Measuring Teacher Expectations: A Generalizability Study

Wallis Malone
wmalone01@bellarmine.edu

Follow this and additional works at: https://scholarworks.bellarmine.edu/tdc

Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation
https://scholarworks.bellarmine.edu/tdc/62

This Dissertation is brought to you for free and open access by the Graduate Research at ScholarWorks@Bellarmine. It has been accepted for inclusion in Graduate Theses, Dissertations, and Capstones by an authorized administrator of ScholarWorks@Bellarmine. For more information, please contact jstemmer@bellarmine.edu, kpeers@bellarmine.edu.
Measuring Teacher Expectations: A Generalizability Study

By Wallis Malone Owens

B.A. in Sociology, May 2001, University of Kentucky
M.A. in Elementary Education, May 2005, Indiana University Southeast

A Dissertation Submitted to

The Faculty of
The Annsley Frazier Thornton School of Education of
Bellarmine University
In partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Education and Social Change

July 24, 2018
Dissertation directed by Dr. Grant Smith
Chair of Doctoral Programs and Assistant Professor of Research Design and Statistics
School of Education
Acknowledgements

Nelson Mandela said, “It always seems impossible until it’s done.” It is with sincere gratitude and appreciation that I publicly acknowledge those who traveled this journey with me. Those who helped me navigate a process that sometimes felt impossible, and who now celebrate its completion alongside me.

I want to thank my committee chair Dr. Grant Smith for his belief in me, and the value of this work. I also want to thank my committee members Dr. David Paige and Dr. Todd Whitney for their encouragement and feedback throughout this process. Also, to Dr. Robert Cooter and Dr. Kathleen Cooter for sharing their passion for change, dedication to all students, and for encouraging me to take that first step and enter this program.

I am also grateful to my family and friends for their love, support, and sacrifices. Without them, this dissertation would not have been possible. I dedicate this work to the memory of my mother, Dona Malone for making me the person I am today, and for believing I could be or do anything. To my husband Dylan for his support daily. This was definitely a team effort. Thanks for being calm when I was hysterical, excited when my enthusiasm faded, and my coach when I wanted to give up and walk away.

Finally, to my daughter Marley. Marking on my papers and sitting on my computer while I tried to type was not helpful at all, but I want you to know that you inspire me to aim higher, be better, and work harder every day. You are my biggest and best blessing and I hope that I can instill in you the same determination and self-confidence that your grandmother instilled in me.
Abstract

For the last fifty years, researchers have studied teacher expectations and their impact on student achievement. A large body of research supports the hypothesis that teachers form expectations for students, (Brophy & Good, 1970; Dusek & O’Connell, 1973; O’Connell, Dusek & Wheeler, 1974; Rist, 1970), these expectations cause teachers to behave differently, (Braun, 1976; Brophy & Good, 1970; Rothbart, Dalfen, & Barrett, 1971, Good & Nichols, 2001) and the differential treatment can affect student achievement (Brophy & Good, 1970; Jussim & Eccles, 1992; Alvidrez & Weinstein, 1999; Goldenberg, 1992, Madon, Jussim, & Eccles, 1997).

Although teacher expectancy theory is widely accepted, it is not without its critics. Hoge (1984) identified a weakness in the research concerning the reliability of the data and encouraged future research in this area. Since that time, little has been done to address this issue. Of the 38 studies on teacher expectations published in the last five years only nine of them made any mention of reliability.

The current Generalizability study examined the reliability of the data produced by the two most common measures of teacher expectations, self-report and observation. Teachers (n = 31) completed a self-report survey designed to examine teacher expectations in the spring semester of 2018. Following completion of the self-report survey, two evaluators simultaneously rated teachers on indicators of student expectations during instruction, with follow-up ratings conducted two weeks later. While self-report data was found to be reliable, teacher observations were not.
# Table of Contents

Acknowledgements .................................................................................................................. iii

Abstract .................................................................................................................................. iv

List of Tables ................................................................................................................................ vi

CHAPTER 1: INTRODUCTION ............................................................................................... 1

Problem Statement ..................................................................................................................... 3

Purpose of the Study ..................................................................................................................... 4

Research Questions .................................................................................................................... 4

Definition of Terms ..................................................................................................................... 5

Summary ..................................................................................................................................... 6

CHAPTER 2: REVIEW OF THE LITERATURE ..................................................................... 7

Expectancy Effects ....................................................................................................................... 8

Teachers Form Expectations for Students .................................................................................... 9

Different Expectations Affect Students ....................................................................................... 16

Reliability .................................................................................................................................... 21

Test-retest Reliability Coefficients ............................................................................................. 22

Classical Test Theory/Generalizability Theory ........................................................................... 22
Statistics……………………………………………………………………………………25

Summary…………………………………………………………………………………26

CHAPTER 3: METHODOLOGY…………………………………………………………….27

Site Selection……………………………………………………………………………..28

Participant Selection……………………………………………………………………29

Teachers………………………………………………………………………………….29

Raters……………………………………………………………………………………30

Rater Training……………………………………………………………………………30

Procedures………………………………………………………………………………31

Measures………………………………………………………………………………..32

Surveys…………………………………………………………………………………..32

Data Collection………………………………………………………………………….36

Data Analysis……………………………………………………………………………36

Summary………………………………………………………………………………….37

CHAPTER 4: RESULTS…………………………………………………………………38

Teacher Observation…………………………………………………………………40
MEASURING TEACHER EXPECTATIONS

Self-report .................................................................................................................. 42

Summary ..................................................................................................................... 45

CHAPTER 5: SUMMARY, RECOMMENDATIONS, AND CONCLUSION ............... 45

Summary of Findings ............................................................................................... 47

Interpretation of Findings ......................................................................................... 50

Limitations .................................................................................................................. 52

Recommendations for Future Research ................................................................. 53

Conclusion ................................................................................................................. 53

APPENDICES .......................................................................................................... 55

Matrix of Current Research, Methodology, and Reliability ................................. 55

Self-report Survey Tool ........................................................................................... 62

Teacher Observation Tool ......................................................................................... 64

REFERENCES ......................................................................................................... 70
List of Tables

1. Descriptive Statistics for Observations using Danielson Framework .........................39
2. Descriptive Statistics for Self-report Surveys ............................................................39
3. Estimated Variance Components in Fully Crossed 2-Facet Design ..........................40
4. Estimated Variance Components in Fully Crossed Single Facet Design ....................42
5. Decision Studies for Fully Crossed 2-Facet Design ...................................................44
Chapter 1
Introduction

According to the U.S. Department of Education (2014), one in seven adults cannot read. Even after years of education reform legislation including No Child Left Behind (NCLB) in 2003, and the Every Child Succeeds Act (ESSA) in 2015, literacy rates have not improved in the last ten years. The benefits of being able to read cannot be overstated, and neither can the risks to those who cannot. Illiteracy affects every aspect of life from access to healthcare, poverty, and even delinquency. According to the Department of Justice, 70% of inmates in the U.S. read at or below the 4th grade level.

