cooking techniques have expanded from your friendships. To my oldest friends, McLane Crane, Devan Smith, and Kaitlyn Stevens, thank you all for your encouragement and for always keeping it real with me in life, parenting, and what it's like to be in an elementary school. To Julia Mahre, the best friend and travel buddy a gal could ask for, thank you for being only a phone call away. Your endless support, affirmations, and positivity about this process have been amazing.

To General Beauregard the cockapoo, we can't imagine our lives without you. Thank you for reminding me that sometimes you do just need to play to feel better. Last, but certainly not least, I am eternally grateful to my husband Thomas Roberts. I'm so thankful to have someone by my side who has been through this process. Your continual honesty, advice, and love have given me the motivation to make it to the finish line. Thank you for sharing this process and life with me.

## TABLE OF CONTENTS

ACKNOWLEDGMENTS .....	iii
LIST OF TABLES.....	vii
LIST OF FIGURES .....	viii
CHAPTER 1. INTRODUCTION .....	1
School Climate Research.....	2
Study Background.....	6
Research Questions.....	10
Organization of Dissertation.....	11
CHAPTER 2. LITERATURE REVIEW .....	12
School Climate and Accountability .....	12
Impact of Accountability on Teachers.....	19
Accountability Impact on Students.....	21
Accountability Policies in Indiana.....	22
Dimensions of School Climate .....	24
Unit of Analysis .....	27
School Climate Outcomes .....	29
Differences Between Schools.....	31
School Climate Instruments.....	33
<i>School Climate Measure (SCM)</i> .....	36
<i>Comprehensive School Climate Inventory</i> .....	37
<i>School Climate Assessment Instrument</i> .....	38
<i>Organizational Climate Inventory</i> .....	40
<i>Delaware School Survey</i> .....	41
<i>5Essentials</i> .....	42
Item Response Theory .....	45
Rasch Modeling .....	48
Rasch Rating Scale Model.....	50
Summary .....	52
Conclusion .....	53
CHAPTER 3. METHODOLOGY .....	54
Pilot Analysis.....	54



Study Overview .....	56
Sampling Procedures .....	57
Survey Data Collection.....	59
School Climate Instrument .....	59
Sample .....	63
<i>3<sup>rd</sup> and 8<sup>th</sup> grade selection</i> .....	63
<i>Rasch RSM Sample</i> .....	66
Data Analysis.....	66
Answering Research Questions .....	69
Conclusion .....	73
CHAPTER 4. RESULTS .....	74
Analysis Procedure .....	74
Sample .....	75
Initial Analysis.....	77
Student Centered Analysis.....	81
School/Teacher Support Analysis.....	87
Conclusion .....	93
CHAPTER 5. CONCLUSION.....	94
Overview.....	94
Findings .....	95
Limitations.....	104
Implications .....	106
Conclusion .....	112
APPENDIX.....	117
REFERENCES .....	119
VITA.....	132

## LIST OF TABLES

Table 1 <i>Breakdown of Participating Teachers</i> .....	59
Table 2 <i>Previous Construct Breakdowns</i> .....	62
Table 3 <i>Descriptive Statistics for Total Sample (n=500)</i> .....	76
Table 4 <i>Initial Analysis Variance Estimates</i> .....	77
Table 5 <i>Initial Loadings for First Contrast and Cluster Group</i> .....	79
Table 6 <i>Item Breakdown for Separate Rasch Analyses</i> .....	80
Table 7 <i>Analysis of Variance Estimates for the Student-Centered Subscale</i> .....	81
Table 8 <i>Item Estimates for the Student-Centered Subscale</i> .....	83
Table 9 <i>DIF Analysis of 3<sup>rd</sup> and 8<sup>th</sup> grade teachers for the Student-Centered Subscale</i> ..	86
Table 10 <i>Analysis of Variance Estimates for the School/Teacher Support Subscale</i> .....	87
Table 11 <i>Item Estimates for the School/Teacher Support Subscale</i> .....	89
Table 12 <i>DIF Analysis of 3<sup>rd</sup> and 8<sup>th</sup> grade teachers for the School/Teacher Support Subscale</i> .....	92

LIST OF FIGURES

Figure 1 *Graphic of the Five Essential Supports*..... 42  
Figure 2 *Initial Standardized Residuals for the First Contrast* ..... 78  
Figure 3 *Standardized Residuals for the First Contrast for the Student-Centered Subscale*  
..... 82  
Figure 4 *Wright Map of Item Difficulty for the Student-Centered Subscale*..... 84  
Figure 5 *Category Probabilities for the Student-Centered Subscale*..... 85  
Figure 6 *Standardized Residuals for the First Contrast for the School/Teacher Support  
Subscale* ..... 88  
Figure 7 *Wright Map of Item Difficulty for the School/Teacher Support Subscale*..... 90  
Figure 8 *Category Probabilities for the School/Teacher Support Subscale*..... 91

## CHAPTER 1. INTRODUCTION

President Obama signed the Every Student Succeeds Act (ESSA) into law on December 10, 2015, replacing the previous reauthorization (No Child Left Behind; NCLB) of the Elementary and Secondary Education Act (ESEA) (Department of Education, n.d.). This represented a shift from the policies of NCLB, particularly in the accountability measures states used to assess performance and progress in schools. For example, state education agencies (SEAs) were no longer required to submit Adequate Yearly Progress (AYP) reports or include student assessment scores in teacher evaluations (O'Brien, 2016). In response to critiques that school and student success measured by student achievement scores was too narrow, a new indicator of school quality and student success provided SEAs the flexibility to broaden a states' definition of success. Changes impacting SEAs and schools were expected as early as the 2016-2017 school year, with ESSA taking full effect in the 2017-2018 school year.

The Indiana Department of Education (IDOE) ESSA plan was approved by the U.S. Secretary of Education in January 2018. In an initial draft of the ESSA plan, the IDOE (2017) stated plans to

begin a pilot of culture and climate surveys with struggling schools, with the goal of producing a proposal for statewide implementation ... [after] wide agreement that those elements of school are both vital and challenging to measure ... either to support struggling schools or for accountability purposes" (p. 20).

However, in the final draft, it was recommended that future consideration was needed prior to implementing a climate or culture measure to the formal accountability system, citing a lack of validated measures. As such, students "demonstrating excellent or improved

attendance rates” will be used as a measure of school quality and/or student success in lieu of a school culture or climate scale (IDOE, 2018).

Indiana is not alone in recognizing the use of a school climate measure but failing to include one in final ESSA plans. In fact, only four states include a school climate measure as a federally reported indicator: Illinois, Maryland, Montana, and New Mexico. Although school climate surveys are popular with both researchers and policymakers, concerns existed both on lack of prior experience in administering large-scale school climate surveys, and that data could be manipulated to game the system (Melnick, Cook-Harvey, & Darling-Hammond, 2017; Stringer, 2017). Indeed, Gangi (2009) suggested that “choosing the most appropriate assessment tool” is key for schools that cannot afford to “waste time, energy, or funds” on invalid and ineffective measures (p. 4). However, just because SEAs do not require a school climate measure as a federal indicator of school quality or student success, does not mean that climate measures are not desired or being used for accountability purposes. As the IDOE’s case described above exemplifies, one of the largest and least addressed problems within the field of school climate is the lack of continuity in the instrument, scaling and treatment of data.

### **School Climate Research**

Over the past 35 years, educational research has demonstrated that school success is in part determined by a school’s climate and should be considered in improvement/reform strategies (Cohen, McCabe, Michelli, Pickeral, 2009; Thapa, Choen, Guffey, & Higgins-D’Alessandro, 2013). In the past, states attempting reform focused on objective features of schools, such as achievement scores, graduation and attendance rates, through additional testing and professional development, often relegating school climate under safety, bullying prevention, and health policies (U.S Department of Education,

2017). School climate is linked to students' social, behavioral, and learning outcomes (Zullig, Koopman, Patton, and Ubbes, 2014) as well as teachers' stress, effectiveness, and retention (Johnson, 2006; Loeb, Darling-Hammond, & Luczak, 2005). A positive school climate is associated with reduced bullying (Kosciw & Elizabeth, 2006; Meyer-Adams & Conner, 2008) and less violence/aggressive behaviors from students in schools (Goldstein, Young, & Boyd, 2008; Gregory et al., 2010) and impacts students' achievement in both the short-term and long-term (Hoy, Hannum, & Tschannen-Moran, 1998). Research suggested a positive school climate promotes student learning through cooperative learning, group cohesion, respect, and mutual trust (Thapa et al., 2013). von der Embse, Pendergast, Segool, Saeki, and Ryan (2016) suggested school climate serves as an indicator and buffer to rapidly changing educational policies on teacher outcomes. For teachers, a positive school climate influenced teachers' feeling of accomplishment and commitment to the profession, which are in turn predictive of teacher stress, burnout, and retention (Grayson & Alvarez, 2008; Johnson, 2006; Thapa et al., 2013).

The factors comprising school climate differ broadly due to the prevalence of multiple theoretical orientations and research purposes. Multidimensionality in operationalizing school climate is evident in the literature (Wang & Degol, 2016). This lack of consensus on the dimensions of school climate is problematic for policymakers and researchers and can undermine school improvement efforts (Cohen et al., 2009). Thapa et al. (2013) described the lack of a clear definition of school climate as a real and systemic issue in the field, which has "stymied and continues to stymie the advancement of school climate research" (p. 371).

For example, Cohen et al (2009) defined school climate as the "quality and character of school life ... based on patterns of people's experiences of school life and

reflects norms, goals, values, interpersonal relationships, teaching and learning practices, and organizational structures” (p. 182). The authors, after reviewing the larger school climate literature, suggested four major areas or elements of school climate are generally recognized: 1) relationships among those in the school, including students, parents, teachers, and administration, 2) physical safety, 3) teaching and learning, and 4) the institutional environment. Johnson, Stevens, and Zvoch (2007) measured school climate perceptions from teachers using five factors: collaboration, decision-making, instructional innovation, student relations, and school resources. Later, Zullig et al (2010) suggested five domains have historically been measured: order and safety, academic outcomes, social relationships, school facilities, and school connectedness (p. 41). In a theoretical paper, Rudasill, Snyder, Levinson, and Adelson (2017) proposed school climate is composed of the social interactions and relationships, safety-physical and emotional, and the values/beliefs of all members of the school, but not the structural aspects of a school. Additional school climate research has included knowledge and fairness of disciplinary policies and discipline problems, peer relationships among students, administrative leadership, and the perceived racial climate (Cheema & Kitsantas, 2014; Martín, Martínez-Arias, Marchesi, & Pérez, 2008; O’Malley, Voight, Renshaw, & Eklund, 2015; Tschannen-Moran, Parish, & DiPaola, 2006).

This lack of consistency extends in whom to survey on school climate. In instances where multiple members of a school community are surveyed, parallel instruments are inconsistent, making larger analyses complicated (Brand, Felner, Seitsinger, Burns, & Bolton, 2008). Yet, studies examining one perspective of school climate (e.g. students or teachers) as indicative of the larger school is problematic, as school climate may differ for certain groups within a school. Students tend to be the predominant way to assess school

climate, but data collection is time consuming and non-representative (Brand et al., 2008). Some researchers assert because teachers are the most directly responsible for instruction in the classroom, teachers' opinions on school climate are most important (Collie et al., 2012). School climate surveys given to principals often focus on organizational management and larger school-level variables (Back, Polk, Keys, McMahon, 2016).

Perhaps most problematic in school climate research is the dearth of empirical validation in climate measures (Brand et al., 2008). Zullig et al (2010) argued the most commonly used and adapted school climate instruments were developed many years ago, with current validation or peer review difficult to find, if not nonexistent. Wolfe, Ray, and Harris (2004) stated that "Many studies of teacher perception of influence, students, and school climate have been based on responses to national surveys ... [using] sums or averages of the Likert-type scale responses to the question clusters as the aggregate variable" (p. 845) without first assessing the degree of quality of the measure. Indeed, in a review of various published scholarly work, it is often unclear how scales were developed and if items have been validated before implementation.

Even though the need for complex statistical models was noted early on (Bryk & Raudenbush, 1992), most of the school climate literature fails to actually address the complex nature of school data, much less item or person validity. Using individual reports to predict individual outcomes violates data independence, but using school level reports to predict individual level outcomes disregards the heterogeneity within schools and within perceptions of climate themselves (Wang & Degol, 2016). Using multilevel models or reporting an average score with a confidence interval could address some concerns of both the nested and independent nature of school climate perceptions. However, this does not



address the validity of school climate measures themselves. Ultimately, the decision to aggregate school climate ratings depends on the research questions and purpose of a study.

In a meta-analysis of published school climate studies, Wang and Degol (2016) found 91% of the studies measured school climate through some empirical means. Of the nearly 300 empirical studies, nearly half (48%) employed a correlational design to associate school climate with other variables. Very few studies focused on the development and validation of a school climate instrument (15%) or used experimental or quasi-experimental methods (9%). To this point, the reliance on correlational data to link teacher and student perceptions of school climate, without indication of valid measures of school climate, to other variables as causal is problematic. As school climate is a complex and multifaceted topic, using correlation or factor analysis as the sole form of validation may overlook fundamental issues at the item level. For instance, a 60-item measure may possess high alpha reliability simply by virtue of having a large number of items. Yet a shorter measure may not accurately represent school climate either. By using an item response theory over a classical approach, analyzing an instrument at the item level takes into account both the individual and the quality of the items themselves as being an observable, operationalized response to an underlying trait or condition, providing a more effective instrument (Bond & Fox, 2015).

### **Study Background**

The School Effectiveness in Indiana (SEI) survey, developed in collaboration with partners at the University of Notre Dame, University of Kentucky, and NORC at the University of Chicago, was given to a sample of elementary and middle school teachers in the state of Indiana during the 2016-2017 school year. The study's purpose was to examine the conditions contributing to school effectiveness using the 5Essentials framework (Byrk

et al., 2010), to compare traditional public, magnet, private, and charter schools. Results aimed to provide guidance or suggestions for school leaders and policymakers for improvement efforts in Indiana schools.

The 5Essentials framework was developed in the mid-1990s by researchers at the University of Chicago Consortium on School Research at the University of Chicago Urban Education Institute in partnership with Chicago Public Schools and is a well-regarded framework for school improvement efforts. The five essentials are: effective leaders, collaborative teachers, involved families, a supportive environment, and ambitious instruction. Research has shown schools with strong scores on at least three of the five essential areas are ten times more likely to show positive gains in student learning over time than schools weaker in the same areas (UChicagoImpact, 2019).

A supportive environment is more than just classroom instruction. First, schools must work to provide a safe and engaging learning environment for students with clear expectations and rules. Second, schools must balance expectations for high academic achievement with support and reasonable goals. Social capital and a learning community mentality are important to develop, particularly in struggling schools. Last, schools should work to develop a cohesive curriculum that is aligned with appropriate standards. Taken together, these four components of a supportive environment in the 5Essentials framework represent an operationalized measure of school climate and can be measured. More than 5 million members of school communities have taken the 5Essentials Survey, including students, teachers, and parents in approximately 6,000 schools spanning the United States (UChicagoImpact, 2019). It is included in the NCSSLE Approved School Climate Surveys.

Using the 5Essentials framework, the Philadelphia school district in partnership with the University of Pennsylvania Graduate School of Education in 2014 developed a

survey for students, teachers, principals and parents/guardians. The five constructs developed were analogous with the 5Essentials framework: climate, instruction, leadership, professional capacity, and parent/guardian community ties (Office of Research and Evaluation, 2016). The Philadelphia school climate scale is important to the school climate items in the SEI survey; both stem from the work of the Chicago Consortium and the 5Essentials framework. As such, overlap exists between the two climate scales. Nearly two thirds (16 of 23) of the items from the Philadelphia scale are included in the SEI school climate portion.

Portions of this instrument have been previously analyzed. First, a 9-item instrument measuring the construct of institutional challenges produced an alpha reliability of 0.85 (Berends, Goldring, Stein, & Cravens, 2010). Second, a 60-item instrument was used to measure school climate in the Philadelphia School District (School District of Philadelphia Office of Research and Evaluation, 2016). In this larger measure, exploratory factor analysis was used to refine the measure. Alpha reliability for teachers on the six subscales ranged from 0.69 to 0.92. An underlying issue with combining the two scales is that both have been presented as measuring different constructs (institutional challenges and various subdimensions school climate). However, all three instruments share the same question stem: “To what extent do you consider each of the following factors *a challenge* to student learning in your classroom and/or school?” It is important to understand this instrument at the item level, to ensure that it effectively measures school climate, and not another construct.

Overall, because school climate research employs a variety of definitions, instruments, and empirical approaches to determining school climate, the literature abounds with findings and conclusions that are often contradictory and difficult to

replicate. Although most researchers and educators recognize the value of non-cognitive measures, the abundance of conceptual frameworks, instruments, and findings have flooded the field with a variety of empirical research that is often misaligned (Rudasill et al., 2017). In recent years, the larger school climate literature has moved to become more self-aware, with researchers addressing the messy nature of the literature (Cohen et al., 2009; Zullig et al., 2014). Ultimately, it is the schools and practitioners that suffer from a lack of consistency in school climate measures and findings (Rudasill et al., 2017). From a policy standpoint, in the case of Indiana and many other state ESSA plans, it is clear a current demand exists for appropriate, cohesive, and validated measures of school climate, and that this demand is going unmet.

Given these significant limitations with current measures, the purpose of this study was to validate an instrument measuring school climate based on the four most commonly accepted dimensions of school climate, using items adapted from a well-regarded and established theoretical framework to provide an effective measure for educators and researchers. The study utilizes teachers' perceptions of challenges to student learning to determine school climate. To validate this instrument, a Rasch analysis will be implemented. A Rasch analysis estimates an item's difficulty and a person's ability on the same continuum in the expressed on a common logit scale (Bond & Fox, 2015). Rasch models are invariant, meaning that a person's ability can be determined from the items, and an item's difficulty can be assessed from a person's ability, regardless of sample, providing validity and stability for future use (Royal & Elahi, 2011). Although the instrument was designed using items adapted from the 5Essentials framework, and items have been used in various contexts of school climate, there is no empirical validation that combining these items together will provide a sound, unidimensional measure.

Validating this scale is timely and has the potential for actual policy implications in the state of Indiana, while also being able to provide a validated school climate measure for states who are seeking school climate measures. Much of the work in school climate has predominantly been done in traditional public elementary schools, which creates a large gap in the perspectives of private, charter, and secondary teachers. It is important to assess to what degree the instrument is stable and can measure opinions of teachers in different school environments appropriately. In addition, there is a clear and stated need in Indiana for a measure of school quality and student success that was not met, stemming from a lack of reliable and validated measures. As the larger scale hopes to provide guidance or suggestions for improvement efforts in Indiana schools, it is critical that valid instruments are used to make informed decisions in school improvement efforts.

### **Research Questions**

The following three research questions were formed to assist in validating this school climate instrument:

- How well does the instrument measure the latent trait of school climate?
- How well do the individual instrument items reflective of school climate fit to the Rasch model?
- Do differences exist in item responses from teachers from 3<sup>rd</sup> and 8<sup>th</sup> grades with similar levels of the latent trait of school climate?

To answer the first research question, unidimensionality was assessed by conducting a principal components analysis (PCAR) of the Rasch residuals. To answer the second research question, item fit and item difficulty of the instrument was analyzed using infit and outfit statistics. To answer the third research question, a differential item functioning

analysis was conducted between teachers in elementary and middle schools, specifically 3<sup>rd</sup> and 8<sup>th</sup> grade teachers.

### **Organization of Dissertation**

This chapter serves as an introduction for this dissertation containing five chapters. Chapter Two presents a review of relevant literature regarding school climate, by first demonstrating how school climate literature has grown increasingly complex over time, and second by highlighting the differences in dimensions, participants, and instrument development and analysis. Differences in school level are briefly described, and the chapter concludes with a description of the method of data analysis used: the Rasch Rating Scale Model. Chapter Three provides the research design used, and includes information about the instrument, participant sampling, and data collection. Results from a pilot study are given and the chapter ends with an explanation of how the research questions in the dissertation are answered. Chapter Four contains the findings from data analysis. This includes descriptive analyses from participants and the results from the Rasch RSM analysis. Dimensionality, item fit, item difficulty, thresholds, and DIF estimates are presented. Chapter Five concludes the dissertation by reviewing the results for each research question. A discussion of the contributions of the study to the larger school climate literature is provided, followed by limitations, and implications from this study for policymakers, researchers, and practitioners.

## CHAPTER 2. LITERATURE REVIEW

This chapter reviews the literature regarding school climate, the policies surrounding school climate and accountability, and methodological literature related to item response theory, specifically the Rasch rating scale model. First, how school climate literature has changed over time will be discussed, situated in the larger education policies of the United States. A discussion on the differences in dimensions, participants, and instrument analysis with a selection of established measures will provide a context for school climate instruments. A brief review of past and current education policies in Indiana will be provided. Then, a history of school climate results is presented before highlighting differences in school level stemming from relevant education policies. This chapter concludes with a description of the primary method of analysis: a one-parameter item response theory model used for polytomous data, specifically the Rasch rating scale model.

### **School Climate and Accountability**

The first written accounts of the culture of school climate occurred near the beginning of the 20<sup>th</sup> century (see Perry, 1908; Dewey, 1927) and tended to focus on the theoretical or observable characteristics of classrooms and schools. School climate reemerged as a method of accountability in the 1950s as researchers began to systematically examine school and student attributes with instruments/assessments influenced by popular organizational theories of the time (Cohen et al., 2009). Urban and Waggoner (2008) wrote that the success of testing for soldiers during World War II led to school systems “developing elaborate bureaus of educational research whose major function was to purchase and administer the standardized tests that were believed to measure the educational potential and achievement of students” (p. 270). The first

educational climate instrument was used to measure the pressures perceived by college students as related to student performance (College Characteristics Index; CCI, Pace & Stern, 1958) with Henry Murray's (1938) theory that students would respond differently to their environments based on their own needs. From here, Murray's theory seemed to be the underlying theory as researchers turned their focus to the K-12 classroom.

In 1965, President Johnson signed into law the Elementary and Secondary Education Act (ESEA), creating the Title I program to address inequities in impoverished children and schools. This in conjunction with the National Defense Education Act (NDEA) ushered in a new era of federal government involvement in education, with a renewed focus on reform and standards. There was a strong belief that education could reduce the social-class divisions in America by reducing poverty (Spring, 2008). Reese (2011) aptly summarized that "seeking security in a world of uncertainty ... Americans in the postwar era again turned to the schools as a source of stability and a fulcrum of change" (p. 252). As a result of the increased federal focus on promoting education, the general public had now come to see education as critical to a strong nation, which increased the responsibilities of teachers and policymakers (Reese, 2011). Expectations for teachers were reflected in their pre-service preparation, who were taught that IQ scores and achievement tests were the best way to assess and track students, different students will require different pedagogies, and a more student-centered learning approach would be the best way to reach all students (Reese, 2011).

By the late 1960s and early 1970s, an increasingly negative view of government and public education developed. The economic situation in the U.S. was dismal and as well as reports of schools being a hotbed for violent activity were prevalent in the news (Reese, 2011). Researchers turned their focus to connecting school climate with student



outcomes, such as student achievement and student perceptions (Zullig et al., 2010). There was a growing sense among academics and reformers that the education system was much too rigid to educate an imaginative and socially conscious society. Pressure from prominent educational reformers at the time encouraged educators to focus more on the holistic perspectives of the child. However, public polls at the time showed that a majority of the public supported better discipline, higher standards, and a back-to-basics curriculum (Reese, 2011). Seeing the right conditions, conservative policymakers heavily promoted the findings of the Coleman Report, a 1966 document arguing students' socioeconomic status and family background were more important in achievement outcomes than school resources, as a way to argue against more federal involvement in education.

However, the 1983 report *A Nation at Risk*, called for urgent reform in schools to compete with a global population outscoring American students in academic achievement. Pulliam and Van Patten (2007) wrote that the report is often “credited with creating the momentum for educational reform, but it did not offer a model for high-quality education... concentrate[ing] on mechanical solutions [with] no means of implementing excellence while maintaining equality” (p. 253). The idea that teachers and schools should be held accountable for student outcomes is not a new idea in American education. But, accountability as a movement is credited to research and the growing negative public perception of school quality during this time period (Pulliam & Van Patten, 2007).

Spring (2008) offered that accountability served as a reaction to place control back into the hands of educators and away from community control. As a result, testing became a way to communicate progress to the community while firmly keeping power

within the realm of educators. Also, corporate leaders and policymakers “had come to believe that education did matter to the economy and (therefore) to them,” (McGill, 2015, p. 23). Corporate reform operates on the underlying logic that test scores measure education, and punishment for low scores will cause teachers to teach better, increasing test scores, making education better. Additionally, it was thought that applying privatization principles to schools (i.e. school choice) would promote innovation and provide a better model of schooling.

In the 1990s, the focus returned back to the idea of “press,” referring to the level of emphasis the members of the school place on values and practices that promote high academic performance and attempting to link school climate to a larger number of student outcomes (Gangi, 2009; Zullig et al., 2010). During this time, state governors and the federal government took on a larger role in setting education policies (Morrison, 2006). The 1994 reauthorization of ESEA introduced a commitment to standards-based reform and the idea of adequate yearly progress. However, consequences for not meeting performance gains were not made clear.

The No Child Left Behind Act (NCLB) was a 2001 reauthorization of the ESEA designed to “improve student achievement and change the culture of America’s schools” (U.S. Department of Education, 2001, p. 9). Simpson, LaCava, and Graner (2004) cited NCLB as “potentially the most significant educational initiative to have been enacted in decades” (p. 67). NCLB is based on four key principles: stronger accountability for achievement, greater financial flexibility for states with federal money, an emphasis on proven teaching methods (as demonstrated by research), and that parental choice [of schools] is good (U.S. Department of Education, 2001). Federal law now required all states to have an accountability system monitoring all public schools and their students.

States were given discretion on developing accountability measures, resulting in a wide array of content standards, assessments, and adequate yearly progress (AYP) standards (Dee et al 2010). Annual testing was required for each state in reading and mathematics for students in grades 3-8. Sanctions were added for chronically underperforming schools that did not meet AYP goals.

AYP goals were set with the ultimate goal for all students to achieve proficiency in 2014. Targets were not calculated by improvement, but rather by meeting set thresholds. Goals were set for both the total student population and for certain demographic subgroups. Failure to meet any of these goals could result in an entire school being labeled as failing. Apple (2004) suggested the policy shift from student needs to student performance had the unintended consequence of reallocating resources from the students that need them the most to those who are already succeeding. Most states reported increases in achievement scores. However, as each state created the standards and metric by which students are measured, it is unclear if true gains have been made. Dee et al (2010) found modest gains in mathematics scores for traditionally disadvantaged students in lower grades, but no clear gains found in reading. Additional criticisms of the high-stakes testing have included states lowering the difficulty of tests, teaching to the test at the expense of other subject areas, and cheating (Koyama, 2012).

Race to the Top (RTT) was a competitive grant program by the Obama administration beginning in 2009 that had a significant impact in school reform efforts. RTTT was seen as a remedy or response to the perceived failures of NCLB, without reauthorizing ESEA. States were awarded grants based on their policies regarding four areas of reform. These areas were: 1) standards and assessments that prepare students for college and the workplace; 2) recruiting, rewarding, and retaining effective teachers and

principals; 3) building informative data systems on instruction by measuring students' success; and 4) turning around low-achieving schools. (U.S. Department of Education, 2009).

