VALIDITY OF STUDENT PERCEPTION SURVEYS AS INSTRUMENTS OF TEACHER QUALITY INDICATORS IN NON-URBAN SCHOOLS

A Dissertation

Presented to

The Faculty of the Education Department

Carson-Newman University

In Partial Fulfillment

Of the

Requirements for the Degree

Doctor of Education

By Trevor Sean Gore

March 2018
Dissertation Approval

Student Name: Trevor Sean Gore

Dissertation Title: Validity of Student Perception Surveys as Instruments of Teacher Quality Indicators in Non-urban Schools

This dissertation has been approved and accepted by the faculty of the Education Department, Carson-Newman University, in partial fulfillment of the requirements for the degree, Doctor of Education.

Dissertation Committee:

Signatures:

Brenda P. Dean, Ed. D., Dissertation Chair

P. Mark Taylor, Ph. D., Methodologist Member

Michael Sobiech, Ph. D., Content Member

Approved by the Dissertation Committee Date: March 20, 2018
I hereby grant permission to the Education Department, Carson-Newman University, to reproduce this research in part or in full for professional purposes, with the understanding that in no case will it be for financial profit to any person or institution.

Signature: [Signature]

Date: 21 February 2018
Abstract

The purpose of this quantitative study was to explore the impact non-urban environments have on the validity of student perception surveys. Previous literature has focused heavily upon participant districts that are primarily urban and contain disproportionately large populations of students from minority backgrounds. Consequently, a number of questions regarding student survey validity could not be properly answered, and there were issues in generalizing previous results to schools and districts outside of urban areas and to populations more closely aligned to national demographic averages. In order to address this issue, student survey scores and individual value-added scores for teachers in a Tennessee school district that contained both urban and non-urban areas were obtained. A Pearson’s coefficient was used to test whether that district’s data were able to support the hypothesis that a correlation exists between student survey scores and student growth, as measured by value-added scores, even when non-urban areas were considered. The data revealed sufficient correlation, regardless of whether the district as a whole was considered or only the urban or non-urban areas. The only breakdowns in correlation occurred when grade-bands were considered. However, those groupings of the data were only presented for the benefit of future researchers and did not serve as a basis for the testing of this study’s hypothesis.
Acknowledgements

First, I wish to thank the members of my dissertation committee for their support and feedback. Dr. Brenda Dean, the chair of my committee, has put in countless hours of feedback and prayer that my data would arrive on time. Dr. P. Mark Taylor has often encouraged and helped me navigate when I was unsure how the methodology for my study would come together. Dr. Michael Sobiech provided immeasurable expertise in helping get my ideas on paper and clearly stated.

Second, I also wish to thank those English teacher friends of mine who agreed to proofread chapters for me: Mrs. Stuart, Mrs. Hosea, and Mrs. Douglas (who took on multiple chapters).

Third, I am grateful to the department of research and performance management and the human resource department of the district pseudonymously referred to as County Schools. Without their efforts pulling the necessary data, this study could never have occurred.

Finally, I wish to thank my wife, Hillary Gore, and my parents, Dr. John and Mary Gore. Without their love and support I would have never finished.
Dedication

This work is dedicated to my parents, Dr. John and Mary Gore, who instilled in me a love of learning, a curiosity to always wonder why, and who set the standard for me to pursue. They made sacrifices that only they will ever fully know so I could reach toward my full potential. If I can arrive at being half the person they are, I will be able to count myself twice as blessed as any other person.

I also wish to dedicate this work to my wife, Hillary Gore. Without her encouragement, support, and love I could never have found the energy to run this race. I dedicate this work to her so it might serve as an encouragement when she begins her doctorate in July.
# Table of Contents

Abstract .................................................................................................................. iv  
Acknowledgements .................................................................................................. v  
Dedication ................................................................................................................ vi  
List of Tables and Figures ....................................................................................... x  

1. **Purpose and Organization** ............................................................................... 1  
   Introduction and background of the study ......................................................... 1  
   Statement of the problem ................................................................................. 2  
   Racial considerations ....................................................................................... 3  
   Aptitude considerations ................................................................................... 5  
   Purpose of the study/significance of study ...................................................... 7  
   Significance of the study .................................................................................. 8  
   Theoretical foundation ..................................................................................... 9  
   Conceptual Framework .................................................................................... 10  
   Research questions and null hypotheses ....................................................... 11  
   Limitations and delimitations ......................................................................... 12  
   Assumptions and definition of terms ............................................................. 14  
   Organization of the document ....................................................................... 17  

2. **Review of Related Material** ......................................................................... 18  
   Literature related to the topic ......................................................................... 18  
   Literature related to college student surveys. ............................................... 19  
   Porter, 2010. ..................................................................................................... 20  
   Literature related to value-added measures. ................................................. 21
Literature related to student surveys ................................................................. 24
Summaries of previous studies ......................................................................... 24
Remedios, & Lieberman, 2008 ......................................................................... 24
Bill and Melinda Gates Foundation, 2010 ...................................................... 25
Desimone, Smith, & Frisvold, 2010 .................................................................. 27
Stehle, Spinath, & Kadmon, 2012 ................................................................. 29
White, 2013 ........................................................................................................ 30
Wagner, Göllner, Helmke, Trutwein, & Lüdtke, 2013 .................................... 31
Ferguson & Danielson, 2014 .......................................................................... 33
Chaplin et al., 2014 .......................................................................................... 35
Fauth, Decristan, Rieser, Klieme, & Büttner, 2014 ........................................... 37
Luppescu, 2016 .................................................................................................. 38
Muñoz, & Dossett, 2016 .................................................................................. 40
Wallace, Kelcey, Ruzek, 2016 ......................................................................... 41
Analysis of theories related to the studies ...................................................... 41
Racial/Genetic Considerations ........................................................................ 42
Age/Aptitude Considerations .......................................................................... 45
Unresolved issues, significant problems and unanswered questions requiring study ..... 47
3. Research Methodology .............................................................................. 49
Introduction ....................................................................................................... 49
Population and sample .................................................................................... 49
Description of instruments(s) ......................................................................... 50
Research procedures and time period of the study ........................................ 50
<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical Test Used</td>
<td>51</td>
</tr>
<tr>
<td>Data Analysis</td>
<td>52</td>
</tr>
<tr>
<td>4. Results of the Data Analysis</td>
<td>58</td>
</tr>
<tr>
<td>Introduction</td>
<td>58</td>
</tr>
<tr>
<td>Participants’ demographics</td>
<td>59</td>
</tr>
<tr>
<td>Organization of the study’s findings</td>
<td>62</td>
</tr>
<tr>
<td>Results regarding the hypothesis</td>
<td>62</td>
</tr>
<tr>
<td>District-wide results</td>
<td>62</td>
</tr>
<tr>
<td>Urban area, all schools</td>
<td>63</td>
</tr>
<tr>
<td>Non-urban area, all schools</td>
<td>63</td>
</tr>
<tr>
<td>Grade-band results</td>
<td>64</td>
</tr>
<tr>
<td>Relevant, yet unexpected, results and findings discovered</td>
<td>67</td>
</tr>
<tr>
<td>Summary</td>
<td>67</td>
</tr>
<tr>
<td>5. Conclusions, Implications and Recommendations</td>
<td>69</td>
</tr>
<tr>
<td>Introduction</td>
<td>69</td>
</tr>
<tr>
<td>Discussion of conclusions drawn regarding the hypotheses</td>
<td>69</td>
</tr>
<tr>
<td>Recommendations for further studies and policy issues</td>
<td>71</td>
</tr>
<tr>
<td>Limitations of this study</td>
<td>71</td>
</tr>
<tr>
<td>Recommendations for future studies</td>
<td>74</td>
</tr>
<tr>
<td>Recommendations for policy issues</td>
<td>75</td>
</tr>
<tr>
<td>Conclusion</td>
<td>76</td>
</tr>
<tr>
<td>References</td>
<td>78</td>
</tr>
<tr>
<td>Appendix</td>
<td>86</td>
</tr>
</tbody>
</table>
List of Tables

Table 2.1 Racial Compositions of Studied Communities ...........................................44
Table 4.1 Racial Composition of County School’s Area...........................................61
Table 4.2 Correlation Between Student Survey Results and Student Growth By Location
and Grade Group.....................................................................................................66
Chapter 1: Purpose and Organization

Introduction and background of the study

Recently, the idea of using student survey data as a method for evaluating teachers has gained in popularity (Heitin, 2012). Tripod Education Partners (2017) reported that over 100,000 teachers have used their student survey instrument. The idea of student surveys is simple: students are given a number of questions or statements regarding their teacher, aspects of the class, etc., and they are asked to rate their agreement on a Likert type scale. A number of states and districts have experimented with using these surveys to evaluate teacher quality (see Luppescu, 2016) and implementation of best practices.

The use of student surveys as an evaluative tool offers several benefits. Firstly, student surveys provide an efficient method of obtaining data on teachers’ practices. Observations by outside observers tend to be periodic and infrequent, but many feel the real indicator of teacher impact is what happens on a daily basis, and sending an administrator into the classroom every few days is prohibitively expensive. Student surveys offer a chance to resolve this dilemma. They often require no more than one class period to administer, yet they involve participants who have knowledge of every class period from that school year. If students are reliably capable of providing the same information about the teacher’s use of best practices that an administrator could document, then the students are the more efficient source for information.

Secondly, student surveys offer inclusivity. Those who benefit (or suffer) the most from instruction are the students. It is logical that they should be consulted and asked to voice their opinions on the process. Furthermore, since Maslow’s Hierarchy of
Needs (Maslow, 2012) stipulates that people tend to focus on satisfying their basest needs first and foremost, student surveys are useful for identifying (and then addressing) lack of satisfaction in basic needs like safety and love/belonging. Involving students in the discussion helps leaders solve these basic needs and push students to focus on higher needs such as self-actualization and the needs which encourage learning and growth.

Thirdly, student surveys offer diversity in measurement. Inasmuch as no one measurement tool is perfect, the use of several measurement tools is one way to prevent deficiencies from becoming a significant factor. Student surveys, value-added scores, and in-person observations all capture related, but not identical aspects of teacher effectiveness (Chaplin, Gill, Thompkins, & Miller, 2014), indicating that, when used together, they are effective in buttressing each other and giving a stronger overall measure of teacher quality.

**Statement of the problem**

Despite the recent increase in use of student surveys, several questions remain regarding the appropriateness of basing teacher evaluation and any subsequent teacher salary or employment decisions upon student surveys. In particular, before student survey evaluations can support such high-stakes decisions, one must ensure that survey-based evaluations are reliable across all contexts. If student surveys are only valid within certain age groups or specific cultural environments, one could risk subjecting teachers to inequalities within the evaluation process. Unfortunately, to date, the majority of studies dealing with student surveys as measures of teacher effectiveness have not explored racial nor aptitude considerations sufficiently so as to produce a high degree of confidence in the ability to generalize study results to the full population.
Racial considerations.

Many existing studies have overrepresented students from minority racial groups in percentages much larger than the national average, limiting the ability to explore adequately the correlation between student survey data and teacher effectiveness. For instance, the Bill and Melinda Gates Foundation (2010) conducted one study comparing student survey results to student growth data. The 2010 census estimated the U.S. population as 72.4% white, 12.6% black, and 16.3% Hispanic / Latino (QuickFacts, 2015). The Gates Foundation study used data from Charlotte (50% white, 35% black, 13.1% Hispanic/Latino), Dallas (50.7% white, 25% black, 42.4% Hispanic/Latino), Denver (68.9% white, 10.2% black, 31.8% Hispanic/Latino), Hillsborough County Florida (the Tampa area; 71.3% white, 16.7% black, 24.9% Hispanic/Latino), and New York City (44% white, 25.5% black, 28.6% Hispanic/Latino). Three of the five participant districts contained Black students at more than twice the national average, and four of the five districts contained Hispanics/Latinos at close to or more than double the national average. Another influential study in the same vein (Chaplin et al., 2014) focused on Pittsburgh schools (66% white, 26.1% black, 2.3% Hispanic/Latino), and once again overrepresented Black students by more than double. For the same time period, Pittsburgh contained a Hispanics/Latino(a) population that was 14.1% of the national average.

Naturally, as far as those districts are concerned, the studies in question fulfilled their objectives, for the data were representative of the student populations for those cities. However, if one wishes to generalize and draw conclusions about the usefulness of student surveys in all contexts, a more representative sample must be found. Not only
is the use of a representative sample good practice, but survey nonresponsiveness is influenced by genetic factors (Littvay, Popa, & Fazekas, 2013; Thompson, Zhang, & Arvey, 2011) and racial composition could be an issue.