Reading is not the only area where students are underperforming. The U.S. ranked 40th in math skills on the Programme for International Student Assessment (PISA) in 2015. This assessment tests the skills and knowledge of 15-year-old students all over the world. The results of the most recent assessment in 2015 represents 72 countries and over 28 million 15-year-olds.

Knowing the importance of education, and the struggle to adequately educate U.S. children, one is forced to ask why the U.S. has failed to improve. Education websites, organizations, and even the rhetoric from politicians indicate that as a country, we are sincerely concerned with improvement, but the statistics are discouraging. One education researcher making interesting claims is Dr. John Hattie. Hattie is the director of the Melbourne Education Research Institute at the University of Melbourne, Australia. His book, Visible Learning (2008), is based on his belief that schools can be improved by examining the empirical evidence, and advocating the use of those educational practices that have been found to be most effective. To determine which strategies are most effective, Hattie synthesizes the results found in meta-
analyses and ranks them by effect size. The average effect size for the more than 1,200 meta-
analysis he analyzed was 0.4 standard deviations. He suggests that the most effective methods
have effect sizes larger than the mean, and those resources, including financial resources, should
go to support those efforts that research indicates are more successful with students.

Hattie is not without his critics. In one commentary, Hattie’s research is taken to task for
many reasons, including failure to control for bias, only using published results, not having
clearly defined variables, and not evaluating the quality of the studies in his meta-analysis
(Snook, O’Neill, Clark, O’Neill & Openshaw, 2009). The authors argue that meta-analysis rarely
controls for bias that can result in unreliable conclusions. In addition, many published studies
demonstrate success of an intervention, or are funded by companies with an interest in proving
the effectiveness of one thing or another. Snook et.al., uses Hattie’s discussion of homework as
one example. The average effect size for homework was $d = .29$, which on Hattie’s scale is not
considered significant. This might lead one to believe that homework does not make a difference.
However, as the authors point out, averaging effect sizes fails to take into account the complex
nature of schools. It fails to look at variables like race, gender, age, and socioeconomic status. In
this case, the age variable was neglected. A closer look at the studies within the meta-analysis
revealed that homework had a small effect size for primary students, ($d = .15$) but had a much
larger effect on the achievement of high school students ($d = 0.64$). Knowing these findings, one
might draw a different conclusion about the effectiveness of homework, at least for older
students.

Another criticism of Hattie’s work is his failure to analyze the quality of the studies. How
large were the sample sizes of the included studies? Was the methodology sound? Was the data
found to be both valid and reliable? This critique lays part of the foundation for the current study,
whose purpose is to take one slice of Hattie’s work, his data on teacher expectations, and examine the reliability of that data.

According to Hattie, (2008) teacher expectations, with an effect size of $d = 0.43$, can have a significant impact on student achievement. Many studies support this conclusion, which suggests that teacher behavior is modified, based on the expectations teachers have for students, and that these differences affect student achievement, either positively or negatively. Those with high expectations are more likely to create classroom climates that lead to increased student engagement, varied learning opportunities, and higher self-esteem. Conversely, those with low expectations are more likely to create a climate that lacks those essential characteristics (Rosenthal & Jacobson, 1968; Brophy & Good, 1974; McKown & Weinstein, 2008; Agirdag, Van Houtte, & Van Avermaet, 2013).

This study is formed out of a desire to improve outcomes for all students, and to strengthen the empirical evidence on the effects of teacher expectations on student achievement. Considering the reliability of data impacts the value of results, the goal of this study is to examine the reliability of the data produced by the most common published measures of teacher expectations.

Problem Statement

Over 30 years ago, Hoge (1984) offered a striking critique of teacher expectancy research with a focus on two primary issues, reliability and validity. He stated, “The issue of reliability must be addressed, and it is a little surprising that relatively few uncontaminated tests of the indices’ reliability have been reported” (p. 220). A review of the research used by Hattie as well as the most current research highlights the continuation of the problem today. Reliability refers
to the ability of an instrument to have consistent results over time. In order for data to be valid, it must be reliable although the converse is not necessarily true. Of the 38 studies on teacher expectations published in the last five years, only nine include how reliability was established.

Purpose

The purpose of this generalizability study is to broaden the research on teacher expectations by examining the reliability of data produced by the most common methods used to measure teacher expectations. To accomplish this goal, the same two raters will observe approximately 30 teachers on the same occasion, with a second observation occurring approximately ten days later. Participating teachers will also complete a survey designed to measure teacher expectations twice within the same period.

Research Questions

The research questions guiding this study are:

1. How reliable are the data from the most common published measures of teacher expectations?

2. What sources of variance affect reliability across the most common published measures of teacher expectations?

3. Do observed measures and self-reported measures of teacher expectations constitute parallel forms of measurement?

This research is significant because of its ability to inform current educational practice. Teachers are not only taught that high expectations are critical to student success, they are evaluated in this area. The Danielson Framework for Teaching and Learning is used as the
model, or one of a handful of approved models in over 20 states (Danielson, 2017). This rubric contains specific domains and indicators meant to assess teacher expectations. In Domain 2B - Establishing a Culture for Learning, Danielson asserts that high expectancy classrooms create an environment where students are taught and come to believe they are capable of high achievement even when work seems hard. She further suggests that in these classrooms there is an expectation of quality work, effort, and participation and that these expectations are shared both verbally and non-verbally. Findings from this research could inform how evidence for this domain is collected and evaluated in the future.

One of the challenges to conducting research is developing methodology and using measurement tools that are both valid and reliable. One would assume that with such a large body of evidence and support for the expectancy phenomenon that an equally robust amount of information would exist concerning validity and reliability, but as stated earlier, a major void exists in this area of research. This study will provide information critical to helping future researchers select measures that demonstrate stability.

**Definition of Terms**

**Teacher Expectations**

According to Hoge (1984) teacher expectations can be defined as the inferences that teachers make about the present and future academic achievement and general classroom behavior of their students” (p. 32).
Interrater Reliability

Interrater reliability is a method used to assess how consistently different raters approximate the same phenomenon.

Generalizability Theory

Shavelson and Webb (1991) define generalizability theory also known as “G theory” as a statistical theory concerned with the dependability of measurements whose strength is its ability to estimate multiple sources of error in a single analysis.

Generalizability Study

Cronbach (1972) defines a generalizability study, also known as “G study” as a collection of data from which, “estimates can be made of the components of variance for measurements made by a certain procedure.” (p. 16)

Reliability

For the purposes of this study, reliability is defined as the consistency of measures over time.