Prior to RTT, grounds for teacher tenure/promotion were generally absent of any effectiveness evaluations and more about length of stay, despite evidence that reforms were necessary (Weisberg, Sexton, Mulhern, & Keeling, 2009). Previous research suggested traditional teacher evaluation systems did not provide differentiation between high and low performing teachers (Donaldson, 2009), nor did they provide constructive or informative feedback to teachers (Sinnema & Robinson, 2007). Steinberg and Donaldson (2016) reported that by the 2015-2016 school year, 88% of states and the District of Columbia had revised and/or created new teacher evaluation systems. Many of these new evaluation systems tied student achievement data directly to teacher evaluations, placing an incredible amount of pressure on teachers to produce high achievement scores. Neal & Schanzenbach (2010) warned that using value-added measures to determine such important decisions could increase teacher decisions to “teach to the test” or even cheat to secure positive outcomes.

Another hallmark of the RTT program was the adoption of Common Core State Standards (CCSS) in mathematics and English/language arts, which provide specific skills and knowledge that students should acquire and teachers with knowledge and guidance on the most important content to be taught. According to the Common Core State Standards Initiative (2010), these standards should provide a more coherent and focused curriculum and establish objective learning benchmarks for teachers and students to reach. Polikoff and Porter (2014) believed that with respect to content coverage in the

classroom, policy effectiveness would be determined by the alignment of teachers' instruction with the standards. As of 2016, more than 40 states had adopted the CCSS.

President Obama signed the Every Student Succeeds Act (ESSA) into law on December 10, 2015, replacing the previous reauthorization (No Child Left Behind; NCLB) of the ESEA (Department of Education, n.d.). Many of the state accountability provisions of NCLB remained. Students would be tested in grades 3-8 and scores broken down by school totals and demographic subgroups. A significant change came with the addition of a new indicator of school quality and student success. This fifth indicator provides greater flexibility for the standards and assessments states may use for accountability and gives states the ability to “reflect on their values and prioritization of characteristics beyond academic achievement” (Gohl, 2018). Measures permitted included those dealing with student or teacher engagement, student access or completion of advanced coursework, attendance or truancy percentages, postsecondary readiness, or school climate and safety (US DOE, 2017).

In the past, states focused on features of schools, such as raising achievement scores, graduation and attendance rates, through additional testing and professional development, (U.S Department of Education, 2017). With the addition of a new accountability indicator, a greater need exists for validated non-cognitive measures in school assessments (West, 2016). For decades, educational research has demonstrated that school success (in many aspects) is related to a school's climate and should be considered in improvement/reform strategies (Cohen et al., 2009; Thapa et al., 2013). Although school climate surveys are popular with both researchers and policymakers, concerns exist that too many frameworks exist for defining school climate measures, administering large-scale school climate surveys is too costly in comparison to other

accountability metrics, and that data could be manipulated (Melnick et al., 2017; Stringer, 2017).

### **Impact of Accountability on Teachers**

Accountability pressures may disproportionately affect some groups of teachers more than others (Grissom, Kalogrides, & Loeb, 2017). For instance, only teachers who teach a tested subject in a tested grade are considered when calculating the academic performance of a school through high-stakes standardized testing. This pressure may impact the types of administrative and instructional decisions principals and teachers make. Principals may structure hiring and firing teachers around their abilities to produce high or positive student achievement scores (Cohen-Vogel, 2011). Feng, Figlio, and Sass (2010) found teachers were more likely to leave schools with low accountability scores, and within those with the highest accountability pressure, higher quality teachers were more likely to leave than lower quality teachers. Ahn and Vigdor (2014) found higher levels of administrative turnover in schools with multiple years of failing scores.

In public elementary schools, students in Grades Pre K-2 are not evaluated annually with the same “high stakes” metric as students in Grades 3-5. In public middle schools, students in all grades are tested annually, although not all subjects are tested on annually. Chingos and West (2011) found that teachers rated as more effective were more likely to remain in grades where high-stakes testing occurred. Grissom et al (2017) suggested “clear incentives” for those assigning teachers to grades to prioritize more effective teachers to grades with more influence in a school’s rating (p. 1084). Mainly that teachers who are considered stronger should be placed in tested grades, so that a school’s rating has the potential to be as high as possible. Teachers who may not be as strong are often assigned to grades or subjects that are not tested. However, this

reassignment of weaker teachers to non-tested grades (K-2) could come at a cost for students in earlier grades. Grissom et al (2017) found lower performing teachers moved to low-stakes tested grades resulted in reduced student score gains on low-stakes measures for one large urban school district. By concentrating weaker teachers in non-tested grades, further implications could be made that students entering 3<sup>rd</sup> grade may not be as prepared as they should be for the high-stakes testing that begins in this grade, increasing the pressure on 3<sup>rd</sup> grade teachers to produce high scores.

Teachers may spend more of the school day on instruction, particularly in the subjects that will be tested, and provide extra support to students who are struggling (Rouse, Hannaway, Goldhaber, & Figlio, 2007). McDuffie et al (2017) found that many teachers expressed familiarity and professional development in mathematics standards. However, a majority of the teachers interviewed admitted having strong feelings of worry and concern their students would not be prepared for the state assessment and score poorly, and that this would impact their teaching evaluations. Many of the teachers expressed concerns with the quick implementation of changing standards, and that teachers had altered their instructional practices deliberately to accommodate and prepare for the tests. This supports earlier research that teachers may “teach to the test” when high-stakes testing is involved, focusing on more direct instruction on certain tested skills (Au, 2007; Palmer & Rangel, 2011). Yet, Edgerton, Polikoff, and Desimone (2017) found little evidence that teachers’ instructional choices are ever fully determined by state standards. For many teachers, the curve to sufficiently adapt instruction to align with evolving state standards can provide additional sources of stress, particularly because achievement results are frequently tied to teachers’ evaluations and personnel decisions.

## **Accountability Impact on Students**

Although the pressure to meet accountability standards primarily impacts teachers and school administrators through the threat of potential sanctions, this pressure is likely to spill over to students (Holbein & Ladd, 2017). Students have reported heightened levels of anxiety, boredom, anger, and motivation loss (Wheelock et al., 2000; Hoffman, Assad, & Paris, 2001). Student engagement, connected with school climate, is negatively impacted over time by accountability pressures (Markowitz, 2018). Student chronic absenteeism is a widespread problem in elementary and secondary schools and impacts student achievement, development, and engagement in school (Balfanz & Barnes, 2012). Gottfried (2019) found chronic absenteeism is damaging academically to both absent students and their classmates. Schools facing accountability pressures are more likely to report higher student attendance, but increased student misbehavior (Holbein & Ladd, 2017).

Carrell and Hoekstra (2009) found that disruptive elementary students have a statistically significant negative effect on the math and reading scores of their classmates. In addition, one disruptive student in a classroom increased the probability that other students would commit an infraction, creating a domino effect. Reported disciplinary incidents also notably increase between elementary and middle school, suggesting different challenges middle school teachers may face than elementary teachers. (Theriot & Dupper, 2010). Figlio (2006) found a relationship between high-stakes testing performance and suspension penalties in high school students. Higher performing students were given reduced suspension times during the testing window, while lower performing students committing the same action were suspended for longer, regardless of gender, race/ethnicity, or SES.



Misbehavior is disproportionately reported for students of color, particularly Black students. Particularly with student suspensions, many school districts have reformed their discipline codes to reduce the use of suspensions and promote more restorative justice practices (Eden, 2017). These policies have drastically reduced the number of student suspensions in many cases. However, what has not been equally measured is the impact that reducing suspensions has had on school climate. Eden (2017) found that under discipline reform, the number of suspensions mattered less for school climate than the dynamics surrounding the new policies. In other words, the change from implementing new policies that misbehaving students would not be suspended for their actions, contributed to an increase in disruptive student behaviors, and a general reduction in perceived school climate. Schools that served a majority minority population experienced the most rapid decline in perceived school climate after the change in discipline policies.

### **Accountability Policies in Indiana**

In the early 2000s (prior to NCLB), the Indiana legislature and the Indiana State Board of Education created and adopted performance-based accountability policies for Indiana schools (Public Law 221-1999). In 2011, this accountability system was revised to place schools along a grading scale from A-F (IDOEf, 2018). Under this accountability model, grades were assigned by a metric of comparing a preliminary testing score in math and English/language arts on the Indiana Statewide Testing for Educational Progress-Plus (ISTEP+) based on a percentage of students passing the annual statewide assessment. This score would then be adjusted based on if a school's growth and participation targets were met; final scores were weighted for comparison. Students are tested annually in English/language arts and Mathematics in 3<sup>rd</sup> through 8<sup>th</sup> grades. Third

grade is the first year of state mandated standardized testing, and serves as the baseline marker for students, while eighth grade is one of the last years of mandatory testing (IDOE, 2018e). Students are also tested in science in 4<sup>th</sup> and 6<sup>th</sup> grades, and social studies in 5<sup>th</sup> and 7<sup>th</sup> grades. A new accountability system was adopted for the 2015-2016 school year, aimed to simplify scoring calculations, incorporate both growth and performance goals, and base accountability more on grade level and less on school type.

Accountability grades are important for Indiana traditional public, charter, and choice schools. For traditional public schools, a potential state intervention is based on the number of consecutive “F” grades a school receives. For charter schools, minimum standards must be met for charter renewal and to avoid closure by not receiving four consecutive “F” grades. For choice schools, choice scholarship payments are suspended for one year if a school receives two consecutive “F” grades and are suspended for a longer period of time if low grades continue to be received by a school. In Indiana, schools that service grades K-2 (known as feeder schools) receive an accountability score taken from the average performance scores from the receiving school where students attend 3<sup>rd</sup> grade.

In 2011, Senate Bill 1 was signed regarding teacher evaluations in Indiana. The bill mandated teachers be evaluated annually using a combination of factors, including seniority, degrees, student achievement, and growth. It also introduced a merit pay system for teachers. Student achievement scores on teachers’ evaluations vary from district to district but could account for up to a third of evaluation scores. Teachers in 4<sup>th</sup> through 8<sup>th</sup> grade will be impacted by merit pay the most, largely because these grades will have data from the ISTEP+ to calculate growth over time (Morello, 2012). Teachers are evaluated by other student learning objectives in grades and subjects where testing

data is not available. von der Embse et al (2016) argued that the growing policy focus on high-stakes assessments could potentially have a negative impact on school climate. Results found the use of student achievement scores to make administrative decisions for teachers was associated with greater levels of teacher stress outside of general teacher stressors. The next section reviews literature related to the multiple perspectives of the dimensions of school climate.

### **Dimensions of School Climate**

A lack of cohesion in operationalizing school climate remains problematic in the literature, stemming from the complex nature of school climate, multiple theoretical orientations, and a variety of dimensions. Research critically examining the literature of school climate has identified a variety of theoretical perspectives school climate research may fall under. Anderson (1982) found three main theoretical perspectives: input-output, sociological, and ecological. Input-output theory (see Glasman & Biniaminov, 1981) suggests an effective school climate results from the correct combination of inputs to produce the right outputs and is seen as an oversimplified view of the school. Sociological theory (Brookover & Erickson, 1969) offers the school as a cultural system of relationships between and among students, teachers, and parents, where the social/cultural environments will directly affect the learning outcomes of a school and its students. Ecological theory (Moos, 1974) acts a mixing of the input-output and sociological theories and views the social/cultural environment as well as the physical attributes (resources, environment) of a school, viewing a large host of variables as potentially indicative of a climate.

Gangi (2009) framed school climate research as falling under one of three theoretical views. First, some view school climate from an individual level, where

individual differences in students influence the larger climate (Miller & Fredericks, 1990; Raudenbush, Rowan, & Kang, 1991). Others believe school climate forms from a school level, where all students are impacted by the same climate (James, 1982). Last, some (see Hoy & Fedman, 1987) believe the health of an organization (i.e. harmony between students/teachers) is important for a school to be successful. Rudasill et al (2017) argue that school climate research has grown from three traditions: school effects, organizational, and psychological. Similar to Gangi (2009), the organization level is built on the assumption that worker' conditions (most often described by teachers' perceptions) will impact the behaviors of those involved. Those viewing school effects as the primary foundation to school climate consider the overall school environment. Last, research from the psychological tradition measures teacher and student opinion, but "often referencing a definition or model from another research tradition but without explicitly testing a theoretical model" (Rudasill et al., 2017). It appears that there is general agreement on the existence of three ideologies to fall under when justifying the importance of school climate; subsequent measures then depend on the theoretical orientation.

A large number of school climate studies adopt Bronfenbrenner's Ecological Systems theory (EST; Bronfenbrenner, 1979; Mitchell et al., 2010, Rudasill et al., 2017, Thapa et al., 2013). EST states that the "environmental contexts around an individual are nested and interactive" (Rudasill et al., 2017, p. 4). Here, characteristics of each system an individual is in remains important and influential (Mitchell et al., 2010). For example, high teacher turnover may suggest the larger school environment is unstable or unsupportive, which could impact the general perceptions of teachers within the classroom environment. This environment may then be projected, purposefully or

unintentionally to the students in a classroom, who recognize that the school climate system they exist in is not ideal. Anderson (1982) stated that few measures of school climate actually encompass the overall climate, and many tend to utilize theories that produce bias and seldom capture the essence of school climate. Nearly 40 years later, a lack of consensus persists in the field of school climate has led some researchers to believe measuring school climate is desirable, but perhaps not attainable.

Wang and Degol (2016) wrote that while researchers now agree on the complex and multidimensional nature of school climate and its importance on student outcomes, there still remains no consensus on what factors form school climate. Gangi (2009) suggested in her dissertation a unified definition of school climate was adopted when practitioners, researchers, and policymakers met from the National Center for Learning and Citizenship, Education Commission of the States, and the National School Climate Center at the Center for Social and Emotional Education in early 2007. Four broad areas related to school climate were put forth: 1) relationships among those in the school, including students, parents, teachers, and administration, 2) physical safety, 3) teaching and learning, and 4) the institutional environment (Collie et al., 2012). Yet, there remains a wide array of dimensions used in school climate research that fall outside these areas.

Wang and Degol use the four-dimension framework in their work, while Thapa et al (2013) add an additional dimension-- the school improvement process. Zullig et al (2010) suggested five domains have historically been measured: order and safety, academic outcomes, social relationships, school facilities, and school connectedness (p. 41). Other additional measures have included knowledge and fairness of disciplinary policies and discipline problems, peer relationships among students, administrative

leadership, and the perceived racial climate (Cheema & Kitsantas, 2014; Martín et al., 2008; O'Malley et al., 2015; Tschannen-Moran et al., 2006).

Johnson et al (2007) measured school climate perceptions from teachers using five factors: collaboration, decision-making, instructional innovation, student relations, and school resources. Rudasill et al (2017) proposed that school climate is composed of the social interactions and relationships, safety-physical and emotional, and the values/beliefs of all members of the school. The authors argue that the structural and contextual components of a school should not be included in a definition of school climate as a true climate stems from people and their relationships. Again, this is often dependent on the researcher's theoretical orientation, as well as the purpose of the overall measure to the study.

### **Unit of Analysis**

An additional complication in school climate research is who should be surveyed. Some researchers believe all members of a school community should be surveyed, as all impact the school climate (Rudasill et al., 2017). Ramsey, Spira, Parisi, and Rebok (2016) argued that because personal beliefs often are subjective to the environment, using multiple sources to measure school climate could provide a more comprehensive picture. This could be students (Brand et al., 2003), teachers, principals, and parents or some combination of these perspectives (Ramsey et al., 2016). However, most research tends to examine only one perspective of school climate as indicative of the entire school's climate.

In particular, organizational climate models often focus on the perspective of the teacher (Hoy & Hannum, 1997). Collie et al (2012) address this head-on, arguing while school climate does affect the community, teachers are most directly responsible for the

instructional environment around their students. As such, teachers' perceptions of their classroom and school context are important and shape classroom decision-making. However, aggregating teacher ratings to form a single school indicator is problematic, particularly with the complex composition of school climate. Because teachers' perceptions of school climate are often used in determining the effectiveness of schools (Koth, Bradshaw, & Leaf, 2008), researchers should be clear in operationalizing and reporting the nuances of school climate.

Brand et al (2008) argued that student-based surveys of school climate are informative, but often tricky, time-consuming, and non-representative of all students. However, student surveys continue to be the predominant way for researchers to assess school climate, regardless of the general agreement on the multifaceted nature of the concept (Berkowitz, Moore, Astor, and Benbenishty; 2017). Higgins-D'Alessandro and Guo (2009) found students and teachers were similar in their responses to school climate within the same school. Brand et al (2008) compared middle school students and teachers' ratings on school climate, finding moderate correlations. Humphrey (1984) found little to no association between ratings of school climate between elementary students and their teachers.

Brand et al (2008) cited the lack of parallel instruments between students, teachers, and administration as contributing to the barrier in a more comprehensive school climate measure. School climate surveys for principals often focus on organizational management and school-level variables (Back et al., 2016). Rudasill et al. (2017) argued that although leadership and organizational effectiveness may influence the development of school climate, it should not be considered a dimension within school

climate. Additionally, principals are often less informed about the subtleties in individual classrooms or the day-to-day mood of the school than teachers or students.

When both students and teachers are surveyed on school climate, teachers are typically more focused on classroom-level factors such as student behaviors and classroom management, while students were more focused on school-level factors such as relationships and teacher turnover (Mitchell, Bradshaw, & Leaf, 2010). Students are more likely to be impacted by individual factors such as family background and education, behavior problems, etc. (Fan, Williams, & Corkin, 2011). Clifford et al (2012) wrote that school principals have a direct influence on school climate through the conditions they create and influence. Poor climate is often connected with teacher stress, reduced job satisfaction, and commitment (Klassen & Chiu, 2010). Teachers are more likely to report being committed to teaching when supported by school leadership, and school climate is an important factor in teacher retention (Fulton, Yoon, & Lee, 2005).

### **School Climate Outcomes**

School climate instruments are often used or adapted to link or support other outcomes. However, both gaps in and overlap exist in what is measured by school climate, and what is linked to it. The theoretical orientation and the purpose of the specific school climate measurement can impact how findings are reported and treated. Additionally, the validity of scales is rarely questioned beyond basic or intermediate statistical procedures, which could provide implications for the reliability of many findings. Research suggests that at the school level, high teacher and student turnover are inversely related to school climate (Mitchell et al., 2010; Plank, Bradshaw, Young, 2009). High levels of administrative and principal turnover are also negatively related to student and parent perceptions of order and discipline within a school (Griffith, 1999). Creating a



positive school climate is important during efforts of school reform (Guo & Higgins-D'Alessandro, 2011) and is influential to teachers' perceptions of the effectiveness of professional development programs (Guo & Yang, 2012). School access to facilities and the infrastructure of a school is often a dimension of school climate. The quality of school facilities (e.g. building condition, classroom size, cleanliness) has been found to impact student achievement (Uline & Tschannen-Moran, 2008).

School climate has been linked to students' social, behavioral, and learning outcomes (Zullig et al., 2014). Kohl, Recchia and Steffgen (2013) found over 70 quantitative studies that used some school climate instrument to focus on the link between school climate and aggression. Wilson (2004) found that the relationship between school climate and aggression/victimization behaviors is dependent on students' perception of connectedness to a school. A positive school climate has been associated with reduced bullying (Kosciw & Elizabeth, 2006; Meyer-Adams & Conner, 2008) and less violence/aggressive behaviors in schools (Goldstein, Young, & Boyd, 2008; Gregory et al., 2010).

Research has identified race/ethnicity as a predictor of student perceptions of school climate and found that minority, low-income, and female students often differ significantly in their perceptions of school climate (see Thapa et al, 2013). Esposito (1999) suggested that a positive school climate had a disproportionately strong impact on minority students' academic outcomes. School climate has been found to impact students' achievement in both the short-term and long-term (Hoy, Hannum, & Tschannen-Moran, 1998). Knowledge of how a school is organized, as well as the norms, behaviors and attitudes can help to increase positive school climate. Berkowitz et al (2017) found, in a review of recent school climate research with respect to achievement,

that positive climate at one point in time contributed to higher achievement at a second point in time. The authors suggest that longitudinal designs would be required for more causal claims, particularly because school climate is not static.

Teachers' beliefs they can have a positive effect on student learning are related to school climate perceptions (Guo & Higgins-D'Alessandro, 2011). von der Embse et al (2016) suggested that school climate serves as an indicator and buffer to rapidly changing educational policies on teacher outcomes. Respect between school community members plus shared expectations between the principals, teachers, and students has been found to increase student engagement (Bryk, 2010; Ennis, 1998). Teacher-student relationships are important for students' motivation, GPA, and test scores (Hoy & Hannum, 1997; Ryan & Patrick, 2001; Wang & Degol, 2016; Wang & Holcombe, 2010). A positive school climate promotes student learning through cooperative learning, group cohesion, respect, and mutual trust (Thapa et al., 2013). Poor school/classroom management has been linked with increased behavior problems and a decreased focus on academics, which is associated with more negative perceptions of climate from students (Koth et al., 2008). A positive school climate has been found to influence teacher stress, burnout, feelings of accomplishment, and commitment to the profession, which in turn are predictive of teacher attrition and retention (Grayson & Alvarez, 2008; Johnson, 2006; Thapa et al., 2013).

### **Differences Between Schools**

A small body of research suggests differences in school climate by school type (Krommendyk, 2007; Lubienski, Lubienski, & Crane, 2008). For example, teachers in private schools have reported a stronger sense of community (Bryk, Lee, & Holland, 1993; Royal, DeAngelis, & Rossi, 1996). In a descriptive review of school climate data

from American and Canadian private and religious schools, Sikkink (2012) found that Catholic schools exhibited more overall positive school climate perceptions than evangelical Protestant schools. Choy (1997) noted that public high school teachers expressed students had poorer attitudes toward and more problems with learning, greater rates of absenteeism, and less involvement from parents; however, charter and magnet schools are not included in her database. In a small survey of teachers from charter and public schools, charter school teachers reported greater support from administrators, parents, and students, suggesting the climate of charter schools was similar to the climate of many private schools (Bomotti, Ginsberg, & Cobb, 1999). Krommendyk (2007) found that school climate was most open and healthy in private religious schools followed by charter and public schools.

The typical teaching and learning environments between elementary and middle school classrooms may also differ substantially. In the typical elementary classroom, a teacher may spend a majority of his/her day with the same group of students, instructing them in various subjects (e.g. English/Language Arts, Mathematics, Science, Social Studies). A middle classroom might also look this way (especially in small private and/or religious schools where one teacher often teaches multiple grades and/or subjects). However, a large portion of middle school teachers are most likely teaching one subject or similar grouping of subjects (e.g. pre-algebra and algebra I) all day with students rotating between teachers for different subjects. Midgley (1995) concluded that because elementary students spend much of their day within one classroom, the larger school culture may not be as pertinent or influential to their experiences as it would to middle school students.

Anderson and Midgley (1997) suggested that the structure of the policies and practices at middle schools emphasize relative ability more and task mastery less than elementary schools. This change in motivation has a significant impact on the classroom environment. Middle school classrooms, in comparison with elementary school classrooms, place a greater emphasis on teacher control, while students at this age are displaying an increased desire for autonomy (Eccles et al 1993). This transition from primary to secondary school can be difficult and a source of stress for many students (Evans, Borriello, & Field, 2018). Building on past research, Cappella et al (2017) concluded: “the social and instructional contexts of middle grade schools are not well aligned with early adolescents’ developmental needs for autonomy” (p. 3). Student disengagement in the classroom, particularly for secondary schools, may occur as a result of this mismatch in student needs and instructional context (Marks, 2000; Strambler & Weinstein, 2010). This tension between student and teacher may manifest in secondary teachers differently than in elementary teachers. Additionally, middle school teachers often report lower levels of support, teaching self-efficacy, and higher teacher burden, teaching more students over the course of the day, than elementary teachers (Eccles et al., 1993; Kim et al., 2014; Capella et al., 2017).

### **School Climate Instruments**

Researchers must address how any school climate instrument contributes to the purpose of their study. Is measuring school climate itself the objective? If so, will a survey be developed, or will it be chosen from the existing hundreds of instruments? Will or has this measure be validated? Is the purpose of the study to analyze the relationship between school climate and student outcomes? Or, is the purpose of the study to determine how school climate functions in an overall school system among other

constructs? Failure to address each of these issues is problematic. Brand et al (2008) stated that unlike many student measures, “studies of teachers' climate ratings have not typically provided a great deal of systematic attention to the dimensional structure and psychometric properties of their assessments” (p. 511).

In 1987, approximately 42 measures of school climate were known to exist (Shindler et al., 2003). In a critical and systematic review of all available instruments since 1990, Gangi (2009) found 102 total measures claiming to measure school climate. From here, nine criteria and eleven selection rounds found a total of 3 comprehensive instruments. Criteria included an assessment of school relationships, safety, teaching and learning, external environment components, were designed for the K-12 setting, and had empirically supported viewable test items. (Gangi, 2009, p. 18). Measures that met criteria were the Comprehensive School Climate Inventory (CSCI), the Tennessee School Climate Inventory-Revised (SCI-R) and the School Climate Assessment Instrument (SCAI). Gangi (2009) noted due to the specific nature of her study, some measures omitted could also be valid and effective for measuring school climate.

In 2012, the American Institutes for Research (AIR; Clifford et al., 2012) conducted a search of all publicly available school climate surveys and found approximately 125 school climate instruments. AIR acknowledged Gangi’s work in establishing criteria for many of the measures. As AIR’s intent was to explore the relationship between principal performance and school climate, only those measures that sampled teacher and principal opinion were considered. After applying criteria, 25 instruments were initially identified, and after a panel of reviewers, 13 were chosen for a more in-depth review. The three measures (CSCI; SCI-R; SCAI) from Gangi (2009) were included in this review. Other popular measures included were the 5Essentials School

Effectiveness Survey (5E; drawn from Bryk, Sebring, Allensworth, Luppescu, & Easton, 2010) and the Organizational Climate Index (OCI; Hoy, Smith, & Sweetland, 2002).

Even though the need for complex statistical models was noted early on (Bryk & Raudenbush, 1992), most of the school climate literature fails to address the complex nature of school data much less the item or person validity of school climate surveys given to students or teachers. Wang and Degol (2016) conducted a meta-analysis of 327 school climate studies. Nearly all (91%) studies measured school climate through some empirical means. Of those 297 empirical studies, nearly half (48%) employed a correlational design to associate school climate with other variables (Wang & Degol, 2016). Only a small percentage focused on the development and validation of a school climate instrument (15%) or used experimental or quasi-experimental methods (9%). The remaining studies (28%) used qualitative methodology to measure school climate. As the validity of scales is rarely questioned beyond basic or intermediate statistical procedures, evidenced by a previous section with popular school climate measures, the reliance on correlational data to link student or teacher perceptions with student outcomes as causal remains problematic in many findings. Wang and Degol (2016) summarized the problematic state of school climate research methodology, concluding using individual reports to predict individual outcomes violates data independence, but using school level reports to predict individual level outcomes disregards the heterogeneity within schools and within perceptions of climate themselves.

The U.S. Department of Education's Office of Safe and Healthy Students (OSHS) and the National Center on Safe and Supportive Learning Environments (NCSSLE) compiles a list of student, staff, and parent surveys that can be used by educators to assess school climate. The compendium is useful for both educators and researchers in

providing a starting place for recognized instruments. Although not comprehensive, the list provides a relatively accurate picture of the most popular and reliable measures to assess school climate. The most current list was last updated October 2018 and includes 23 student-level measures, 20 teacher-level measures, and 11 parent-level measures (Office of Safe and Healthy Schools, 2018). Many of these are versions of the same instrument for different populations. For instance, the Comprehensive School Climate Inventory has a student, teacher, and parent version. The section below highlights a selection of school climate measures recommended with validation reported.