Granted, it is possible that student surveys are not impacted by racial issues. While the previously mentioned research indicates that genetics impacts surveys, the discovery is recent and there is no evidence that propensity toward nonresponsiveness is unequally distributed between racial groups. It is possible that samples that do not align with the national racial averages may not introduce any more probability of nonresponsiveness than any other random sample might. Student surveys are generally conducted at school with an adult unobtrusively monitoring the children, so forgetfulness, lost mail, and other common problems that might increase nonresponsiveness are less of a threat than they would be to other surveys. Students taking surveys in groups and seeing other students responding could create peer pressure in favor of responding to the questions. On the whole, one could reasonably speculate that nonresponsiveness will decrease. There is even evidence that the inclusion of reluctant respondents generally increases the accuracy of the sample (Olson, 2006), and the use of subtle persuasive techniques does not negate these benefits.

However, simply getting kids to respond may not always improve validity. The most common hypothesis explaining why genetics skews response rates is that some people have a genetic predisposition to fear retaliation. Olson’s study did not involve surveys administered by authority figures with the power to retaliate, so one cannot assume the results can be generalized. But even if Olson’s results are true, the suggested impact of fear leads to the idea that social, psychological, or political forces might impact
survey responses in the same way that a genetic predisposition to fear retaliation might. Might students from minority groups feel that their opinions will not be valued by teachers and administrators simply because of their minority status? If so, minority students who wish to give their teachers a poor rating could feel that (in the best case) their opinion will change nothing, and (in the worst case) their opinion will spark retaliation from those in power.

The best way to counter this concern is to explore whether a sample closer to the national racial composition can generate a correlation between student survey data and student academic growth. One study (White, 2013) compared student survey data, administrator observations, and student growth scores using a district that, based on 2015 data, appears closer to the national average (50% white, 13% black, 25% Hispanic/Latino; Newton-Conover City Schools, 2015). It found no correlation between student survey results and student growth data. Therefore, there is a pressing need for another study that uses a sample of participants with a racial makeup closer to the national average. Only then can one know whether White’s results were anomalous or whether racial factors are significant hurdles to survey validity when measuring teacher quality.

**Aptitude considerations.**

The other variable, which previous studies have not sufficiently explored, is student aptitude, especially aptitude deficiencies as a result of youth and immaturity. Most studies examining correlations between student survey results and student academic growth have only included survey participants from the fourth grade or higher (Chaplin et al., 2014; Bill and Melinda Gates Foundation, 2010; Luppescu, 2016).
Luppescu (2016) sought to test the idea that fourth and fifth-grade survey participants can produce the same level of validity as older students. That study found that, while the results were still within statistical limits and appeared valid, there was a slight decrease in measures of validity when fourth and fifth-grade results were compared to sixth-grade results. The decrease was not enough to reject the hypothesis that fourth and fifth-graders can produce accurate survey results, but it may have signaled the need for further verification before one can confidently utilize surveys from young children as teacher evaluation tools.

There is evidence that districts are using surveys from children as young as kindergarten (Parent memo, 2014; Tripod, 2017), which means that if existing studies into survey validity are to be generalized, there must be more data regarding whether correlations hold true when dealing with children from kindergarten to third grade. By including children from this age range in this study’s sample, it will be easier to verify if low literacy levels are skewing results.

Tripod (2017) uses age-appropriate wording for a kindergarten through second grade specific instrument, yet it is also known that people taking surveys are willing to answer even questions that are beyond their reading comprehension level (Al-Tayyib, Rogers, Gribble, Villarroel, & Turner, 2002). Given that Tripod (2017), like most survey instruments, tests for inconsistent results and discards them, it could be possible for those in the lower grades who struggle academically to have slightly more of their responses discarded for lack of the literacy skills needed to provide consistent answers. This would leave those who are succeeding and benefiting from the teacher’s methods with a stronger voice, and could reward teachers for helping a few students outpace their peers.
in the class. Therefore, in order to generalize the results from existing studies of student survey data and its relationship to student growth, there is a need for research that includes participants as young as kindergarten.

**Purpose of the study/significance of study**

The purpose of this study is to verify that student survey results are equitable, valid, and useful measures of teacher quality. To date, most existing studies into the validity of student survey data have displayed a heavy focus on urban districts with a high population of students belonging to racial minorities. Therefore, any impact which population density or race might exert over the validity of survey results could skew researchers’ understanding of student surveys.

In order to account for the potential impact of population density and race, this study will gather student survey results and value-added scores from a school district located in Tennessee. This particular district is run at the county level, so for the sake of anonymity it will be referred to as County Schools within this study. The candidate will compare student survey results to value-added scores using a Pearson’s Product Moment correlation coefficient. The researcher selected County Schools because of its large size and diversity. At the time County Schools generated the data used in this study, the district contained 7 cities (1 urban area and 6 non-urban areas) as well as unincorporated and fairly rural areas. In the urban area there was an environment similar to that used in previous studies in terms of racial composition as well as the cultural influences cities tend to create. The non-urban areas contained racial and cultural compositions different from the urban area. This diversity was so strong that the year following the collection of these data all cities other than the urban area pulled out of County Schools and formed
their own districts (which will be referred to as District A through District F). This highly diverse and volatile cultural environment is the ideal place to explore whether student surveys can maintain reliability regardless of the environment around the school.

By examining the correlation between student survey results and value-added scores in the mixture of urban, suburban, and rural schools contained within County Schools, it will be possible to draw more informed conclusions regarding whether student surveys reliably indicate teachers’ abilities to generate academic growth in students, or whether previously identified correlations are the result of favorable environmental factors.

**Significance of the study.**

The majority of existing literature focused on urban, high minority districts (Chaplin et al., 2014; Bill and Melinda Gates Foundation, 2010; Luppescu, 2016) and supported a strong correlation between student survey data and student growth. This has encouraged some districts, like the one that participated in this study, to use student surveys as part of the evaluation process for teachers. Furthermore, this literature has inspired a large number of recommendations for other actions. However, some data from non-urban districts showed no correlation (White, 2013) between student survey data and student growth.

This study could be a significant step forward in resolving this conflict. By examining districts similar to those White studied, but with a larger sample size and more districts, this study could help validate White’s conclusions (indicating that student surveys are not accurate enough to generalize to all contexts and suggesting the need for more research), or this study could provide evidence that White’s conclusions may be
outliers resulting from random chance. In either case, this study has the potential to bring increased focus to literature on the use of student surveys for the purpose of teacher evaluation.

**Theoretical foundation**

Because there are two potential sources of bias which existing literature on student surveys has not fully addressed (racial and aptitude considerations) there are two ways the theoretical framework must be used. Maslow’s hierarchy of needs is sufficient to handle both considerations.

The need to explore the racial considerations requires using Maslow’s Hierarchy of Needs (Maslow, 2012) in the most straightforward sense. If students from minority groups believe that honest criticism of their teacher on student surveys could bring retaliation, then the child will likely pursue the course of action with the greatest probability of satisfying his or her safety needs. The child is not likely to consider esteem needs until safety is assured. If the threat to the child’s safety (from retaliation) is greater than the potential for the child to satisfy esteem needs by expressing his or her opinion, the child will avoid any answer which could spark retaliation, even if it means reduced feelings of esteem.

The need to address aptitude considerations requires viewing Maslow’s Hierarchy of Needs with an understanding that meeting basic needs is prerequisite to growth. Among young students who are learning to read, their concern may not be with the fear of retaliation, but rather with performing acceptably and fear of the shame of failure. Students who are illiterate or barely literate may focus simply on answering questions in a way that hides their struggle (arriving safely at their destination). This may mean
selecting an answer because it is understood, not because it represents the child’s feelings. On the other hand, the more advanced students in the class will focus on using their answers for self-expression and conveying the truth (the self-actualization level).

**Conceptual Framework**

As nations, states, and school districts have increased their focus on teacher effectiveness and accountability, the need to find the right balance between high precision of measurement and minimal invasiveness/intrusion into the process has grown into a pressing concern. Similarly, districts have had to better define expectations for teachers in order to determine measurement methods, and many have come to the conclusion that no one instrument can adequately measure instructional quality (Bill and Melinda Gates Foundation, 2012). It is necessary to use a combination of instruments. Consequently, many districts must find an evaluation system that not only quantifies quality of instruction, but also helps distinguish between teachers and allows leaders to identify which teachers are strongest in which areas.

A common system that is able to distinguish between teachers in this way (Chaplin et al., 2014) is an evaluation model combining observations, value-added (or test score) data, and student surveys. This evaluation system uses the idea that there is strength in examining several measures of daily practice (observations, which show a teacher’s level of skill with content and delivery, and student surveys, which show a teacher’s interpersonal skills and emotional intelligence) and also factors in the end result (student growth in the form of value-added scores).

This combination allows for districts to consider and better understand situations where a teacher strong in interpersonal skills might overcome his or her weaknesses in
instructional practice, or visa-versa, in order to achieve strong gains (or where weakness in one skill drags down strength in another area). Nonetheless, it is important to verify that there is not too much deviation between each measure and student growth data. Some difference in scores is anticipated since different people have different skills, but it is important to verify that all measures used have some level of correlation to student growth (Chin, & Goldhaber, 2015). If no correlation exists, there is the danger that one is not looking at a complex diversity of skill, but rather that one is looking at both useful and useless measurements with no way to tell the difference.

This problem is best summed up by the well-known balance between correlation and causation. While correlation does not imply causation, causation will create correlation. For this reason, it is imperative that one expect some level of correlation. With several factors causing increases in student growth and all factors having the ability to compensate for weaknesses in other factors, correlation may not be as strong as a researcher might hope, but it must be present if one is to conclude that student survey scores capture a variable which impacts student growth.

**Research questions and null hypotheses**

The central question guiding this study is, “Do student survey results correlate strongly with student academic growth within non-urban environments?”

- **H₁**: There is a direct correlation between a teacher’s score on student surveys and the average amount of student growth that a teacher produces within a non-urban setting.

- **H₀**: There is no correlation between a teacher’s student survey scores and the student growth produced within a non-urban setting.
Limitations and delimitations

The primary purpose of this study is to partially overcome the limitations faced by previous studies resulting from bias in participant selection. Naturally, any selection of participants involving anything less than 100% participation carries some level of bias, so this study will not be completely able to overcome this issue, but it plans to fill in gaps left by previous works. The largest of these gaps is that most work to date has focused on urban districts with a high percentage of students belonging to racial groups that would be considered minorities nationally. To balance existing participant bias, this study will maintain its primary focus on schools within the parts of District Schools that fell in non-urban areas. The researcher will explore the validity of the survey results obtained from schools within an urban area as well, in order to verify that the urban schools conform to the expectations previous studies have created. Then the non-urban schools (most of which have racial and cultural demographics closer to the national average) will be the primary focus of analysis in order to test whether those schools also produce valid survey results.

The researcher will utilize only pre-existing data from districts which have used student surveys. The primary purpose of this delimitation is to ensure findings align with practical considerations. This study’s central concern is to ensure teacher evaluation measures are fair, accurate, and useful. If data districts generate in the course of teacher evaluation do not yield valid results, then is it important to discover that situation. Using an instrument other than that which districts use in the name of standardization would undercut this study’s ability to speak to the concerns facing districts in regard to evaluation methods.
This decision to only use found data and to look for an opportunity to sample suburban school comes with the challenge that the researcher must use older data. In particular, the researcher will only use data from the 2013-14 school year within County Schools. This was the only year in which all schools within that county participated in student surveys. It was also a year in which the state calculated value-added scores for all teachers. After the 2013-14 school year, Tennessee began a transition to a new state-testing model that had several challenges. As a result, there were several years in which some grade levels did not have any state testing, or in which state-test results were not used as part of the students’ grades (potentially impacting a students’ incentives to do their best). By using the 2013-14 school year, this study can obtain a sample from before this disruption.

An additional limitation is most student surveys are administered in large urban areas. Therefore, discovering smaller districts using student surveys is challenging. This is one of the reasons why the researcher selected County Schools, since it contains urban, suburban, and rural areas. The advantage is that the influence of the urban area within the county caused the district to choose to administer student surveys. The disadvantage is that the cultural influence of the urban area may have been sufficient to cause the non-urban areas to behave more like an urban area than they might have had they been more remote. There is no way to guarantee that isolated rural areas would behave the same way as the areas in this sample.

The use of value-added scores to quantify student growth further limits the sample. The use of value-added calculations within a district or state may come from cultural or social trends that might bias the sample without the researcher’s knowledge.
States and districts generating value-added scores may have political or cultural features distinct from those not using value-added and this could limit generalizability. Finally, while value-added calculations come from reputable sources, and most experts generally accept them as accurate, they are not 100% accepted, and different states use different formulas with slightly different variables. If Tennessee’s value-added score formula turns out to lack validity, or even turns out to not be the most reliable of all the various formulas in use, this study would lose all or part of its validity.

**Assumptions and definition of terms**

This study attempts to test whether student perceptions of teacher behaviors and effectiveness correlates to actual student growth. Given that none of those constructs can be directly tested in a perfect way, the candidate will approximate student perception measures via professionally developed student survey instruments and will measure student growth using value-added calculations.

The term “student perception instrument” can have a variety of meanings and must be clarified. Bernhardt (1998) pointed out that perception data can come from virtually any stakeholder (students, parents, teachers, etc.) and is used to determine the environment of a school. For the purpose of this paper, the term “student survey” will be used to mean the same as “student perception survey,” for brevity. Additionally, while perception surveys can have as many uses as a leader could imagine, this study will only explore the use of student perception surveys as a measure of teacher effectiveness and surrogate for observational data.