Summary

“Evidence does not supply us with rules for action but only with hypotheses for intelligent problem solving, and for making inquiries about our ends in education.” (John Dewey, quoted in Hattie, 2008, p.147)

The complex issues surrounding education reform require researchers to conduct studies that use sound methodologies, and as Dewey suggested, provide us with evidence to help school leaders and teachers make sound educational decisions. To achieve this, researchers must
provide practitioners with valid information by using reliable measures to gather and interpret data. This work furthers those goals by informing current educational practice, education reform and policy, and future research.

**Chapter 2: Review of the Literature**

This literature review examines significant research findings and concepts related to this study. Many online databases were used to conduct this literature review including ProQuest, EBSCOhost, JSTOR, and Google Scholar. These databases were searched for academic literature, and scholarly journal articles, using search terms such as teacher expectations, Pygmalion, expectancy effects, achievement gap, achievement, equity, instructional environment, and class level effects. In an attempt to view and synthesize both the wide range of research stemming from the seminal study conducted by Rosenthal (1968) and the research most current and relevant to this study, this review includes significant studies from 1968 to the present. Research is then organized by topic.

The goal of this study is to investigate the reliability of teacher expectancy measures. In order to achieve this goal, a number of studies were examined for evidence of reliability. The data collected were then used to both rank and quantify an optimal way of measuring teacher expectations. The questions that guide this research are:

1. How reliable are the data from the most common published measures of teacher expectations?

2. What sources of variance affect reliability across the most common published measures of teacher expectations?
3. Do observed measures and self-reported measures of teacher expectations constitute parallel forms of measurement?

**Expectancy Effects**

Finding a specific and focused definition for teacher expectancy effects in the literature proved to be difficult. This is because many researchers failed to discuss the variable explicitly and instead left the reader to infer the intended definition after examining the hypothesis, research questions, and/or the instruments used in each study. The vagueness of the definition has led to a diverse use of the concept in empirical literature (Hoge, 1984). For example, in Rosenthal and Jacobsen’s (1968) seminal research study, students from Jacobson Elementary were given what they had been led to believe was an intelligence test. Teachers were falsely told that the test identified students who had an “unusual potential for intellectual growth” and would likely “bloom” within the year. When the students were re-tested at the end of the year, they found that those identified as bloomers scored significantly higher than the other students did. Their findings suggested that when teachers expected high performance out of their students, students would perform as expected. A year later, the students were re-tested and the results were still consistent. Although there is no direct discussion of the expectancy variable, based on the methods and hypothesis used in this study, it can be inferred that the intended definition of teacher expectancy was future performance based on current information. West and Anderson (1976) used a similar definition. They defined teacher expectations as, “assumed teachers’ attitudes about students which are a function of some information supplied by the investigator” (p. 616). Brophy and Good’s (1974) definition reflects the change to a more naturalistic approach to studying teacher expectations. In their definition, they remove the need for the teacher to be supplied with incorrect information and suggest that the inferences teachers make could be based
on other factors. They define teacher expectations as, “inferences that teachers make about the present and future academic achievement and general classroom behavior of their students” (p.32). Hoge (1984) argued that these definitions were a good start, but points out that they may be too vague. He asserts that more work is needed to determine the focus of the inferences, beliefs, or attitudes in question. The lack of focus in the current research seems to support his criticism. Using the same or a similar definition, some teacher expectancy research focuses on specific academic areas while others examine overall IQ (Fischbach, Baudson, Preckel, Martin, & Brunner, 2013; Baker, Tichovolsky, Kupersmidt, Voegler-Lee, & Arnold, 2015). Some examine student behavior or self-esteem. Some studies look at teacher beliefs through self-report while others attempt to identify teacher beliefs through observation. Some research focuses on individual student and teacher relationships while others look at whole group interactions.

Although questions remain concerning the definition of teacher expectations, a large body of empirical research has been developed to support the hypothesis that teachers form expectations for students (Brophy & Good, 1970; Dusek & O’Connell, 1973; O’Connell, Dusek & Wheeler, 1974; Rist, 1970), that the formation of these expectations leads teachers to treat students differently, (Braun, 1976; Brophy & Good, 1970; Rothbart, Dalfen, & Barrett, 1971, Good & Nichols, 2001) and these differences in treatment can affect student achievement (Brophy & Good, 1970; Jussim & Eccles, 1992; Alvidrez & Weinstein, 1999; Goldenberg, 1992, Madon, Jussim, & Eccles, 1997). The next sections will examine the relevant studies that support this theory of action.

**Teachers Form Expectations for Students**

Research has shown that teachers form expectations of students based on a variety of reasons. Teacher expectations can be formed based on information gathered about the student
MEASURING TEACHER EXPECTATIONS

(Brophy, 1983; Brophy and Good, 1970; Clifton, 1981; Cooper & Tom, 1984; Raudenbush, 1984; and Wineburg, 1987), based on student attractiveness (Dusek and Joseph, 1985), race (Wigfield, Galper, Denton, & Seefeldt, 1999), gender (Page and Rosenthal, 1990), and previous information or labels placed on the student (Harris, Milich, Corbitt, Hoover, & Brady, 1992).

The central phenomenon of self-fulfilling prophecy on the surface is very logical. Teachers with high expectations will inspire higher learning by encouraging students and using multiple teaching and learning strategies. While teachers with low expectations might discourage learning by their teaching methods, interactions, and choices of strategies (Goldenberg, 1992). The relationship that seems simple is actually much more complicated. For example, although some variance in teacher expectations is due to bias, some variance in teacher expectancy can be explained by actual student performance (Brophy, 1983; Tenenbaum & Ruck, 2007). Research suggests that teachers have the ability to predict performance on a variety of tests, across subject areas, and at multiple grade levels with some consistency (Hoge & Coladarci, 1989; Hoge and Butcher, 1984; Svanum & Bringle, 1982; Egan & Archer, 1985; Stoner & Purcell, 1985; Hopkins, George, & Williams, 1985). If student achievement is the basis for the formation of teacher expectations, then might differentiated teacher behavior be warranted? Researchers have attempted to answer this question by investigating how teacher expectations are formed.

Some research suggest that teachers can form expectations for students based on students’ race and ethnicity (Wigfield, Galper, Denton, & Seefeldt, 1999). Research conducted by Wigfield and colleagues examined first grade teachers’ beliefs about former Head Start and non-Head Start children. The sample for this study consisted of 83 first grade students who were previously enrolled in Head Start and 55 students who were not. The students were given tests to determine their academic proficiency including the Peabody Picture Vocabulary Test, which is a
test of receptive language, and the Woodcock-Johnson Achievement Test, which was used to assess each student’s progress over time in the academic areas of reading and math. To assess teacher beliefs about the students’ academic capabilities, efforts, and values the researchers adapted an instrument they had used in previous research (Wigfield & Harold, 1992). The instrument asked teachers to evaluate each student’s current academic ability, compare them to peers, assess the child’s level of enjoyment for each activity, and assess the child’s effort. An analysis of these measures suggested that differences in teachers’ attitudes were based on race rather than Head Start involvement (Wigfield, Galper, Denton, & Seefeldt, 1999). They conclude that teachers rated white students higher than they rated black students in the areas of ability, making friends, and expectations for how students would achieve in subsequent grades regardless of their involvement in a Head Start program.