#### *School Climate Measure (SCM)*

Most of the “historically common school climate measures” were “developed approximately 20 years ago with no reported psychometrics” or peer review (Zullig et al., 2010, p. 148). Indeed, in a review of various published scholarly work, it is often difficult to determine how scales were developed and if items have been validated before implementation. Many times, various instruments may be combined to form a larger measure, as in the case of Zullig et al (2010), who combined items from four separate measures described as psychometrically sound but lacking in various climate dimensions. Sampling students from Grades 6-12, the authors used exploratory and confirmatory principal components analysis (PCA) and structural equation modeling (SEM).

Thirty-seven items loaded on to eight factors: positive student relationships, school connectedness, academic support, order and discipline, school physical environment, school social environment, perceived exclusion/privilege, and academic satisfaction. Factor loadings ranged from .42 to .87. Factors condensed into three larger areas: social environment, positive student-teacher relationships and perceived exclusion/privilege for a 36-item final scale. The authors concluded positive student-

teacher relationships are highly correlated with student perceptions of academic outcomes, and student's social structure should be considered when determining climate. Limitations include a relatively homogenous sample of students from three districts, no comparison instrument for teacher/principal comparison, and no comparison to any established measures of school climate.

In a follow-up study, Zullig, Collins, Ghani, Patton, Huebner, and Ajamie (2014) sought to address these limitations. Four subscales from the larger SCM were used as they aligned within the study's priorities and explained 36% of the overall (45.7%) variance in the previous study. Confirmatory factor analysis confirmed the four domains with loadings from .45 to .92. The SCM, also known as the S3 School Climate Survey is publicly available. The scale was validated again in 2015 with the addition of two dimensions: parental involvement and opportunities for student engagement (Zullig et al., 2015). Similar statistical procedures were used as before (Zullig et al, 2010; Zullig et al., 2014). The process eliminated some of the original items after further review, ending with a 42-item measure for secondary students.

#### *Comprehensive School Climate Inventory*

The Comprehensive School Climate Inventory (CSCI) is a measure developed by the National School Climate Center (NSCC) in 2002. The U.S. Department of Education's Safe and Supportive Schools Technical Assistance Center, Gangi (2009), and Clifford et al (2012) all recognize the CSCI as being a strong measure of school climate. The CSCI has two versions for students and one each for school personnel and parents and has undergone continuous validation and testing since its inception. This has included multiple rounds of pilot testing, focus groups, and expert panel reviews. The CSCI can only be accessed through the NSCC, who offer the survey with consultation



and analysis of data. As this measure is quite expensive, most results are not publicly accessible, and are often used within states/districts/schools as part of reform efforts.

The CSCI measures five broad categories of school climate, with thirteen subdimensions within (NSCC, n.d.). The most recent validation of data for Version 3.0 occurred in 2011 with a nation-wide sample of elementary, middle and high school students, teachers and school staff (Guo, Choe, & Higgins-D'Allesandro, 2011). An EFA found 70 items loading on 10 factors. After a CFA and the deletion of 7 items, a 10-factor model was finalized with 63 items. Means and standard deviations for each of the Likert scale items were given for each of the subsamples. Cronbach's Alpha ranged from .47 to .90. Root mean square error of approximation (RMSEA) was borderline acceptable for high school data (0.054) and acceptable for middle school data (0.043). Guo, Choe, and Higgins-D'Allesandro (2011) had not completed validation of the parent and school staff at the time of the review but suggested "EFA results demonstrate good construct validity and Cronbach's alphas show strong internal consistency at the factor level" with a majority of factors parallel between the all three measures (p. 19).

#### *School Climate Assessment Instrument*

The School Climate Assessment Instrument (SCAI) is a school climate measure created in conjunction with the Alliance for the Study of School Climate (ASSC). According to Clifford et al (2012), the purpose of the SCAI was to devise surveys for teachers, staff, parents, and students to capture a detailed understanding of a school's health, function, and performance. Shindler (2016) wrote the ASSC SCAI is the only instrument to date that possesses a "unique analytic trait design that is a contrast to most surveys that use a Likert scale" and is the only survey instrument "whose data can be

mapped onto a conceptual road-map of function and effectiveness (p. 1). The surveys are available for use by individual schools, districts, and states with permission.

The authors operate under a “success psychology” framework, which suggests that students have an orientation toward success or failure according to three essential factors: fixed or growth mindset, feelings of belongingness, and locus of control (Shindler, n.d.). Analytic levels in the SCAI provide the respondent with three concrete statements representing a range of conditions or phenomena. Shindler (2016) wrote this design provides better accuracy of ratings and reliability between raters, clearer results and more useful practical interpretations for schools to use in reform efforts than a traditional Likert scale measure. Dimensions include Physical Environment, Teacher Relations, Student Interactions, Leadership and Decisions, Management and Discipline, Learning and Assessment, Attitude and Culture, and Parents and Community.

Six of the dimensions on the SCAI allow for direct comparison of responses. Reliability was not available for any early pilot measures, and the measure has not been published in a peer-reviewed journal. Intra-rater reliability (around .9) and subscale reliability (a range of .7 to .8) was reported as high in Gangi (2010). Shindler (2016) reported the Cronbach’s alpha reliabilities of the eight scales for students (.83 to .93, n = 327), teachers (.83 to .91, n = 208), and parents (.87 to .94, n = 89). The SCAI was used to assess if a difference occurred in student perceptions of school climate using ANOVAs between a small group of public (1) and private (2) high schools (Buening, 2014). Buening (2014) did find significant differences in the mean scores between school types and students (freshman vs. seniors). The measure has also been used in a Texas school district, in California by an independent group, in North Carolina by the DRIVE

consulting group, and a modified version was used in Michigan from 2011 to 2014 as part of a Safe and Supportive Schools (S3) federal grant initiative.

### *Organizational Climate Inventory*

The Organizational Climate Inventory (OCI) is a short climate survey taken by teachers that measures four dimensions: collegial leadership, achievement pressure, institutional vulnerability, and professional teacher behavior (Hoy, Smith, & Sweetland, 2002). There are different measures for teachers in elementary, middle, and secondary schools. No student or family version exists. The organizational climate of a school is determined by the health and the openness in the work environment. Healthy schools have positive and supportive relationships between students, teachers, and administration. The OCI is developed from the OCDQ (Halpin & Croft, 1963); many items were taken from previous work (Hoy, Hannum, Tschannen-Moran, 1998).

To norm the 30-item measure, Hoy et al (2002) distributed the instrument to 97 Ohio high schools with 15 or more faculty members as part of a larger study to determine suitability and reliability. After three items were dropped (unexpected loadings), a PCA was conducted and found a four-factor solution explaining 69.84% of the variance. Alpha coefficients for each of the factors ranged from .87 to .94. The authors concluded that OCI is “a short, reliable, and valid measure of the climate of schools, which taps four critical dimensions of organizational life in high schools” (Hoy et al., 2002).

Douglas (2010) used the OCI in conjunction with another measure to examine the relationship between school climate and teacher commitment in Alabama elementary school teachers. Pearson’s correlations and multiple regression found that teacher behavior was the strongest predictor of teacher commitment. The OCI is not a recommended measure on the OSHS approved school climate surveys. However, the

OCDQ is, and has remained a popular instrument for measuring the organizational climate of schools. Janken (2011) found school climate to be positively correlated with student outcomes (student math and reading growth) in charter schools based on correlations. Other findings suggested that schools with similar climates produced similar growth, but schools with similar growth scores did not necessarily have similar school climate scores.

#### *Delaware School Survey*

The Delaware School Survey has five scales regarding school climate, bullying, student engagement, student behavior, and social and emotional scales. Bear, Yang, Pell, and Gaskins (2012) use authoritative discipline theory and a social-ecological perspective built around the social relationships that exist within the school. The Delaware School Climate Scale (DSCS) consists of five subscales with 31 items and is given to students, teachers, and parents (Bear et al., 2016). The five subscales are teacher-student relationships, student-student relationships, school safety, clarity of expectations, and fairness of rules. Summing the scores across all subscales from the three surveys produces a total school climate score. The DSCS is one of few measures published in peer-reviewed journals.

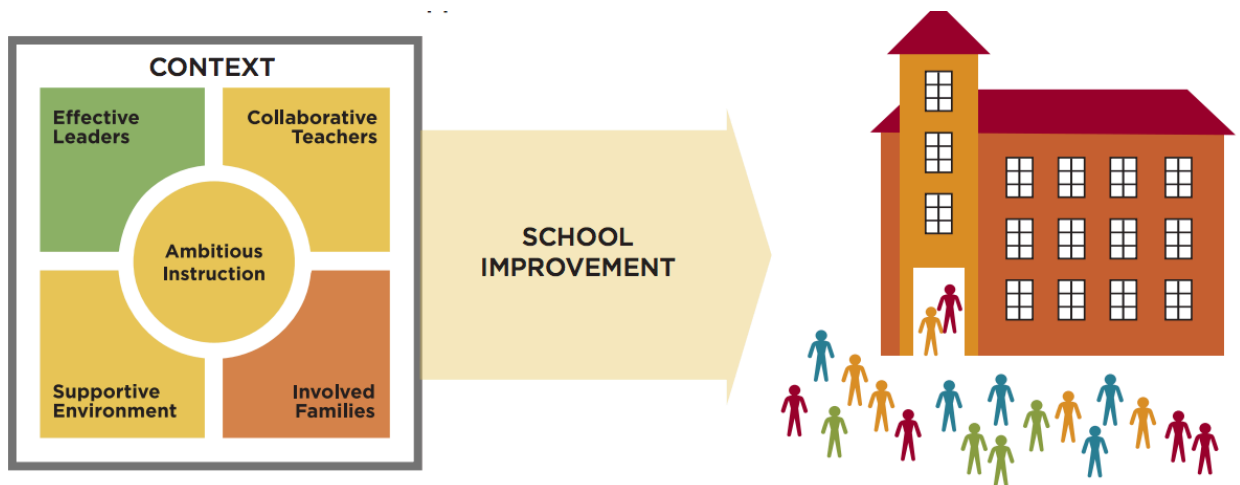
The original sample for the DSCS-S was comprised of over 24,000 public school students in Delaware (Bear et al., 2016). EFA was used to verify factor structure, which resulted in testing a variety of models including a second-order, one-, bi-, and seven-factor models. Measurement invariance was tested in a hierarchical sequence to determine if the factor structure was equivalent across groups and was adequate across grade level, gender and race/ethnicity. The second-order model was chosen as the final model as it had acceptable fit and was most in line with the underlying theoretical

framework (CFI = .925; RMSEA = .036; Bear et al., 2016). Reliability coefficients for the subscales ranged from .70 to .88 (Bear et al., 2016). Similar procedures were used to validate the DSCS-T/S, which aligns with the student version, with a population of over 5,000 Delaware teachers and school staff. Here, a seven-factor model was chosen as the preferred model (CFI = .950; RMSEA = .048; Bear et al., 2014). A seven-factor structure was also validated in the parent sample (Bear, Yang, Pasipanodya, 2012).

### *5Essentials*

University of Chicago Consortium on School Research at the University of Chicago Urban Education Institute in partnership with Chicago Public Schools and is a well-regarded framework for school improvement efforts. Anthony Bryk and colleagues spent 15 years following Chicago public schools that allowed them to “develop, test, and validate a framework of essential supports for school improvement” (Bryk, 2010). The five essentials are presented in Figure 1 below:

Figure 1 *Graphic of the Five Essential Supports*



Taken directly from Klugman, Gordon, Sebring, & Sporte (2015, p. 5).

The 5Essentials include effective leaders, collaborative teachers, a supportive environment, ambitious instruction, and involved families. The 5Essentials are the

“foundation for improving student learning ... anchored within a context unique to each school—a climate of relational trust, a school organizational structure, and resources of the local community (Sebring et al., 2006, p. 19). A student-centered learning climate (or supportive environment) is more than just instruction. First, schools must work to provide a safe and engaging learning environment for students with clear expectations and rules. Second, schools must balance expectations for high academic achievement with support and reasonable goals. Social capital and a learning community mentality are important to develop, particularly in struggling schools. Last, schools should work to develop a cohesive curriculum that is aligned with appropriate standards.

The 5Essentials survey can be given to students, teachers, and parents. Scores are reported as three different, but related scores: a measure score, an essential score, and a 5Essentials score. The school environment measure score is “a summary indicator describing how teachers or students responded to the specific questions making up each Measure” (5Essentials, 2018). This is calculated by combining responses using Rasch analysis that are then compared with a benchmark on a scale (from 1-99). When data is reported to schools, it appears in a percentile format. No reliability scores, or item difficulty levels are given for any of the items publicly.

Building on this research from the Chicago Consortium, the Philadelphia school district in partnership with the University of Pennsylvania Graduate School of Education in 2014 created a survey for students, teachers, principals and parents/guardians. Focus groups, cognitive interviews, a panel of experts, and a variety of pilot studies were conducted to validate this measure. The five constructs are: climate, instruction, leadership, professional capacity, and parent/guardian community ties; these constructs are analogous with the 5Essentials framework (School District of Philadelphia Office of

Research and Evaluation, 2016). In an EFA, a minimum loading value of 0.3 was specified, and the Kaiser criterion was used, all of which confirm the validity of the topics and subtopics within. Subscale alpha reliabilities were calculated from data from the 2015-2016 school year for students (.86), teachers (.95), principals (.91), and parents/guardians (.72). Climate was composed of six subscales for teachers (with corresponding alpha reliability): bullying (0.75), respect (0.69), student-centered learning climate (.85); classroom-level challenges (.78); school-level challenges (.92); and external challenges (.82). From here, items were refined on the scales if they did not load properly.

The Philadelphia school climate scale is important to the school climate items in the SEI survey, as both stem from the work of the Chicago Consortium and the 5Essentials framework. There is much overlap between the two climate scales--16 items from the Philadelphia scale--are included in the 23-item SEI school climate series. In addition, portions of this instrument have been previously validated. First, a 9-item instrument measuring the construct of institutional challenges produced high alpha reliability (referred to as the IC scale from here on). Second, a 60-item instrument was used to measure school climate in the Philadelphia School District (referred to as the Philly scale from here on). In this larger measure, exploratory factor analysis was used to refine the measure. An underlying issue with combining the two scales is that both have been presented as measuring different constructs (institutional challenges and school climate). However, all three instruments share the same question stem: "To what extent do you consider each of the following factors *a challenge* to student learning in your classroom and/or school?" While the instrument was designed using the 5Essentials framework, and items have been used in various contexts of school climate, there is no

empirical validation that combining these items together will provide a sound, unidimensional measure.

The NCSSLE states that all scales listed have undergone some sort of reliability and validity testing. However, a majority of reports provided to show validation are unpublished and unlinked documents. Where methodology is accessible through further research, the primary way to demonstrate validation is factor analyses, which focuses more on the scale as a whole and less at the individual parts. Using factor analyses and correlational data in quantitative school climate research continues to be the norm (Wang & Degol, 2016). As school climate is a complex notion with a variety of operationalized definitions, using correlation or factor analysis as the sole form of validation might overlook fundamental issues at the item level. For instance, a 50-item measure could produce an acceptable factor solution, with high alpha reliability simply by virtue of having a large number of items. Yet a shorter measure may not accurately represent the nuances of school climate either. By using item response theory over a classical approach, analyzing an instrument at the item level takes into account both the individual and the quality of the items themselves as being an observable, operationalized response to an underlying trait or condition, providing a more effective instrument (Bond & Fox, 2015).

### **Item Response Theory**

Item Response Theory (IRT) models use the latent trait characteristics of individuals as predictors of observed responses on a scale and is a modern approach to measurement analysis (De Ayala, 2009; Toland, 2014). Instead of using a total score from scale items to determine the amount of the latent trait, IRT calculates the probability that a given response is reliant on both the amount of latent trait a person possesses (person ability) and the amount of latent trait an item requires for endorsement (item



difficulty) (De Ayala, 2009; Setari 2016). A one-parameter model (1PL) is the simplest IRT model and conceptualizes each item on a scale in terms of its location on the latent continuum (De Ayala, 2009). Central to the IRT model is the focus on the item itself, which enables more precise optimization on scale development (Baker, 2001; Toland, 2014).

Item difficulty and person ability levels are expressed in log odds units (logits). Logits are the transformed raw scores of the ordinal data to log odds ratios that can then be placed on a consistent interval scale (Bond & Fox, 2015). While the theoretical range of values is from  $-\infty$  to  $\infty$ , values typically range from -3.0 to 3.0 on a logit scale (Bond & Fox, 2015; De Ayala, 2009). For item difficulty, negative values (below 0.0 points) indicate a question is easier to endorse, and positive values (above 0.0 points) more difficult. Items that hover right around the 0.0-point mark are suggested to be of average difficulty (De Ayala, 2009). Items that are considered easier are those that respondents with less of the latent tend to endorse, and items that are considered more difficult are more likely to be endorsed by persons with more of the latent trait being measured (De Ayala, 2009). As both people and items reside on the same logit continuum, person locations and item locations share the same scale and the above range.

The mathematical equation for the IRT 1PL unidimensional dichotomous model is (De Ayala, 2009, p. 18):

$$p(x_j = 1 | \theta, \alpha_i, b_j) = \frac{e^{(\theta_s - b_j)}}{1 + e^{(\theta_s - b_j)}}$$

The left side of the IRT 1PL equation is the logistic function of the model estimating an item response of 1. Here, the probability (p) of the response of 1 ( $x_j = 1$ ) for an item ( $i$ ) and participant ( $s$ ) given a person's location/ability level ( $\theta_s$ ) and item's

location/difficulty level ( $b_j$ ) as related to how well the item discriminates between respondents ( $\alpha$ ) (De Ayala, 2009; Bond & Fox, 2015). The item discrimination parameter ( $\alpha$ ) is related to the item response function's slope and demonstrates how well an item can differentiate between people with varying ability (De Ayala, 2009).

The item characteristic curve (ICC) is a graphical representation of the probability of an item at a certain difficulty level being endorsed by a person with a certain ability level (De Ayala, 2009). In the 1PL model, the slope or discrimination parameter  $\alpha$  is a constant. When  $\alpha$  values are increased, the slope is steeper, and able to predict with more certainty a person's ability level. When  $\alpha$  values are decreased, the slope is flatter, and able to predict with less certainty a person's ability level. A 1PL IRT model with a constant  $\alpha$  of 1 is known as the Rasch model for dichotomous data.

Hambleton, van der Linden, and Wells (2010) note that the usefulness of IRT depends on being able to accurately estimate item parameters and person ability. When both sets of information are not present, statisticians utilize a method of estimation to find the unknown parameter estimations. One common strategy in IRT analysis is to use a joint maximum likelihood estimation (JMLE). This technique uses an iterative two-stage process. First, item locations are estimated using person ability locations. These item parameters are estimated independently of one another. The rationale behind this first step is because there are typically more survey respondents than items (De Ayala, 2009). In the second step, item locations are treated as known and used to estimate person ability locations. After the second step, the new person ability locations are repeated, using the updated figures from each stage, until a convergence occurs (i.e., estimates are consistent or change a minimal amount during the two stages) (Hambleton, van der Linden, &

Wells, 2010). A JMLE technique is commonly used in Rasch analysis programs, but other estimation techniques exist.

### **Rasch Modeling**

The Rasch model is a framework with which to examine the properties of a survey or other psychometric measure. Mathematically, both the Rasch model and the 1PL model are equivalent as the Rasch model is a 1PL IRT model with  $\alpha = 1$  for the discrimination parameter (De Ayala, 2009). Although the 1PL IRT and Rasch models are the same mathematically, the important distinction behind the two is in the philosophy. An IRT model is focused on fitting the data to the model, given the model's constraints (De Ayala, 2009). A Rasch model is seen as a standard by which to create a measurement tool and is able to provide a means by which to measure persons and items on the same ruler (Linacre, 2005). A Rasch model allows for the purposeful examination of an instrument to provide validity and stability for future use. For data to be useful, it must fit to the Rasch model.

Rasch (and IRT) models rely on the same three assumptions: unidimensionality, local independence, and equal discrimination (Bond and Fox, 2015). First, an instrument should measure one latent trait, and that latent trait exists on a continuum that is directly and solely responsible for how a person responds on an item. Second, a person's response to an item is independent of any other item, conditional on that person's ability location. Last, in a Rasch model analysis, it is assumed that item discrimination is uniform, an "average of the discrimination of all items, and under discrimination of any items relative to this average suggests multidimensionality, while over discrimination relative to this average suggests response dependence" (Andrich, 2005).

Bond and Fox (2015) wrote that though no items or people will ever fit a model perfectly, it is more valuable and informative to identify those that deviate more than expected, which impacts the overall statistics. Fit statistics are derived from the summarized residuals of the actual response ( $x_{ni}$ ) from the model expectation ( $E_{ni}$ ) and are written as either a mean square or standardized value ( $t$  or  $Z$ ) (Bond & Fox, 2015; Wright & Masters, 1982). Fit statistics are then categorized as either infit (those that emphasize unexpected responses near a person or item measure; INFIT MNSQ) or outfit (those that emphasize unexpected responses away from a person or item measure; OUTFIT MNSQ) (Bond & Fox, 2015). De Ayala (2010) suggests values from 0.5 to 1.5 are acceptable, but those greater than 2 warrant closer inspection. Linacre (1999) suggests outfit means values greater than 2 could be indicative of more noise in the category. For a more stringent interval, Smith, et al. (1998) suggested that an acceptable item infit is  $1 \pm 2/\sqrt{n}$  and outfit is  $1 \pm 6/\sqrt{n}$ . Items that are considered misfitting should be checked for problematic verbiage, removed and/or edited for future versions of the measure.

Differential item functioning (DIF; Bond & Fox, 2015) is used within Rasch analysis to determine if variation in responses exists between groups that could lead to a source of potential bias in measurement (Tennant & Pallant, 2007). Groups of individuals are “stratified into matching ability levels and their relative performance on each item is quantified” (Badia, Prieto, & Linacre, 2002). DIF estimates the item difficulty for two groups and compares the difference between the two using a t-test (Setari, 2016). Two types of DIF can be examined: uniform DIF (which is constant across ability level) and non-uniform DIF (which varies across ability level) (Tennant & Pallant, 2007).

WINSTEPS also identifies Differential Group Functioning (DGF) between groups of persons.

When interpreting DIF results, the magnitude, sample size, and comparison of different groups should be considered (De Ayala, 2009). A Welch's t-statistic is used to estimate the difference between the two means (Linacre, 2018a, p. 671):

$$t = \frac{DIF\ Contrast}{Joint\ S.E} = \frac{(M_1 - M_2)}{\sqrt{(SE_1^2 + SE_2^2)}}$$

where  $M$  represents the mean item difficulty for the reference group, with subscripts representing the reference or focal group, divided by the square root of the standard errors of the mean item difficulties. The difference between the two means should be noticeable, and significant enough to happen outside of chance ( $\sim |t| > 2$ ; Linacre, 2018a).

WINSTEPS also identifies Differential Group Functioning (DGF) between groups of persons. A positive DIF contrast indicates an item is more difficult to endorse for the reference group after adjusting for participants' overall scores (Linacre, 2018). According to Zwick, Thayer, and Lewis (1999) criteria, items with an absolute DIF contrast score from 0.43 to 0.63 logits can be considered slight to moderate presence of DIF. Items with an absolute DIF contrast score greater than or equal to 0.65 logits can be considered moderate to large. Items that exhibit significant DIF results could be the results of a bias in the item, an issue not relevant to the trait being measured by the instrument, or a true difference in the latent trait between members of two groups.

### **Rasch Rating Scale Model**

The Rasch model can be extended to polytomous data by way of two models: the Rating Scale and the Partial Credit models. The Likert-type categories teachers choose to endorse in the school climate items are static in their meanings in the measure (Sinnema,

Ludlow, Robinson, 2015). In other words, “not a challenge” and “a great challenge” mean the same throughout the 23-item measure, which indicates a Rasch RSM is appropriate (Linacre, 2000). Second, the size of the dataset is large enough to ensure at least 10 observations in each of the four categories used for estimation, which suggests the estimation will be robust against any accidents in the data. Following these two considerations, I will utilize the Rasch Rating Scale model in analysis of the school climate instrument data. The Rating Scale Model (RSM; Andrich, 1978) is typically presented as (De Ayala, 2009, p. 181)

$$p(x_j|\beta, \delta_j, \tau) = \frac{\exp[-\sum_{h=0}^{x_j} \tau_h + x_j(\beta - \delta_j)]}{\sum_{h=0}^m \exp[-\sum_{h=0}^k \tau_h + k(\beta - \delta_j)]}$$

where  $p(x_j|\beta, \delta_j, \tau)$  is the probability that a person at  $\beta$  passing  $m$  number of thresholds on an item ( $j$ ) located at  $\delta_j$  will respond at a threshold  $\tau$ . Authors note: similar to De Ayala (2009),  $\exp [z]$  is used in equations instead of  $e^z$ .

According to Bond and Fox (2015), ordered categories separated from one another by thresholds occur at “the level at which the likelihood of being observed in a given response category is exceeded by the likelihood of being observed in the next higher category” (p. 116). This is the rate at which the probability of choosing one response over the following response is 50/50. Threshold parameters, referred to as Rasch-Andrich thresholds, have the same values for all items, as each person can only have one of four potential responses (e.g. not a challenge, a slight challenge, a moderate challenge, a great challenge). De Ayala (2009) suggested that one implication of a common set of thresholds across an instrument’s items is that the thresholds only need to be estimated once for the item set. However, as Rasch model items on the instrument

may have different locations, threshold locations are determined by a combination of an item's location and the threshold's value (De Ayala, 2009).

Threshold parameters are calculated from the assumption that moving between response categories may not be equal. As an example, a teacher may find choosing between “not a challenge” and “a slight challenge” difficult to endorse for student tardiness but find it much easier to endorse “a great challenge” over “not a challenge” with the pressure to perform well on standardized tests. According to Bond and Fox (2015), thresholds that do not increase in a monotonic fashion are labeled as disordered and are often more pronounced on probability curves and should be examined more in depth. Linacre (1999) recommended thresholds increase by a minimum of 1.4 logits but a maximum of 5 logits.

### **Summary**

The field of school climate research has had persistent problems in a variety of capacities since inception. First, school climate has been utilized in different capacities as federal education policies have changed over time. This could partially explain the large variety of dimensions and definitions of school climate, reflective of the “implicit assumption toward including factors in the definition of school climate only if they are predictive of critically important outcomes (Rudasill et al., 2017, p. 7). In other words, school climate is only important inasmuch as it serves as a predictor of a desired outcome, situated in the context of previous and current educational policies. While most researchers and educators recognize the value of non-cognitive measures, agreeing upon any one framework, instrument, or sampling frame has not happened. A selection of popular school climate instruments demonstrated that validation efforts typically begin and end with factor analysis techniques, which could miss complexities in individual

items. Only one theoretic framework including school climate, 5Essentials, recognized the importance of item level validation.

It is important to determine if the new scale possesses good psychometric properties before continuing with any larger analysis. Although it is promising that the 5Essentials explicitly states an importance of validation at the item level, items on the original 5Essentials instrument have been adapted for use in the study at hand and must be examined. Work to validate any school climate measure beyond a factor analysis is critical to further any potential cohesiveness of the field. In addition, there is a clear and stated need in Indiana for a measure of school quality and student success that was not met, stemming from a lack of reliable and validated measures. This scale provides a unique opportunity to validate data already within the specific population that is of interest, but also to researchers and policymakers interested in using a school climate measure on a large scale.