The reason for this narrowed definition of student survey uses is simple. A number of school districts/states have implemented student perception surveys as
measures of teacher quality (see Luppescu, 2016), making the question of validity as a teacher evaluation tool an important concern. This is also why the author has chosen to use found data rather than administering a uniform survey tool to all participant schools. By using found data, the validity of real life attempts to measure teacher effectiveness with student surveys will come under scrutiny, and all results will be more relevant to the question at hand.

This study also faces limitations stemming from the need to make data manageable. While the candidate desires to produce a work that is perfectly generalizable and explores the usefulness of student survey results in any situation that a reader may face, the diversity of cultures, geographic regions, etc. within the United States alone makes it doubtful that any one study can achieve this aim. Therefore, some narrowing must occur. To date, the vast majority of research into the use and validity of student surveys has focused on large cities with a larger than normal number of Black, Hispanic/Latino(a), and (in some cases) Asian students than one might find within the United States as a whole. Therefore, the author has decided that the best approach is to concentrate this work upon the validity of student surveys within non-urban locations with a more typical racial composition so as to fill a gap in the literature.

For the purposes of this work, the definition of what constitutes a representative racial composition will be based on the 2010 census. This is the most recent nationwide census available, and is also the census that is closest in time to the largest number of studies into the validity of student surveys. The limit of what constitutes “excessive” overrepresentation will usually be defined as twice the national average, but some discussions and flexibility in this definition are sometimes required. To illustrate,
increasing the number of Asian participants from 4.8% to 9.6% is likely to have less impact than increasing the number of Hispanic/Latino(a) participants from 16.3% to 32.6%, although both represent a change from the national average to twice that average. Some flexibility may be needed at times, but, in general, “excessive overrepresentation” will refer to twice the national average.

The term “non-urban” is harder to define than racial representation, since some rural areas are close enough to large cities so as to be impacted by the culture of that city. In the interest of simplicity, “non-urban” will refer to locations not classified as an “urbanized area” by the U. S. Census Bureau (Qualifying Urban Areas, 2012). Areas classified as “urban clusters” will, for the purposes of this work, be considered as “non-urban.” The formula determining which areas are “urbanized areas,” which are “urban clusters,” and which areas are on neither list is complex. Speaking generally, major cities are usually “urbanized areas” and their suburbs are often “urban clusters.” Naturally, a third category exists as a consequence of the other two: areas that are too small to qualify for any existing categorization. There is no official “rural” category within the census list, so references in this work to “rural” areas refer to areas found on neither list.

This study assumes that student surveys and value-added scores have internal consistency. Similarly, it assumes that a correlation should exist between the two variables if they are both valid measures of teacher performance and have external validity. This leads to the assumption that if no correlation exists, either student surveys and/or value-added calculations are not valid measures of teacher performance and lack sufficient external validity.
One assumption that led to the implementation of student surveys is that different teachers have different strengths and a variety of measurement tools are appropriate for evaluating this diversity of skill. In particular, one can assume there is merit in seeking to measure teachers’ interpersonal skills, because student-teacher relations and student feelings of acceptance play a significant role in determining whether and how well a student learns (Allan & Pileičikienė, 2010). Consequently, this study’s hypothesis assumes that if students provide honest and open feedback regarding teachers, there should be a correlation between those attitudes and measurable growth by the students.

**Organization of the document**

This study begins with an examination of the need for further research of student survey validity (Chapter 1). Chapter 2 will focus on reviewing the related literature. This will include previous studies related to student surveys, as well as reviewing literature related to the validity of value-added measurements in order to verify that value-added is an appropriate indicator of student growth. There will also be an exploration of literature related to best practices for teacher-evaluation models so as to establish a framework for judging whether sufficient correlation exists. Chapter 3 will outline the methodology and procedures for the study. Chapter 4 will present the findings and analysis results, followed by a conclusion and commentary regarding findings and limitations of the study.
Chapter 2: Review of Related Material

The use of student surveys is a well-established practice in higher education (Stehle, Spinath, & Kadmon, 2012) but was not widely used in K-12 education until the early 2000s (Dillon, 2010; Ripley, 2012; Tripod Education Partners, 2017) when Tripod Educational Partners released their instrument. Since that point, researchers have conducted several studies to examine the validity of student surveys (Bill and Melinda Gates Foundation, 2010; Chaplin et al., 2014; Luppescu, 2016; White, 2013; and others), with the majority of studies focusing on the use of surveys in large urban centers.

Literature related to the topic

The foundations for using student survey data in teacher evaluations came from the requirements of the No Child Left Behind Act (NCLB) (2002) and later from the Race to the Top (Race to the Top Act of 2011) for identifying, hiring, and retaining first highly qualified teachers and later highly effective teachers. The most common outgrowths of this search included the use of increased standardized testing, implementation of value-added scoring, and increased use of teacher observations, but various districts and states have used several other potential sources of data.

Another concern embodied by the Race to the Top grants was with closing the achievement gap between demographic groups. This concern was one of the issues that motivated Ronald Ferguson to conduct research designed to help educators hear the opinions and concerns of students from racial and ethnic minorities and at-risk demographics that became the basis of the Tripod student survey instrument (LaFee, 2014). Concurrently, student surveys were being used for taking the temperature of
school-wide climates (Schulz, Sud, & Crowe, 2014). As these events interacted, the idea of surveys regarding specific teachers gained momentum.

Based on the belief that students are able to recognize when an activity helps them learn, the financial and temporal constraints incurred by increases in teacher observations and other factors, several districts and educational leaders began seeking alternatives. Student surveys offered a solution to many issues. Compared to observations, they offered a lower cost, were less time intensive, and could potentially produce the same information.

**Literature related to college student surveys.**

Within the college setting, student surveys have been one of the most common methods for understanding outcomes (Porter & Umbach, 2006). One strong selling point has been their convenience. In 2013, a faculty group with no formal training in survey design was able to produce an instrument with inter-rater reliability strong enough that it satisfied the general faculty (Hunt, Baldwin, Tsui, & Matthews) and convinced the faculty congress to recommend its use. This Hunt, Baldwin, Tsui, and Matthews felt this approach saved a good amount of time and money as opposed to hiring outside contractors.

However, some researchers have expressed a variety of concerns. Non-responsiveness to surveys is one factor (Porter & Umbach, 2006), and there appear to be many potential sources for non-response bias. Lower-ability students are less likely to respond to a survey than high-ability students.

There are factors impacting researchers’ ability to gather high-quality survey data that correlate strongly with various genetic indicators. Women respond to surveys more
frequently than men, and among racial groups, White students are more likely to respond. Several studies into the behavior of twins have also identified that genetic factors not linked to sex or race can predispose some people to non-responsiveness (Littvay, Popa, & Fazekas, 2013; Thompson, Zhang, & Arvey, 2011).

Religious background also impacts willingness to disclose one’s true attitudes through survey replies (Sherkat, 2007). Sherkat’s study was inspired by the noticeable difference between pre-election polls and the election results from the Bush/Kerry election. The even larger difference in results that occurred during the Trump/Clinton election appeared to further highlight the study’s contributions to the field.

*Porter, 2010.*

In 2010, Porter examined the validity of college surveys further and, based on available literature, argued that nonresponsiveness is not the only issue facing surveys attempting to gather data by asking students to recall an entire school year.

Porter first discussed the problem of comprehension. For any survey to be accurate, one must be certain that all participants understand the wording in the same way. Even everyday words of a low Lexile level, such as “you,” can prompt a variety of responses depending upon whether participants understand the word to be singular or plural. If the word is understood as a plural, there can be further skewing due to differing definitions of whom to include in the plural group. Similarly, frequency words such as “occasionally” and “very often” are subject to inconsistent interpretations.

The second problem Porter discussed was retrieval. This especially impacts surveys attempting to gather information regarding periods of time as long as a school year. Porter referenced a 2002 study (Garry, Sharman, Feldman, Marlatt, & Loftus) that
showed that college students were not accurately able to recall the details of their sexual encounters, including frequency, condom usage, or number of partners. Only reports of anal sex were reported with a high degree of accuracy, prompting Porter to comment “that unless students’ contact with faculty and other academic experiences yield as vivid memories as having anal sex, it is unlikely they will be able to report accurately on them” (Porter, 2010, pp. 18).

Porter also discussed the inherent need for judgment calls within student surveys. Based upon evidence demonstrating the inaccuracy of memory and the changeability of word definitions, Porter argued that many participants asked to report frequency would do one of two things. First, they might recall small snippets of recent events and extrapolate from that small sample. Therefore, if an action has been frequent (or not) within the previous week, it will be reported as having had that frequency for the entire year, regardless of whether that is the true annual frequency. Second, participants report what they feel is normal. Survey design changes as simple as moving a choice from the middle of a scale to an extreme side has been shown to reduce how often people select that answer, even among the same population for the same activity.

**Literature related to value-added measures.**

The Race to the Top initiative also caused many existing evaluation techniques to move from the area of experimentation into a central role. For instance, Tennessee’s value-added calculation (TVAAS) was developed in 1992 (Knox County Board of Education, 2014) but stayed largely on the fringes of educational policy. There was no standard protocol for its use, and it had little influence over high stakes educational decisions until 2011 when Tennessee’s Race to the Top grant requirements caused the
legislature to set TVAAS scores as 35% of the teacher evaluation calculations, making value-added a significant player in educational policy within the state.

One advantage that has helped value-added measures stand out when compared to other options, such as raw test score data, has been the ability to make success equally achievable for all teachers. Because value-added measures compare recorded growth to expected growth for all students in similar situations rather than judging based on the number of correct answers on a standardized test, students in difficult-to-increase situations are no less likely to move their value-added scores than any other student (Balch, & Koedel, 2014).

One common concern is that high achieving students may have gotten so many questions correct in previous years that there is no room for test scores to grow, causing their teacher to appear ineffective. Value-added calculations allow a district to define “significant growth” in whatever way makes sense for the given student and his or her peers in substantially similar circumstances. In the same vein, students missing significant portions of class due to hardships such as cancer or homelessness have to grow less in order for the teacher to receive a high value-added score since it is reasonable to expect that any child in those circumstances would learn less than a child attending every class free from hardships.

This relative scoring gave value-added calculations the advantage of being able to ensure that teachers of high/low achieving students were no more or less likely to have strong value-added scores (National Center on Scaling Up Effective Schools, 2014). Value-added serves as a good method of preventing unequal evaluations due to bias in
student ability or demographics (Chetty, Friedman, & Rockoff, 2014a; Imberman, & Lovenheim, 2015; National Center on Scaling Up Effective Schools, 2014).

Another finding that has encouraged the use of value-added scores within evaluation models was that students of high-value-added teachers could be more likely to experience increased educational and economic success later in life (Chetty, Friedman, & Rockoff, 2011; Chetty, Friedman, & Rockoff, 2014b). There have been some criticisms suggesting these data were not correctly interpreted (Adler, 2013) and do not predict the substantial gains the researchers claimed occurred. However, the publicity value-added experienced from those studies pushed it into the spotlight and helped establish it as one of the best measures of teacher impact currently available.

As with any evaluation measure, there are likely to be concerns, and the use of value-added measures is a contentious topic. Nonetheless, there is consensus that value-added scores correlate to achievement (Koedel, Mihaly, & Rockoff, 2015) and correlate with other evaluation methods as well (Chaplin et al., 2014). While some debate remains, even skeptical scholars rarely advocate for its abandonment. Most are satisfied with encouraging states and districts not to use value-added as the only evaluation measure (Warring, 2015), but rather as one of several, interlocking measures, so as to allow the multitude of measurements to correct for any single instrument’s failings. A better measure of student growth may one day emerge, but until that time, value-added calculations have been proven to have the precision needed to produce reliable and valid results.
**Literature related to student surveys.**

Because value-added is best used in conjunction with other tools, standardized test results are deemed controversial, and sporadic and periodic observations by administration are considered not highly efficient by some researchers (Dillon, 2010; Ripley, 2012), student survey data have become an alternative source for gathering some of the supporting data needed in order to corroborate the other sources.

However, as surveys emerged and attempted to provide a balance for value-added, observations, and standardized testing scores, they also created some of their own controversy. In particular, teachers of younger grades were especially skeptical about the validity of student surveys (Dretzke, Sheldon, & Lim, 2015), and some concern existed about the role of literacy skills in gathering quality data from younger students (Luppescu, 2016). Even in environments with high initial support for student surveys, confidence levels decreased in all grades after the first administration (Dretzke, Sheldon, & Lim, 2015). Consequently, several high-profile studies have examined the validity of student surveys.