Of the three measurement items used, the study only reports reliability for one, the Woodcock-Johnson Achievement Test ($r = .94$ for kindergarten to twelfth grade). Although not reported in this research article, reliability data was available from the Community-University Partnership for the Study of Children, Youth, and Families (2011) for the Peabody Picture Vocabulary Test. They reported high reliability with a coefficient of $r = .92$. The third measure, which was included to report teacher belief or expectation, did not include a reliability coefficient. The article did however, state that the measurement had been used previously but neither cited article reported reliability of the instrument nor of the previous instruments that had been used with different populations than in the current study.

Rubie-Davis, Hattie, and Hamilton (2006), also examined the effect of race and ethnicity on teacher expectations. They conducted a study to examine the relationship between student achievement and varying expectations for students in four different ethnic groups (Maori, Pacific
Island, Asian and New Zealand European). The methods used for this study were similar to the previous study. Teachers were asked to rate student potential and these ratings were compared to actual student performance. The researchers provided teachers with definitions of on level, above level, and below level performance and asked them to predict where each student might fall. The academic measure used was a running record. The authors make no mention of reliability of the expectancy measure used.

The results of the study suggest that teachers’ expectations of Maori students were much lower than their actual performance while conversely, expectations for Pacific Island, Asian, and New Zealand European students were higher than their actual performance indicated. It was a common belief amongst teachers that Maori families did not value education. Researchers found that although Maori students scored similarly on the beginning of the year pre-test assessments, their gains were the least on the end of year assessment. Because performance and social class are similar for Maori and Pacific Island students, researchers concluded that these results were due to teachers’ expectations.

Tenenbaum and Ruck (2007) also found varied teacher expectations due to race. They conducted four separate meta-analysis to examine expectations, referrals (discipline, special education, and gifted), and speech patterns toward different minority groups as compared with white students. The measurement tools in each study were examined. Some researchers used standardized measures such as the Connor’s Teacher Rating Scale (Epstein, March, Conner, & Jackson, 1998) or the Teacher-Rating Scale (Sbarra & Pianta, 2001) while others, 26 of 32 used self-created measures. Self-created measurement tools included items that asked teachers to rate students’ talent, performance, and exertion of effort (Jussim & Eccles, 1995), academic ability and seriousness (Chang & Sue, 2003), and ability to verbalize and predict grade point average.
(Mazer, 1971). They found significant evidence that teachers held more positive expectations, wrote more positive referrals, and used more positive or neutral speech with European American students than African American or Latino students. Teachers held the highest expectations for Asian American students. The authors fail to report actual reliability coefficients but state that all studies “assessed and achieved appropriate levels of reliability” (p. 256).

Teachers can also develop expectations based on gender (Page & Rosenthal, 1990; Tiedemann, 2000; McKown & Weinstein, 2002; Blömeke, Dunekacke & Jenßen, 2017). Many studies on teacher expectations and gender focused on math instruction. Page and Rosenthal’s (1990) study had a unique design in which they asked teachers to design math lessons for students of different races and genders. They examined the lessons and found that teachers designed math lessons for Asian male students that were more rigorous compared to those designed for white female students.

Tiedemann (2000), also studied gender as it related to achievement in math and like many researchers used a self-created questionnaire. Fifty-two teachers participated in the study. They were asked to identify six high, medium, and low performing math students. One boy and one girl in each category and then complete a questionnaire based on their expectations of performance for these students. They were asked to estimate the child's competence, current performance, potential for improvement and deterioration, and predict future performance. There was no mention of reliability for these data.

McKown and Weinstein (2002) studied student’s susceptibility to teacher expectancy based on gender and race. In this study, teachers identified their expectations for student by ranking them from highest to lowest. Students’ actual achievement was measured using a standardized test and these scores were compared to teacher ranking. The results of this study
suggest that stigmatized groups were more likely to be influenced or harmed by teachers’ underestimation of ability. Although the method of asking teachers to rank students based on achievement is quite common in studies on teacher expectations, the added step of creating over and underestimates and using those to assess teacher expectations was unique, however, there was no mention of reliability of the measurement.

A study of pre-service preschool teachers, Blömeke, Dunekacke, and Jenßen, (2017) found that gender bias in mathematics may develop during teacher preparation programs. They found that participants at the beginning of the program actually demonstrated less bias than those near completion. As a means of assessing expectations, participants completed a questionnaire. The questionnaire used a Likert type scale and asked participants to respond to various statements including whether or not girls need more support and are less competent than are boys in math. It also contained positive or neutral statements that suggested boys and girls had the same ability to solve math problems. Although the report contained a very detailed explanation of the items and included reliability coefficients, closer examination revealed that the coefficients presented were actually for the sections of the instrument that measured knowledge of content and pedagogy and not the sections on teacher expectations.

Clifford and Walster (1973) studied the effect that physical attractiveness might have on teacher expectations. Four hundred and four elementary teachers were given folders containing a picture of an attractive or unattractive child and a grade report that summarized the child as a B student. Teachers were told that the purpose of the questionnaire was to assess which information needed to be included in a child’s permanent file. After reviewing information, teachers completed the questionnaire. Teachers estimated student IQ, social status, parent’s attitude toward school, and potential for future academic success. There was no mention of
reliability of the measurement. The study found that attractive students were rated as more intelligent, having more potential to learn, and having parents who valued education, more than their less attractive peers.

Student labels also have been found to impact teacher expectations (Stinnett, Crawford, Gillespie, Cruce, & Langford, 2001; Darley and Gross, 1983). In both studies, similar methods were used to assess how labeling a student might affect teacher expectations. Stinnett et al. (2001) tested this hypothesis by providing 144 pre-service teachers descriptions of elementary school children. All of the descriptions were the same but one of the following was used within each: no label, Attention Deficit Hyperactivity Disorder (ADHD), special education, or on Ritalin. When the student had a label, pre-service teachers believed the student had more difficulty paying attention. The teacher rating scale was tested for reliability and found to have a good internal reliability ($\alpha = .79$).

In addition to vignettes and labels used in other studies, Darley and Gross (1983) added a section in which the student answered performance-related questions. The exact same video was shown for each condition, with one exception. The background was manipulated to reflect a middle or lower-income background. Like Stinnett, they also found that the label elicited a bias that affected teacher expectations. In this particular study, participants whose video background suggested low socio-economic background rated the child’s potential academic achievement as lower than those whose video suggested a middle-class background. There was no mention of reliability of the measurement used in this study.

As cited above, many studies support the theory that teachers develop expectations for students. However, is merely having different expectations enough to affect student achievement
or outcome? The research suggests that when teachers act upon these expectations students either benefit or suffer.