## **Conclusion**

This chapter provided a review of literature regarding school climate, the foundation for the instrument validated in this study. A review on how school climate literature has changed over time, followed by a discussion on the differences in dimensions, participants, and instrument analysis was offered. Then, an account of school climate results was presented before highlighting appropriate differences in school level. A brief history of the current context of Indiana education policies was provided. Concluding this chapter was a description of the primary method of analysis: a one-parameter item response theory model used for polytomous data, specifically the Rasch Rating Scale Model. The next chapter will provide the methodology used to conduct scale validation for this dissertation: the Rasch Rating Scale Model.



## CHAPTER 3. METHODOLOGY

This chapter begins with results from a pilot study, followed by the study's research questions. Then, information about the instrument, participant sampling, and data collection will be stated. The final portion of this chapter provides an explanation of how each research question will be addressed in the study.

### **Pilot Analysis**

Prior to this study, pilot analyses were run using a pilot group of 200 teachers. Stratified random sampling from three school sectors (public, private, charter, other private) were used to sample 200 teachers from the preliminary data ( $n = 4974$ ) of a state-representative sample of schools. This sampling frame was chosen purposefully as Linacre (1994) suggests that Item Calibrations are stable within  $\pm 0.5$  logits when a sample size range of 108-243 is used. The Rasch RSM analysis results produced a person reliability estimate of 0.87 and an item reliability estimate of 0.98 with the Rasch dimension explaining 40.2% of the variance in the data.

Pilot item difficulty level estimates ranged between -1.23 and 1.94 and (see Table 3), indicating a range of difficulty for participants to endorse. This spread of difficulty suggests most items fall between moderately-easy and moderately challenging to endorse and the instrument includes no very-difficult or very-easy items. The most challenging item for teachers to endorse was "Threat(s) to your safety or safety of students" and the least challenging item was "Pressure to perform well on standardized tests." INFIT values in this series of items ranged from 0.66 to 1.44 indicating good item fit for all 23 items. OUTFIT values behaved similarly within a range of 0.66 to 1.40 for all 23 items.

To determine if differential item functioning occurred (DIF), a F-test was performed see if the items functioned similarly for different groups. There was some

evidence that magnet and charter school teachers responded differently to items, however as only one magnet teacher was included in comparison to eight charter teachers (of the 200 total), these differences are inconclusive. One item was also flagged as having a possible different meaning for the school sectors: “Students with special needs (e.g., hearing, vision, speech impairment, physical disabilities, mental or emotional/psychological impairment).” Private and charter schools may lack the infrastructure or dedicated special education services that public schools typically are required to have, which may be evidenced here by the group difference in responses. However, the pilot sample excluded many important teacher groups (specifically charter and magnet teachers) and did not take into account grade level, which is the primary focus of this dissertation. Both of these could be significant in the larger sample and impact the stability of the instrument once analyzed.

This school climate scale provides a unique opportunity to validate an instrument already within the specific population that is of interest. To validate this school climate instrument, a Rasch analysis will be implemented. A Rasch analysis estimates an item’s difficulty and a person’s ability on the same continuum in the expressed on a common logit scale (Bond & Fox, 2015). Applying a Rasch model allows for the purposeful examination of a measure at the item level to provide validity and stability for future use. While the instrument was designed using items adapted from the 5Essentials framework, and many items have been used in various contexts of school climate, there is no empirical validation that combining these items together will provide a sound, unidimensional measure.

The purpose of this study was to validate a school climate instrument based on the four most commonly accepted dimensions of school climate, using measures adapted

from the 5Essentials framework to provide an effective instrument for educators and researchers. This study is guided by the following research questions:

- How well does the instrument measure the latent trait of school climate?
- How well do the individual instrument items reflective of school climate fit to the Rasch model?
- Do differences exist in item responses from teachers from 3<sup>rd</sup> and 8<sup>th</sup> grades with similar levels of the latent trait of school climate?

### **Study Overview**

The School Effectiveness in Indiana study (SEI) methodology and instrumentation was developed through collaboration between researchers at the University of Notre Dame, University of Kentucky, and NORC at the University of Chicago. The study was made possible by a Lyle Spencer Research Award from the Spencer Foundation with additional funding from the Walton Family Foundation. The study's purpose was to examine what conditions contribute to school effectiveness, using a comparative framework with traditional public, private, and charter schools. Results aim to provide guidance or suggestions for improvement efforts in Indiana schools.

Survey items were selected through a process of reviewing existing school survey instruments, consulting experts in the field, and feedback from members of the research team. A driving force behind the constructs chosen is the “predictive validity and relationships to student achievement”, specifically the “organizational conditions that enable [and promote] student achievement” (Berends & Waddington, 2015).

Collectively, these survey items hope to address the following research question: “How do schools of choice (charter or private schools) differ from traditional public schools in terms of organizational and instructional conditions, school leadership, professional

capacity, school learning climate and funding conditions, and parent involvement and support that promote achievement?” (SEI Methodology Report, 2017). The Institutional Research Board (IRB) from both the University of Notre Dame and the University of Kentucky approved the study. The survey was given electronically and comprised of four primary sections: Section I: School and Classroom Climate, Section II: Professional Development, Section III: Parent Involvement, Section IV: Teaching Assignment and Background.

### **Sampling Procedures**

The school population (n=1844) included Indiana public, private, and charter schools that served, (but was not exclusively limited to) students from Kindergarten through 8<sup>th</sup> grade found through Indiana Department of Education (IDOE) state databases. Among the private schools in this sample, only those that participated in the state testing program (ISTEP+; Indiana Statewide Testing for Educational Progress-Plus) were included. Indiana is unique in that a majority of private schools opt in to take the statewide student standardized assessment and have so for years (Waddington & Berends, 2018). The purpose of this specific group of private schools is to allow for comparisons between public and private schools based on student achievement scores. Schools in the study were selected specifically to utilize student achievement data, and as such, (most) teachers have data (ISTEP+) from the state that can be linked, providing a unique opportunity to compare different school sectors within the same state. However, that is outside the scope of this dissertation.

A selection of 600 schools (55 charter, 200 private, and 345 public) served as the initial sample for the study in February of 2016. As one of the primary objectives of the overall study was to explore the differences between types of schools, stratification and

efficient allocation was implemented in line with study objectives to produce a balanced sample. Schools were stratified first by school type. Then, schools were selected with an equal probability after being sorted by enrollment, location, and school level. The sorting order was locked in this manner to provide a consistent way for school replacement if necessary. If a school was flagged for replacement (through refusal or other reasons), it was replaced with the next school of that type in the file.

Once a school was selected, a roster was compiled using data from the IDOE, individual or district school websites, or from contacts within the particular school. Teachers employed full-time from grades 3-8 (who specialized in either mathematics or English/language arts) were prioritized, followed by grades K-2. A maximum of 10 teachers from each school were chosen. For a school with 10 or fewer teachers, all teachers were asked to participate. For those schools with 10 or more teachers, teachers from grades 3-8 (who specialized in either mathematics or English/language arts) were selected first, followed by any remaining teachers in grades 3 and above, then any remaining teachers.

Efforts to recruit schools and teachers took place from August 2016 to April 2017 (SEI Methodology Report, 2017). From the initial school sample (n=600), 266 schools remained active in participation, 212 were refused at the district level, 86 were refused at the school level, 15 schools had closed, and 22 schools were removed from the sample because they were an alternate school or a school with a small staff or student body (SEI Methodology Report, 2017). After the school replacement procedure, 577 schools agreed to participate. From these schools, 577 principals and 5399 teachers received the survey.

## Survey Data Collection

The survey was sent electronically to a school email address to teachers and took an average of 44 minutes to complete; a \$25 Amazon gift code was provided for teachers who completed the entire survey. From this number, 5031 completed the full survey, 20 surveys were at least 40% complete, 246 refused participation, 7 were ineligible, and 65 were unavailable for a response rate of 93.7% (SEI Methodology Report, 2017). Of the 5031 teachers who completed the survey, data was received from 4974 teachers. Approximately 84% (4184) of teachers were women, 15.8% (785) were men, and 10% (5) were missing/not selected. Table 1 provides a breakdown of the teachers participating at each school type.

Table 1 *Breakdown of Participating Teachers*

<b>School Type</b>	<b>Total Teachers Participating</b>	<b>Percentage of Total</b>
Catholic	969	19.5%
Charter	194	3.9%
Magnet	25	0.5%
Other Private	626	12.6%
Traditional Public	3160	63.5%
Total	4974	100%

Raw data was delivered via a Secure File Transfer Protocol (SFTP) in a MS Excel file.

All subsequently created files were stored on a secure cloud file through the University of Kentucky. The files will be exported into SPSS Statistics Version 24 and STATA Version 14.2 for cleaning and descriptive analysis. The data will then be exported into Winsteps 4.3 for Rasch analysis (Linacre, 2018b).

## School Climate Instrument

The scale measuring school climate is located in the first section of the larger survey, consisting of 23 Likert-scale items (Appendix A) each with four categories to endorse (not a challenge, a slight challenge, a moderate challenge, a great challenge).

Items in this series were designed to ascertain the degree to which teachers endorse certain items as being more or less problematic to them with regard to student learning in their classroom. The initial stem reads as: “To what extent do you consider each of the following factors *a challenge* to student learning in your classroom?” with 23 individual challenges following (Table 1; lettered from a-w). Topics included in this measure include teachers’ perceptions of challenges to student learning related to students (student/teacher ratio, uninterested students, etc.), the perceived support teachers receive (time to prepare, pressure to perform, professional support staff, etc.), and larger school related challenges (noise level, inadequate facilities, access to technology, etc.).

These items were adapted from the 5Essentials framework and survey, developed from research by the University of Chicago Consortium on School Research at the University of Chicago Urban Education Institute in partnership with Chicago Public Schools. Building on this research, the Philadelphia school district in partnership with the University of Pennsylvania Graduate School of Education in 2014 created a survey for students, teachers, principals and parents/guardians modeled on the 5Essentials framework. The constructs measure five aspects related to school improvement: climate, instruction, leadership, professional capacity, and parent/guardian community ties (School District of Philadelphia Office of Research and Evaluation, 2016). In an EFA, a minimum loading value of 0.3 was specified, and the Kaiser criterion was used, all of which confirm the validity of the topics and subtopics within. Subscale alpha reliabilities were calculated from data from the 2015-2016 school year for students (.86), teachers (.95), principals (.91), and parents/guardians (.72).

Climate was composed of 6 subscales for teachers (with corresponding alpha reliability): bullying (0.75), respect (0.69), student-centered learning climate (.85);

classroom-level challenges (.78); school-level challenges (.92); and external challenges (.82). Additional use of a selection of 9 items (labeled in Appendix A) designed to measure the construct of institutional challenges produced overall alpha reliabilities of .819 and .850 and were used primarily for descriptive purposes (Berends et al., 2010). There is a great deal of overlap between the Philadelphia measure for school climate and the measure used in this study (15 of 23 items; view Appendix A for item breakdown). Three items on the instrument were not included on any previous measures and will be examined specifically for fit (uninterested students, low morale among students, and inadequate physical facilities). Table 2 provides a breakdown of items categorized by the previous measures.



Table 2 *Previous Construct Breakdowns*

<b>Institutional Challenges</b>	<b>Classroom Level (P)</b>	<b>School Level (P)</b>	<b>External Challenges (P)</b>	<b>Student Centered Learning Climate (P)</b>
a. Low morale among fellow teachers/administrators.	i. Students with different academic abilities	s. Teacher turnover in this school	l. Parents uninterested in their children's learning progress	a. Low morale among fellow teachers/administrators
b. Students who come from a wide range of backgrounds	k. Disruptive Students	e. Amount of professional support staff		
		t. Student absenteeism		
d. The noise level in the school building		u. Student tardiness		
f. Students with special needs		p. Lack of teacher planning time built into the school day		
g. Amount of time to prepare for class		n. Pressure to perform well on standardized tests		
h. High student/teacher ratio		v. Lack of guidance or support for teaching special education students		
i. Student with different academic abilities		w. Lack of guidance or support for teaching English Language		
s. Teacher turnover in this school		o. Lack of school resources to provide the extra help for students who need it		
		m. Access to technology		
		c. Threat(s) to your safety or safety of students		

## **Sample**

### *3<sup>rd</sup> and 8<sup>th</sup> grade selection*

Third and eighth grade teachers were selected for the DIF analysis because they are located at the beginning and end of state mandated testing, and have different experiences in content knowledge, teaching environments, and accountability pressures. The requirements for teacher licensure for 3<sup>rd</sup> and 8<sup>th</sup> grades are different in the state of Indiana (IDOE, 2018b), which translates into different content knowledge. The requirements for licensure for a 3<sup>rd</sup> grade teacher are typically generalist and require a broad understanding of a variety of content areas, including English language arts, mathematics, science, and social studies, among others (IDOE, 2018c). This type of degree allows for the licensure of teachers from Kindergarten to 6<sup>th</sup> grade. The requirements to teach 8<sup>th</sup> grade (and others beyond the elementary school grades) are more specialized to a specific area, i.e. English/language arts, mathematics, science-chemistry. Here, an 8<sup>th</sup> grade teacher is certified to teach mathematics from 5<sup>th</sup> to 12<sup>th</sup>, which could include grade specific content, or more broad math concepts like geometry or algebra (IDOE, 2018d). As such, the content knowledge between 3<sup>rd</sup> and 8<sup>th</sup> grade teachers is different and could contribute to different perceptions of challenges to student learning.

In addition to the certification requirements, the typical teaching and learning environments between 3<sup>rd</sup> and 8<sup>th</sup> grade classrooms may differ substantially at an organizational level. In the typical 3<sup>rd</sup> grade classroom, a teacher may spend a majority of his/her day with the same group of students, instructing them in various subjects (e.g. English/Language Arts, Mathematics, Science, Social Studies). An 8<sup>th</sup> grade classroom might also look this way (especially in small private and/or religious schools). However,

a large portion of 8<sup>th</sup> grade teachers are most likely teaching one subject or similar grouping of subjects (e.g. pre-algebra and algebra I) all day with students rotating between teachers for different subjects. Middle school classrooms place a greater emphasis on teacher control, while students at this age have an increased desire for autonomy (Cappella et al., 2017). Middle school teachers report lower levels of support from the school, teaching self-efficacy, and a higher teaching burden than elementary teachers (Eccles et al., 1993; Kim et al., 2014; Capella et al., 2017). Reported disciplinary incidents also notably increase between elementary and middle school (Theriot & Dupper, 2010). Here, context matters when teachers in 3<sup>rd</sup> and 8<sup>th</sup> grade perceive school climate and its effect on student learning in their classroom.

Achievement scores are a key component of performance-based accountability models and are important for schools in the state. Principals may structure hiring and firing teachers around their abilities to produce high or positive student achievement scores (Cohen-Vogel, 2011). Third grade is the first year of state mandated standardized testing, and serves as the baseline marker for students, while eighth grade is one of the last years of testing (IDOE, 2018e). It is possible these two sets of teachers may face different pressures for their students to perform well. Third grade teachers may face the unique pressure of navigating students' through their first high stakes testing experience. Third grade teachers may also face students who are less prepared due to the assignment of perceived weaker teachers to the non-tested grades. Eighth grade teachers may face more disciplinary issues with students, and decreased student motivation and engagement from a student population who have taken many years of high-stakes testing.

Students are tested on English/language arts and Mathematics annually from 3<sup>rd</sup> through 8<sup>th</sup> grade, in science in 4<sup>th</sup> and 6<sup>th</sup> grades, and social studies in 5<sup>th</sup> and 7<sup>th</sup> grades.

Students in 3<sup>rd</sup> grade also take the IREAD-3, a measure of reading skills. Students who fail this test may not be promoted to 4<sup>th</sup> grade the following year. This places an additional burden on third grade teachers. Schools and teachers in the larger study were selected specifically to utilize student achievement data. At the time of data collection, the assessment used was ISTEP+, and those teaching grades 3-8 were prioritized.

As mentioned previously, Indiana is unique in that a majority of private and religious schools opt to participate in the state testing program (ISTEP+). Many of these schools use the resulting scores and the accountability grade given by the state as part of their accreditation process. In addition, schools who wish to receive funding for voucher students and/or student athletics are required to participate in the state testing program. All students in participating schools must take ISTEP+, regardless of their status as a voucher student or not (Waddington & Berends, 2018). Accountability grades, as determined by the state, rely heavily on state test scores. Although accountability grades do not exert the same pressure on private schools as public schools (with sanctions and state involvement), scores are often used as a marketing tool for private schools to demonstrate excellence or improvement. The decision was made to retain teachers from both traditional public, charter, and private or religious schools to provide a more holistic view of schooling. As the purpose was to validate this school climate instrument for use in all schools, including teachers from a variety of schools allows for direct comparisons and the ability to use the instrument in a variety of school capacities.

The sample selected for this study was teachers who indicated teaching 3<sup>rd</sup> or 8<sup>th</sup> grade as their primary teaching assignment (n=1960) from the larger study sample. Teachers were split into two groups, those who selected 3<sup>rd</sup> grade (n=1050) as the primary grade taught and those who indicated 8<sup>th</sup> grade (n=938). From here, descriptive

statistics were run to further narrow the sample. Those that selected 3<sup>rd</sup> grade and any grades between 6-12<sup>th</sup> were dropped from the sample. Those that selected 8<sup>th</sup> grade and any grades between pre-k-4 were dropped from the sample. Next, any of these teachers who selected a main teaching assignment field outside tested subjects were dropped from the sample. In most cases, these were teachers who selected a main teaching assignment in subjects such as science, social studies/history, special education, or religion that are not assessed with standardized testing in these grades. After this process, 967 3<sup>rd</sup> grade and 621 8<sup>th</sup> grade teachers remained.

### *Rasch RSM Sample*

There is some debate on an appropriate sample size for Rasch RSM, and so multiple methods exist. For item calibrations or person measures stable within  $\pm 0.5$  logits, Linacre (1994) suggested between 108 and 250 (or  $20 \times \text{test length} = 20 \times 23 = 460$ ) participants, and at minimum, at least 10 observations per category. De Ayala (2009, p. 199) suggests a ratio of 5 persons per every parameter estimated equaling 345 participants ( $23 \text{ instrument items} \times 3 \text{ item parameters} \times 5 = 345 \text{ total participants}$ ). Considering the large number of participants in this sample, 250 teachers from the 3<sup>rd</sup> and 8<sup>th</sup> grade samples will be randomly selected from their respective groups for a total of 500 teachers.

### **Data Analysis**

One potential issue with combining the two scales is that the scales have been operationalized as measuring both institutional challenges and school climate. Portions of the 23-item school climate instrument have been validated. First, a 9-item instrument measuring the construct of institutional challenges produced high alpha reliability (the IC scale). Second, a 60-item instrument measuring school climate in the Philadelphia School

District (the Philly scale). In this larger measure, exploratory factor analysis was used to refine the measure. Yet all three instruments share the same question stem: “To what extent do you consider each of the following factors *a challenge* to student learning in your classroom and/or school?”

In a field where confusion and a lack of continuity in both defining and measuring school climate exists, there are concerns these items may behave differently or are redundant. It is important to assess to what degree the survey is stable and can measure opinions of teachers in different school environments appropriately. In addition, there is a clear and stated need in Indiana for a measure of school quality and student success that was not met, stemming from a lack of reliable and validated measures. Validating this scale is timely and has the potential for actual policy implications in the state of Indiana. As the larger scale hopes to provide guidance or suggestions for improvement efforts in Indiana schools, it is critical that valid instruments are used to make informed decisions in school improvement efforts. Work to validate any school climate measure beyond alpha reliability or factor analysis is critical to further any potential cohesiveness of the field.

For this instrument, it is appropriate to utilize a Rasch Rating Scale Model (Rasch RSM; Andrich, 1978) model for a number of reasons. A Rasch model allows for the purposeful examination of a measure to provide validity and stability for future use by placing item and person measures on the same continuum. Rasch models are invariant, meaning that a person’s ability can be determined from the items, and an item’s difficulty can be assessed from a person’s ability, regardless of sample (Royal & Elahi, 2011). Sussman, et al (2012) paraphrasing Embretson and Reise (2000) wrote that Rasch models offer “some of the strongest empirical justification for interval scaling” (p. 137).

While the instrument was designed using the 5Essentials framework, and items have been used in various contexts of school climate, there is no empirical validation that combining these items together will provide a sound, unidimensional measure. Previous work with a subset of 9 items in this measure has been categorized as a construct of institutional challenges, and not as school climate. Moreover, there are three items that appear related to school climate but have no justification for being on the current instrument. All of these factors could be problematic to unidimensionality, an important assumption of the Rasch model. Model analysis may also flag items to remove or detect thresholds functioning poorly.

Considering the design of the items, the Likert-type categories teachers chose to endorse in the instrument are static in their meanings for all items in the measure, which indicates a RSM should be applied (Sinnema, Ludlow, Robinson, 2015) Second, the size of the dataset is large enough to ensure at least 10 observations in each of the four categories used for estimation, which suggests the estimation will be robust against any accidents in the data. Following these considerations, I will utilize the Rasch RSM in analysis of the school climate instrument data. The equation used for this study is (De Ayala, 2009, p. 181):

$$p(x_j|\beta, \delta_j, \tau) = \frac{\exp[-\sum_{h=0}^{x_j} \tau_h + x_j(\beta - \delta_j)]}{\sum_{h=0}^m \exp[-\sum_{h=0}^k \tau_h + k(\beta - \delta_j)]}$$

where  $p(x_j|\beta, \delta_j, \tau)$  is the probability that a person at  $\beta$  passing  $m$  number of thresholds on an item ( $j$ ) located at  $\delta_j$  will respond at a threshold  $\tau$ . Authors note: similar to De Ayala (2009),  $\exp [z]$  is used in equations instead of  $e^z$ .

Ordered categories separated from one another by thresholds occur at the level where the likelihood of choosing a certain response is exceeded by the likelihood of

choosing the next response in a higher category (Bond & Fox, 2015). Threshold parameters, referred to as Rasch-Andrich thresholds, have the same values for all items, as each person can only have one of four potential responses (e.g. not a challenge, a slight challenge, a moderate challenge, a great challenge). However, as Rasch model items on the instrument may have different locations, threshold locations are determined by a combination of an item's location and the threshold's value (De Ayala, 2009). According to Bond and Fox (2015), thresholds that do not increase in a monotonic fashion are labeled as disordered and are often more pronounced on probability curves and should be examined more in depth. Ordered thresholds should increase by a minimum of 1.4 logits, but not above 5 logits (Linacre, 1999).

### **Answering Research Questions**

The Rasch RS model results will be examined for unidimensionality, item fit, and differential item functioning (DIF). To answer the first research question, *How well does the instrument measure the latent trait of school climate*, unidimensionality will be assessed by conducting a principal components analysis (PCAR) of the Rasch residuals. A PCAR is not interpreted in the same way as a common factor analysis, as the PCAR does not show loadings on any factor. Instead, the least amount of contrast is preferred as it suggests that the maximum amount of variance is being explained (Linacre, 2000).

Simulation studies have suggested eigenvalues less than 1.4 and up to 2.0 are considered random noise (Linacre, 2000). If not, the loadings suggest a contrasting pattern in the residuals that requires closer examination of the content of the items. A "second" factor must have an eigenvalue greater than 3 to be considered multidimensional (Linacre, 2000b). It is important to note, as commonly accepted criteria have not been determined for PCAR analysis at this time, a PCAR is not definitive, but is



highly suggestive of multidimensionality (Linacre, 2000). If a second factor has an eigenvalue above this value, the standardized residuals for the positive and negative loading items for this contrast will be examined for any sort of patterns (Linacre 2018). If multidimensionality is present after examining the content of the loadings, it may suggest the instrument is measuring more than one dimension (Bond & Fox, 2015). If items are generally clustered and no particular theme can be determined, this could be a result of the larger dimension of school climate being multifaceted. If this is the case, there should be no concern of violating assumptions and analyses can continue.

During the analysis of residuals on the loadings, items are anchored at their difficulties and clustered into subtests by which persons are measured on. The person measures from the items are correlated with each other and each cluster of items produces a standard error. Linacre (2018a, p. 396) suggested that correlations below 0.57 indicate person measures on the two clusters have half the variance in common as they do independently of one another. High correlations suggest that the person measures on item clusters being compared share a majority of variance, which is essential in a unidimensional instrument. In these suggested categories, Linacre (2018a) proposed the cutoff point for different latent variables is 0.57, which would suggest a cluster of items is measuring different sub-dimensions of school climate. If a common theme is found between items in the clusters, an argument can be made to drop certain items, or split the instrument in two or more subscales that are then analyzed individually with the Rasch RSM or analyzed as a multidimensional Rasch model. A PCAR will be conducted as it best aligns with the goal of the Rasch analysis: to confirm the dimensionality of the scale and determine the residual variance under the common underlying factor of school climate.

To answer the second research question, *How well do the individual instrument items reflective of school climate fit to the Rasch model?*, item fit and item difficulty of the instrument will be analyzed. Fit statistics are derived from the summarized residuals of the actual response ( $x_{ni}$ ) from the model expectation ( $E_{ni}$ ) and are written as either a mean square or standardized value ( $t$  or  $Z$ ) (Bond & Fox, 2015; Wright & Masters, 1982). Fit statistics are then categorized as either infit (those that emphasize unexpected responses near a person or item measure; INFIT MNSQ) or outfit (those that emphasize unexpected responses away from a person or item measure; OUTFIT MNSQ) (Bond & Fox, 2015). De Ayala (2010) suggests values from 0.5 to 1.5 are acceptable, but those greater than 2 warrant closer inspection. Linacre (1999) suggests outfit means values greater than 2 could be indicative of more noise in the category. For a more stringent interval, Smith, et al. (1998) suggested that an acceptable item infit is  $1 \pm 2/\sqrt{n}$  and outfit is  $1 \pm 6/\sqrt{n}$ . Any items that fall outside of this range will be flagged for a more in-depth examination of content.

The instrument should contain a range of items that vary in difficulty of endorsement in order to accurately represent a wide range of ability in participants. For item difficulty, negative values (below 0.0 points) indicate a question is easier to endorse, and positive values (above 0.0 points) more difficult. Items that hover right around the 0.0 point mark are suggested to be of average difficulty (De Ayala, 2009). Items that are considered easier are those that respondents with less of the latent tend to endorse, and items that are considered more difficult are more likely to be endorsed by persons with more of the latent trait being measured (De Ayala, 2009). As both people and items reside on the same logit continuum, person locations and item locations share the same scale and the above range.

It is important to assess the degree to which categories are functioning properly for each response option. Category frequencies summarize the distribution of responses to each item in the instrument (Bond & Fox, 2015). Both the shape of the distribution and the frequency of responses per category should be examined. Ideally, Linacre (1999) recommends a minimum of ten responses per category. A low frequency of item responses will not provide a stable estimation of threshold values (Bond & Fox, 2015). In cases where categories are not functioning properly, it may be ideal to collapse certain categories.

Threshold parameters are calculated from the assumption that moving between response categories may not be equal. According to Bond and Fox (2015), thresholds that do not increase in a monotonic fashion are labeled as disordered and are often more pronounced on probability curves and should be examined more in depth. Thresholds should increase by a minimum of 1.4 logits, but no more than 5 logits (Linacre, 1999). Thresholds that are problematic will appear graphically as flattened curves that cover less of the measured variable (Bond & Fox, 2015). This suggests that response categories may not provide distinct definitions on the variable, and that respondents are not using the rating scale as the model would predict (Bond & Fox, 2015).