**Summaries of previous studies**

**Remedios, & Lieberman, 2008**

In 2008, Remedios and Lieberman conducted a study of 479 psychology students at a Scottish university by administering a survey during registration and again after the beginning of the following semester (resulting in 610 pairs of surveys). There were more pairs of surveys than students due to some students taking more than one psychology course and completing one pair of surveys per course. The purpose was to attempt to discover whether college survey results depend more upon the professor teaching well or
upon the combination of easy class/little homework. The researchers identified factors they considered to be potential sources of bias. These included high grades earned, few hours spent studying, and low perceived difficulty. The research team also identified two indicators of positive views toward the class: the rate at which students said they would recommend the course to a friend and the rate at which students said they enjoyed the course. No correlation existed between potential sources of bias and measures of student course enjoyment. The biggest determiner of positive responses appeared to be the degree of involvement the student felt.

This finding had its limitations since results came from a population comprised exclusively of individuals who had chosen to continue their education as opposed to the obligatory nature of K-12 enrollment, but the trend of student engagement best predicting high survey scores was promising. There seems to be no reason to assume the potential biasing factors are less tempting to college students than to primary and secondary school students.


The Bill and Melinda Gates Foundation (2010) conducted the first major study into student survey validity, the one establishing student surveys as a strong option for teacher evaluations. It was a nation-wide study of fourth through eighth-grade math and English language arts teachers and not only included data on student surveys, but also explored value-added scores, standardized tests (both state tests and some administered by the researchers), observations, pedagogical knowledge, and teacher perceptions of school environment. The study’s purpose was to examine a wide range of variables in order to explore the effectiveness of each for measuring student outcomes and teacher
capacities. Upon compiling the results, both value-added measures and student survey measures correlated well with the student achievement data the researchers derived from alternate measures of student and teacher outcomes.

Despite being a nation-wide and comprehensive study, all participant districts were drawn from large urban areas: Charlotte, Dallas, Denver, Hillsborough County Florida (the Tampa area), and New York City. The study also attempted to include Memphis, but data from Memphis were not available in time to be included in the study. This potential limitation is compounded by the fact that researchers did not use all schools in those districts and that participant school selection was not a random process. Schools focusing on special education programs, alternative schools, community schools, autonomous dropout and pregnancy programs, returning educational schools, and others were automatically excluded from the study (Wallace, Kelcey, & Ruzek, 2016), further opening the door for skewing effects.

While this study was significant in proving that student surveys can correlate with gains, its largest limitation came from its exclusive reliance upon major urban centers for data. Racial and ethnic minorities were routinely over-represented and there was the potential that cultural and social elements impacting survey validity might have skewed results. For this reason, it is difficult to generalize these results to predict survey validity in other parts of the nation.

Another potential limitation of this study was that it used Tripod survey results. The use of Tripod was not inherently a flaw: some instrument had to be used, and there is no clear evidence that Tripod is any better or worse than any other instrument. Nonetheless, using Tripod’s instrument as the only student survey instrument made it
difficult to separate validity of student surveys from validity of Tripod. Based on this study alone, it was not possible to conclude with certainty that Tripod’s instrument was valid because student surveys are valid in general. It was possible that Tripod’s instrument had some feature(s) that made it valid even if student surveys as a general concept are invalid. Wallace, Kelcey, and Ruzek (2016) have even suggested that Tripod’s strong correlations do not prove validity and several other alternate hypotheses could explain even these strong results.


Desimone, Smith, and Frisvold (2010) sought to explore the utility of student surveys by comparing teacher and student responses to the same question. Their sample came from those involved in eighth-grade math education (students and teachers). The research team assumed that teacher surveys would have adequate reliability and validity based upon the findings of pre-existing literature which found that teachers’ self-reported results tend to agree with external measurements when it comes to questions regarding frequency. According to the literature the team reviewed, teacher self-scores on questions related to quality are not as reliable. The hypothesis that a correlation should exist (despite the bias teacher self-scoring could introduce) also rested upon the fact that many of the questions in the survey were more objective and factual in nature, rather than subjective. For example, the survey asked about frequency with which tools were used in the class (calculators, computers, etc.). The team assumed these objective questions would generate more factual answers.

This study did find that questions regarding concrete items (use of calculators and textbooks in particular) provided stronger correlation between teacher and student
responses than questions related to more general concepts (writing and discussion of mathematics in particular). The authors of this study did not feel their data supported attempts to explain why this difference occurred.

Regarding in-class variation, a small amount of variation was due to student characteristics (achievement, race, socio-economic factors, etc.) and none of the within-class variation could be explained by class or teacher level covariates. The authors propose that the primary cause of in-class variation in survey results was from measurement errors. In results similar to Porter and Umbach (2006), the student surveys that were the most likely to match the teachers’ were surveys coming from females in advanced classes who considered success in math to be important (and both they and their class as a whole were doing well in math) and who had parents with higher education. Conversely, factors such as eligibility for free lunch and students in classes marked as “other” (the researchers considered that to be a euphemism for SPED classes) predicted lower agreement between student and teacher survey results. Being Black seemed to predict a gap, but that effect disappeared when researchers controlled for student achievement within the sample of all Black student participants. This study, along with Ferguson & Danielson’s (2014), suggests that student demographics may impact survey results, and controlling for student characteristics may be a worthwhile step in increasing validity.
Stehle, Spinath, & Kadmon, 2012

Stehle, Speinath, and Kadmon’s (2012) study set out to resolve some of the inconsistent results that some existing literature had produced in regard to correlating scores from surveys of college students and measures of student growth. Their hypothesis was that the inconsistencies were a function of differing definitions and measurements of student growth, not a failure of the construct of student perceptions to correlate to the construct of student growth.

In order to test the hypothesis, the researchers gave a survey to a population of third-year medical students and examined the mean overall rating given to the course, the mean rating given to the instructor, and the mean rating of their perceptions of learning. These scores were compared to student performance on a practical and then a multiple-choice test. A correlation existed between better practical exam scores and student survey scores, but no correlation existed between multiple-choice test results and student survey scores. The researchers speculated that these results were a consequence of practical exams requiring a higher level of learning. For this reason, better teachers earned better survey scores and produced students with better practical exam scores. The researchers speculated that multiple-choice, on the other hand, was a lower-level task and relied as much on student study habits as on teacher quality, disrupting the correlation.

The primary limitation this study faced was its own definition of student growth. The impetus for the study was a hypothesis that the inconsistent results of previous studies came from inconsistencies and inaccuracies within definitions of student growth, yet this study used exam scores as the only definition of growth. Granted, the researchers did use two different exam styles, but the assumption that exams best measure learning
was not tested for accuracy. By not examining a range of different definitions and theories for how to measure student learning, this study did not show that different instruments administered to the same group of students can produce the inconsistencies hypothesized to impact correlation. Relative to the topic at hand for this dissertation, the validity of student surveys in primary and secondary schools, this study’s focus on medical students limits generalizability.

**White, 2013.**

The next major study occurred in 2013 (White) and examined correlations between value-added scores (as a measure of student growth), administrator evaluations, and student surveys (using the Tripod instrument). This study was limited to third-through eighth-grade teachers of subjects where there was a state standardized test. This study’s data came from a small North Carolinian school district.

This study failed to demonstrate a correlation between student surveys and student growth scores and failed to demonstrate a correlation between administrator evaluations and student-growth scores. It was able to show a correlation between student survey results and administrator evaluation results, as well as a correlation between the composite of all three measures (surveys, value-added scores, and observational scores) and the total student survey score for that teacher. White’s findings were especially interesting because, despite bringing the validity of both student surveys and administrator observations into question, they supported the conclusion other researchers made that student surveys can be effective when combined with a wider range of measurement techniques (Warring, 2015).
The primary limitation in White’s work was, as the study admitted, sample size. By limiting the study to only one district and only 51 teachers from that district, it was difficult to generalize results to draw conclusions about the usefulness of student surveys at the national or international level. Just as in the Gates Foundation study, White’s work was limited in that it used the Tripod instrument as the student survey, making it hard to conclude that student surveys did not correlate to value-added scores. It may be most accurate to conclude that this study gave evidence that Tripod survey scores in the Newton-Conover City School system did not correlate to value-added scores.

Wagner, Göllner, Helmke, Trutwein, & Lüdtke, 2013.

In 2013, Wagner, Göllner, Helmke, Trutwein, and Lüdtke conducted a study in Germany comparing teachers of German and English courses. Their data came from the results of a 2003-04 school year survey. The team analyzed those data to verify the ability of students to recognize quality teaching and to test whether results could be generalized across subjects and classes. The sample included 6,909 students, representing 280 classes. The authors did not provide information regarding the geographic or racial demographics of the sample.

Within the context of this study, the criteria for confirming the hypothesis that students can recognize quality teaching was the presence of consistency among student perceptions and whether a factor analysis showed that consistent student views were a significant factor in explaining differences in teacher ratings. There was no comparison between student opinions expressed and metrics such as observations scores or evidence of student growth. Consistency in student viewpoint was considered evidence of student recognition of quality teaching. Per this standard, the study verified the hypothesis that
students can recognize quality teaching. Student agreement was strongest with regard to questions related to classroom management and was weakest with questions about lesson structure.

Regarding the hypothesis that these results can be generalized to other classes, results were mixed. Once again, the results regarding classroom management questions seem to indicate that it is safe to generalize that students in other classes would be able to report quality of classroom management with accuracy. However, questions related to student orientation (motivation, understanding, ableness, and involvement) generated lower levels of confidence in the ability to generalize and the researchers concluded that school-to-school comparability was limited. The researchers pointed out that one of this study’s primary limitations was its focus on language arts classes. While it sought to answer the question of generalizability, the exclusive use of language-based courses limited the findings to certain contexts.

Another limitation the researchers discussed was the inability to control for students’ ability to separate teacher ability from subject difficulty when choosing answers. Within the realm of student surveys and the ever-changing level of maturity among the survey participants, there is certainly a need for more research into whether teachers are rated lower by their students in more difficult courses than they are by their students in easier courses.

Within the context of the literature on student surveys, this study had the unique characteristic that it did not use the Tripod instrument. Therefore, this study is able to offer a glimpse into whether student surveys as a whole are valid or whether correlations come from the validity of Tripod alone. The consistent levels of student agreement
indicated some aspects of student surveys as a whole having internal validity, while the lower confidence levels within survey topics where Tripod has performed well (student orientation categories) indicates a need for further research.

Ferguson & Danielson, 2014.

As a follow-up to the 2010 Bill and Melinda Gates Foundation study, in 2014 Ferguson and Danielson reexamined the data from the Gates study but looked in more detail at the correlation between the student survey measure and the observational measure. Because it came from the 2010 study, the sample was drawn from fourth-through eighth grade math and English teachers. Ferguson is the primary author of the Tripod survey instrument the Gates Foundation used for its study. Danielson is a principal author of Framework for Teaching model that inspired and informed most of the observation rubrics the Gates Foundation used in its 2010 study. Whereas the Gates Foundation had focused on finding the right combination of the seven or so variables under consideration, Ferguson and Danielson asked the question of whether the student survey results correlated with the adult conducted observations utilizing rubrics from the Framework for Teaching.

This study’s first step was to scale the observational and Tripod scores with an algorithm similar to that used by the value-added formula to attempt to control for race, ethnicity, and free lunch status. There was also a variable introduced to account for the effect of schools on scores.

The various categories and components of the two systems were then matched into conceptually related pairs (for instance, the major category of “establishing a culture for learning” within the Framework for Teaching model was paired with the “challenge”
and “confer” categories from the Tripod instrument). Even though the two models were developed independently, both were based on studies of elements of effective teaching. As a result, every category from one model was able to be paired with a conceptual partner from the other.

Ferguson and Danielson then performed multiple regression analyses upon each pairing to determine whether the students and the observers were reporting the same findings for each class, even though neither knew what the other had said. They tested both the Framework for Teaching categories and the Tripod categories to see if either could serve as a predictor for the partner category. The majority of pairings passed at the .05 level, though some pairings did not predict their partner score. Ferguson and Danielson explained those failures to predict as representing areas where the two paired categories were not well aligned conceptually. In those situations, administrators and students measured a slightly different aspect of the classroom and may not have come to the same conclusions. The most predictive pairing was the one based upon measures of classroom management. Ferguson and Danielson suggested that both measurement tools were the most conceptually aligned in the area of classroom management, making for the strongest agreement between adult and student observers.

Regarding student survey validity, Ferguson and Danielson posit that any measure used to inform official evaluations can become subject to attempted manipulation. The Gates Foundation study collected data from districts that were not using any of the measures in question for official purposes, and, by extension, the Ferguson and Danielson study used data unattached to significant consequences for the teacher since their study used the same data as the Gates Foundation’s. Therefore, this study cautioned districts
using either observations and/or student surveys for formal evaluations to keep a watchful eye on results to ensure tampering does not occur.