**Different Expectations Affect Students**

"It is not what we say or feel that makes us what we are. It is what we do, or fail to do" (Thompson & Austen, 1996). In the same vein, many researchers have found that it is not only what teachers believe about a student’s ability, but also how they act upon that belief that makes the difference (Brophy & Good, 1978; Babad, 1998; Goldenberg, 1992). The next section will explore the research on teacher behaviors and their effect on student achievement.

Brophy and Good (1978) conducted a study to learn how teachers demonstrate their expectations to students in a way that actually affects student achievement. In this study, they were only concerned with individual student and teacher interactions. They first asked teachers to rank students from high to low without giving teachers any specific criteria. From these ratings, children were selected from each end of the spectrum to observe. Based on the observations Brophy and Good were able to identify 17 behaviors that they believed represented a teacher’s level of expectations. These behaviors include variability of assignments, providing clues or paraphrasing, amount of praise, and preferential seating for students for whom they hold high expectations. The study reported that interrater reliability was achieved.

Rubie (2008) also examined teacher behavior based on expectations and described teacher behaviors or interactions as either proximal or distal. Proximal interactions were defined as verbal and nonverbal interactions teachers have with students while distal interactions were those interactions that students have little control over like lesson planning and instructional environment. Her findings suggest that teachers may structure the learning environment
differently and plan lessons that are less rigorous for students for whom they hold lower expectations. While research has shown that engaging tasks result in improved progress, low expectation students are more often given assignments that are structured and repetitive (Gamoran, 1992). In addition, research has shown that low achieving students are not exposed to higher level questioning or chances to extend thinking; this all but ensures that these students will not develop these skills (Zohar, Degani, & Vaaknin, 2001).

The effects of differing teacher expectations are also seen within teacher grading procedures and in their assessment of student effort. Jussim and Eccles (1992) found teachers underestimate the amount of effort it takes certain students to complete assignments. Although struggling students spend more time completing homework assignments than their higher achieving peers they were not recognized for this effort. They found that even if the work that was turned in was mediocre, students for whom the teacher held higher expectations would be given a higher grade. Although this might be seen as potentially more harmful to high expectation students, the opportunities extended to students based on inflated grades were thought to outnumber the risks. In this study, both students and teachers completed questionnaires. The article gives three sources to support the reliability of the measures but does not include reliability coefficients. Two of the sources do explain how the student questions and questionnaire were created and the third source could not be located.

Teacher expectations can also lead to differences in the amount and quality of emotional support, classroom support, and pressure placed on students by high and low expectation teachers (Babad, 1998). The study results show that teachers provided more emotional support to students for whom they held higher expectations, and that students not only noticed, they were resentful of the differential treatment (Babad, 1995; Babad, Babad & Rosenthal, 2003).
The majority of the research on teacher expectations and student achievement is focused on individual student-teacher interactions. Although Brophy (1983) proposed that, “differential teacher treatment of intact groups and classes may well be a much more widespread and powerful mediator of self-fulfilling prophecy effects on student achievement than differential teacher treatment of individual students within the same group,” (p. 312). However, only a few studies exist on the class or group level effects of teacher expectations.

Research conducted in New Zealand schools supported Brophy’s idea that a class or group effect might exist and have a greater effect on student achievement (McKown & Weinstein, 2002 and Weinstein et al., 1982). Weinstein and colleagues classified teachers as either high or low differentiating teachers. It is important to note that although both sets of teachers held differing expectations for their students, they also appeared to have a different teaching philosophy. High differentiating teachers were of the belief that students need separate learning tasks. This belief led them to create classrooms where students were grouped by ability, favoritism of students was obvious, there was a focus on extrinsic rewards, opportunities for student recognition was limited, they believed that intelligence was fixed, they limited student choice, and had differing levels of communication with parents. Low differential teachers behaved in an opposite manner. They grouped students in mixed ability groups, encouraged intrinsic motivation, and developed relationships with parents, etc. This was significant because although both types of teachers held different expectations for their students, low differential teachers attempted to treat all students the same, which by year-end resulted in a decrease in the achievement gap between high and low performing students.

Two studies published within the last five years have built upon this work by also studying distal interactions and class level effects. Sedova and Salamounova (2016) explored
whether students categorized as weak by their teachers could effectively participate in classroom
dialogue during literacy instruction. Researchers chose to focus on two low expectancy students
within a classroom and analyze their classroom interaction on videotaped lessons. Results
suggest that a students’ choice of how they participate can be influenced by teacher behavior and
that this effect can be mediated when teachers focus on the goals of their curriculum as opposed
to student characteristics. In this study, the two teachers’ expectations were measured by self-
report. Teachers were asked to rank their students by what they believed would be the likelihood
of each to attend college. Next, they videotaped their classrooms and analyzed the interactions
between the teachers and students who they ranked the lowest. There is no mention of reliability.

Tedeschi (2016) examined how race and gender might affect a teacher’s decisions to
write a discipline referral. Using an online questionnaire, participants were asked to complete
both demographic information and respond to a scenario describing a behavior and how they
would respond to that specific behavior. Results suggest that teachers’ decisions were not
impacted by the race of the teacher or race and gender of the student. However, teacher gender
did have an impact on whether or not teachers made a decision to write an office referral.

Another study of interest by Donahue, Weinstein, Cowan and Cowan (2000) introduced
the idea of stratification of teachers’ perceptions of student competence. Teachers rated
student’s competence and this score was used to create a classroom dispersion score. This score
was used to describe how teachers perceived the levels of intelligence within their classrooms.
The study results suggested that greater dispersion meant the teacher perceived less child
competence and these teachers’ perceptions remained stable throughout the school year.
Comparisons with measures of student achievement found that there was no evidence that these
teachers’ perceptions were more accurate. This study used a combination of a researcher created
survey and results of the Child Adaptive Behavior Inventory (CABI) to assess teacher expectations. The article made no mention of reliability of either instrument.

In the 1970’s research moved from the experimental approach where researchers attempted to manipulate teacher expectations toward the use of naturalistic studies (Hoge, 1984). Researchers began to question whether or not the time of year played a role in the development of teacher expectations. Two studies draw an interesting conclusion about teacher expectations and their development over time, and inform this study (Cooper, 1979; Good, Cooper, and Blakey, 1980). In these studies, data was collected multiple times during the year, and although data supported the conclusion that teacher behavior differed for groups of students at various times, researchers believed that beginning of the year teacher behavior was an attempt to teach routines and procedures to students and less likely related to differing expectations. They concluded that collecting data later in the year minimized these issues.

Although it is widely accepted that teacher expectations affect student achievement these results are not strongly linked among all groups. Jussim, Eccles, and Madon (1996) found that teacher expectations and student achievement more closely matched for African American students than European American students. McKown and Weinstein (2002) likewise found that African American elementary school students were more susceptible to negative teacher expectations than their white peers were. Both studies suggest African American students might be more vulnerable to teacher expectancy effects than are other student groups.