To answer the third research question: *Are there differences in item responses from teachers from 3<sup>rd</sup> and 8<sup>th</sup> grades with similar levels of the latent trait of school climate?*, a differential item functioning analysis will be conducted between teachers in elementary and middle schools, specifically 3<sup>rd</sup> and 8<sup>th</sup> grade teachers. Differential item functioning (DIF; Bond & Fox, 2015) is used within Rasch analysis to determine if variation in responses exists between groups that could lead to a source of potential bias in measurement (Tennant & Pallant, 2007). Groups of individuals are “stratified into

matching ability levels and their relative performance on each item is quantified” (Badia, Prieto, Linacre, 2002). DIF estimates the item difficulty for two groups and compares the difference between the two using a t-test (Setari, 2016). When interpreting DIF results, the magnitude, sample size, and comparison of different groups should be considered (De Ayala, 2009). According to Zwick, Thayer, and Lewis (1999) criteria, items with an absolute DIF contrast score from 0.43 to 0.63 logits can be considered slight to moderate presence of DIF. Items with an absolute DIF contrast score greater than or equal to 0.65 logits can be considered moderate to large.

Items that exhibit significant DIF results could be the results of a bias in the item, or an issue not relevant to the trait being measured by the instrument. Specifically, 3<sup>rd</sup> and 8<sup>th</sup> grade teachers will be selected for the DIF. The DIF analysis will determine if participants from different groups exhibit variation in responses on specific items. If differences exist, this could suggest that respondents could be viewing items differently, or that a true difference exists in the latent trait between members of certain groups.

## **Conclusion**

This chapter began with a restatement of the purpose and significance of the study’s research questions. Following this, a pilot study was detailed prior to information about the instrument, participant sampling, and data collection. The last portion of this chapter provided a review of the Rasch Rating Scale Model analysis used for the study in addition to explanations of how each research question will be addressed in the study. The following chapter provides the results of the Rasch RSM analysis.

## CHAPTER 4. RESULTS

This chapter describes the results of the study. It begins with a brief review of the purpose of the study and analysis techniques used. Then, descriptive details about the study participants will be provided. Finally, the results from the Rasch RSM analysis, including dimensionality, item fit, item difficulty, thresholds, and DIF estimates will be presented.

### **Analysis Procedure**

The purpose of this study was to validate a school climate instrument based on the four most commonly accepted dimensions of school climate, using items adapted from a well-regarded and established theoretical framework, to provide an effective measure for educators and researchers. This instrument, adapted from the 5Essentials framework, provides a unique opportunity to validate an instrument already within the specific population that is of interest. For this study, a Rasch model is appropriate for validating this instrument. A Rasch model is seen as a standard by which to create a measurement tool and is able to provide a ruler by which to measure persons and items on the same ruler (Linacre, 2005). A Rasch model allows for the purposeful examination of a measure to provide validity and stability for future use. Rasch models are invariant; a person's ability can be determined from the items, and an item's difficulty can be assessed from a person's ability, regardless of sample (Royal & Elahi, 2011).

The Likert-type categories teachers chose to endorse in the instrument suggest a Rasch RSM is appropriate (Sinnema, Ludlow, Robinson, 2015). Second, the size of the dataset is large enough to ensure at least 10 observations in each of the four categories used for estimation, which suggests the estimation will be robust against any accidents or missing values in the data. Following these considerations, I will utilize the Rasch Rating

Scale model (Rasch RSM) in analysis of the school climate instrument data. After the Rasch RSM analysis has run, the validation process will consist of examining output for unidimensionality, item fit to the model, and differential item functioning (DIF). This includes considering item and person reliability, the principal components analysis (PCAR) of the Rasch residuals, item infit and outfit statistics, the range of item difficulties, threshold values, and the differential item functioning (DIF) between 3<sup>rd</sup> and 8<sup>th</sup> grade teachers.

### **Sample**

From the 3<sup>rd</sup> grade sample (n=250), 187 taught at traditional public schools, 7 at charter schools, 38 at Catholic schools, and 18 at other private schools (Table 3). For participant gender: 230 were women, 20 were men. For participant race/ethnicity: 8 identified as Hispanic or Latino, 7 as Black or African American, 230 as White, 1 as Asian or Pacific Islander, 2 as Biracial/Multiethnic, 1 as Other, and 1 did not answer. Nearly all teachers (96.4%) reported general elementary education as their primary teaching assignment field. Teachers, on average, reported teaching at their current school for 9.3 years (SD = 8.8), with a range from first year teachers to teachers of 43 years. However, 48% of 3<sup>rd</sup> grade teachers had taught in their current school for 5 or fewer years, suggesting the mean years taught is inflated.

From the 8<sup>th</sup> grade sample (n=250), 117 taught at traditional public schools, 18 at charter schools, 66 at Catholic schools, and 49 at other private schools (Table 3). For participant gender: 187 were women, 63 were men. For participant race: 1 identified as Hispanic or Latino, 5 as Black or African American, 236 as White, 4 as Asian or Pacific Islander, and 4 as Biracial/Multiethnic. Nearly all teachers (94.8%) reported Reading, English/Language Arts, or Mathematics as their primary teaching assignment field.

Teachers, on average, reported teaching at their current school for 8.4 years (SD = 8.3), with a range from first year teachers to teachers of 43 years. However, 51.2% of 8<sup>th</sup> grade teachers had taught in their current school for 5 or fewer years, suggesting the mean years taught is inflated.

Table 3 *Descriptive Statistics for Total Sample (n=500)*

Descriptive	3 <sup>rd</sup> Grade (n=250)	8 <sup>th</sup> Grade (n=250)
School Type	%	%
Traditional Public	74.8	46.8
Charter	2.8	7.2
Catholic	15.2	26.4
Other Private	7.2	19.6
Gender		
Male	8.0	25.2
Female	92.0	74.8
Race/Ethnicity		
White	92.0	94.4
Non-White	8.0	5.6

Overall, the sample of teachers was predominantly female and White. The breakdown of school sector for teachers was similar to the overall breakdown of schools sampled in the larger study, with slightly more charter schools represented and slightly less traditional public school teachers in this sample. Elementary teachers overwhelmingly reported general elementary education as their primary teaching assignment, and middle school teachers reported a single subject (reading, ELA, or math), aligning with the typical structure of each respective group. Difference in average years taught at current school between the grades was not significant ( $p=0.39$ ). However, nearly half (49.6%) of all teachers had been in their current school for 5 or fewer years, with 28.8% of those teachers teaching less than 2 years at their current school.

## Initial Analysis

The Rasch RSM analysis results produced a person reliability estimate of 0.85 and a separation of 2.38 and an item reliability estimate of 0.99 with a separation of 11.63. Reliability and separation from person and item measures indicate the instrument is doing reasonably well at distinguishing between levels of the latent trait and the sample of persons is large enough to confirm item difficulties (Linacre, 2018). Both estimates provide reasonable evidence the Rasch results can be further examined. To determine dimensionality of the tool, a Principal Components Analysis (PCAR) was run using the Rasch residual estimates. Table 5 below shows the results:

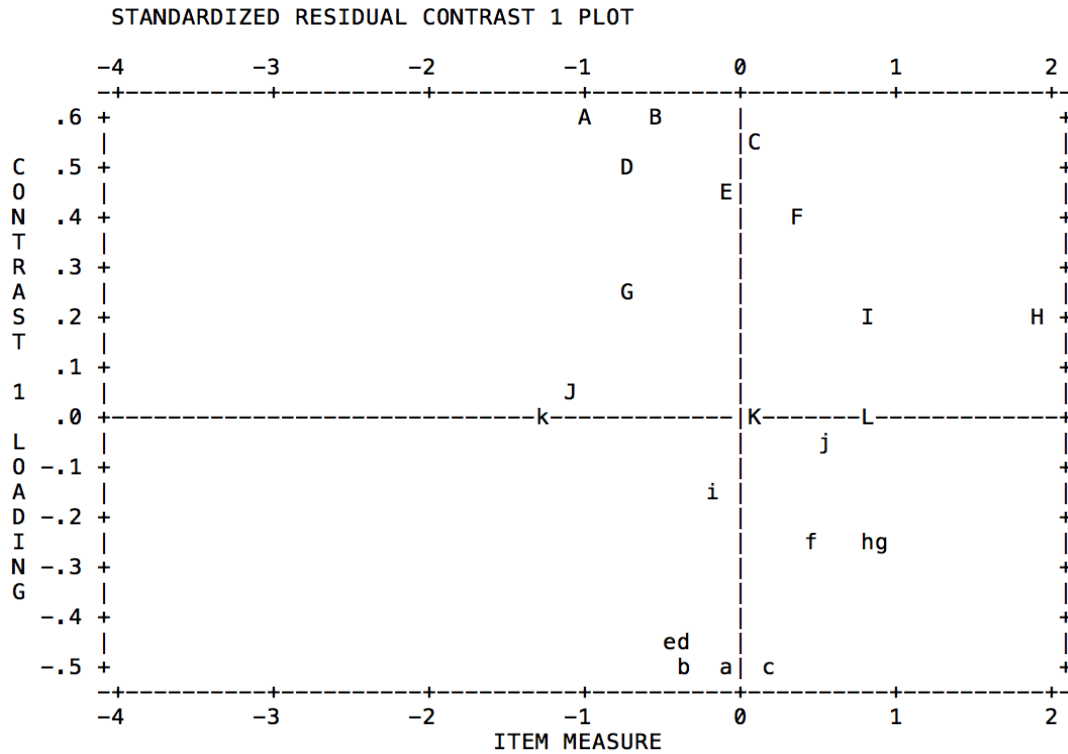
Table 4 *Initial Analysis Variance Estimates*

<i>Identifier</i>	<i>Eigenvalue</i>	<i>%</i>
Total Raw variance in observations	38.143	100.0
Raw variance explained by measures	15.143	39.7
Persons	3.826	10.0
Items	11.317	29.7
Raw unexplained variance (total)	23.000	60.3
1 <sup>st</sup> contrast	3.077	8.1
2 <sup>nd</sup> contrast	2.128	5.6
3 <sup>rd</sup> contrast	1.631	4.3
4 <sup>th</sup> contrast	1.524	4.0
5 <sup>th</sup> contrast	1.428	3.7

Results show the Rasch dimension explained 38.1% of the variance in the data. The item difficulties (29.7%) explain nearly three times the percentage in variance in the data than the person measures (10%). The first contrast had an eigenvalue of 3.08, the equivalent of 3 items, suggesting that the tool could be multidimensional. The variance explained by the first contrast is 8.1%, which is not larger than the variance explained by the item difficulties. Figure 2 provides a plot of the item residual loadings for the first contrast. Table 6 provides the standardized residual loadings for the first contrast.



Figure 2 *Initial Standardized Residuals for the First Contrast*



Of the 23 items, 10 items had positive loadings, 12 items had negative loadings, and one item loaded at zero. Examining the items in the positive loading, a majority of the items include student factors, suggesting a common theme. The items in the negative loading shared a general focus on teacher and school support related factors, suggesting a common theme. One item, “high student/teacher ratio” did not load on either. These results suggest multidimensionality may be present. Further tests were run to confirm the presence of multidimensionality. Table 6 provides the standardized residual loadings for the first contrast with the addition of the cluster each item falls under.

Table 5 *Initial Loadings for First Contrast and Cluster Group*

Item	Loading	Contrast
j. Uninterested students	.59	1
l. Parents uninterested in their children's learning progress	.59	1
r. Low morale among students	.55	1
k. Disruptive students	.49	1
t. Student absenteeism	.46	1
u. Student tardiness	.41	1
b. Students who come from a wide range of backgrounds	.25	2
c. Threat(s) to your safety or safety of students	.20	2
d. The noise level in the school building	.19	2
i. Students with different academic abilities	.03	2
h. High student/teacher ratio	.00	2
f. Students with special needs	-.15	2
a. Low morale among fellow teachers/administrators	-.05	2
n. Pressure to perform well on standardized tests	-.02	2
s. Teacher turnover in this school	-.02	2
e. Amount of professional support staff	-.49	3
p. Lack of teacher planning time built into the school day	-.49	3
v. Lack of guidance or support for teaching special education students (i.e., students with IEPs)	-.48	3
o. Lack of school resources to provide the extra help for students who need it	-.46	3
g. Amount of time to prepare for class	-.45	3
m. Access to technology	-.25	3
w. Lack of guidance or support for teaching English Language Learners	-.24	3
q. Inadequate physical facilities	-.23	3

During the analysis of residuals, items are anchored at their difficulties and clustered into subtests by which persons are measured on. The person measures from the items are correlated with each other and each cluster of items produces a standard error. High correlations suggest that the person measures on the item clusters being compared are sharing a majority of variance, which is essential in a unidimensional instrument. When viewing the clusters in the data, the disattenuated correlation between person measures on items in Clusters 1 and the person measures on items in Cluster 3 is 0.4668.

Linacre (2018, p. 396) reported that correlations below 0.57 indicate person measures on the two clusters have half the variance in common as they do independently of one another. Clusters 1 and 2 are highly correlated (.870), which is roughly three times as dependent than independent, and Clusters 2 and 3 are moderately correlated (.735), which indicates that the person measures share more than half their variance. In these suggested categories, Linacre (2018) proposed the cutoff point for different latent variables is 0.57, which would suggest that the cluster of items are measuring different sub-dimensions of school climate.

Following these results and the suggestions in Linacre (2018), the instrument will be split into two scales from the positive and negative loadings produced from the PCA of Rasch residuals. The high student/teacher ratio item had neither a positive or negative loading and will be dropped from the analyses. The 10 items focused on student-centered aspects of school climate will be analyzed, and the 12 items that focus on school/teacher support will be analyzed. Table 7 provides a breakdown of which items will be analyzed together.

Table 6 *Item Breakdown for Separate Rasch Analyses*

Student-Centered	School/Teacher Support
b. Students who come from a wide range of backgrounds	a. Low morale among fellow teachers/administrators
c. Threat(s) to your safety or safety of students	e. Amount of professional support staff
d. The noise level in the school building	f. Students with special needs
i. Students with different academic abilities	g. Amount of time to prepare for class
j. Uninterested students	m. Access to technology
k. Disruptive students	n. Pressure to perform well on standardized tests
l. Parents uninterested in their children's learning progress	o. Lack of school resources to provide the extra help for students who need it
r. Low morale among students	p. Lack of teacher planning time built into the school day
t. Student absenteeism	q. Inadequate physical facilities
u. Student tardiness	s. Teacher turnover in this school
	v. Lack of guidance or support for teaching special education students (i.e., students with IEPs)
	w. Lack of guidance or support for teaching English Language Learners

## Student Centered Analysis

The Rasch RSM analysis results produced a person reliability estimate of 0.82 and a separation of 2.11 and an item reliability estimate of 1.00 with a separation of 15.97. Reliability and separation from person and item measures indicate that the instrument is doing a reasonably good job at distinguishing between levels of the latent trait and the sample of persons is large enough to confirm item difficulties (Linacre, 2018). Both estimates provide reasonable evidence the Rasch results can be further examined. To determine dimensionality of the tool, a Principal Components Analysis (PCAR) was run using the Rasch residual estimates. Table 8 shows the results below:

Table 7 *Analysis of Variance Estimates for the Student-Centered Subscale*

<i>Identifier</i>	<i>Eigenvalue</i>	<i>%</i>
Total Raw variance in observations	22.684	100.0
Raw variance explained by measures	12.684	55.9
Persons	4.650	20.5
Items	8.034	35.4
Raw unexplained variance (total)	10.000	44.1
1 <sup>st</sup> contrast	1.928	8.5
2 <sup>nd</sup> contrast	1.638	7.2
3 <sup>rd</sup> contrast	1.454	6.4
4 <sup>th</sup> contrast	1.232	5.4
5 <sup>th</sup> contrast	1.063	4.7

The Rasch dimension explains 22.7% of the variance in the data. The item difficulties (8.0%) explain nearly twice the percentage in variance in the data than the person measures (4.7%). The first contrast had an eigenvalue of 1.93, less than the 2.0 criteria established by Linacre (2000b) for multidimensionality. Figure 3 provides a plot of the item residual loadings for the first contrast. Table 9 provides the standardized residual loadings for the first contrast. As multidimensionality is not an issue from the results of the PCA of the Rasch residuals, the item infit and outfit t-statistics were

examined. Both statistics are provided in Table 9. Infit values ranged from 0.73 to 1.36 and outfit values ranged from 0.70 to 1.37 for the 10 items.

Figure 3 *Standardized Residuals for the First Contrast for the Student-Centered Subscale*

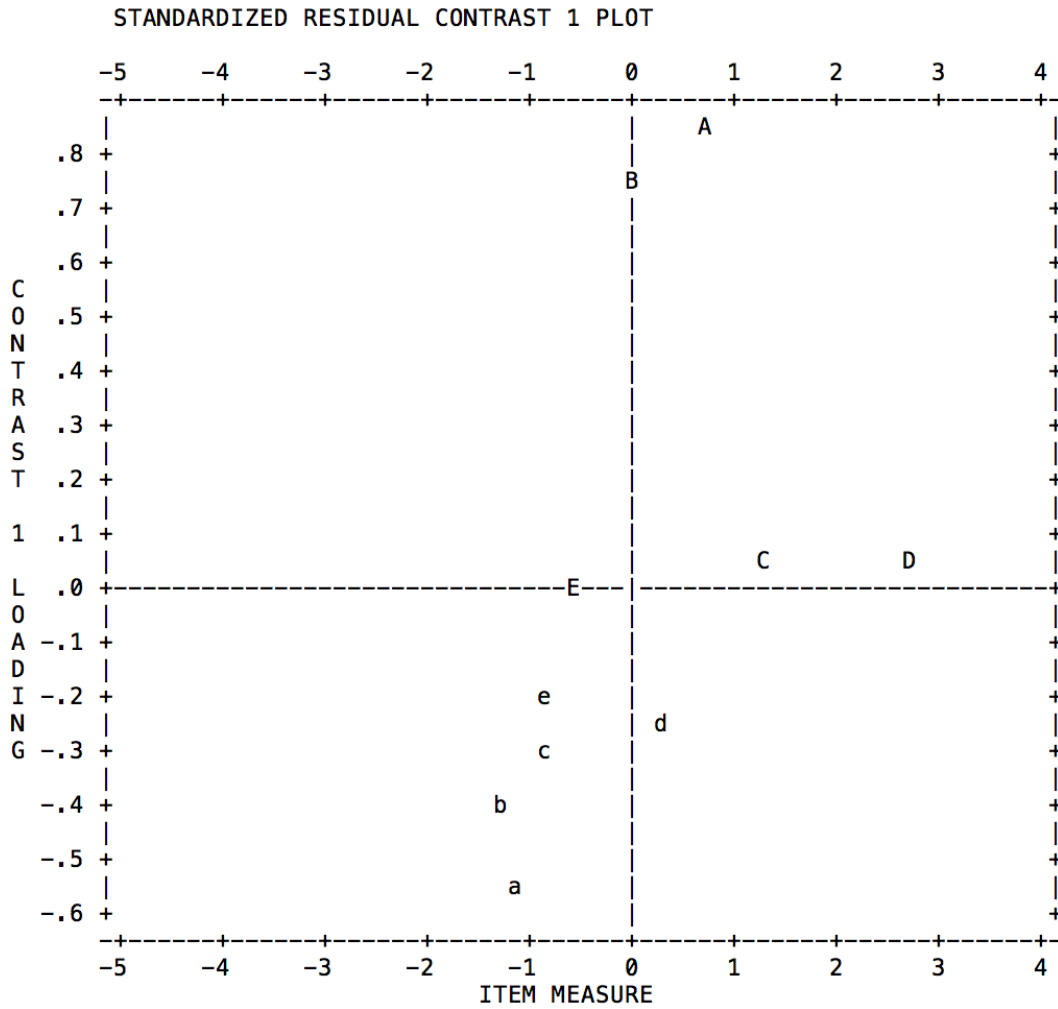
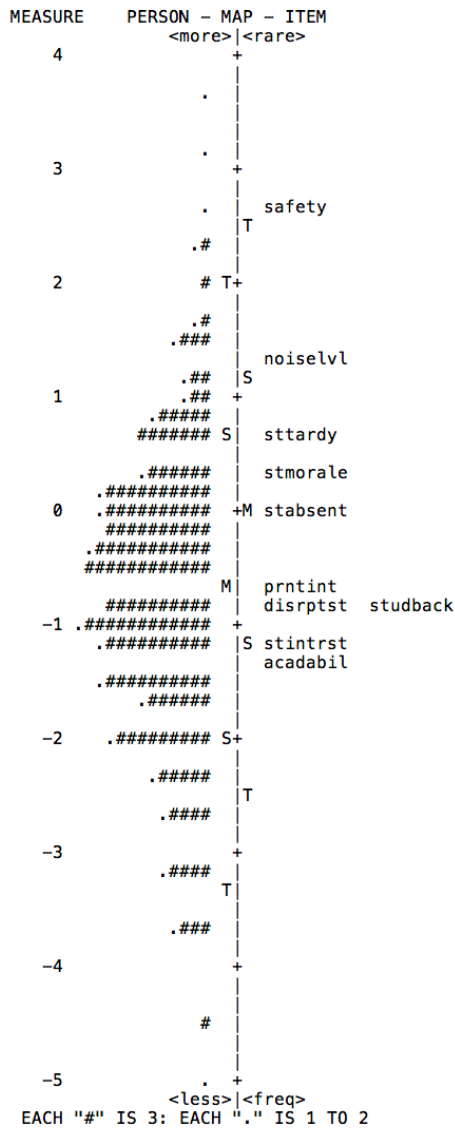


Table 8 *Item Estimates for the Student-Centered Subscale*

<i>Item</i>	<i>First Contrast Loading</i>	<i>Difficulty</i>	<i>Model S.E.</i>	<i>Infit</i>		<i>Outfit</i>	
				<i>MNSQ</i>	<i>ZSTD</i>	<i>MNSQ</i>	<i>ZSTD</i>
studback	-.18	-.85	.06	1.09	1.5	1.17	2.7
safety	.03	2.71	.10	1.22	2.4	1.14	.9
noiselv1	.05	1.30	.08	1.36	4.8	1.37	4.1
acadabil	-.42	-1.34	.06	1.07	2.3	1.18	2.6
stintrst	-.56	-1.21	.06	.73	-5.4	.70	-5.2
disrptst	-.29	-.89	.06	.87	-2.3	.85	-2.6
prntint	.00	-.63	.06	.84	-2.8	.83	-2.9
stmorale	-.26	.25	.07	.92	-1.3	.86	-2.1
stabsent	.75	.02	.07	1.08	1.3	1.10	1.5
sttardy	.83	.65	.07	1.13	2.0	1.06	.8

Item difficulty level estimates ranged between -1.34 and 2.71 (see Table 9 for values), indicating a range of difficulty for participants to endorse. The spread of difficulty of items suggests that most items fall between moderately-easy (-1.34) and moderately challenging (1.30) to endorse, with the exception of one item that is much more difficult (item c). The most challenging item for teachers to endorse was “Threat(s) to your safety or safety of students” and the least challenging item was “Students with different academic abilities.” There does not seem to be a redundancy of items at any difficulty level. A Wright map depicting the items and their difficulty is included below in Figure 4.

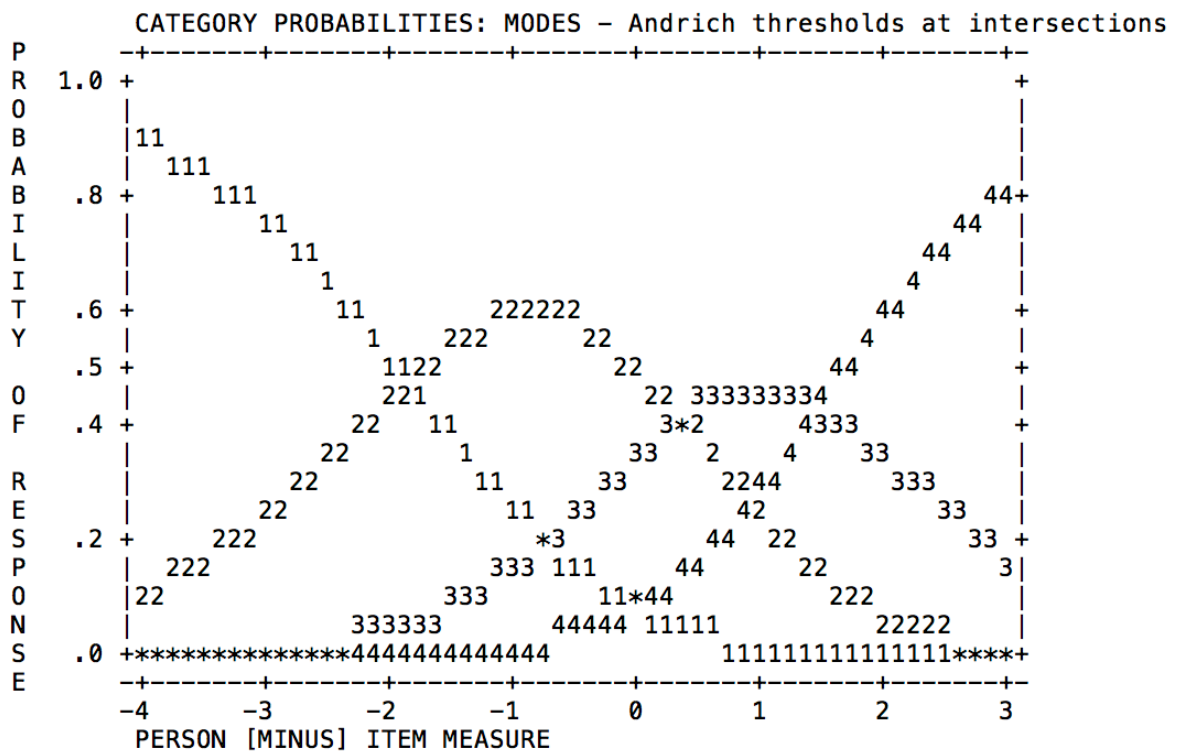
Figure 4 *Wright Map of Item Difficulty for the Student-Centered Subscale*



Thresholds that do not increase in a monotonic fashion are labeled as disordered and often more pronounced on probability curves (Bond & Fox, 2015). Linacre (1999) recommends thresholds increase by a minimum of 1.4 logits but a maximum of 5 logits. For the student-centered items, the Andrich thresholds for all ten items increased in a monotonic fashion. Nine items in the instrument have at least ten responses for each category. The one item that does not have ten responses deals with student and teacher safety in the classroom. Overall, the Andrich thresholds do not fully meet Linacre's

(1999) criteria; the distance between the first and second does, but the distance between the second and third does not meet the minimum of 1.4 logits difference. Graphically, the probability curves are somewhat flatter than ideal (Figure 5). This suggests response categories may not provide distinct definitions on the variable (Bond & Fox, 2015). However, with only four response categories and three thresholds, it would not be advisable to collapse any response categories.