Another limitation in the area of generalizability that this study faced was the attempt to control for race, ethnicity, school, and free lunch status. For the purposes of Ferguson and Danielson’s study, that modification of the data was useful. It helped remove skewing introduced by school-specific expectations (some schools having a culture that demands more of teachers than others) and made it possible to determine whether the two measures correlated. However, when researchers wish to generalize these results and argue that this correlation means data supporting the external validity of classroom observations also support the external validity of student surveys, this scaling presents a problem. Not every district scales student survey scores according to student demographic data, and not every student survey collects enough detail about participants to make such scaling possible. Genetics also plays a role in survey response behavior (Littvay, Popa, & Fazekas, 2013; Thompson, Zhang, & Arvey, 2011), as does religious background (Sherkat, 2007). If any of these factors correlate with race, ethnicity, or socio-economic status, the Ferguson and Danielson study’s use of scaling to eliminate those effects would have made it impossible to compare a scaled and standardized theoretical world to the real world.

Chaplin et al., 2014.

In 2014, Chaplin et al. conducted a study that focused on the Pittsburgh area’s three measures of teacher evaluation (student surveys, administer observations, and value-added scores) among math, reading, and science teachers of fourth through twelfth grades. Its data discovered a correlation between observation scores and student survey
scores, observation scores and value-added scores, and student survey scores and value-added scores. Every combination of the pairing between the three variables showed statistically significant correlation. As with other studies, generalization of this study is difficult as it focuses on only one city (heavily urban, just as in the sample from the Bill and Melinda Gates Foundation study, 2010).

Additionally, Chaplin et al.’s study occurred in the first year that Pittsburgh implemented a new evaluation system, and many teachers could not be included in the sample. One-third of teachers within the district had been assigned to be part of a group receiving supported growth plans and were only rated on one part of the observational component, meaning they had to be excluded from the study due to a lack of complete data points. Teachers with unsatisfactory performances in the previous year (about 5%-7% of teachers in the district) were given no observational scores (the district did not wish to base further actions regarding this group of teachers on an untested pilot evaluation program). This lowest performing group of teachers also had to be excluded from the study. Consequently, only 329 of the 2,082 teachers in the district had a complete set of data in all three areas.

Basing the study on 15.8% of teachers in the district is understandable, as any study must use a sample smaller than the population, and 329 is an acceptable sample size (provided researchers chose the sample via random selection). Nonetheless, removing the lowest performing teachers before participant selection introduced selection bias and hindered efforts to apply this study’s findings to a larger group. This study was also limited in that it, too, used the Tripod student survey instrument and did not
contribute to the literature’s capacity to generalize that Tripod results represented student survey results in general.

**Fauth, Decristan, Rieser, Klieme, & Büttner, 2014.**

In 2014, Fauth, Decristan, Rieser, Klieme, and Büttner published a study examining the validity of student surveys in younger grades (German third graders) in order to examine whether any biasing factors impacted the validity of results. In particular, they explored the impact of teaching quality, teacher popularity, and the connection between student interest and the subject taught.

The research team found a substantial correlation between teacher quality and teacher popularity. This study found classroom management to be predictive of student learning (just as Ferguson & Danielson, 2014 did). The categories of support and cognitive activation predicted student interest in the subject matter.

Because many reservations regarding student survey validity hinge on the belief that surveys may result in popularity contests, the correlation between teacher popularity and student survey scores is one of the most interesting discoveries of this study. As the authors pointed out, however, this correlation has many possible causes. Well-rated teachers could be more popular because the students appreciate quality teaching, or inspiring student affinity may be a tool some skilled teachers use to motivate students to learn. It is, of course, also possible that the popularity caused the high ratings and the surveys may have had a popularity contest aspect to them. This study did not discover data that supported any conclusion beyond that of claiming a correlation exists.

The final unique characteristic of this study was that trained staff administered the surveys verbally, rather than having teachers administer written surveys. Just as
Luppescu (2016) pointed out, the younger the participants in the survey, the greater the potential for reading and language difficulties to impact results. Therefore, all correlations from the Fauth, Decristan, Rieser, Klieme, and Büttner study should be interpreted with the understanding that they made provision to control for reading and language difficulties.

**Luppescu, 2016.**

Luppescu’s (2016) primary focus was not to evaluate the validity of student surveys, but rather to determine whether surveys of fourth and fifth graders could yield the same level of internal validity that surveys involving sixth through twelfth graders had previously demonstrated. The question of survey validity was not totally ignored. Luppescu’s study came on the heels of a 2015 study (Klugman, Gordon, Sebring, & Sporte) that found a correlation between positive survey results and positive student outcomes but did not have adequate data to feel confident in hypothesizing any particular causal relationship.

Luppescu was primarily interested in the question of survey participant age and was attempting to solve a problem facing Illinois elementary schools: in 2011, the Illinois legislature mandated the use of the 5Essentials survey for all schools in the state. The 5Essentials survey had been used within the Chicago area for a number of years but had never been tested on students below sixth grade. Chicago elementary schools housed kindergarten through sixth grade, and all student data for elementary schools came from sixth graders alone. Outside of Chicago, most elementary schools contained kindergarten through fifth grade, so Luppescu examined whether fourth and fifth graders could produce the same level of internal validity that sixth through twelve graders displayed.
The external validity of the 5Essentials survey was assumed and not within the scope of the study.

Luppescu found a slight drop in internal validity among the surveys completed by fourth and fifth graders when compared to those completed by sixth through twelfth graders. Nonetheless, all data still fell within the defined parameters for statistical significance and supported the hypothesis that fourth and fifth graders can produce sufficient internal validity to make their survey responses usable.

The primary limitation of this study was its assumption that the 5Essentials survey had external validity in all parts of the state. Luppescu supported this assumption by citing Klugman et al., but Klugman et al.’s study clearly stated that its data did not support a claim of causal relationship, only one of correlation. Klugman et al. were clear in reminding readers that by sampling only one year’s data, one could not assume correlation was a consistent pattern. All they could demonstrate was that correlation occurred in that one year.

For researchers, it is also important to note that Luppescu did not examine grades younger than fourth. A focus on fourth and fifth grade served the purposes at hand (providing non-Chicagoan elementary schools with grade bands that could complete the survey), but one key question is how the slight decrease in internal validity seen in fourth and fifth graders would progress if younger grades participated in the study. In some areas such as Memphis, students as young as kindergarten participate in surveys evaluating teachers (Parent Memo, 2014), so this question of validity for younger grades has become a pressing issue.
Muñoz, & Dossett, 2016.

Muñoz and Dossett compared student surveys (in the form of Kentucky’s version of the 7C indicators that inspired the Tripod instrument), observations (in the form of Danielson’s Kentucky Framework for Teaching Domains), and value-added scores to determine which aspects of surveys and observations were most predictive of student achievement. Teachers recruited as participants came from across Kentucky (194 math teachers, 191 reading teachers). This study occurred during the pilot phase of a new teacher evaluation system, so teachers from rural areas, towns, suburbs, and urban areas were all included, though not all teachers in the state were part of the pilot and the team could only select participants from regions used for that program. Muñoz and Dossett’s data found the indicators most predictive of reading gains to all be observational (though four of the seven survey categories also correlated with reading gains at statistically significant levels). The indicators most predictive of math gains were spread between observations and student surveys.

The primary limitation of this study was its use of pilot year data, which meant that all data was collected without the influencing factors that come with consequences for outcomes. Additionally, while no evidence of biased selection criteria exists, the researchers had no control over which districts within the state participated in the pilot. This study’s data set did not include any teachers rated as “ineffective” (the lowest level within the state’s classification system). This absence could have impacted results, making it hard to generalize.
Wallace, Kelcey, Ruzek, 2016.

In 2016, Wallace, Kelcey, and Ruzek evaluated the underlying structure of the Tripod student survey instrument, using the 2010 MET study (Bill and Melinda Gates Foundation, 2010). The limitation that the team claimed Tripod had was in the lack of normed and trained observers. While the authors conceded that the Tripod results reported by the Bill and Melinda Gates Foundation study were accurate and appeared to be consistent regardless of which teacher’s class completed the survey, the authors contended that the theories supporting the effectiveness of the 7 Cs and the Tripod survey in general were ill-defined and not well enough developed for one to conclude that Tripod scores were valid measurements of teacher quality.

Analysis of theories related to the studies

For any use of student surveys, the most pressing and obvious theoretical assumption is that students are able to recognize good teaching (assuming the survey asks the right questions) and are mature enough to be honest in their responses and avoid bias if a teacher has recently done something unpopular. This theoretical assumption is so obvious that Tripod Educational Partners Inc. even addresses it on the “Frequently Asked Questions” page of its website (Tripod, 2017).

Yet, one less frequently considered question is, “Are students willing to take the risk of being honest?” Considering Maslow’s Hierarchy of Needs (Maslow, 2012), one could see how students able to distinguish good teaching, able to set aside their feelings about the teacher, and able to avoid the temptation to lie to exact revenge might still not be willing to tell the truth about a teacher if retaliation is a concern. Student surveys are focused on improving student access to satisfaction of the safety, belonging, and esteem
needs of Maslow’s hierarchy. The assumption is that by asking the students about their teachers, they will feel a sense of ownership. Belonging and esteem will be enhanced and school leaders might even discover new ways to improve the students’ feelings of safety and security.

But if students fear the wrath of the teacher, the threat to their safety needs (in the form of retaliation) might outweigh their ability to increase the overall safety by sharing their thoughts honestly. In such a case, the students may feel they must choose between long-term security (provided by honest answers on the survey) and short-term security (threatened by honesty, if honesty causes negative comments regarding a teacher willing to retaliate).

**Racial/Genetic Considerations.**

Given that fear of retaliation and related concerns regarding Maslow’s hierarchy of needs can impact human behavior, it is important to consider that some people have genetic predispositions toward survey non-responsiveness (Littvay, Popa, & Fazekas, 2013; Thompson, Zhang, & Arvey, 2011) and that the current theory is that this non-responsiveness stems from genetic predispositions toward fear of retaliation. Little to no literature exists exploring whether this genetic predisposition is disproportionately distributed among some races or ethnicities, but both males and racial minorities have increased propensities toward non-responsiveness on surveys (Porter & Umbach, 2006). For this reason, researchers should be cautious in using studies about survey validity to predict the validity of student surveys in environments with substantially different racial or ethnic compositions.
This lack of confidence is especially important given the existing literature’s heavy overrepresentation of Blacks within the communities used for the majority of studies. As an example, see table 2.1. Of the seven cities used in Bill and Melinda Gates Foundation, 2010; Chaplin et al., 2014; Ferguson & Danielson, 2014; and White, 2013, only three (Denver, Hillsborough County, and the Newton-Conover City Schools) did not have Black populations in excess of twice the national average based on the 2010 census or comparable data.
### Table 2.1
*Racial Compositions of Studied Communities*

<table>
<thead>
<tr>
<th>Racial group</th>
<th>U.S.</th>
<th>Charlotte</th>
<th>Dallas</th>
<th>Denver</th>
<th>Hillsborough County</th>
<th>NY City</th>
<th>Newton-Conover</th>
<th>Pittsburgh</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>72.4%</td>
<td>50%</td>
<td>50.7%</td>
<td>68.9%</td>
<td>71.3%</td>
<td>44%</td>
<td>50%</td>
<td>66%</td>
</tr>
<tr>
<td>Black</td>
<td>12.6%</td>
<td>35%</td>
<td>35%</td>
<td>25%</td>
<td>16.7%</td>
<td>25.5%</td>
<td>13%</td>
<td>26.1%</td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>16.3%</td>
<td>13.1%</td>
<td>13.1%</td>
<td>42.4%</td>
<td>24.9%</td>
<td>28.6%</td>
<td>25%</td>
<td>2.3%</td>
</tr>
<tr>
<td>Asian</td>
<td>4.8%</td>
<td>5%</td>
<td>2.9%</td>
<td>3.4%</td>
<td>3.4%</td>
<td>12.7%</td>
<td>6%</td>
<td>4.4%</td>
</tr>
</tbody>
</table>

*Note.* Data collected from (Newton-Conover City Schools, 2015; QuickFacts, 2015)
This frequent use of non-representative racial compositions is even more concerning when considering that the area with the Black population closest to the national average (Newton-Conover) is the area with data that returned a finding of “no correlation” between student surveys and value-added scores. In college surveys, both the institute’s location in an urban area and an increase in non-white participants correlate to lower response rates (Porter & Umbach, 2006). One must consider the possibility that a school containing many students from racial or ethnic minorities might contain a disproportionately large number of students fearing retaliation (and willing to lie on the survey in order to protect themselves). Alternately, the presence of a larger percentage of racial or ethnic minority members in the community might encourage students from these groups to feel safer in sharing their views. This last possibility is not likely to occur in a location with more typical racial demographics. Because of these questions, it is imperative to conduct further studies and explore whether student surveys are valid within non-urban contexts containing more typical racial compositions.

**Age/Aptitude Considerations.**

One reality which all surveys must consider is that participants will often answer survey questions even if they do not understand the questions (Al-Tayyib, Rogers, Gribble, Villarroel, & Turner, 2002). Therefore, understanding the impact of literacy and maturity upon survey results is imperative when young students are part of the sample.

Questions of literacy have generally been minimized within the literature because most studies have focused on grades four and above with only a few studies venturing as low as third grade. However, some districts have students as young as kindergarten participating in surveys judging teacher effectiveness (Parent memo, 2014), and Tripod
Educational Partners Inc. appears to have received enough interest in surveys of students in kindergarten through third grade that they have developed an instrument especially for that age range (Tripod, 2017).