The amount of research on teacher expectations and education topics in general can be overwhelming. A Google search on the terms “teacher expectations” yields over 78 million potential results. A narrower search of research articles using Google Scholar yields 1.5 million articles. Realizing the overwhelming nature of education research and observing how few best
practices were actually making it to the school level, Hattie (2009) synthesizes over 800 meta-analyses on various educational topics and ranks them based on effect size. Hattie examined eight meta-analyses containing over 600 studies on the topic of teacher expectations for which he reports a moderate effect size of $d = 0.43$. A closer review of these meta-analysis and a sampling of the studies used to create these analyses is concerning. Although most studies used one of the following methods: a questionnaire, an observation, or a combination of both, few studies mentioned any test of reliability for the instrument they used. The absence of reliability coefficients questions the validity of the study and raises doubt as to the accuracy of the results.

The studies that Hattie used span from 1978 to 2007. A more recent examination of the literature exposes the same concern. For example, over the last five years, 38 studies have been published on teacher expectations and how they affect student achievement. Of those studies, 24 used a survey, three rank students according to specific criteria, two compared ratings with actual data and looked for differences between teacher rating and actual test data, three interviewed teachers, five combined surveys with either interviews or observations and one used teacher recommendations. Only nine studies made any mention of reliability and included the reliability coefficient.

**Reliability**

The goal of reliability studies is to approximate the stability of scores across repeated measures (Webb, Shavelson, & Haertel, 2006; Gravetter & Forzano, 2012). Reliability coefficients are used as a means of quantifying multiple measurements using a scale from 0-1.0 (Webb, Shavelson, & Haertel, 2006). The higher the coefficient the more reliable the score. A score of .80 is considered sufficiently reliable for decision-making purposes; however, if the decisions have major significance, a higher score is desirable.
Test-Retest Reliability Coefficients

Reliability study designs are specific to the type of instrument being tested. For example, test-retest reliability is a study design that requires a minimum of two scores. Each participant is to take the same assessment on two separate occasions. The correlation of these scores provides a test-retest reliability coefficient (Webb, Shavelson, & Haertel, 2006). Many conditions can contribute to variability between scores in this study design. Cronbach, Glesser, Nanda, and Rajaratnam (1972) refer to these conditions as either lasting or temporary. Lasting conditions, as their name suggest, contribute to consistency in scores over time. Some examples of these attributes are the test-taker’s ability and testwiseness. The more times the participant takes the test the better they might become at taking the test. Temporary conditions provide inconsistency in scores across conditions. Examples of temporary conditions might include noises that disrupt the testing environment, temperature changes, and random guessing.

Another reliability design of interest is inter-rater reliability. Inter-rater reliability is used to determine how different raters assess the same phenomenon (Webb, Shavelson, & Haertel, 2006). This is important when using humans to collect data as human tend to be less consistent.

Classical Test Theory/Generalizability Theory

The theoretical frameworks providing guidance to this study are Classical Test Theory (CTT) and Generalizability Theory (G theory). CTT is commonly used in social and behavioral sciences (Webb, Shavelson, & Haertel, 2006). CTT is based on the assumption that researchers are only able to obtain an observed score, which is defined as the true score plus error. If little variability exists between a person’s scores over varied conditions, then the correlation would be quite high. Inversely, large amounts of variability between true and observed scores across
conditions would indicate a lack of correspondence between these scores due to the presence of large amounts of error. Although CTT provides a way to decompose an observed score into a true score with an error term, it cannot provide details about the source of variability (Feldt & Brennan, 1989).

Because the goal of this study is to identify sources of variability, G theory is used. Shavelson and Webb (1991) define G theory as, “the statistical theory about the dependability of behavioral measurement” (p. 1). They further describe dependability as the accuracy to which one is able to generalize a person’s observed score over a number of measurement conditions.

CTT and analysis of variance (ANOVA) are considered the foundation for G-theory (Brennen, 2000). The combination of ANOVA and CTT creates the ability to partition sources of variance, which was previously unavailable in CTT. Messick (1989) offers two different perspectives on G-theory. He suggests that generalizability can be interpreted as either reliability or transfer. Reliability is defined as “the consistency of performance across the tasks, occasions, and raters and transfer as the kinds of tasks that performance on an assessment might be predictive of” (Messick, 1996, p. 250). While CTT is limited to the examination of the primary variable compared to error variance, G theory has the capability of examining various sources of error within the method of measurement.

Brennan (2001) suggests that the theoretical framework is the most important and unique aspect underlying G theory. The framework consists of the universe of admissible observations, which refers to all of the conditions that the researcher would find acceptable in a measurement situation. These conditions, tests, raters, and occasions are called facets of measurement and all their possible combinations comprise the universe of admissible observations (Shavelson & Webb, 1981).
Facets can be defined as either fixed or random (Shavelson & Webb, 1991). Facets are considered random when a small sample of the universe and is drawn randomly. For example, if several items were chosen from an assessment and the researcher would be willing to exchange them for any other items on that assessment the facet could be considered random. A fixed facet is comparable to a fixed factor in ANOVA. A facet is considered fixed if any of the following are satisfied: the researcher does not wish to generalize beyond the group, the population is so small that all facet conditions have been considered, or generalizing beyond the condition would be unreasonable (Shavelson & Webb, 1981).

Another area of importance concerning the G-theory is creating a “complete” and “balanced” design. Balanced and complete studies ensure that all interactions have been considered and all facets have the same number of items. This design minimizes overall error but can be challenging because it has the potential to decrease the data size in order to meet the requirements.

The current study’s design was broken into two G studies that allowed for calculations of variance within different conditions. The results can be used to investigate the reliability of outcomes under varied measurement scenarios and to construct more efficient measurement procedures and improve decisions concerning measurement (Webb, Shavelson, & Haertel, 2006). The first was a 2-facet fully crossed study where facets were raters and occasion and the unit of measurement was teachers. A 2-facet fully crossed design has six other sources of variability. A facet is considered crossed when every level of the facet is observed with every other facet in a data set. This condition was satisfied in this portion of the study, because the same raters were used for each observation on the same occasions. The second piece of the
design used a single facet g-study where the facet was time/occasion and the unit of measurement was again teachers.

Using a fully crossed design has both advantages and disadvantages. The ability to assess every level of a facet against every other level of a facet makes for a robust study. The disadvantage is the difficulty of creating the conditions for this type of study within a large sample size. It can prove to be time consuming and expensive (Brennan, 2001).