Figure 5 *Category Probabilities for the Student-Centered Subscale*





To determine if differential item functioning occurred (DIF), a F-test was performed see if the items functioned similarly for 3<sup>rd</sup> and 8<sup>th</sup> grade teachers. DIF results are reported in Table 10 below:

Table 9 *DIF Analysis of 3<sup>rd</sup> and 8<sup>th</sup> grade teachers for the Student-Centered Subscale*

<i>Item</i>	<i>DIF Contrast</i>	<i>t</i>	<i>df</i>	<i>p</i>
studback	-.38	-3.02	496	.003
safety	-.82	-3.80	476	<.001
noiselyl	-.22	-1.39	496	.16
acadabil	-.21	-1.65	495	.10
stintrst	.58	4.56	496	<.001
disrptst	-.50	-3.92	496	<.001
prntint	-.20	-1.55	496	.12
stmorale	.76	5.52	492	<.001
stabsent	.69	5.14	493	<.001
sttardy	-.17	-1.17	495	.24

There is a significant amount of DIF in a majority of the items on the student-centered subscale between the reference group (3<sup>rd</sup> grade) and the focal group (8<sup>th</sup> grade). In fact, six of the ten items are significant at the  $\alpha = .01$  level. Items with an absolute DIF contrast score from 0.43 to 0.63 logits can be considered slight to moderate presence of DIF. Items with an absolute DIF contrast score greater than or equal to 0.65 logits can be considered moderate to large (Zwick, Thayer, & Lewis, 1999). Following these criteria, three items have moderate to large DIF and two items show slight to moderate DIF. The three items with moderate to large DIF contrasts: item c, item r, and item t. Item c is “Threat(s) to your safety or safety of students,” item r is “Students with low morale,” and item t is “Student absenteeism”. The two items with slight to moderate DIF: item j and item k. Item j is “uninterested students” and item k is “disruptive students”. The remaining five items fall outside the criteria and are considered to have slight to no DIF.

## School/Teacher Support Analysis

The Rasch RSM analysis results produced a person reliability estimate of 0.76 and a separation of 1.79 and an item reliability estimate of .99 with a separation of 10.10. Reliability and separation from person and item measures indicate that the sample of persons is large enough to confirm item difficulties, but the instrument may not be distinguishing between levels of the latent trait well following Linacre (2018) recommendation of a person reliability greater than 0.80. Both estimates provide reasonable evidence the Rasch results can be further examined. To determine dimensionality of the tool, a Principal Components Analysis (PCAR) was run using the Rasch residual estimates. Table 11 shows the results below:

Table 10 *Analysis of Variance Estimates for the School/Teacher Support Subscale*

<i>Identifier</i>	<i>Eigenvalue</i>	<i>%</i>
Total Raw variance in observations	20.574	100.0
Raw variance explained by measures	8.574	41.7
Persons	2.76	13.4
Items	5.82	28.3
Raw unexplained variance (total)	12.000	58.3
1 <sup>st</sup> contrast	2.012	9.8
2 <sup>nd</sup> contrast	1.412	6.9

Results show the Rasch dimension explains 20.6 % of the variance in the data. The item difficulties (5.8%) are explaining over twice the percentage in variance in the data than the person measures (2.8%). The first contrast had an eigenvalue of 2.1, which is greater than the random noise criteria established by Linacre (2000b) for multidimensionality. Figure 6 provides a plot of the item residual loadings for the first contrast. Table 12 provides the standardized residual loadings for the first contrast.

Figure 6 *Standardized Residuals for the First Contrast for the School/Teacher Support Subscale*

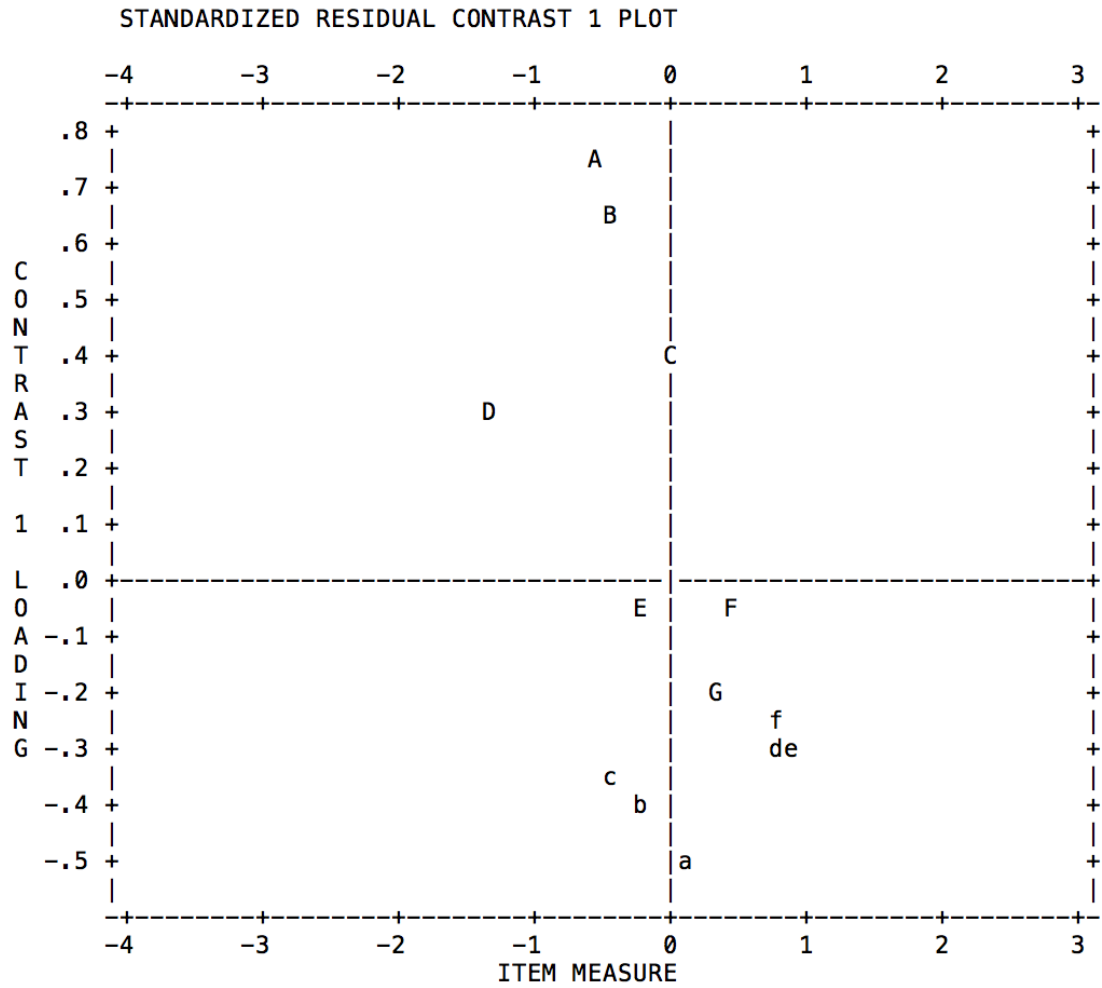


Table 11 *Item Estimates for the School/Teacher Support Subscale*

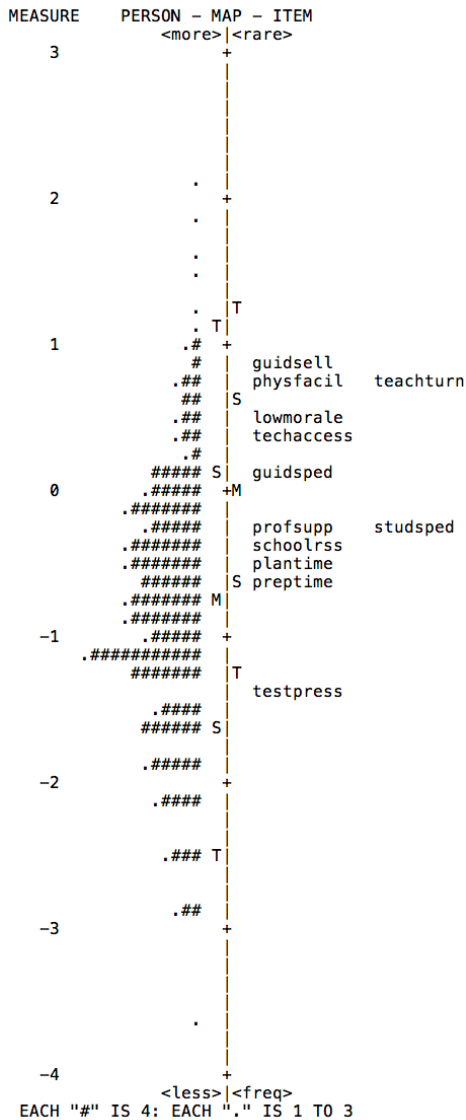
<i>Item</i>	<i>First Contrast Loading</i>	<i>Difficulty</i>	<i>Model S.E</i>	<i>Infit</i>		<i>Outfit</i>	
				<i>MNSQ</i>	<i>ZSTD</i>	<i>MNSQ</i>	<i>ZSTD</i>
preptime	.82	-.58	.05	.97	-.5	.99	-.2
plantime	.78	-.49	.05	.99	-0.2	.97	-0.5
testpress	.26	-1.36	.06	1.12	2.0	1.13	2.0
guidsped	-.47	.13	.06	.95	-0.8	.92	-1.1
profsupp	-.33	-.22	.06	.78	-4.1	.77	-2.1
guidsell	-.33	.84	.07	1.23	3.0	1.13	1.5
teachturn	-.29	.75	.07	1.36	4.7	1.24	2.7
schoolrss	-.28	-.42	.05	.64	-7.4	.64	-6.7
physfacil	-.20	.76	.07	1.09	1.3	1.16	1.9
techaccess	-.14	.36	.06	1.36	5.2	1.38	4.7
studped	.04	-.25	.06	.88	-2.1	.92	-1.3
lowmorale	-.03	.48	.06	.98	-.3	1.00	.0

Of the twelve items, four items had a positive loading, and eight items had a negative loading. When examining the items closer, although there is an imbalance in items loading on each side, it appears there is no indication similar items are grouped together under a certain theme within the instrument. This suggests multidimensionality should not be of concern and is likely a result of the larger facet of school/teacher support factors in school climate. As multidimensionality is not an issue from the results of the PCA of the Rasch residuals, the item infit and outfit t-statistics were examined. Both statistics are provided in Table 12. Infit values ranged from 0.64 to 1.36 and outfit values ranged from 0.64 to 1.38 for the twelve items.

Item difficulty level estimates ranged between -1.36 and .84 (see Table 12 for values), indicating a range of difficulty for participants to endorse. A Wright map depicting the items and their difficulty is included below in Figure 7. The spread of difficulty of items is clustered together and suggests that most items fall between easy and moderately-easy to endorse, with the exception of one item that is much easier than others (testpress). The most challenging item for teachers to endorse was “Lack of

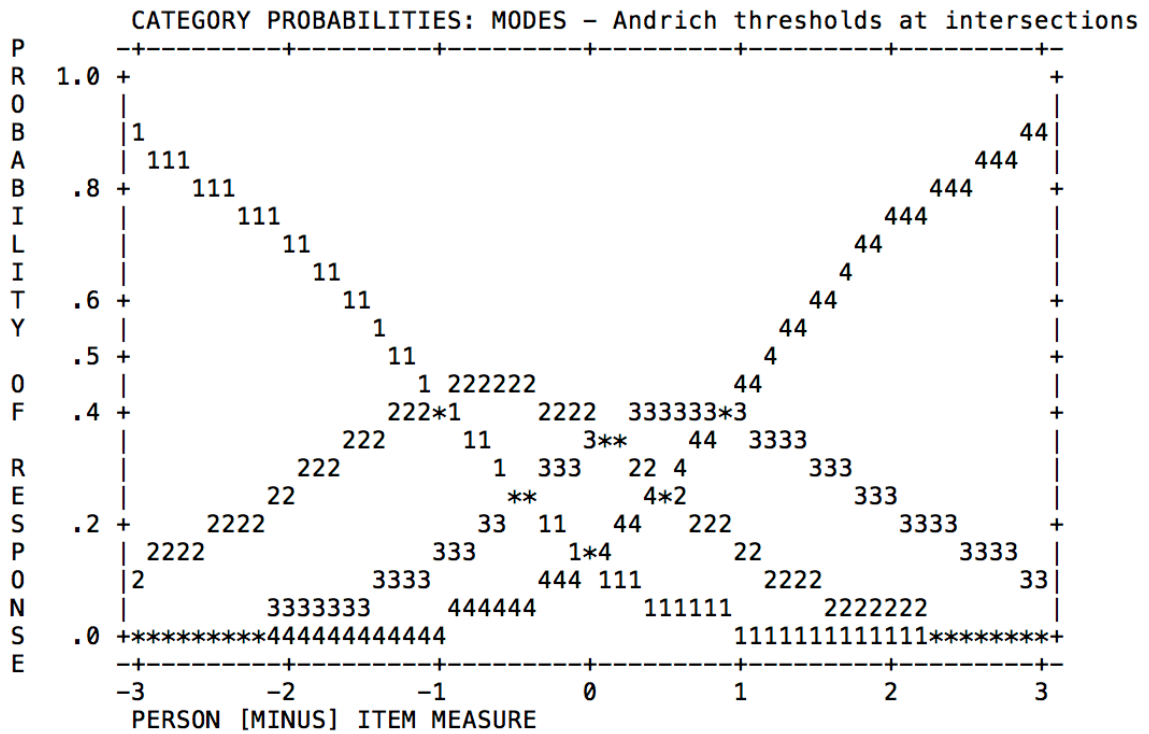
guidance or support for teaching English language learners” and the least challenging item was “Pressure to perform well on standardized tests.” Most of the items on this scale fall within -.05 to 1 logits of the latent trait. In addition, there is some overlap in difficulty for two pairs of items. The preptime and plantime items are very similarly worded, suggesting some level of repetition.

Figure 7 *Wright Map of Item Difficulty for the School/Teacher Support Subscale*



For the teacher support/school items, thresholds for all thirteen items increased in a monotonic fashion. All categories have at least ten responses for all items. However, the Andrich thresholds do not meet Linacre's (1999) criteria. Graphically, the probability curves are much flatter than ideal, which could be problematic (Figure 8). With only four response categories and three thresholds, it would not be advisable to collapse any response categories.

Figure 8 *Category Probabilities for the School/Teacher Support Subscale*



To determine if differential item functioning occurred (DIF), a F-test was performed see if the items functioned similarly for 3<sup>rd</sup> and 8<sup>th</sup> grade teachers. DIF results are reported in Table 13 below:

Table 12 *DIF Analysis of 3<sup>rd</sup> and 8<sup>th</sup> grade teachers for the School/Teacher Support Subscale*

<i>Item</i>	<i>DIF Contrast</i>	<i>t</i>	<i>df</i>	<i>p</i>
preptime	-.16	-1.43	497	.15
plantime	-.26	-2.36	497	.02
testpress	-.33	-2.90	396	.004
guidsped	..09	.75	497	.46
profsupp	-.06	-.52	497	.61
guidsell	-.15	-1.08	495	.28
teachturn	.62	4.58	487	.00
schoolrss	.08	.71	497	.48
physfacil	.32	2.37	496	.02
techaccess	-.08	-.68	497	.50
studped	-.05	-.46	497	.65
lowmorale	.27	2.20	497	.03

As demonstrated by the table, there is a significant presence of DIF in almost half of the items on the school/teacher support subscale between the reference group (3<sup>rd</sup> grade) and the focal group (8<sup>th</sup> grade). Five items are significant at the  $\alpha = .05$  level and two of these items are significant at the  $\alpha = .01$  level. Following Zwick, Thayer, and Lewis (1999) criteria, one item possessed moderate to large DIF: teachturn. This item deals with teacher turnover in the school. The remaining eleven items fall outside the criteria and are considered to have slight to no DIF. The results for the school/teacher support subscale suggest DIF is present for teachers in elementary and middle school, but not to a great degree, with the exception of one item.

## **Conclusion**

This chapter detailed the results of the study. It began with a brief review of the purpose of the study and analysis techniques used. Descriptive details about the study participants were provided. Finally, the results from the Rasch RSM analysis, including dimensionality, item fit, item difficulty, thresholds, and DIF estimates were presented. As the PCAR results suggested multidimensionality, the instrument was split into two subdimensions and a Rasch RSM was run on each. The following chapter will address the research questions of this study with the results presented in this chapter and discuss possible implications and limitations.



## CHAPTER 5. CONCLUSION

### Overview

The purpose of this study was to validate a school climate instrument based on the four most commonly accepted dimensions of school climate, using items adapted from the 5Essentials framework to provide an effective measure for educators and researchers. The study began by providing a background for school climate and school climate measures as well as a brief introduction to the current federal and state accountability policies. Then, a review of literature regarding school climate, the foundation for the instrument validated in this study was provided. How school climate literature has changed over time, the differences in dimensions, participants, and instrument analysis, and an account of school climate results was presented before highlighting appropriate differences in school level were addressed. Concluding this chapter was a description of the primary method of analysis: a one-parameter item response theory model used for polytomous data, specifically the Rasch Rating Scale Model.

Following the literature review, a description of the methodology used to validate the school climate instrument was provided. Information from a pilot study was given prior to information about the instrument, participant sampling, and data collection. The last portion of provided a review of the Rasch Rating Scale Model analysis used for the study in addition to how each research question was addressed in the study. After descriptive details about study participants were provided, the results from the Rasch RSM analysis were presented.

For the validation of this school climate instrument with the Rasch RSM, dimensionality was first assessed. Recognizing that assumptions of unidimensionality were violated, the instrument was split into two subdimensions: student-centered and

teacher support/school factors. From here, analyses were run for each subdimensions, including dimensionality, item fit, item difficulty, threshold parameters, and DIF estimates between 3<sup>rd</sup> and 8<sup>th</sup> grade teachers. The following chapter will address each of the research questions of this study with results presented in the previous chapter. Concluding this chapter will be a discussion of the contributions of the study to the larger school climate literature, followed by limitations, and implication from this study for policymakers, researchers, and practitioners.

## **Findings**

1. How well does the instrument measure the latent trait of school climate?

To answer the first research question, unidimensionality was assessed by conducting a principal components analysis (PCAR) of the Rasch residuals. A PCAR does not show loadings on any factor like a common factor analysis. Instead, the least amount of contrast between residuals is preferred as it suggests that the maximum amount of variance is being explained (Linacre, 2000a). A “second” factor must have an eigenvalue greater than 3 to be considered multidimensional (Linacre, 2000b). If not, the loadings suggest a contrasting pattern in the residuals that requires closer examination of the content of the items.

If a second factor has an eigenvalue above this value, the standardized residuals for the positive and negative loading items for this contrast are examined for any sort of patterns (Linacre 2018). A PCAR factor loading above 3 is not definitive, but is highly suggestive of multidimensionality (Linacre, 2000). If items are generally clustered and no particular theme can be determined, this could be a result of the larger dimension of school climate being multifaceted. If this is the case, there should be no concern of violating assumptions and analyses can continue.

Results from the PCAR demonstrated the school climate instrument measured two distinct subdimensions. This argument is based on a variety of diagnostic tests. First, the eigenvalue of the first contrast is above the cutoff of 3. When viewing loadings of the 23 items, 10 items loaded positively, 12 items loaded negatively, and one item loaded at zero. Items with a positive loading shared a common theme of student related factors. Items with a negative loading shared a common focus on teacher and school support related items. The item, “high student/teacher ratio” loaded at zero and was dropped from further analyses. The third metric used to determine unidimensionality was the correlations between the clusters of item loadings. High correlations suggest person measures on the item clusters being compared are sharing a majority of variance, which is essential in a unidimensional instrument. Linacre (2018) proposed the cutoff point for different latent variables is 0.57. The correlation between person measures on items in Clusters 1 and the person measures on items in Cluster 3 is 0.4668, suggesting the clusters of items are measuring different sub-dimensions of school climate.

Following these diagnostics and the suggestions in Linacre (2018), the instrument was split into two scales from the positive and negative loadings produced from the PCAR of the Rasch residuals: Student Centered and School/Teacher Support. Unidimensionality was assessed for both the Student Centered and School/Teacher Support subdimensions of school climate. For the Student Centered subdimension, the first contrast (1.9) was below the 2.0 criteria established by Linacre (2000b). For the School/Teacher Support subdimension, the first contrast (2.0) met the 2.0 criteria established by Linacre (2000b). When examining the items closer, there was no indication similar items were grouped together under a certain theme within the two subdimensions.

In assessing the unidimensionality of the instrument, the decision was made to split the instrument into two subdimensions of school climate and drop one item: Student Centered and School/Teacher Support. As described in Chapter 3, one issue in combining items from other instruments is that both had been operationalized as measuring different constructs (institutional challenges and school climate). Factor analysis of the Philly school climate instrument resulted in a number of factors. The institutional challenges scale possessed a high alpha reliability. There was also the addition of two items that had no prior scale or validation. However, these items all shared the same question stem: “To what extent do you consider each of the following factors *a challenge* to student learning in your classroom and/or school?” If analysis had suggested the combined instrument was not unidimensional, it was hypothesized items may cluster around their original scale measures. However, this did not occur. The IC and Philly scales were evenly split between the two subdimensions found, suggesting that it was not the combination of any prior scale impacting the dimensionality. The loadings, rather, reflected an overall split in items that dealt with student issues and those that dealt with teacher/school support issues.

2. How well do the individual instrument items reflective of school climate fit to the Rasch model?

To answer the second research question, item fit and item difficulty of the instrument were analyzed. Fit statistics need infit or outfit values less than 2.0 to be considered acceptable in fitting the model. For the Student-Centered subscale, infit values ranged from 0.73 to 1.36 and outfit values ranged from 0.70 to 1.37 for the 10 items. For the School/Teacher Support subscale, infit values ranged from 0.64 to 1.36 and outfit values ranged from 0.64 to 1.38 for the twelve items. Overall, the infit and outfit statistics

for both subscales fell within the acceptable range of 0.5-1.5, suggesting the items are effective in measuring teachers' latent ability levels (Linacre, 2018).

The instrument should contain a range of items that vary in difficulty of endorsement in order to accurately represent a wide range of ability in participants. For item difficulty, values below 0 indicate a question is easier to endorse, and values above 0 are more difficult. Items that are considered more difficult are more likely to be endorsed by persons with more of the latent trait being measured (De Ayala, 2009). Items that hover right around the 0 mark are suggested to be of average difficulty (De Ayala, 2009). As both people and items reside on the same logit continuum, person locations and item locations share the same scale and the above range.

For the Student-Centered subscale, item difficulty level estimates ranged between -1.34 and 2.71, indicating a range of difficulty for participants to endorse. The spread of difficulty of items suggests that most items fall between moderately-easy (-1.34) and moderately challenging (1.30) to endorse, with the exception of one item that is much more difficult (safety). The most challenging item for teachers to endorse was "Threat(s) to your safety or safety of students" and the least challenging item was "Students with different academic abilities." From a practical point, it is a positive sign the item "Threat(s) to your safety or safety of students" is so difficult to endorse, as it means a majority of teachers find it difficult to strongly endorse threats as a significant challenge to student learning in their classrooms. It is also logical that many teachers would believe a range in academic abilities among students could pose as a challenge to student learning. Viewing the range in difficulty of the ten items, there does not seem to be a redundancy of items at any difficulty level.

Item difficulty level estimates ranged between -1.36 and .84, indicating a range of difficulty for participants to endorse. The spread of difficulty of items is clustered together and suggests that most items fall between easy and moderately-easy to endorse, with the exception of one item that is much easier than others (testpress). The most challenging item for teachers to endorse was “Lack of guidance or support for teaching English Language Learners” and the least challenging item was “Pressure to perform well on standardized tests.” Item guidsell, dealing with the lack of support for teaching English Language Learners (ELL), suggested teachers find it difficult to positively endorse this item as a challenge to student learning. This might indicate teachers are receiving adequate support for ELL students, or that such support is not required for their student base. Placing testpress within the context of current state and federal accountability policies, it is not surprising teachers find it relatively easy to strongly endorse the pressure to perform well as a challenge to student learning and a contributor to a more negative school climate. Promotion and assignment to grades, merit pay, and the accountability grade a school receives are contingent on students’ standardized state test scores. There does seem to be some redundancy in the difficulty of items on this subscale, as most items are clustered between -0.5 and 1.0 logits. Additionally, the two items dealing with teacher planning time and teacher prep time are very similarly worded, suggesting some level of redundancy.

Category frequencies summarize the distribution of responses to each item in the instrument (Bond & Fox, 2015). Both the shape of the distribution and the frequency of responses per category should be examined. Ideally, Linacre (1999) recommends a minimum of ten responses per category. A low frequency of item responses will not provide a stable estimation of threshold values (Bond & Fox, 2015). In cases where

categories are not functioning properly, it may be ideal to collapse certain categories. Threshold parameters are calculated from the assumption that moving between response categories may not be equal. Thresholds that do not increase in a monotonic fashion are labeled as disordered and should be examined more in depth (Bond & Fox, 2015). Thresholds should not be too close, yet also not too far away from one another to cover the span of a respondent's ability. Linacre (1999) recommends thresholds increase by a minimum of 1.4 logits but a maximum of 5 logits. Thresholds that are problematic will appear graphically as flattened curves that cover less of the measured variable (Bond & Fox, 2015). Flatter curves indicate response categories may not provide distinct definitions on the variable, and that respondents are not using the rating scale as the model would predict (Bond & Fox, 2015).

For the student-centered items, the Andrich thresholds for all ten items increased in a monotonic fashion. Nine items in the instrument have at least ten responses for each category. The one item that does not have ten responses deals with student and teacher safety in the classroom. Overall, the Andrich thresholds do not fully meet Linacre's (1999) criteria; the distance between the first and second does, but the distance between the second and third does not meet the minimum of 1.4 logits difference. Graphically, the probability curves were flatter than ideal, suggesting that response categories might not provide distinct steps between response categories (Bond & Fox, 2015). For the teacher support/school items, thresholds for all twelve items increased in a monotonic fashion. All categories have at least ten responses for all items. However, the Andrich thresholds do not meet Linacre's (1999) criteria for any item. Graphically, the probability curves are much flatter than ideal, which could be problematic with response categories not providing distinct categories for respondents to choose from (Bond & Fox, 2015).

However, with only four response categories and three thresholds, no response categories were collapsed. Overall, an effective instrument should contain a range of item difficulties to accurately represent the varying levels of ability in participants. Collectively, the two subscales do a reasonably well job spanning a variety of difficulty levels, from easy and moderately easy to moderately challenging.

3. Are there differences in item responses from teachers from 3<sup>rd</sup> and 8<sup>th</sup> grades with similar levels of the latent trait of school climate?

To answer the third research question, a differential item functioning (DIF) analysis was conducted between 3<sup>rd</sup> and 8<sup>th</sup> grade teachers. The DIF analysis determines if participants from different groups exhibit systematic variation in responses on specific items. Items that exhibit significant DIF results could be the results of a bias in the item, or an issue not relevant to the trait being measured by the instrument. If differences exist, this could suggest that respondents could be viewing items differently, or that a true difference exists in the latent trait between members of certain groups.

There is a significant amount of DIF in a majority of the items on the student-centered subscale between the reference group (3<sup>rd</sup> grade) and the focal group (8<sup>th</sup> grade). In fact, six of the ten items are significant at the  $\alpha = .01$  level. The DIF contrast score indicates the difference in item difficulty between the groups (Linacre, 2018). Items with an absolute DIF contrast score from 0.43 to 0.63 logits can be considered slight to moderate presence of DIF (Zwick, Thayer, & Lewis, 1999). Items with an absolute DIF contrast score greater than or equal to 0.65 logits can be considered moderate to large. Following these criteria, two items would be considered to show slight to moderate DIF and three items would be considered to have moderate to large DIF. The remaining five items fall outside the criteria and are considered to have slight to no DIF.