There is a need for more research to verify the validity of using survey participants in third grade or younger, given that peer-to-peer literacy differences could be very high. Teachers of students who begin the year with higher academic levels (relative to their peers) do tend to receive higher observational scores (Warring, 2015). This discovery, combined with the high inter-rater reliability of teacher observations (Sporte, Jiang, & Luppescu, 2014), reveals the potential for student aptitude to impact teacher effectiveness. Among very young children, it is certainly probable that those students benefiting the least from classroom practices and the most educationally disadvantaged would frequently lack the literacy needed to express their viewpoints in a coherent and consistent manner. As most well designed survey instruments elicit the same information in multiple ways and check for inconsistencies, the inability to read and respond reliably could skew results. Survey administration teams might disregard a disproportionate number of answers from educationally disadvantaged students due to the students’ inability to express their opinion consistently. This would create overrepresentation of results from the most educated students (those best served by the existing system).

Age limitations on aptitude aside, all groups appear to suffer from aptitude limitations impacting their abilities to report academic behaviors in survey format (Porter, 2010). Common issues include finding definitions for even the commonest of terms, such as whether readers should understand the word “you” to be singular or plural; the
related need for value judgments in even seemingly factual statements (for instance, if the student is asked whether the teacher pushes the class to work hard, the extent of the rigor needed to qualify as a push must be defined by the student); and struggles to accurately assess the frequency of activities.

**Unresolved issues, significant problems and unanswered questions requiring study**

For the reasons listed above, the main unanswered question within the field of student surveys remains: “Are student surveys valid in non-urban environments with racial compositions closer to the national averages?” The Bill and Melinda Gates Foundation (2010) provided a robust and extensive examination of student surveys within urban districts with high numbers of students from racial or ethnic minorities. Combined with corroboration from Chaplin et al. (2014), there is a growing body of evidence that student perception surveys can be effective tools within urban centers.

However, White’s (2013) failure to correlate student survey results and value-added scores when studying a smaller and non-urban district is troubling. Muñoz and Dossett (2016) produced different findings. One of these studies must, therefore, represent either an anomaly or a type I or II error regarding the hypothesis. These results might also signal that student survey validity breaks down in the absence of certain population or racial compositions. There is a need for further work to explore which of these is the case.

It would be easy to assume that if student surveys are known to be valid for urban districts with large numbers of students from racial or ethnic minorities, there is nothing wrong with allowing that kind of district to use surveys. One might assume any problem could be solved by simply encouraging smaller districts not to adopt surveys for
evaluation measures. The catch comes when one considers that within urban districts containing many members of racial or ethnic minorities, not all of the schools have large numbers of students from racial or ethnic minorities. Some of the schools may not even be urban if the district includes suburbs. Given that teacher evaluation scores in some districts determine salaries and non-renewal decisions, the stakes are too high to consider a measure that is only accurate within the majority of the district. Measures used for evaluation must be reasonably accurate in all potential situations, barring random chance anomalies.

It is, therefore necessary to determine where the limits of validity rest in terms of population density and percentage of the population belonging to racial or ethnic minorities. This study proposes to answer that question. If it can be demonstrated that student surveys are valid regardless of population or racial factors and that White’s results were outliers, then much of the debate can be put to rest. On the other hand, if White’s results find support from a new and more comprehensive study, such findings would have extensive repercussions for educational policy and future studies.
Chapter 3: Research Methodology

Introduction

This study stands on the shoulders of preexisting studies about student survey validity and seeks to fill in some gaps that remain. Most of the existing literature on student surveys has focused on large urban districts with a high minority population (Chaplin et al., 2014; Bill and Melinda Gates Foundation, 2010; Luppescu, 2016). This use of consistently similar environments has been a threat to generalizability, which this study attempts to resolve by focusing on non-urban schools and testing whether student survey validity remains despite the change of environment.

Population and sample

The sample came from schools belonging to a Tennessee school district that, for anonymity, will be called County Schools. The County school district at the time these data were produced contained one urban area and a number of incorporated non-urban areas. This study examined these schools’ data to see if they produced results similar to those of previous studies. While previous studies make it logical to hypothesize that surveys conducted within an urban area should be valid, this study tested this assumption so as to lay a firm foundation for drawing conclusions from the non-urban schools’ data.

Since student survey and value-added data must come from the same school year (thus, from a past school year), all data utilized were preexisting data County School supplied. For that reason, the teachers used in the study were limited to those whose students participated in student surveys and for whom value-added scores were calculated individually. The tests used to generate Tennessee’s value-added scores begin in third grade, so teachers of kindergarten through second grade were automatically excluded.
No teacher was used if the only value-added score available was school-wide, since that could not be assumed to necessarily correlate to student perceptions of that one particular teacher. The researcher included all teachers meeting those two criteria (having a student survey score and having an individual value-added score), regardless of grade or subject taught.

**Description of instruments(s)**

The student survey measure utilized the instrument developed by Tripod Education Partners, Inc. from the 2013-14 school year. The researcher chose to reject data from more recent years because shortly after 2013-14, Tennessee began a transition to a new state testing model and experienced a number of administrative issues. As a result, there were several years in which value-added scores could not be calculated for all grade levels and others in which the newness of the test made reliance upon that data problematic.

The researcher used Tennessee’s value-added score (TVAAS) as a measure of student growth to ensure all teachers were scored using the same model. This further increased this study’s alignment with existing evaluation data, helping to test whether each measure within a teacher’s evaluation score contributed to a fair and equitable evaluation of teacher quality.

**Research procedures and time period of the study**

Because value-added scores often cannot be calculated until the late summer or fall of the following school year, and it was important to ensure the student survey data came from the same school year as its corresponding value-added calculations, all data were provided by County Schools. Due to the inconsistent publication of TVAAS data
within the past few years, as well as the fact that 2013-14 was the only school year in which all schools within the county participated in student surveys, the 2013-14 school year was the only one represented. Using this preexisting data, it took eight months to complete this study.

**Statistical Test Used**

Within this study, the independent variable is student growth (as measured by TVAAS scores) and the dependent variable is the student survey scores (as measured by Tripod). The assignment of these variables rests on the assumption that students whose teachers help them grow will answer positively to the Tripod survey’s questions and generate higher scores for their teachers. Because both the TVAAS and Tripod scores used in this study came from the average score teachers earned across the population of students participating in the process generating each score, decimal scores were possible, though County Schools rounded these results to the nearest whole number before sending the data to the researcher. For this reason, both TVAAS and Tripod scores were measured in interval scales.

This study’s hypothesis assumes a direct correlation between the growth a teacher produces and the answers students give on the survey, but is unable to postulate causation between the two variables. For these reasons, the researcher concluded that a Pearson’s product moment correlation coefficient would be the best statistical test as it measures the correlation between pairs of data using an interval scale. A Pearson’s coefficient assumes the normality of variables, and TVAAS’s use of student growth compared to peers makes TVAAS scores normally distributed. If a correlation exists between TVAAS and student survey scores, both should be normally distributed.
Prior to testing the data, the Pearson’s coefficient’s assumptions of absence of outliers, linearity, and homoscedasticity could not be verified. However, the researcher did not deem this a problem because violations of any of those assumptions would be likely to argue for a rejection of the hypothesis and/or discontinuation of the practice of using student surveys to evaluate teachers. If a significant number of outliers were to emerge, one would find it difficult to argue that student surveys are equitable measures of teacher quality, given that a statistically significant number of teachers would be judged to be substantially better or worse than their skills deserve. If the data were non-linear or heteroscedastic, the same problem would emerge since student survey scores factor into teacher observations. Lack of either linearity or homoscedasticity would mean that teachers could drastically increase their evaluation scores without drastically improving student growth, or teachers could fail to see great gains in evaluation scores even after producing great improvements in student growth. As it addressed the largest number of considerations, Pearson’s coefficient was the most appropriate test for measuring whether the data supported the hypothesis.

**Data Analysis**

Because all data came from the same district and has been used in teacher evaluation calculations, most data standardization had already occurred. Student survey data had already been processed and converted to a final score, which was used as part of each teacher’s evaluation for the 2013-14 school year. Tripod surveys contain seven indicators, which may be combined to create a composite score. This composite is what County Schools used for determining the student survey portion of the educator’s evaluation calculation, so the researcher only explored the composite score in an attempt
to ensure that this study tested the equity of existing practice using the actual data informing that practice. The researcher did not explore the scores from each of the seven indicators in depth nor their individual correlation to value-added scores because, aside from the researcher’s desire to focus this study on testing real-life situations, this study’s central question focused on whether student surveys serve as indicators of teacher quality. Focusing on each indicator’s individual correlation could divert the study’s focus into questions of what makes student surveys valid, rather than asking if the practice under consideration is valid. If failure to examine each of the seven indicators introduced bias and prevented these data from showing proper correlation, the same bias would have occurred when County Schools determined each teacher’s evaluation score. If such bias existed, overcoming it would have served only to mask the impact of continued bias within actual practice.

Value-added scores were provided to the researcher already standardized by the state and processed by the district. Within Tennessee, value-added scores are reported on a scale of 1-5, with a score of 3 indicating an average growth meeting expectations and maintaining the student’s same relevant position as compared to the state average (SAS EVAAS, 2017). A score of 2 or 4 indicates results one standard error below or above expectations, respectively, and scores of 1 and 5 show results two or more standard errors below or above expectations, respectively. The data, as the district conveyed them to the researcher, were on this same scale. Value-added scoring is generally discussed in terms of whole numbers, and County Schools provided the researcher with TVAAS scores already rounded to the nearest whole number.
Once student survey scores and value-added scores were collected, the researcher used a Pearson’s product moment correlation coefficient to determine whether this study’s hypothesis of a correlation between student survey scores and value-added scores could be supported by the data with a p-value of .05. The researcher selected a p-value of .05 as opposed to more stringent criteria on the basis that, despite the multitude of variables for which value-added and student surveys control, there is still a human element. Stricter criteria ran excessive risk of a type II error. Additionally, while student survey data should correlate to student growth, few would object to personable, supportive, and encouraging teachers performing disproportionally well on a student survey. Therefore, it is best to use the looser p-value of .05.

This study, using previous literature, has identified three possible threats to the validity of student survey results: racial considerations, aptitude considerations, and population density (urban vs. non-urban). This study’s primary concern was with testing the impact of population density, with a secondary concern related to exploring the effect of previous studies’ use of racially atypical samples. This study drew its primary conclusions from data related to those two topics.

However, while the authors of several previous studies have expressed concern regarding the impact of age and aptitude on student survey results, few studies have explored this issue in depth, and the role of age upon student surveys remains a primarily unanswered question. This study was similarly unprepared to fully explore the impact of age and aptitude, but the researcher agreed that age and aptitude are potential concerns and realized that County Schools’ inclusion of students in grades as low as kindergarten could have impacted the validity of the conclusions drawn in this study. For this reason,
this study analyzed its data along grade level divisions as well as along urban vs. non-urban lines in order to both lay the foundation for this study to generate more informed interpretations and to lay a small foundation from whence future studies might address the important question of how age and aptitude impact survey results.

To address the questions of whether population density (urban vs. non-urban) and/or racial considerations impacted student survey results, the researcher performed three tests. A Pearson’s correlation coefficient was calculated to determine the strength of correlation between value-added scores and student survey scores for all county schools (location not considered). This process helped determine whether the sample as a whole conformed to the expectations set by the majority of past studies, or whether these data represented an unexpected situation more akin to that discovered by White’s 2013 study.

Next, the researcher performed the same statistical test, but only on county schools located within the urban area. This served much of the same purpose as the previous test, helping determine whether the urban area followed the expected pattern. The researcher knew that if deviations turned up within the non-urban areas, it would be important to know whether the urban area and the district as a whole conformed to expectations, or whether deviations may be district-wide and not the results of differing population densities or racial compositions.

These previous two tests laid the groundwork for the researcher to calculate the Pearson’s correlation for all county schools not within the urban area. Because there is a strong correlation between the racial composition of a county school and whether that school is urban or non-urban, those tests were not able to precisely pinpoint whether any lack of validity discovered resulted from racial composition or from population density
(or pure chance). Nonetheless, because a similar correlation between racial composition and population density exists across the country, the researcher decided that this inability to precisely pinpoint the cause of any potential change in validity was acceptable. Should the urban vs. non-urban distinction prove to impact survey validity, it would be more expedient for a future study to explore the causes in more depth. This study’s focus on only one district was too limiting for conclusions involving race to be generalizable to the country as a whole, so future research would be the better course.