**Statistics**

Brennan (2010) describes a G-study as a way to decompose an observed score into components and estimate variability for each. The following section is described in Brennan (2010) and Shavelson and Webb (1991) and explains the statistical process used to decompose scores using G-theory. The current G-study uses a repeated measures factorial ANOVA with person, rater, and occasion as main effects, person x rater, person by occasion, and rater by occasion as interaction effects and person x rater x occasion as the residual error term. It can be expressed in a linear model:

\[ X_{pio} = \mu + \nu_p + \nu_i + \nu_o + \nu_{pi} + \nu_{po} + \nu_{io} + \nu_{pio} \]

Where \( \mu \) represents the grand mean and \( \nu \) is the effect of the component. It can also be expressed in a linear model as:

\[ \sigma^2 (X_{pio}) = \sigma^2_p + \sigma^2_i + \sigma^2_o + \sigma^2_{pi} + \sigma^2_{po} + \sigma^2_{io} + \sigma^2_{pio} \]
D – Studies

The D-study variance components are calculated using the components from the G-study with a selected sample size for each component. It can be expressed using the following linear model:

\[ X_{pl0} = \mu + \nu_p + \nu_I + \nu_0 + \nu_{pl} + \nu_{p0} + \nu_{I0} + \nu_{pI0} \]

Summary

The importance of teacher expectations and their role in student achievement is well documented. Research supports the theory that teachers form expectations for students (Brophy & Good, 1970; Dusek & O’Connell, 1973; O’Connell, Dusek & Wheeler, 1974; Rist, 1970,). the formation of these expectations leads teachers to treat students differently, (Braun, 1976; Brophy & Good, 1970; Rothbart, Dalfen, & Barrett, 1971, Good & Nichols, 2001) and these differences in treatment can affect student achievement (Brophy & Good, 1970; Jussim & Eccles, 1992; Alvidrez & Weinstein, 1999; Goldenberg, 1992, Madon, Jussim, & Eccles, 1997).

Although this body of work that spans decades seems to provide a sufficient amount of support for this phenomenon, the reliability of the results is of concern. Hattie’s (2009) and Hoge’s (1983) work acts as the springboard for this thinking. Hattie quantifies the impact or effect teacher expectations have on student achievement. He used eight meta-analyses that encompassed over 600 studies on the topic. A random review of the studies used in these meta-analyses showed a lack of reliability data. A review of work over the last five years indicates that although this is still a heavily researched area reliability is still an issue. It is the goal of this study to add to, and strengthen the work in this area by evaluating the reliability of the data of the most common methods used to measure teacher expectations. The underlying theoretical
framework that informs this research and study design is Generalizability Theory. This theory allows one to partition error within a score.

Chapter 3 will outline the process that will be used to obtain empirical evidence. It will describe the methods of analysis and attempt to provide clarity around the research questions. The methodology was selected and constructed based on the existing published literature and theoretical framework.

Chapter 3:
Methodology

This study applies generalizability theory to identify sources of variance in the data produced by the most common methods used to evaluate teacher expectations. A review of the most current research indicates that teacher expectancy studies rely heavily on two areas, self-report and observation. Two coders (raters) were invited to observe 31 teachers on two separate occasions (spring, 2018) to evaluate teacher expectations using domains 2B and 3A of the Danielson framework. In addition, the same teachers were invited to complete a survey used to assess teacher expectations.

Observation data collected by raters will be analyzed using generalizability theory in a two-facet fully crossed design (Shavelson and Webb, 1991) and survey data will be analyzed using a single facet design. Generalizability studies will be used because of their usefulness in both identifying various sources of error and in the design of more efficient procedures (Brennan, 1992; Shavelson & Webb, 1991). The sample for this study is mainly a convenience sample, however, based on the research, a number of criteria will be observed: 1) only elementary schools with higher populations of minority students and students eligible for free or reduced prices lunches were invited to participate, and 2) the invited raters have educational
background and some experience working with teachers in a supervisory or coaching role.
Neither raters nor teachers were compensated for their time.

**Site Selection**

This study was conducted in a large urban school system located in the south. It consists of about 172 schools, 6,600 teachers and over 100,000 students. Eighty percent of the children who reside in the city attend public schools according to data provided by the district. The district has 91 elementary schools. Several considerations were made concerning site selection. Using G-power, it was determined that a sample of 30 maintains alpha at .05, beta at .20 and power adequate to detect medium effect sizes \((d = .50)\). Most elementary schools in the District have less than 30 classroom teachers on staff so a minimum of two school sites would be needed. A combination of research and convenience sampling informed the choice of sites. Although this school district uses a unique student assignment plan whose aim is to help balance both economic and ethnic diversity within the schools, many schools still have large populations of at-risk students defined as non-white students and students who qualify for free or reduced-price lunch. Research suggests that younger students as well as at-risk students are more likely to be impacted by teacher expectations, therefore the ability to accurately measure this phenomenon is crucial (McKown & Weinstein, 2002; Hinnant, O'Brien, Ghazarian, 2009). These findings influenced the selection of sites. The decision was made to use only elementary schools with at least 50% of students eligible to receive free and reduced lunch prices and at least 50% minority. Although the studies cited above found teacher expectations to be impactful to minority students and students living in poverty, neither specifically sought participants in these categories. Instead, each study invited volunteers to participate, recorded the demographic data of students, and then analyzed the effect controlling for demographics. Since participants in this study were
actually teachers, not students, an attempt was made to increase the probability of each teacher’s classroom having students who fit the desired demographics by choosing schools with higher amounts of each. From there convenience sampling was used. Because it would not be efficient to choose schools where only one or two teachers might participate, principals were contacted first via email and asked if the school was willing to participate. It was anticipated that having principals who wished to participate would ensure a larger number of teachers would volunteer. Two schools fitting the criteria agreed to participate in the study.

**Participant Selection**

**Teachers**

Once the schools were identified, teachers listened to a brief presentation about the study goals and requirements during a faculty meeting. Consent forms containing the same information were left at the school and anyone wishing to participate was asked to complete one and turn into the office if interested by the next day. Once these forms were returned, participants received an email containing a link to the online survey. Thirty-two teachers agreed to participate. Each participant was sent the link to the survey to complete. A reminder was sent to participants who failed to complete the survey after three days. After surveys were completed, each participant was emailed information on when the observation team would come to complete the first set of walk-throughs. Principals also sent the master schedule ahead of time to help raters determine the order of classroom visits. The same process was completed two weeks later. One teacher’s data was removed because the teacher was absent during the first round of observations (n=31).
Of the final participants, 28 of the 31 were white. Thirty of the 31 participants were female and they were between the ages of 31 and 46. Twenty-six of the teachers had between 6 and 17 years of teaching experience.

**Raters**

The two raters were also carefully chosen. Each rater had an elementary background and had worked in a number of educational settings. One currently works as a resource teacher and has more than 27 years of experience in public education. Although, not an official teacher evaluator, persons in this role are expected to provide staff training and assist teachers by modeling effective instructional and classroom management strategies. The other rater is also a highly qualified professional educator who has her Master of Science in Elementary Education and has worked for several years in both public and private schools in Kentucky and Florida.