Three items exhibited moderate to large DIF contrasts: safety, stmorale, and stabsent. Safety is “Threat(s) to your safety or safety of students”; stmorale is “Students with low morale”; and stabsent is “Student absenteeism”. For items r and t, the value of the contrast and the t-value were positive, indicating that 3<sup>rd</sup> grade teachers in this sample systematically find this item more difficult to endorse (and see this as less of a challenge to student learning/a more positive school climate) when compared with 8<sup>th</sup> grade teachers having the same amount of the latent trait. Students with low morale and absenteeism rates seem to be more negatively impacting school climate for 8<sup>th</sup> grade teachers. Student disengagement in the classroom continues to be a problem, particularly for secondary schools, although disengagement also occurs in primary schools as well (Marks, 2000; Strambler & Weinstein, 2010). Gottfried (2019) found that chronic absenteeism is damaging academically to both absent students and their classmates.

For safety, the value of the contrast and t-value were negative, meaning that the 3<sup>rd</sup> grade teachers in this sample, on average find this item less difficult (and see this as a greater challenge to student learning/a more negative school climate) when compared with 8<sup>th</sup> grade teachers with the same amount of the latent trait. Item c is by far the hardest item to positively endorse across both subscales. It has the largest DIF contrast, with a DIF measure of 2.34 for 3<sup>rd</sup> grade and 3.16 for 8<sup>th</sup> grade teachers. It is also one of the only items between both subscales lacking 10 responses for each category. Looking descriptively at the data, a significant difference exists between teachers endorsing the “not a challenge” (71% of 3<sup>rd</sup>; 95% of 8<sup>th</sup>) and “slight challenge” (24% of 3<sup>rd</sup>; 14% of 8<sup>th</sup>) options. This finding is somewhat contradictory to literature findings that episodes of misbehavior and violence increase in secondary schools (Theriot & Dupper, 2010). Perhaps there was some ambiguity in the phrase “Threat(s) to your safety or safety of

students” as conceptualized by the two groups of teachers. For example, teachers may perceive threats to themselves and threats to students as two different concepts. In addition, threats could refer to those coming inside the school, from or by students to the teacher or other students, or threats to safety could refer to a hypothesized outside occurrence.

The two items demonstrating slight to moderate DIF are *stintrst*, uninterested students, and *disrptst*, disruptive students. *Stintrst* had a positive DIF contrast, meaning that 3<sup>rd</sup> grade teachers in this sample systematically find this item more difficult (and see this as less of a challenge to student learning/a more positive school climate) than 8<sup>th</sup> grade teachers. As discussed previously, student disengagement is a problem, particularly in secondary grades. This phenomenon could be a result of the tension between middle school instruction and the students’ needs for autonomy (Cappella et al., 2017) It is not surprising that teachers’ rating students with low morale as a significant challenge would also endorse an item dealing with uninterested students as a significant challenge to student learning. *Disrptst* had a negative DIF contrast, meaning that 3<sup>rd</sup> grade teachers on average see disruptive students as more of a challenge to student learning. This difference in item scores between 3<sup>rd</sup> and 8<sup>th</sup> grade teachers could be linked to 3<sup>rd</sup> grade teachers’ higher scores in item c, the safety of students and the teacher.

Taken together, these findings suggest teachers in 8<sup>th</sup> grade perceived student absenteeism, uninterested students, and low morale as connected, and a problem to student learning in their classrooms and negative school climate at a higher average rate than 3<sup>rd</sup> grade teachers. This finding is particularly relevant as the indicator Indiana chose for school quality and student success is a measure of student attendance. As the 8<sup>th</sup> grade teachers in this sample, on average, perceived student absenteeism and student

disengagement as more problematic to student learning than 3<sup>rd</sup> grade teachers, this could translate into additional stress for teachers.

DIF was found in almost half of the items on the school/teacher support subscale between the reference group (3<sup>rd</sup> grade) and the focal group (8<sup>th</sup> grade). Five items are significant at the  $\alpha = .05$  level and two of these items are significant at the  $\alpha = .01$  level. Following Zwick, Thayer, and Lewis (1999) criteria, one item showed moderate to large DIF. The item (teachturn), dealing with teacher turnover, has a positive DIF contrast, indicating 3<sup>rd</sup> grade teachers have a lower item difficulty (and see this as more of a challenge to student learning/a more positive school climate) on this item when compared with 3<sup>rd</sup> grade teachers having the same amount of the latent trait. This could indicate middle school teachers do not perceive turnover as impacting their school climate as much, or in a different way than elementary school teachers. The remaining eleven items fall outside the criteria and are considered to have slight to no DIF. The results for the school/teacher support subscale suggest that DIF is present for teachers in elementary and middle school, but marginally for most items.

### **Limitations**

There are some limitations to this study. First, while both principals and teachers were surveyed in the larger study, only teachers were given the school climate instrument. This measure does not include the perspectives of administrators, students, or parents, which could provide a more comprehensive picture of school climate (Ramsey et al., 2016). However, research has suggested students and teachers are similar in their perceptions of school climate within the same school (Higgins-D'Alessandro & Guo, 2009). In addition, the survey was given to teachers once, which can only provide teachers' perceptions of school climate at one point in time. A longitudinal or repeated

measures study could provide information on how school climate perceptions may change over time, and in response to policy changes.

Teachers in 3<sup>rd</sup> and 8<sup>th</sup> grades were specifically selected for both the Rasch RSM analysis and the DIF analysis as differences were hypothesized between the two groups, stemming from relevant literature (e.g. classroom environment, student development) and policies (e.g. certification requirements, curriculum, merit pay). Results showed DIF played a significant role in the student-centered subscale of items but did not have an impact on teachers' perceptions of teacher support and other organization items as related to school climate. To this end, DIF in the instrument indicated the variation in teachers' perceptions on students impacting school climate in their schools in a fundamentally different way than the variation in responses for school and teacher support items. Although a random sampling of teachers was used in the study, it is unknown if differences would remain if other grades were included or if a different or larger sample of 3<sup>rd</sup> and 8<sup>th</sup> grade teachers were used.

The pilot analysis indicated slight DIF between teachers in different school types on certain items with a small sample. However, the decision was made to include teachers from all school types in this study as to provide an instrument that could measure teachers' perceptions of school climate, regardless of employment at a public or private school. Future data analysis should consider school type as a potential source of DIF. Additionally, combining both school type and grade level could provide additional information on how teachers within certain grades within certain schools (e.g. traditional public vs. private) may differ in their perceptions of school climate.

The sample taken is reflective of the larger study sample and the general teacher population in the state, is relatively homogenous both in gender and race/ethnicity.

Underrepresented teachers in the sample may provide valuable evidence that could be covered by more robustly represented groups. In addition, a majority of teachers in the Rasch RSM analysis had less than 5 years of teaching in their current school. Teachers with more experience could be more acclimated to the norms of teaching and perceive different factors contributing to school climate. Comparing teachers with varying levels of teaching experience could illustrate other differences.

If comparing the breakdown of items on this measure to the four most commonly recognized areas or elements of school climate, relationships among those in the school, school safety, teaching and learning, and the institutional environment, there is a clear abundance of items dealing with teaching and learning, and the institutional environment, and a clear lack of items dealing with relationships and school safety. A single item addressed perceptions of school safety in the measure, and no items addressed the relationships between members of the school community, a substantial limitation of this measure. As it stands, the 22 items measured in this study represent an incomplete version of school climate. Additional items dealing with these two aspects should be included in this measure to provide a more comprehensive picture of teachers' perceptions of school climate.

### **Implications**

School climate is a complex topic, with many facets (Thapa et al., 2009; Rudasill et al., 2017). When taken together as a measure of school climate, the instrument did not meet the criteria for unidimensionality, as determined by the results from the analysis of the Rasch residuals, the residuals loading on common themes, and the disattenuated correlations of the clustered loadings, and the extreme presence of DIF between 3<sup>rd</sup> and 8<sup>th</sup> grades. From these metrics, the decision was made to separate the school climate

instrument into two subscales, retain all items, and analyze each subscale with a Rasch RSM. From here, all items had appropriate infit and outfit statistics. As latent analyses are rarely used in school climate validation, there was little precedent as to what could happen as data was analyzed. However, as multidimensionality in school climate is a hallmark in the literature, it is not surprising the scale did not pass unidimensionality criteria as defined in the Rasch analysis (Wang & Degol, 2016). Any subsequent use of this measure should acknowledge the multifaceted nature of school climate, and not treat the measure as unidimensional. Results from this sample of 3<sup>rd</sup> and 8<sup>th</sup> grade teachers suggest teachers' underlying perceptions of school climate are impacted by grade level.

This measure was constructed using items from at least two sources that shared the same question stem: "To what extent do you consider each of the following factors a *challenge* to student learning in your classroom and/or school?" The previous usage of items had been operationalized as measuring two different constructs using some overlapping items (institutional challenges and school climate). Many of the items focused on teachers' perceptions of certain events as being a challenge to student learning in their classroom. Table 2, which breaks down the constructs previous items were under, indicated that a majority of the initial 23 items were recognized as institutional or school level challenges. This construct carried through to the teacher/school support subscale. Items that dealt with students themselves did not have a clear previous construct.

It is unique that a measure, formed from the same theoretical framework, using the same question stem, split not by the two previously used instruments, not by the dimensions commonly identified as part of school climate, but along the lines of student focused items, and teacher/school focused items. It is likely the difference between the student-centered and teacher/school support subdimensions is due to the overall

broadness of the concept of school climate. Perhaps the student-centered items are measuring the concept of classroom climate, which is often a component of school climate. For instance, all of the student-centered items include the word “student” and many of the teacher/school support items include phrases like “lack of guidance or support”, “amount of”, “teacher”, or “teaching”. Linacre (2018) illustrated a similar phenomenon when describing math assessments with a variety of problems:

“Mathematics includes arithmetic, algebra, geometry and word problems. Sometimes we want a ‘geometry test’. But, for most purposes, we want a ‘math test’.” Still, it is important these two areas are distinct, because one focuses on student-related issues, things perhaps teachers may have more direct control over, and the other teacher/school support related issues that teachers may not have control over. Both are assessing the climate of a school, in different ways through the perspective of teachers.

However, as Zieky (1993) wrote, “It is important to realize that DIF is *not* a synonym for *bias*” (p.340) and that raw differences cannot alone determine if an item is unfair to certain groups. When interpreting DIF, potential differences between two groups should be considered. Is a difference in item functioning due to item bias or an issue beyond the scope of the item or instrument (Bond & Fox, 2015)? DIF may occur as the result of one group having less experience or background knowledge related to the item or items in an instrument (Setari, 2016). It has been suggested one general cause of DIF is multidimensionality in items measuring some concept outside of the primary latent variable (Roussos & Stout, 1996). However, significant DIF may also be explained from the general complexity of items that may represent one of more latent variables simultaneously (McDonald, 2000). Lippincott, Williams, and Wilkins (2006) discussed the “relative bias” that may occur if individuals are rating themselves as relative to others

in the same setting. For instance, a teacher may perceive her classroom to be a difficult environment, but perhaps not as difficult as a fellow teacher, and over- or under-compensate ratings relative to another teacher.

The most significant differences are in the student-centered subscale. Here, the items with moderate to large DIF and the items with slight to moderate DIF are connected. The 3<sup>rd</sup> grade teachers in this sample perceive both disruptive students and threats to teacher and student safety as a larger challenge to student learning and more impactful to negative school climate ratings. Teachers in 8<sup>th</sup> grade perceive student absenteeism and student disengagement as more challenging than 3<sup>rd</sup> grade teachers in the sample. This finding may suggest the addition of the indicator of student attendance policy for Indiana schools may disproportionately and negatively affect 8<sup>th</sup> grade teachers' perceptions of school climate and accountability pressures.

From the overwhelming presence of significant DIF in the student subscale items, it would not be surprising if similar levels of DIF were present in other grades. In relation to accountability pressures, teachers in this sample were chosen from grades 3<sup>rd</sup> and 8<sup>th</sup> so that achievement scores could be linked in subsequent data analysis. Teachers in later grades who are responsible for English/language arts, mathematics, and either science or social studies/history, could face additional pressures that manifest in different perceptions. In addition, teachers in non-tested grades were not included in this sample. Although these teachers are not accountable for student achievement in the same way that tested grades are, accountability pressures do exist, and their perceptions of school climate could provide more insight into a measure that could be used on teachers in both tested and non-tested grades.



The difference in item functioning could simply be reflective of the typical teaching and learning environments between elementary and middle school classrooms. As an 8<sup>th</sup> grade teacher may see several groups of students rotating throughout the day in their classroom, making one decision on multiple classes may be difficult. However, seeing multiple students may have less impact when discussing school/teacher support items, and as such this subscale shows less substantial DIF in its items. From knowledge of previous school climate measures, it seems more reasonable to conclude the DIF analysis has not detected a bias in the items, but an overall group difference that reflects the true nature of the construct, that school climate is complex and multifaceted, can vary from school to school, and that a teachers' grade level does matter when considering their perceptions of challenges to student learning and school climate.

The DIF analysis indicated items where 3<sup>rd</sup> and 8<sup>th</sup> grade teachers responded differently as a group after adjusting for overall scores. Many items exhibited DIF between the two groups, but it is equally as important to highlight items that indicated little to no DIF, as these may have implications of their own. Items with an absolute DIF contrast of less than .43 may still produce a significant p-value to indicate a difference in item difficulty between the two groups. Table 10 provides the DIF analysis for the Student-Centered subscale. Item b, students who come from a variety of backgrounds, has a significant p-value ( $p=.003$ ), but a DIF contrast considered to be negligible (.38) (Zwick, Thayer, & Lewis, 1999). Here, the model has detected a difference in perception between the 3<sup>rd</sup> and 8<sup>th</sup> grade teachers, mainly the perception that teaching diverse students is difficult for teachers, specifically more difficult for 3<sup>rd</sup> grade teachers in this sample. However, the magnitude of the contrast was just slightly below the cutoff.

Difference in item functioning may become more or less significant among different samples of teachers.

Four items in the School/Teacher Support subscale in Table 13 also demonstrated a significant p-value, but indicated slight to no DIF. One item in particular, testpressure to perform well on standardized tests, was predicted to have a significant difference between the two groups, due to the policies surrounding testing and accountability. However, the DIF contrast was considered as negligible. Although the DIF in these items was not considered as impacting item difficulty, mean differences between the variables' raw scores in all six cases are significant via a two-sample t-test. Here, differences between the groups are not due to an item difficulty functioning differently, rather a true difference in the ratings between the two groups.

The purpose of this study was to validate an instrument based on the four most commonly accepted dimensions of school climate, with items adapted and previously used in other contexts from the same framework. The two previous uses of items had varying operationalizations and purposes. A 9-item measure of institutional challenges produced a high alpha reliability. The remaining portion of items used in the instrument for this study were a selection of items from a 60-item school climate instrument constructed for the Philadelphia School District with six subscales and high alpha reliabilities. In the construction of this instrument, a majority of items were taken from the school-level challenges subscale.

Results from the Rasch RSM demonstrated the instrument does do an appropriate job in differentiating between the amount of latent trait teachers have from both grades. It is probable the significant presence of DIF is a reflection of the multidimensionality of the instrument, school climate as a construct, and the differences in students in both

grades. Items should not be dropped between elementary or middle school teachers from these differences. However, because of the significant presence of DIF with this sample of 3<sup>rd</sup> and 8<sup>th</sup> grade teachers, any direct comparisons with teachers outside of their own grade level could be misleading without additional qualitative and quantitative work. One larger implication for survey research with teachers from a variety of grades is that items may function differently for different groups of teachers that may not be accurately represented by a t-test or other classical test theory technique.

Although school-level factors are an important part in determining a comprehensive school climate instrument, items dealing with relationships between school members and questions on respect and safety are equally as important and should not be neglected. If this instrument is to claim to be a valid measure of school climate, it should address at least all four of the most commonly accepted dimensions of school climate. This instrument would benefit from including more of the items from the 60-item Philly scale, particularly those dealing with safety and the relationships between students, teachers, administration, and perhaps community members. Expanding the selection of items would require running a new Rasch RSM to determine the validity of both new and old items.

## **Conclusion**

With the additional requirement of a new indicator of school quality and student success under ESSA accountability provisions, states have the ability to include an academic indicator that isn't determined by a high-stakes achievement test. However clear in theory, non-cognitive measures like socio-emotional learning and school climate have been much more difficult to implement in reality, particularly on such a large scale. For the state of Indiana, initial ESSA drafts included a measure of culture and climate in

schools, citing these elements as vital to measure school accountability and success. Final plans did not include such a measure. Many other states expressed initial support for such a measure but did not ultimately include one. Other states, including Kentucky, have passed subsequent laws requiring a measure of school climate or learning culture.

A contribution of this study is that it provides a validated instrument containing a student-centered and a teacher/school support subscale within one instrument. However, these two scales are not fully representative of school climate, as they overemphasize institutional and learning environments and underemphasize the social and safety aspects of school climate. With the items that are included, the significant presence of DIF in both subscales indicates the instrument behaves differently for 3<sup>rd</sup> and 8<sup>th</sup> grade teachers. This finding, indicative of fundamental differences in how teachers from 3<sup>rd</sup> and 8<sup>th</sup> grade report their perceptions of school climate factors, is also indicative of the multidimensional nature of school climate. Additionally, it suggests that summarizing the subscales or the instrument into one or more factor scores, or instrument averages loses both the individual nature of respondents and neglects the shifts in importance teachers place on various factors over others between the two groups.

The potential for DIF to occur in other groups of teachers is unknown, but highly likely. Comparing teachers by a grade level cluster [lower elementary (K-2), upper elementary (3-5), middle (6-8), high (9-12)] could provide additional clarity beyond 3<sup>rd</sup> and 8<sup>th</sup> grade. For instance, 4<sup>th</sup> and 5<sup>th</sup> grade teachers may be more similar to one another because they face the additional pressure of being tested in a subject outside of English/language arts and mathematics and less similar to 3<sup>rd</sup> grade. Or 4<sup>th</sup> and 5<sup>th</sup> grade teachers may be similar to 3<sup>rd</sup> grade because they share a similar instructional environment and licensure requirements. Each of the smaller groups of teachers should

be analyzed and addressed prior to any future use of the measure for broad comparisons. Further work utilizing a mixed methods approach to examine differences between other grade levels in terms of sources of accountability pressure and how these vary by grade level could provide additional insight into how teachers perceive school climate and accountability pressures.

A wider implication of this study is that it reinforces the need to validate all school climate instruments, particularly at the item level, especially when combining previous measures, regardless of the similarities or previous validation efforts. Many of the most commonly used and adapted school climate instruments were developed many years ago, with validation at the item level unpublished or nonexistent (Zullig et al., 2010).. In the case of this school climate instrument, Rasch RSM analyses demonstrated the scale was not unidimensional, each subdimension had acceptable person and item fit, and that items possessed significant DIF between 3<sup>rd</sup> and 8<sup>th</sup> grade teachers. The items that displayed DIF were often connected to one another and related to research findings. For instance, 8<sup>th</sup> grade teachers expressed low student morale and disinterested students were more of a challenge to student learning, aligning with research demonstrating the middle school years are often difficult and a source of stress for students (Evans, Borriello, & Field, 2018). But, the two subscales do not fully encompass the four dimensions of school climate, as they overemphasize institutional and learning environments and underemphasize the social and safety aspects of school climate. As it stands, this instrument is an incomplete measure of school climate and should not be used in its current form to make conclusions on school climate. Conclusions and implications drawn from this instrument can only claim to reflect the institutional and learning environments through the school/teacher support and student-centered subscales.

School climate research has suffered from persistent problems in a variety of capacities, from operationalization to validation. Although most researchers and educators recognize the value of non-cognitive measures, the abundance of conceptual frameworks, instruments, and findings have flooded the field with a variety of empirical research that is often misaligned (Rudasill et al., 2017). In the case of Indiana and other state ESSA plans, it is clear a current demand exists for appropriate, cohesive, and validated measures of school climate, and that this demand is going unmet or being met with instruments that have not been carefully validated. Given that many school districts survey teachers with similar materials, one critical finding from this validation is that one measure may not behave similarly for different groups of teachers, which could impact any straight comparisons and decisions coming from those comparisons. This instrument has the potential to be a valuable resource for policymakers, researchers, and educators interested in measuring school climate and this study was the first step in providing a school climate instrument validated at the item level using Rasch analysis to provide a stable ruler to measure persons and items on. However, much more work must be done before it could be utilized as an effective and comprehensive measure of school climate.

From the results of this study, the 22-item measure analyzed would not qualify as an effective full measure of school climate. However, it is very possible that an instrument exists comprehensively measuring school climate that has been analyzed at the item level with a state representative sample and using differential item functioning techniques to ensure validity. It is also very possible with the addition of items related to relationships between members of the school and more safety items, that a comprehensive and valid school climate instrument could result. The review of instruments in a previous chapter was by no means comprehensive, but a quick

representation of the state of school climate instruments-messy, complex, and not clearly validated.

In the summer/fall of 2018, it was reported that education officials in Indiana were creating a school improvement model based on the 5Essentials framework, independent of this study (Cavazos, 2018). This model hopes to replace the previous turnaround model for struggling schools by using the five areas of the model to address issues in struggling schools. Many of the components from the previous school improvement model align with the 5Essentials framework but separating supportive environments (the school climate portion) as its own area of focus is an important step for identifying ways struggling schools can become more successful (Lindsay, 2018). Although a school climate measure did not make the final draft of Indiana ESSA protocols as a fifth indicator of school quality or student success, a measure of supportive environments for students will likely be used to identify areas in which schools could improve. The two subscales of school climate validated in this study could be an important component of measuring school effectiveness and provide a more in depth picture of teachers' perceptions of school climate as Indiana replaces their school improvement model with the very framework the study is aligned with. A revised version of this instrument could be used as a valuable resource for states, educators, and researchers who are looking for a validated instrument to measure school climate that can be used to support school and student success. By using data from teachers in both public and private settings, and spanning a variety of grades, a revised instrument could serve as a broad instrument allowing for comparisons across by grade and school type, and a pathway by which states could be confident to utilize

## APPENDIX

### *CHALL Series Survey Items*

To what extent do you consider each of the following factors *a challenge* to student learning in your classroom? (Select one option in each row.)

	<b>Not a challenge</b>	<b>A slight challenge</b>	<b>A moderate challenge</b>	<b>A great challenge</b>
a. Low morale among fellow teachers/administrators*+	1	2	3	4
b. Students who come from a wide range of backgrounds*+	1	2	3	4
c. Threat(s) to your safety or safety of students*+	1	2	3	4
d. The noise level in the school building+	1	2	3	4
e. Amount of professional support staff (e.g., counselors, specialists)*	1	2	3	4
f. Students with special needs (e.g., hearing, vision, speech impairment, physical disabilities, mental or emotional/psychological impairment)+	1	2	3	4
g. Amount of time to prepare for class+	1	2	3	4
h. High student/teacher ratio+	1	2	3	4
i. Students with different academic abilities*+	1	2	3	4
j. Uninterested students	1	2	3	4
k. Disruptive students*	1	2	3	4



l. Parents uninterested in their children's learning progress*	1	2	3	4
m. Access to technology*	1	2	3	4
n. Pressure to perform well on standardized tests*	1	2	3	4
o. Lack of school resources to provide the extra help for students who need it*	1	2	3	4
p. Lack of teacher planning time built into the school day*	1	2	3	4
q. Inadequate physical facilities	1	2	3	4
r. Low morale among students	1	2	3	4
s. Teacher turnover in this school*+	1	2	3	4
t. Student absenteeism*	1	2	3	4
u. Student tardiness*	1	2	3	4
v. Lack of guidance or support for teaching special education students (i.e., students with IEPs)*	1	2	3	4
w. Lack of guidance or support for teaching English Language Learners*	1	2	3	4

\*Those included in Philadelphia District Teacher Survey

+Those included in previous work with Berends and colleagues

## REFERENCES

- Ahn, T., & Vigdor, J. (2014). The impact of No Child Left Behind's accountability sanctions on school performance: Regression discontinuity evidence from North Carolina. NBER Working Paper No. 20511. Cambridge, MA: National Bureau of Economic Research.
- Amrein, A. L., & Berliner, D. C. (2003). The effects of high stakes testing on student motivation and learning. *Educational Leadership*, 60(8), 32-38.
- Anderson, C. (1982). The search for school climate: A review of the research. *Review of Educational Research*, 52, 368–420. doi:10.3102/00346543052003368
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.
- Andrich, D. (2005). The Rasch model explained. In S. Alagumalai, D. D. Durtis, & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars* (pp. 308-328). Berlin, Germany: Springer-Kluwer.
- Back, L. T., Polk, E., Keys, C. B., McMahon, S. D. (2016). Classroom management, school staff relations, school climate, and academic achievement: Testing a model with urban high schools. *Learning Environment Research*, 19, 397-210. doi: 10.1007/s10984-016-9213-x
- Badia, X., Prieto, L., Linacre, J. M. (2002). Differential item and test functioning (DIF & DTF). *Rasch Measurement Transactions*, 16(3), 889.
- Baker, F. B. (2001). *The basics of item response theory (2<sup>nd</sup> edition)*. ERIC Document Reproduction Service No. ED 458 219. College Park, MD: Eric Clearing House on Assessment and Evaluation.
- Bear, G., Yang, C., Harris, A., Mantz, L., Hearn, S., & Boyer, D. (2016). Technical manual for the Delaware School Survey: Scales of school climate; bullying victimization; student engagement; positive, punitive, and social emotional learning techniques; and social and emotional competencies. Retrieved from <http://wh1.oet.udel.edu/pbs/wp-content/uploads/2011/12/Delaware-School-Survey-Technical-Manual-Fall-2016.pdf>
- Bear, G., Yang, C., Pell, M., & Gaskins, C. (2014). Validation of a brief measure of teachers' perceptions of school climate: Relations to student achievement and suspensions. *Learning Environments Research*, 17, 1–16.
- Berends, M., Goldring, E., Stein, M., & Cravens, X. (2010). Instructional conditions in charter schools and students' mathematics achievement gains. *American Journal of Education*, 116(3), 303-335. DOI: 10.1086/651411

- Berends, M., & Waddington, R. J. (2015). *School choice in Indiana: An examination of impacts and the conditions under which choice is effective*. Unpublished grant proposal.
- Berkowitz, R., Moore, H., Astor, R. A., & Benbenishty, R. (2017). A Research Synthesis of the Associations Between Socioeconomic Background, Inequality, School Climate, and Academic Achievement, *Review of Educational Research*, 87(2), 425-469. DOI:10.3102/0034654316669821
- Bomotti, S., Ginsberg, R., & Cobb, B. (1999). Teachers in charter schools: A comparative study. *Education Policy Analysis Archives*, 7(1), 1-27.
- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3<sup>rd</sup> Ed.). New York, NY: Routledge.
- Brand, S., Felner, R. D., Seitsinger, A., Burns, A., & Bolton, N. (2008). A large scale study of the assessment of the social environment of middle and secondary schools: The validity and utility of teachers' ratings of school climate, cultural pluralism, and safety problems for understanding school effects and school improvement. *Journal of School Psychology*, 46, 507-535. doi:10.1016/j.jsp.2007.12.001
- Bronfenbrenner, U. (1979). *The ecology of human development*. Cambridge, MA: Harvard University Press.
- Brookover, W. B., & Erickson, E. L. (1969). *Society, schools and learning*. Boston: Allyn & Bacon.
- Bryk, A. S. (2010). Organizing schools for improvement. *Phi Delta Kappan*, 91(7), 23-30. <https://doi.org/10.1177/003172171009100705>
- Bryk, A. S., Lee, V. E., & Holland, P. B. (1993). *Catholic schools and the common good*. Cambridge, MA: Harvard University Press.
- Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). *Organizing schools for improvement: Lessons from Chicago*. Chicago, IL: University of Chicago Press.
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.
- Buening, J. G. (2014) *Non-Academic Differences Between Public and Private High Schools: The Importance of School Climate*. (Electronic Thesis or Dissertation). Retrieved from <https://etd.ohiolink.edu/>
- Carrell, S. E., & Hoekstra, M. L. (2009). Domino effect. *Education Next*, 9(3), 59-63. Accessed from [https://www.educationnext.org/files/domino\\_effect.pdf](https://www.educationnext.org/files/domino_effect.pdf)