Because County Schools’ data involved children from third through twelfth grade, any lack of validity discovered could have stemmed from the impact of age and not from the urban vs. non-urban distinction. To help the researcher identify all possible causes for the results, and to lay the foundation for any future studies seeking to explore the role of age upon survey results, the researcher calculated twelve more Pearson’s correlation coefficients to explore the various permutations of grade groups and locations. The researcher calculated the coefficient for all county schools within the third through fifth grade range, the coefficient for all schools within the sixth through eighth grade range, and then for the ninth through twelfth grade band. County Schools also reported a category which included teachers serving students in fifth grade (or below) as well as students in sixth grade (or above). This group was labeled “crosses elementary and middle.” This process was repeated using the same grade bands and looking only at schools within the urban area, and then again with the same bands and only looking at schools in the non-urban areas. As the impact of age does not directly relate to this study’s hypothesis, the researcher included these data simply to aid future studies which
might consider the impact of age and to help this study’s researcher develop informed conclusions regarding the impact of the urban vs. non-urban distinction.
Chapter 4: Results of the Data Analysis

Introduction

Chapter 4 presents the results of the data pertaining to student perception survey scores and the analysis of whether there is a correlation between those scores and value-added scores of teachers. The majority of pre-existing literature supports the idea that student surveys are valid measures of teacher quality but is also drawn heavily from urban environments, limiting generalizability to non-urban areas. In order to test the hypothesis that there is a direct correlation between a teacher’s score on student surveys and the average amount of student growth that a teacher produces within a non-urban setting, this study examined a district containing a mixture of urban and non-urban schools.

Student survey results were calculated using an instrument developed by Tripod Education Partners. The Tripod instrument offers a student survey version designed for students in kindergarten through second grade, a version designed for students in third through fifth grade, and one for students in sixth grade and beyond. In all three of these versions, students are asked to rate teachers in seven areas of teacher behaviors that the Gates Foundation along with Tripod Education Partners have identified as categories of behaviors indicative of strong teachers. These areas are: Care, Confer, Captivate, Clarify, Consolidate, Challenge, and Classroom Management. Naturally, there is also a scoring category that some districts use which is just the average of all seven scores. This study used that composite score because that is the score County Schools used for teacher evaluations.
This measure quantified student growth using Tennessee’s value added score (TVAAS) for each teacher. Like most value-added score calculations, TVAAS utilizes an algorithm that considers each student in relationship to what are deemed peers for that same student. Peers are selected by finding other students in the state who were roughly equal to the student in a variety of factors, such as attendance rate, economic situation, past test performance, etc. The given student is then ranked compared to these peers and the TVAAS for each student depends upon how that student grew in relationship to peers. A score of 3 represents a student who remained in the same position compared to those peers at the end of the year. A student moving one standard deviation when compared to peers gives the teacher a score of either a 2 (if the move was negative) or a 4 (if the move was positive). Moves of more than two standard deviations produce scores of either a 1 (for negative moves) or of 5 (for positive moves). The state averages the scores for all students under a given teacher and that becomes the teacher’s TVAAS and, for the purpose of this study, represented how much growth that teacher produced. Inasmuch as both the Tripod and TVAAS instruments generate results in an interval scale, the researcher used a Pearson’s coefficient to look for a relationship between student survey scores and student growth.

**Participants’ demographics**

The data for this study were supplied by a school district within the southeastern portion of the United States. The district encompassed all public schools within the county and, for the sake of anonymity, is referred to as County Schools within this study. The district’s enrollment was 149,928 students (Tennessee Department of Education) during the year these data were produced. There were 6.6% of students identified as
English Learner Students, 68.9% identified as economically disadvantaged, and 13.3% identified as having a disability.

The researcher chose this district because it encompassed schools from an urban area, but also schools from outside that urban area, some of which had populations from rural areas and most of which came from one of the six municipalities surrounding the urban area which were classified as non-urban based on the U. S. Census Bureau’s definitions. Literature to date has focused heavily upon urban areas, so this combination of school locations allowed the researcher to explore validity within more diverse settings. The additional benefit of County Schools as a study participant was that most of the studies to date have focused on areas with racial compositions greatly different from the national averages. However, the non-urban suburbs of the district contained demographic compositions much closer to the national averages, helping ensure this study’s findings can be generalized with greater confidence. Table 4.1 presents the racial composition of the district as a whole, as well as the compositions of each non-urban area. These numbers represent the population of each geographic area, so the actual composition of the schools involved would vary slightly due to school zoning, differing age distributions, and private school attendance rates.
Table 4.1
Racial Composition of County School’s Area (U. S. Census Bureau)

<table>
<thead>
<tr>
<th></th>
<th>U. S. County</th>
<th>Urban Area</th>
<th>Non-Urban Area #1</th>
<th>Non-Urban Area #2</th>
<th>Non-Urban Area #3</th>
<th>Non-Urban Area #4</th>
<th>Non-Urban Area #5</th>
<th>Non-Urban Area #6</th>
<th>Non-Urban Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>72.4%</td>
<td>41.4%</td>
<td>29%</td>
<td>90%</td>
<td>80%</td>
<td>79%</td>
<td>81%</td>
<td>83%</td>
<td>65%</td>
</tr>
<tr>
<td>Black</td>
<td>12.6%</td>
<td>54%</td>
<td>63%</td>
<td>4%</td>
<td>11%</td>
<td>16%</td>
<td>14%</td>
<td>9%</td>
<td>26%</td>
</tr>
<tr>
<td>Hispanic / Latino</td>
<td>16.3%</td>
<td>6%</td>
<td>7%</td>
<td>2%</td>
<td>3%</td>
<td>3%</td>
<td>3%</td>
<td>3%</td>
<td>6%</td>
</tr>
</tbody>
</table>

March 2018
The data set consists of 2,457 teachers of all grades from County Schools who had both an individual TVAAS and a Tripod score for the 2013-14 school year. Of these teachers, 1,999 worked in schools located within the urban area, 458 worked outside of the urban area.

Organization of the study’s findings

The first portion of this chapter focuses on the hypothesis of this study and explores whether student survey scores and value-added scores correlated within the district as a whole, within the urban area alone, and then within the non-urban parts of the district as well. The second portion of this chapter explores whether a correlation exists within various grade-based subdivisions of the data (see table 4.2). The current paucity of information dealing with student survey use within lower grades made it difficult for this study to adequately test any hypotheses related to the impact of grade level and maturity. However, it is the researcher’s opinion that the impact of grade level upon student survey validity is the next step in the process of refining and improving survey-based measures of teacher quality. By providing this additional information regarding what the data show when divided into grade-based groups, the researcher wishes to give future studies a stronger foundation for considering these issues.

Results regarding the hypothesis

District-wide results.

Before the researcher could test the hypothesis, it was necessary to determine whether County Schools as a whole showed a correlation between student survey results and student growth. Given that the majority of schools and students fell within the urban area and over 81% of data pairs came from the urban area, the existing literature
predicted there would be a correlation. If a correlation failed to exist, the probability of these data representing an anomaly would have been too high for the researcher to draw any firm conclusions. The district-wide data contained 2,457 pairs of data and had a Pearson’s coefficient of .085 and a p-value of 0.00002. This study has established a p-value of .05 as the threshold for retaining the hypothesis. Therefore, the district-wide results show a statistically significant correlation exists between student survey scores and student growth. These data support the assumption that these data can provide insight into presence of correlation in the urban and non-urban environments.

Urban area, all schools.

The next step was to verify that the data coming only from schools within the district’s urban area showed a correlation between student survey scores and value-added scores. Previous studies in this field, the majority of which focused on urban areas and found a correlation, support the hypothesis that urban areas produce valid student survey results. There were 1,999 pairs of data from the urban schools of the district, and the Pearson’s coefficient was .09. This produced a p-value of 0.00006. These data agreed with previous studies that urban areas can produce strong levels of correlation between student survey scores and student growth. This set the stage for verifying that non-urban schools could also show statistically significant correlation.

Non-urban area, all schools.

Finally, the hypothesis could be tested by looking at the data from all county schools located in non-urban areas. There were 458 pairs of data drawn from these parts of the district. The hypothesis to test was:
• $H_1$: There is a direct correlation between a teacher’s score on student surveys and the average amount of student growth that teacher produces within a non-urban setting.

• $H_0$: There is no correlation between a teacher’s student survey scores and the student growth produced within a non-urban setting.

The group of all non-urban schools had a Pearson’s coefficient of .134 and a p-value of 0.004. This study tests at the .05 level. Consequently, the null-hypothesis that student survey scores do no correlate to student growth is rejected. These data supported the conclusion that non-urban areas produce student survey scores which correlate to student growth.

**Grade-band results.**

This study’s hypothesis deals with the validity and generalizability of student survey results when comparing urban to non-urban environments. However, during the course of the literature review, the researcher found several studies suggesting that age and related aptitude issues might also play a significant role in determining the validity of student survey results and the level of accuracy those results have for measuring teacher quality.

With this knowledge, the author of this study faced the problem that County Schools surveyed students younger than any other study. Without knowing whether this introduced unidentified bias or limitations, it appeared unwise to interpret the study’s results without considering the potential impact of grade and aptitude. However, even those authors expressing concern over the issue of grade level did not focus their studies on participants young enough to provide a peer-to-peer comparison with participants in
this study’s data. There were simply not enough studies into the impact of grade level to allow the researcher to consider both location and grade level in a fair and accurate manner that could rely upon the testing of a hypothesis.

The researcher determined that the best solution was to test a hypothesis only based on the consideration of location, which has already been presented above. Nonetheless, in order to produce a more informed discussion of the implications of this data, and in the hope of equipping future researchers to address the unanswered questions related to the impact of grade level upon student survey validity, the researcher calculated Pearson’s coefficients and p-values for grade-based cross-sections of the data (see table 4.2). For the purposes of this section, the term “elementary school” refers to fifth grade and below. “Middle school” refers to sixth through eighth grade. “High school” refers to ninth grade through twelfth. County Schools’ data also included points, which it referred to as “crosses elementary and middle,” and those are also included. The researcher was not provided with an explanation of what caused the overlap, and this group is likely to be a hodgepodge of unique situations, so readers should view this category with extreme caution.
Table 4.2  
*Correlation Between Student Survey Results and Student Growth By Location and Grade Group*

<table>
<thead>
<tr>
<th>Location/Grade Group</th>
<th>Pearson’s Coefficient</th>
<th>N=</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Elementary School Data Points (grades 5 and under)</td>
<td>.049</td>
<td>1082</td>
<td>.13</td>
</tr>
<tr>
<td>All Urban Elementary School Data Points</td>
<td>.021</td>
<td>868</td>
<td>.54</td>
</tr>
<tr>
<td>All Non-Urban Elementary School Data Points</td>
<td>.179</td>
<td>214</td>
<td>.01</td>
</tr>
<tr>
<td>All Data Points Crossing Elementary and Middle School Grade Bands</td>
<td>.179</td>
<td>140</td>
<td>.03</td>
</tr>
<tr>
<td>All Urban Data Points Crossing Elementary and Middle School Grade Bands</td>
<td>.194</td>
<td>100</td>
<td>.05</td>
</tr>
<tr>
<td>All Non-Urban Data Points Crossing Elementary and Middle School Grade Bands</td>
<td>.119</td>
<td>40</td>
<td>.46</td>
</tr>
<tr>
<td>All Middle School Data points (grades 6-8)</td>
<td>.069</td>
<td>655</td>
<td>.08</td>
</tr>
<tr>
<td>All Urban Middle School Data points</td>
<td>.068</td>
<td>552</td>
<td>.11</td>
</tr>
<tr>
<td>All Non-Urban Middle School Data points</td>
<td>.122</td>
<td>102</td>
<td>.22</td>
</tr>
<tr>
<td>All High School Data points (grades 9-12)</td>
<td>.149</td>
<td>581</td>
<td>0.0003</td>
</tr>
<tr>
<td>All Urban High School Data points</td>
<td>.192</td>
<td>479</td>
<td>0.00002</td>
</tr>
<tr>
<td>All Non-Urban High School Data points</td>
<td>.005</td>
<td>102</td>
<td>.96</td>
</tr>
</tbody>
</table>
Relevant, yet unexpected, results and findings discovered

In regard to the hypothesis, these data support the idea that urban and non-urban areas are able to produce statistically significant correlation between student survey scores and value added scores. There is a noticeable difference between the p-value for each (.00006 for urban as opposed to .004 for non-urban). That difference could be attributed to the larger sample size for the urban area schools.

A more noticeable and unexpected result is that, among the grade level groups, there were vastly different p-values between so many of what should be comparable groups. Urban elementary schools’ results had a p-value of .54, but results from non-urban elementary schools produced values of only .03. Urban data crossing elementary and middle school produced a p-value of .05, but the non-urban set of this data produced a p-value of .46. Once again, it should be noted that the category of teachers crossing elementary and middle school is the least reliable, due to the variety of situations it encompasses. Among high schools the trend was similar with urban high school data producing a p-value of .00002 and non-urban high school data producing a p-value of .96. It would be reckless to look at these data post hoc, formulate a hypothesis, and assume they prove anything. Yet, knowing that the overall data were valid and the urban vs. non-urban distinction alone does not appear to impact validity, the strong difference between these groups of data does seem to suggest unidentified factors are impacting validity and a detailed study into the impact of grade level is in order.