- Cavazos, S. (2018, September). Indiana officials didn't have to go far to find a new model for improving schools. *Chalkbeat*. Retrieved from <https://www.chalkbeat.org/posts/in/2018/09/04/indiana-officials-didnt-have-to-go-far-to-find-a-new-model-for-improving-schools/>
- Cheema, J. R., & Kitsantas, A. (2014). Influence of disciplinary classroom climate on high school student self-efficacy and mathematics achievement: A look at gender and racial ethnic differences. *International Journal of Science and Mathematics Education, 12*, 1261–1279. doi:10.1007/s10763-013-9454-4
- Chingos, M. M., & West, M. R. (2011). Promotion and reassignment in public school districts: How do schools respond to differences in teacher effectiveness? *Economics of Education Review, 30*, 419–433.
- Choy, S. P. (1997). Public and private schools: How do they differ? Findings from “The condition of education, 1997.” Washington, DC: Office of Educational Research and Improvement, National Center for Education Statistics.
- Clifford, M., Menon, R., Condon, C., & Hornung (2012). Measuring school climate for gauging principal performance: A review of the validity and reliability of publicly accessible measures. Retrieved from [http://www.air.org/files/school\\_climate2](http://www.air.org/files/school_climate2).
- Cohen, J., McCabe, E. M., Michelli, N. M., & Pickeral, T. (2009). School climate: Research, policy, practice, and teacher education. *Teachers College Record, 111*, 180–213.
- Cohen-Vogel, L. (2011). Staffing to the test: Are today's school personnel practices evidence based? *Educational Evaluation and Policy Analysis, 33*(4), 483–505.
- Collie, R. J., Shapka, J. D., & Perry, N. E. (2012). School climate and social-emotional learning: Predicting teacher stress job satisfaction, and teaching efficacy. *Journal of Educational Psychology, 104*(4), 1189-1204. doi: 10.1037/a0029356
- De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.
- Dee, T. S., & Jacob, B. A. (2010) The impact of No Child Left Behind on students, teachers, and schools [with comments and discussion]. *Brookings Papers on Economic Activity*. Accessed from <http://www.jstor.org/stable/pdf/41012846.pdf>
- Donaldson, M. L. (2009). *So long, Lake Wobegon?: Using teacher evaluation to raise teacher quality*. Washington, DC: Center for American Progress.
- Douglas, S. M. (2010). Organization climate and teacher commitment (doctoral dissertation). Accessed from [http://libcontent1.lib.ua.edu/content/u0015/0000001/0000519/u0015\\_0000001\\_000519.pdf](http://libcontent1.lib.ua.edu/content/u0015/0000001/0000519/u0015_0000001_000519.pdf)

- Eccles, J. S., Midgley, C., Wigfield, A., Buchanan, C. M., Reuman, D., Flanagan, C., et al. (1993). Development during adolescence. The impact of stage environment fit on young adolescents' experiences in schools and in families. *American Psychologist*, *48*, 90–101. doi: 10.1037/0003-066X.48.2.90
- Eden, M. (2017). School discipline reform and disorder: Evidence from New York City public schools, 2012-2016. New York, NY: Manhattan Institute.
- Edgerton, A., Polikoff, M., & Desimone, L. (2017). How is policy affecting classroom instruction?. (Executive Summary). *Brookings Evidence Speaks Reports Vol. 2*, 14. Accessed from <https://www.brookings.edu/research/how-is-policy-affecting-classroom-instruction/>
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Ennis, C. D. (1998). Shared expectations: Creating a joint vision for urban schools. In J. Brophy (Ed.), *Advances in research on teaching* (Vol. 7, pp.151–182). Greenwich, CT: JAI Press Inc.
- Esposito, C. (1999). Learning in urban blight: School climate and its effect on the school performance of urban, minority, low-income children. *School Psychology Review*, *28*(3), 365–377.
- Evans, D., Borriello, G. A., & Field, A. P. (2018). A review of the academic and psychological impact of the transition to secondary education. *Frontiers in Psychology*, *9*, 1-18. doi: 10.3389/fpsyg.2018.01482
- Fan, W., Williams, C. M., & Corkin, D. M. (2011). A multilevel analysis of student perceptions of school climate: The effect of social and academic risk factors. *Psychology in the Schools*, *48*, 632–647. doi:10.1002/pits.20579
- Feng, L., Figlio, D., & Sass, T. (2010). School accountability and teacher mobility. Working paper 47. Washington, DC: CALDER, The Urban Institute.
- Figlio, D. N. (2006). Testing, crime and punishment. *Journal of Public Economics*, *90*, 837-851.
- 5Essentials Support Center. (2018). How scores are calculated. Accessed from <http://help.5-essentials.org/customer/en/portal/articles/94413-how-scores-are-calculated>
- Fulton, I. K., Yoon, I., & Lee, C. (2005). Induction into learning communities. Washington, DC: National Commission on Teaching and America's Future. Accessed from <https://eric.ed.gov/?id=ED494581>

- Gangi, T. A. (2009). School climate and faculty relationships: Choosing an effective assessment measure (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3388261)
- Glasman, N. S. & Biniaminov, I. (1981). Input-output analyses of schools. *Review of Educational Research, 51*, 509-540.
- Gohl, K. (2018, October). The importance of what ESSA plans do not include. *Getting Smart*. Retrieved from <https://www.gettingsmart.com/2018/10/sel-the-importance-of-what-essa-plans-do-not-include/>
- Goldstein, S. E., Young, A., & Boyd, C. (2008). Relational aggression at school: Associations with school safety and social climate. *Journal of Youth and Adolescence, 37*, 641–654.
- Grayson, J. L., & Alvarez, H. K. (2008). School climate factors relating to teacher burnout: A mediator model. *Teaching & Teacher Education, 24*, 1349–1363.
- Gregory, A., Cornell, D., Fan, X., Sheras, P., Shih, T., & Huang, F. (2010). High school practices associated with lower student bullying and victimization. *Journal of Educational Psychology, 102*, 483–496.
- Griffith, J. (1999). School climate as “social order” and “social action”: A multi-level analysis of public elementary school student perceptions. *Social Psychology of Education, 2*, 339–369.
- Grissom, J. A., Kalogrides, D., & Loeb, S. (2017). Strategic staffing? How performance pressures affect the distribution of teachers within schools and resulting student achievement. *American Educational Research Journal, 54*(6), 1079-1116. DOI: 10.3102/0002831217716301
- Guo, P., Choe, J., & Higgins-D’Alessandro, A. (2011). Report of construct validity and internal consistency findings for the Comprehensive School Climate Inventory. Retrieved from [https://www.schoolclimate.org/climate/documents/Fordham\\_Univ\\_CSCI\\_development\\_review\\_2011.pdf](https://www.schoolclimate.org/climate/documents/Fordham_Univ_CSCI_development_review_2011.pdf)
- Guo, S. & Yang, Y. (2012). Project-based learning: An effective approach to link teacher professional development and students learning. *Journal of Educational Technology Development and Exchange, 5*(2), 41-56.
- Halpin, A. W., & Croft, D. B. (1963). The organizational climate of schools. Chicago, IL: Midwest Administration Center. Retrieved from <http://donpugh.dyndns.org/Education/questionnaires/THE%20ORGANIZATION%20CLIMATE%20OF%20SCHOOLS.pdf>
- Hambleton, R. K., van der Linden, W. J., & Wells, C. S. (2010). IRT models for the analysis of polytomously scored data: Brief and selected history of model

- building advances. In M. Nering, & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 21-42). London: Routledge Academic.
- Higgins-D'Alessandro, A., & Guo, P. (2009). School culture: can it be adequately operationalized as a school-level variable? Paper presented at the meeting for the Association for Moral Education, Utrecht, Netherlands.
- Hoffman, J. V., Assaf, L. C., & Paris, S. G. (2001). High-stakes testing in reading: Today in Texas, tomorrow?. *The Reading Teacher*, 54(5), 482-492.
- Holbein, J. B., & Ladd, H. F. (2017). Accountability pressure: Regression discontinuity estimates of how No Child Left Behind influenced student behavior, *Economics of Education Review*, 58, 55-67.
- Hoy, W. and Feldman, J. (1987), "Organizational health: the concept and its measure", *Journal of Research and Development in Education*, 20(4), 30-7.
- Hoy, W. K., & Hannum, J. W. (1997). Middle school climate: An empirical assessment of organizational health and student achievement. *Educational Administration Quarterly*, 33, 290-311.
- Hoy, W. K., Smith, P., & Sweetland, S. R. (2002). The development of the Organizational Climate Index for high schools: Its measure and relationship to faculty trust. *High School Journal* 86(2), 38-49
- Hoy, W., Hannum, J., & Tschannen-Moran, M. (1998), Organizational climate and student achievement: a parsimonious view and longitudinal view. *Journal of School Leadership*, 8, 336-59.
- Humphrey, L. L. (1984). Children's self-control in relation to perceived social environment. *Journal of Personality and Social Psychology*, 46, 178-188.
- Indiana Department of Education. (2017). Every Student Succeeds Act (ESSA) [First Draft]. Retrieved from <https://www.doe.in.gov/sites/default/files/essa/essa-plan-draft-one.pdf>
- Indiana Department of Education (2018a). State Template for the Consolidated State Plan Under the Every Student Succeeds Act. Retrieved from <http://www.doe.in.gov/sites/default/files/essa/essa-consolidated-plan.pdf>
- Indiana Department of Education (2018b). REPA Educator Standards. Retrieved from <https://www.doe.in.gov/licensing/rep-a-educator-standards>
- Indiana Department of Education (2018c). Indiana Content Standards for Educators-Elementary Generalist. Retrieved from <https://www.doe.in.gov/sites/default/files/licensing/elementary-generalist.pdf>

- Indiana Department of Education (2018d). Indiana Content Standards for Educators-Mathematics. Retrieved from <https://www.doe.in.gov/sites/default/files/licensing/math.pdf>
- Indiana Department of Education (2018e). ISTEP+ Grades 3-8, 10. Retrieved from <https://www.doe.in.gov/assessment/istep-grades-3-8-10>
- Indiana Department of Education (2018f). History of Indiana's Accountability System. Retrieved from <https://www.doe.in.gov/accountability/history-indiana%E2%80%99s-accountability-system>
- Janken, B. P. (2011). An examination of the relationship between school climate and student growth in select Michigan charter schools (doctoral dissertation). Paper 355. Accessed from <http://commons.emich.edu/cgi/viewcontent.cgi?article=1355&context=theses>
- Johnson S. M. (2006). The workplace matters: Teacher quality, retention, and effectiveness (NEA Research Best Practices Working Paper). Washington, DC: National Education Association. Retrieved from: <http://files.eric.ed.gov/fulltext/ED495822.pdf>
- Johnson, B., Stevens, J. J., & Zvoch, K. (2007). Teachers' perceptions of school climate: A validity study of scores from the Revised School Level Environment Questionnaire. *Educational and Psychological Measurement*, 67(5), 833-844 doi:10.1177/0013164406299102
- Klassen, R. M., & Chiu, M. M. (2010). Effects on teachers' self-efficacy and job satisfaction: Teacher gender, years of experience, and job stress. *Journal of Educational Psychology*, 102, 741-756. doi: 10.1037/a0019237
- Klugman, J., Gordon, M. F., Sebring, P. B., & Sporte, S. E. (2015). A first look at the 5Essentials in Illinois schools. (Executive Summary). Retrieved from <https://consortium.uchicago.edu/sites/default/files/publications/Statewide%20E%20Executive%20Summary.pdf>
- Kohl, D., Recchia, S., & Steffgen, G. (2013) Measuring school climate: an overview of measurement scales, *Educational Research*, 55(4), 411-426, DOI: 10.1080/00131881.2013.844944
- Kosciw, J. G., & Elizabeth, M. D. (2006). The 2005 National School Climate Survey: The experiences of lesbian, gay, bisexual and transgender youth in our nation's schools. New York, NY: GLSEN.
- Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2008). A multilevel study of predictors of student perceptions of school climate: The effect of classroom-level factors. *Journal of Educational Psychology*, 100, 96-104.



- Koyama, J. P. (2012). Making failure matter: Enacting No Child Left Behind's standards, accountabilities, and classifications. *Educational Policy*, 26(6), 870-891. DOI:10.1177/0895904811417592
- Krommendyk, M. (2007). The association between school choice and school climate: Comparing school climate in private religious, charter, and public schools. Retrieved from <https://scholarworks.wmich.edu/dissertations/885>
- Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, 7(4), 328. Accessed from <https://www.rasch.org/rmt/rmt74m.htm>
- Linacre, J.M. (1999). Investigating Rating Scale Category Utility. *Journal of Outcome Measurement*, 3, 103-122. Retrieved from [http://www.jampress.org/JOM\\_V3N2.pdf](http://www.jampress.org/JOM_V3N2.pdf)
- Linacre, J.M. (2000). Comparing and choosing between "Partial Credit Models" (PCM) and "Rating Scale Models" (RSM). *Rasch Measurement Transactions*, 19(3), 768.
- Linacre, J.M. (2005). Rasch dichotomous model vs. One-parameter Logistic Model. *Rasch Measurement Transactions*, 19(3), 1032.
- Linacre, J. M. (2018a). Winsteps® Rasch measurement computer program User's Guide. Retrieved from <https://www.winsteps.com/manuals.htm>
- Linacre, J.M. (2018b). Winsteps® (Version 4.3.1) [Computer Software]. Beaverton, Oregon: Winsteps.com.
- Lindsay, J. (2018, September). State shifting focus of quality reviews toward school leaders, environment. *WFYI Indianapolis*. Retrieved from <https://www.wfyi.org/news/articles/state-shifting-focus-of-quality-reviews-toward-school-leaders-environment>
- Loeb, S., Darling-Hammond, L., & Luczak, J. (2005). How Teaching Conditions Predict Teacher Turnover in California Schools, *Peabody Journal of Education*, 80(3), 44-70. DOI: 10.1207/s15327930pje8003\_4
- Lubienski, S. T., Lubienski, C., & Crane, C. C. (2008). Achievement differences and school type: The role of school climate, teacher certification, and instruction. *American Journal of Education*, 115, 97-138.
- Markowitz, A, J. (2018). Changes in school engagement as a function of No Child Left Behind: A comparative interrupted time series analysis. *American Educational Research Journal*, 55(4), 721–760. DOI: 10.3102/0002831218755668
- Martín, E., Martínez-Arias, R., Marchesi, A., & Pérez, E. M. (2008). Variables that predict academic achievement in the Spanish compulsory secondary educational system: A longitudinal, multi-level analysis. *Spanish Journal of Psychology*, 11, 400–413. doi:10.1017/S113874160000442X

- Meyer-Adams, N., & Conner, B. T. (2008). School violence: Bullying behaviors and the psychosocial school environment in middle schools. *Children & Schools, 30*, 11–221. doi:10.1093/cs/30.4.211
- McDuffie, A., Drake, C., Choppin, J., Davis, J. D., Magaña, M. V., & Carson, C. (2017). Middle school mathematics teachers' perceptions of the Common Core State Standards for Mathematics and related assessment and teacher evaluation systems. *Educational Policy, 31*(2), 139-179. DOI: 10.1177/0895904815586850
- McGill, M. V. (2015). *Race to the bottom*. New York City: Teachers College Press.
- Miller, S. I. & Fredericks, J. (1990). The false ontology of school climate effects. *Educational Theory, 40*(3), 333-342. doi:10.1111/j.1741-5446.1990.00333.x
- Mitchell, M. M., Bradshaw, C. P., & Leaf, P. J. (2010). Student and teacher perceptions of school climate: A multilevel exploration of patterns of discrepancy. *The Journal of School Health, 80*, 271–279. <http://dx.doi.org/10.1111/j.1746-1561.2010.00501.x>
- Moos, R. H. (1974). Systems for the assessment and classification of human environments: An overview. In R. H. Moos & P. M. Insel (Eds.), *Issues in social ecology*. Palo Alto, Calif.: National Press Books
- Morello, R. (2012, May). How districts are preparing for state-mandated teacher evaluations. *StateImpact Indiana*. Retrieved from <https://indianapublicmedia.org/stateimpact/2012/05/16/how-school-districts-are-preparing-for-mandatory-teacher-evaluations/>
- Murray, H. A. (1938). *Explorations in personality*. New York: Oxford University Press
- National School Climate Council (2007). *The school climate challenge: narrowing the gap between school climate research and school climate policy, practice guidelines and teacher education policy*. Retrieved from <http://www.ecs.org/school-climate>
- Neal, D. & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics, 92*(2), 263-283. doi: 10.1162/rest.2010.12318
- NORC at the University of Chicago. (2017). Methodology report: 2017 School Effectiveness in Indiana (SEI) study. Unpublished report.
- O'Malley, M., Voight, A., Renshaw, T. L., & Eklund, K. (2015). School climate, family structure, and academic achievement: A study of moderation effects. *School Psychology Quarterly, 30*, 142–157. DOI:10.1037/spq0000076

- Pace, C. R., & Stern, G. G. (1958). An approach to the measurement of psychological characteristics of college environments. *Journal of Educational Psychology*, 49, 269-277.
- Pallant, J. F., & Tennant, A. (2010). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology*, 46, 1-18. DOI:10.1348/014466506X96931
- Palmer, D. & Rangel, V. S. (2011). High stakes accountability and policy implementation: Teacher decision making in bilingual classrooms in Texas. *Educational Policy*, 25(4), 6614-647. DOI: 10.1177/0895904810374848
- Perry. A. C. (1908). *The management of a city school*. New York: The Macmillan Company.
- Plank, S. B., Bradshaw, C. P., & Young, H. (2009). An application of “Broken-windows” and related theories to the study of disorder, fear, and collective efficacy in schools. *American Journal of Education*, 115(2), 227-247. doi: 10.1086/595669
- Polikoff, M. S. & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis*, 36(4), 399-416.
- Pulliam, J. D. & Van Patten, J. J. (2007). *History of education in America* (9th ed.). New Jersey: Pearson.
- Ramsey, C. M., Spira, A. P., Parisi, J. M., & Rebok, G. W. (2016) School climate: perceptual differences between students, parents, and school staff, *School Effectiveness and School Improvement*, 27:4, 629-641, DOI: 10.1080/09243453.2016.1199436
- Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. high-school data. *Journal of Educational Statistics*, 16, 295–330.
- Reese, W. J. (2011) *America's public schools* (2nd ed). Baltimore, MD: The Johns Hopkins University Press.
- Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. (2007). *Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure*. Cambridge, MA: National Bureau of Economic Research.
- Roussos, L. & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, 20(4), 355-371. DOI: 0146-6216/96/040355-17\$2.10
- Royal, M., D'Angelis, K., & Rossi, R. (1996). *Teachers' sense of community: How do public and private schools compare?* Washington, DC: Office of Educational Research and Improvement, National Center for Educational Statistics.

- Royal, K. D., & Elahi, F. (2011). Psychometric properties of the Death Anxiety Scale (DAS) among terminally ill cancer patients. *Journal of Psychosocial Oncology*, 29(4), 359-71.
- Rudasill, K. M., Snyder, K. E., Levinson, H., & Adelson, J. L. (2017). Systems view of school climate: A theoretical framework for research. *Educational Psychology Review*, 1-26. DOI: 10.1007/s10648-017-9401-y
- Ryan, A. M. & Patrick, H. (2001). The classroom social environment and changes in adolescents' motivation and engagement during middle school. *American Educational Research Journal*, 38(2), 437-460. DOI: 10.3102/00028312038002437
- School District of Philadelphia Office of Research and Evaluation. (2016). District-wide surveys technical report. Retrieved from <http://schoolsveys.philasd.org/files/reports/2016/TechnicalReport.pdf>
- Sebring, P. B., Allensworth, E., Bryk, A. S., Easton, J. Q., & Luppescu, S. (2006). The essential supports for school improvement. Chicago: Consortium on Chicago School Research.
- Setari, A. P. (2016) Construction and validation of a holistic education school evaluation tool using Montessori Erdkinder principles. (Doctoral Dissertation). *Theses and Dissertations-Education Science*. 12. Retrieved from [http://uknowledge.uky.edu/edsc\\_etds/12](http://uknowledge.uky.edu/edsc_etds/12)
- Shindler, J. (2016, May). Explanation and comparison of the ASSC SCAI. Retrieved from [http://web.calstatela.edu/centers/schoolclimate/assessment/Comparison\\_and\\_Efficacy\\_of\\_the\\_ASSC\\_SCAI.pdf](http://web.calstatela.edu/centers/schoolclimate/assessment/Comparison_and_Efficacy_of_the_ASSC_SCAI.pdf)
- Shindler, J., Taylor, C., Cadenas, H., & Jones, A. (2003, April). Sharing the data along with the responsibility: Examining an analytic scale-based model for assessing school climate. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Sikkink, D. (2012). Religious School Differences in School Climate and Academic Mission: A Descriptive Overview of School Organization and Student Outcomes. *Journal of School Choice*, 6, 20-39. DOI: 10.1080/15582159.2012.651394
- Sinnema, C., Ludlow, L., & Robinson, V. (2016). Educational leadership effectiveness: A Rasch analysis, *Journal of Educational Administration*, 54(3), 305-339. doi: 10.1108/JEA-12-2014-0140
- Sinnema, C. & Robinson, V. M. J. (2007). The leadership of teaching and learning: Implications for teacher evaluation. *Leadership and Policy in Schools*, 6, 319-343. doi: 10.1080/15700760701431603

- Smith, R. M., Schumacker, R. E., & Bush, J. M. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurements, 2*, 66–78.
- Steinberg, M. P. & Donaldson, M. L. (2016). The new educational accountability: understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy, 11*(3), 340-359. doi: 10.1162/EDFP\_a\_00186
- Stringer, K. 3 States Cite School Climate Surveys in Their ESSA Plans. Why Don't Others Use Culture for Accountability? <https://www.the74million.org/article/3-states-cite-school-climate-surveys-in-their-essa-plans-why-dont-others-use-culture-for-accountability/>
- Sussman, J., Beaujean, A. A., Worrell, F. C., & Watson, S. (2012). An analysis of Cross Racial Identity Scale scores using classical test theory and Rasch item response models. *Measurement and Evaluation in Counseling and Development, 46*(2), 136-153. doi: 10.1177/0748175612468594
- Tschannen-Moran, M., Parish, J. and DiPaola, M.F. (2006), School climate and state standards: how interpersonal relationships influence student achievement, *Journal of School Leadership, 16*(4), 386-415.
- Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A. (2013). A review of school climate research. *Review of Educational Research, 83*, 357–385.
- Theriot, M. T., & Dupper, D. R. (2009). Student discipline problems and the transition from elementary to middle school. *Education and Urban Society, 42*(2), 205-222. DOI: 10.1177/0013124509349583
- Toland, M. D. (2014). Practical guide to conducting an item response theory analysis. *The Journal of Early Adolescence, 34*, 120-151.
- Uline, C. & Tschannen-Moran, M. (2008) The walls speak: the interplay of quality facilities, school climate, and student achievement. *Journal of Educational Administration, 46*, 55-73.
- Urban, W. J. & Wagoner, J. L. (2009). *American education: A history* (4th ed). New York: Routledge.
- U.S. Department of Education, National Center for Education Statistics. (2017). Back to school statistics. Retrieved from <https://nces.ed.gov/fastfacts/display.asp?id=372>
- U.S. Department of Education, Office of Safe and Healthy Students. National Center of Safe Supportive Learning Environments. (2018). Summary table of Office of Safe and Health Students approved school climate surveys. Retrieved from [https://safesupportivelearning.ed.gov/sites/default/files/Summary%20Table%20of%20OSHS%20Approved%20School%20Climate%20Surveys\\_10.22.18\\_0.pdf](https://safesupportivelearning.ed.gov/sites/default/files/Summary%20Table%20of%20OSHS%20Approved%20School%20Climate%20Surveys_10.22.18_0.pdf)

- von der Embse, N., Pendergast, L. L., Segool, N., Saeki, E., & Ryan, S. (2016). The influence of test-based accountability policies on school climate and teacher stress across four states. *Teaching and Teacher Education, 59*, 492-502. DOI: 10.1016/j.tate.2016.07.013
- Waddington, R. J. & Berends, M. (2018). Impact of the Indiana Choice Scholarship Program: Achievement effects for students in upper elementary and middle school. *Journal of Policy Analysis and Management, 37*(4), 783-808.
- Wang, M. T., & Degol, J. L. (2016). School climate: A review of the construct, measurement, and impact on student outcomes. *Educational Psychology Review, 28*(2), 315–352.
- Wang, M. T., & Holcombe, R. (2010). Adolescents' perceptions of school environment, engagement, and academic achievement in middle school. *American Educational Research Journal, 47*, 633–662. doi:10.3102/0002831209361209
- Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: The New Teacher Project.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press
- Zullig, K. J., Collins, R., Ghani, N., Patton, J. M., Scott Huebner, E., & Ajamie, J. (2014). Psychometric support of the school climate measure in a large, diverse sample of adolescents: A replication and extension. *The Journal of School Health, 84*, 82–90. <http://dx.doi.org/10.1111/josh.12124>
- Zullig, K. J., Ghani, N., Collins, R., & Matthews-Ewald, M. R. (2015). Preliminary Development of the Student Perceptions of School Safety Officers Scale. *Journal of School Violence, 16*, 104–118. <http://dx.doi.org/10.1080/15388220.2015.1116994>
- Zullig, K. J., Koopman, T. M., Patton, J. M., & Ubbes, V. A. (2010). School climate: Historical review, instrument development, and school assessment. *Journal of Psychoeducational Assessment, 28*, 139–152.
- Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*, 1-28.

## VITA

### **Audrey Conway Roberts**

#### EDUCATION

- 2013-2015      **University of Kentucky**  
M.S. in Educational Psychology
- 2008-2012      **Centre College**  
B.S. in Psychology

#### PROFESSIONAL EXPERIENCE

- 2016-Present      Research Assistant, University of Kentucky
- Summer 2018      Quantitative Data Analyst, Bowling Green State University
- Summer 2018      External Evaluator, University of Kentucky
- 2015-2016      Research Assistant, University of Kentucky
- 2013-2014      Research Assistant, University of Kentucky

#### TEACHING EXPERIENCE

- 2015-2016      Teaching Assistant, University of Kentucky
- 2014-2015      Primary Instructor, University of Kentucky

#### PUBLICATIONS

- Roberts, T., Maiorca, C., & **Roberts, A. C.** (submitted for initial review). Positively influencing preservice elementary teachers' mathematics conceptions and content knowledge. *Ohio Journal of Teacher Education*.
- Conway, A. E.**, & Sampson, S. (2018). *Appropriate for All? Rasch Analysis of a Scale for Teachers' Perceived Challenges*. Proceedings of the 2018 Annual Meeting of the American Educational Research Association.