Summary

This study sought to explore the hypothesis that student survey results correlate to student growth within non-urban environments. Most of the existing literature addressed
the question of validity within the context of urban environments, so this study was a way to verify the generalizability of those previous findings. These data supported the hypothesis that a statistically significant correlation does exist when surveys are administered in either urban or non-urban environments.

Yet, this study’s findings also recommend two courses of action. First, at least one more study into the impact of urban vs. non-urban environments needs to occur. There are a number of studies considering urban areas, which support the validity of surveys within an urban environment. White’s 2013 study was one of the few to consider a non-urban environment and found a lack of validity. As this study has also considered non-urban areas and found validity, it would appear prudent to conduct at least one more study to ascertain whether White’s or this study best represents the reality of non-urban area’s student survey validity.

Second, these data suggest that the impacts of age and aptitude upon the validity of student survey results require further study. Given the lack of evidence to suggest the urban vs. non-urban distinction could account for the drastically different results among the grade-based groupings, the most logical assumption as to the cause is that age, aptitude, or other factors are causing the variation. Further research is needed to explore this possibility and help districts make informed choices regarding the use of student surveys.
Chapter 5: Conclusions, Implications and Recommendations

Introduction

The rise of the teacher accountability movement and teacher evaluation models has inspired many to search for the most expedient and efficient methods of determining teacher quality. Student surveys have emerged as a promising way to combine high levels of direct observation of practice with low levels of intrusion. Yet, the idea of surveying students also faces some challenges. To date, most studies into the validity of student surveys have focused on urban areas only, leaving open the question of whether the cultural and social differences present in non-urban areas would impact validity. Additionally, age and aptitude can impact results since participants have been shown to answer questions even if they do not understand what is asked (Al-Tayyib, Rogers, Gribble, Villarroel, & Turner, 2002).

In light of the current literature, this study attempted to determine whether student surveys can be valid in non-urban areas. It left the question of age and aptitude for future studies, but chapter 4 included information of what the data revealed when grouped according to grade-level bands in the hope of helping future studies explore the role of age and aptitude. For the purposes of this study, the only hypothesis tested was that student survey results correlate to student growth within non-urban environments. Looking only at that question, the hypothesis was retained and these data do suggest that student surveys are valid in non-urban settings.

Discussion of conclusions drawn regarding the hypotheses

Regarding the hypothesis that non-urban areas can produce student survey scores which correlate to student growth (as measured by value-added scores), these data
supported the hypothesis. This is especially significant because this finding can serve as a basis for better understanding student survey behaviors. The researcher’s primary theoretical framework for questioning whether the urban / non-urban distinction served as a potential source of bias was that of Maslow’s Hierarchy of Needs.

According to Maslow’s theory, people prioritize the fulfillment of some needs over others. Of particular relevance to the issue of student surveys is the position of safety needs within the order of priorities. Student surveys partially rely upon a person’s need to fulfill the desire for love / belonging and the need for esteem. By asking students for their opinions, schools are expecting that students who feel loved and accepted by their teachers will respond to the meeting of that need by sharing their joy. Those who feel their teachers do not listen to them, esteem them, etc. are likely to respond to the lack that they feel by sharing their thoughts in the hope that the school administration will respond to the feedback and help students meet those needs.

This study’s researcher was concerned that a possible barrier to this expected progression of events could come from the fact that Maslow identified the need to ensure one’s safety as more immediate and more pressing than the needs driving honest student survey responses. If a student belongs to a minority group and believes the majority group will not value his or her opinion as a result, the fear of retaliation could increase. The possibility exists that such a student could feel that if he or she shares an unpleasant truth about teacher performance one of two things would happen. If the majority group ignores the survey, no change will occur. If the majority group becomes angry that a member of a minority group gave such a low score, the teacher or administration might retaliate. When given this situation, the student may feel it is safest to simply lie and say...
nice things about the teacher, since the best possible outcome of sharing the truth is that it will be ignored.

The researcher decided to investigate the validity of student surveys in non-urban environments because those locations typically have racial compositions drastically different from those of urban areas. Table 4.1 has already illustrated that this pattern holds true within the non-urban municipalities located within the region served by County Schools. If students’ fears of retaliation really did impact survey behavior and Maslow’s Hierarchy was a factor, an examination of non-urban data could reveal that impact.

Based upon the data provided by County Schools, it would appear that fear of retaliation either did not exist or was not sufficient to impact the validity of the survey results. This result, however, should be viewed in light of the limitations discussed below.

**Recommendations for further studies and policy issues**

**Limitations of this study.**

The primary limitation this study faced was in geographic and temporal representation. Geographically, these data are limited by the use of only one district. A much more complete picture of the role of population density and the influence of urban areas could be found by incorporating a wider portion of the country and a wider range of non-urban areas. While this study was able to access data drawn from non-urban participants, all participants resided in the same county and could have been impacted by the close proximity of the urban area. A more accurate picture would emerge if more
districts were involved, drawn from areas closer of varying distances to urban areas, and if multiple states and regions could be included.

In terms of temporal representation, these data were limited by the fact they are only from one school year, which was also the first school year in which the non-urban areas participated in student surveys. It is possible survey administration in subsequent years could skew future data in some way. As students become accustomed to the practice of taking these surveys, they may become more comfortable, more honest, and better able to respond in a way that allows researchers to better access students’ opinions, increasing accuracy and validity. If students see the results of these surveys impacting school decisions, it could also embolden them to be more honest.

It is also possible that student survey validity diminishes the longer districts use student surveys. Students who feel their responses did not cause change may cease to take the surveys seriously, and access to honest answers may diminish. Teachers whose evaluation scores are tied to student surveys may be motivated to find ways to artificially inflate scores. As Ferguson and Danielson (2014) pointed out, any instrument that impacts performance rankings can become subject to attempts at manipulation. A more thorough picture would emerge if it were possible to test for correlation for several years in a row to verify that the correlation survives any behavioral changes the new policies cause. Multi-year data including non-urban areas of County Schools is not available, since the year after collecting these data most of the non-urban areas pulled out of the district and ceased conducting student surveys.

Another issue this study faced was the rounding of all data points. Because both student survey scores and value-added scores come from the average of many students’
individual ratings or growth, both of the numbers in the data pair should include decimal values. The more decimal places involved, the more accurate the data. However, when County Schools transmitted the data to the researcher, all scores had been rounded to the nearest integer. Given the sample size, the effects of this rounding are likely to be minimal, but greater accuracy would have been possible had the rounding not occurred and this rounding should be seen as a limitation.

The final limitation this study faced was that of instrumentation. In chapter 2, this study identified the over-reliance of previous studies on the Tripod instrument as a limiting factor in the educational community’s understanding of student survey validity. It was the researcher’s hope to find participant districts that used a wide range of student survey instruments so as to explore the hypothesis that student surveys can have validity, regardless of which survey instrument is used. For a variety of reasons, this goal could not be met. Therefore, this study also suffers from the limitation that all conclusions are not certain to be generalizable to instruments beyond the Tripod measure.

Similarly, this study also is limited by its reliance upon the Tennessee Value-Added Assessment System (TVAAS) as a measure of student growth. Value-added calculations are inherently controversial and hard to verify, inasmuch as they rely upon a peer-to-peer comparison to measure student growth. Defining the term “peer” can be difficult since each person’s life is unique. While a sufficient number of studies support the use of value-added calculations, the ambiguity of defining the term “peer” is likely to add some level of inaccuracy. There is ample evidence that the level of inaccuracy is small and probably negligible, but one cannot assume there is no impact. Any level of inaccuracy introduced by use of the TVAAS score would have impacted this study’s
findings as well, so the prudent course is to recognize the potential for limitations based on this decision.

**Recommendations for future studies.**

By exploring validity within non-urban areas, this study has taken an important step in addressing the question of whether student surveys in general are valid. However, White (2013) used a similar set of data drawn from another part of the country and found a lack of correlation. At best, this present study should serve as evidence suggestive that one of the two data sets (White’s or this study’s) could be anomalous. Future studies in this area are still needed to determine which set represents the anomalous results or whether there are unknown factors causing the two groups of data that seem similar to behave differently. Further focus on student surveys in non-urban settings is needed to settle this question.

As mentioned above, future studies would be able to contribute much to the field if they are able to explore data from a variety of student survey sources. The practice of young students evaluating teacher performance has shown much success, yet existing literature identifies myriad variables impacting survey behavior that would threaten validity. The success of student surveys in spite of these variables is a bit of a paradox and should be explored in more depth. One potential reason for this unlikely success rate could lie within the instruments used, and the majority of studies use the Tripod instrument. By conducting similar studies with other instruments, it will be possible to test the idea of student surveys as a construct and determine whether any well-designed instrument can elicit valid responses or whether Tripod’s instrument has discovered something improving its validity. Either result could benefit many other fields as
understanding the reasons for student survey validity will help any researcher wishing to survey young participants.

Based on this study’s data, the most promising next step within the study of student surveys as teacher evaluations tools seems to be an exploration into the impact of age and aptitude on survey validity. Examining this study’s data based on the permutations of participant’s grade level and location revealed that all grade groups contained one member which passed at the .05 level and one member which failed at that level. The only exception were middle school data, where both urban and non-urban data failed to produce p-values lower than .05. Consequently, future studies could contribute much to the educational community’s understanding by exploring the role of age and aptitude upon survey validity.

**Recommendations for policy issues.**

This study’s findings suggest it is possible for non-urban districts to implement student surveys as one measure of teacher evaluation, as long as it is done with caution. Existing literature has already suggested student surveys should be used as just one part of a larger system of evaluations (Bill and Melinda Gates Foundation, 2013), and this study supports that conclusion. This is especially true when one examines the inconsistencies among the various grade and location groups. Given that non-urban high schools had a p-value of .96, it becomes clear why other elements must be considered so as to avoid random chance harming a teacher’s evaluation score.

This identification of this significant fluctuation in validity among comparable grade-bands also recommends a policy of continual verification for any district implementing student surveys as part of the evaluation system. From a policy and equity
standpoint, it is justifiable to implement a policy if the overall trends show validity but some ways post hoc of slicing the data produce results slightly above the predetermined limit for validity. Data can be divided up many ways, and random chance will make some groupings less favorable than others. That alone does not mean the data is invalid. However, when data from seemingly comparable groups causes one slice to have a p-value of .00002 and the other a p-value of .96, school leaders need to be cautious and proactive in looking for potential anomalous situations warranting special consideration. If such a trend were to continue, there could be serious threats to equity of evaluations. One advisable action is to conduct annual investigations to ensure student survey scores and other established measures of quality continue to correlate, both on the whole and within predetermined data groupings. Annual verification will help protect teachers from anomalous data as well as help school leaders monitor for signs that an instrument is being artificially manipulated, either by people or circumstances.

Conclusion

The purpose of this quantitative study was to verify that the demographic, racial, socio-economic, and other differences that often occur between urban and non-urban areas are not threats to the validity of student survey results. Based on the data available to the researcher, both the urban and non-urban areas showed considerable correlation between student survey scores and value-added measures. However, when the researcher examined grade level groups, large fluctuations in validity emerged. This is a sign that there is still much to learn about what makes student surveys valid and that validity may be more a result of equal bias in all direction than of consistent accuracy. Future studies are needed in order to identify whether this fluctuation is the result of unidentified
variables or whether it is a temporary condition resulting from the unique situation of the district and time period in which these data were collected.
References


Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). *Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the*


doi:10.1016/j.learninstruc.2013.07.001


Parent Memo. (2014, November 6). Retrieved from

Why do they vary? *Research in Higher Education, 47*(2), 229-247. doi:
10.1007/sl162-005-8887-1

Porter, S. R. (2010). Do college student surveys have any validity?. Paper presented at
the 2010 meeting of the Association for Institutional Research, Chicago.

“Qualifying Urban Areas for the 2010 Census; Notice,” 77 Federal Register 59 (27
March 2012), pp. 18652-18669.


Remedios, R., & Lieberman, D. A. (2008). I liked your course because you taught me


https://bellwethereducation.org/sites/default/files/Bellwether_StudentSurvey.pdf

Sporte, S. E., Jiang, J. Y., & Luppescu, S. (2014). Teacher evaluation in practice:


Stehle, S., Spinath, B., & Kadmon, M. (2012). Measuring teaching effectiveness:


*Learning and Instruction, 28*, 1-11. doi:10.1016/j.learninstruc.2013.03.003


Appendix

Permission to Use Data
February 2, 2018

To: Trevor Gore

Re: Research Proposal

After consideration of your proposal, "A Study into the Validity of Student Survey Data in Non-Urban Areas," we have approved your request to conduct this study using data from [redacted] County Schools. You should use this letter as official notification of approval for your study.

The district-level approval being granted with this letter does not obligate any school or any person to participate in this project. Approval by the principal of any participating school is still needed before the study can begin at those schools. Also, individuals must be given the option of not participating.

Approval is contingent on your agreeing to use findings only for the purpose of the study described in the proposal.

We look forward to working with you in the completion of this project.

Sincerely,

Jeffery A. Shive