**Rater Training**

The protocol for evaluators in the district studied has recently changed. Previously, evaluators in this district were required to complete 40 hours of initial training using an online system called Teachscape. This training provides an explanation of each domain and requires participants to complete modules that outline issues around observations, including recognizing and eliminating bias and collecting evidence. Each new evaluator must also complete a 12-hour, face-to-face training where the legalities of teacher evaluation are taught. This year, a new calibration procedure was introduced, and was designed to ensure that evaluators had a common lens through which teachers would be evaluated, a cohesive approach to applying the Danielson Framework, and that evaluators maintained consistent and high expectations. A similar approach was used with the raters for this study. Each rater was oriented on the use of the Google form. Once familiar with, and proficient on the items the researcher conducted calibration practice with
both observers. In this practice, the group completed a walk-through and then had a dialogue about the standards and the evidence collected. Once observers agreed on ratings and all questions were answered observers were considered trained. During actual observations, raters were asked not to compare notes or debrief after classroom visits. However, they were permitted to talk about what they saw once all evaluations were completed and submitted for the day.

Procedures

As discussed above G-power is a free program that can be used to calculate power for a number of statistical tests (Faul, Erdfelder, Buchner, & Lang, 2009). Using this software it was determined that a minimum of 30 participants were needed to complete the survey and observation. Informed consent was obtained from each participant. To maintain participant anonymity, names were dissociated during the data collection and recording processes.

Participants were sent a link to complete the online survey which was housed in a Google Form. Only participants with the link could access the form and only the researcher could see answers. Each participant was allowed to take the survey one time within each window. The online survey collected demographic information and the participants answers to the expectations questions that came from the literature (Cowan, Cowan, Heming, & Miller, 1995). The Google forms application was used to administer the survey and collect observation data. The survey results were then gathered over a two-week period during the spring semester of 2018. Reminders to the subjects were sent out after the first three days to increase participation.

Once the survey data was collected phase two of data collection began, observations. Two raters went into each classroom and observed instruction for a minimum of 10 minutes.
During this time, they answered two questions meant to assess teacher expectations using the Danielson framework. The walkthrough form also allowed raters to collect evidence to support their rating.

**Measures**

For this study, the Danielson Framework for Teaching and Learning along with questions adapted from the Child and Adolescent Behavior Inventory (CABI) were used to evaluate teacher expectations using trained raters and teacher surveys on two occasions. The Danielson Framework is designed to identify teacher behaviors that have been empirically studied and found to impact student achievement and the questions found on the CABI align with similar survey questions used in research to evaluate teacher expectations of students. Data was collected in the spring, which is well into the school year.

Much attention was paid to the selection of instrumentation. Although commonalities could be found amongst methodology, i.e. surveys and observations, no consistent protocol exists within the literature concerning specific instruments. Survey questions and observation tools varied across studies. Due to this obvious gap in the literature, it was necessary to develop or modify an instrument.

**Surveys**

A thorough review of the current literature revealed one commonality among survey instruments. Most of the surveys were researcher created and asked teachers to judge or rate student future achievement. In some studies, participants were given academic criteria on which to base the judgment (Ready & Chu, 2015; Garrett, Rubie-Davies, Alansari, Peterson, & Flint, 2015) while in other cases they were not (Riley, Foster, & Serpell, 2015). Many items must be considered when developing a survey instrument of this type. The most important being social
desirability. Social desirability describes the tendency for participants to answer questions in a way that presents a favorable image of themselves even if those answers do not actually represent their true feelings (Johnson & Fendrich, 2005). All teachers are expected to have high expectations, so asking participants about expectations would likely lead to some false answers. The second issue was that none of the instruments in the literature asked questions that would reflect the teachers expectations for his or her whole class and instead focused on individual students. The current instrument is borrowed from the work of Donohue, Weinstein, Cowan, and Cowan (2000). In their study, teachers rated individual children using the CABI. The CABI is an instrument designed to assess teachers’ perceptions of student ability. Two scales were considered, one for intelligence and one for social abilities. High scores on the intelligence portion of the instrument indicated that the teacher believed that the student was intelligent while high scores on the social abilities section indicated that the teacher felt the child struggled socially. They used this data to create two scores, one describing the degree to which a teacher saw individual children in a class as varying in intelligence and social difficulties. Of all the surveys present in the literature this seemed to be the most promising measure of class expectations. In the present survey, participants will be asked to answer the same questions used in the study above, but instead of assessing individual students, participants will be asked to identify the number of students in their class that fit into each category. For example, participants will be asked, “How many students in your class fit the following description; this child catches on quickly, e.g., is quick at learning new games.” As in the Donahue et.al. study, high scores on the intelligence portion of the instrument will indicate that the teacher believes that many students in their class are intelligent.
The study will also include teacher observations. Observation instruments varied just as much as survey instruments within the literature but all included some form of either a follow-up survey or asked teachers to rate students prior to the observation. Some observations were of students while others focused on the teacher’s behavior (Metzger, 2016; Sedova, & Salamounova, 2016). Other observations focused on how the classroom environment or student experience was impacted by information gathered in the survey. For example, Sedova, & Salamounova (2016) explored whether students viewed as weak by their teachers could productively participate in classroom discourse in literacy lessons.

Brophy (1983) identified 17 behaviors teachers may exhibit that demonstrate low expectations and further impede learning for low students from the literature. This list included characteristics like allowing less wait time, less praise, seating low students further from the teacher etc. Rubie-Davies’ work focused more on class level effects and suggested that it is possible to identify class level differences in teacher expectations, and that these differences impact student achievement. Pellegrini and Blatchford (2000) argued that students spend more time interacting with teachers as part of a class than individually. These findings suggest that teachers can have a significant impact on students on the class level by designing and structuring the class based on their level of expectation for that class (Trusz & Babel, 2016). Teachers with high expectations create an environment with many opportunities to learn while conversely, teachers without high expectations can make decisions that affect the learning opportunities for whole classes.

It was important that the observation tool selected for this study identify as many of these whole class characteristics as possible. Two well-known observation instruments were examined, the Danielson Framework for Teaching and the Classroom Assessment Scoring System.
(CLASS). The CLASS is a comprehensive observation system developed at the Center for Advanced Study of Teaching and Learning University of Virginia. It is designed to measure three domains, Emotional Support, Classroom Organization, and Instructional Support. The domains act as the overall category or umbrella and under each domain there are dimensions (specific category), indicators (ratings), and behavior markers (evidence that supports the rating). Although it has been thoroughly tested for validity and reliability, a decision was made not to use this tool because the scope of the tool was not aligned with the outcomes of this study and the training and time that would be required for each observation was also a consideration. For example, CLASS training is a two-day training and the observation protocol suggested that observation times could be up to two hours. It is believed that this would be far too intrusive and a burden on participants. Also, CLASS focuses on observing and assessing the effectiveness of individual interactions among teachers and students in classrooms but the domains associated with the observation tool fail to specifically highlight the behaviors associated with whole class high expectations although one could argue that they are implied.

The Danielson Framework for Teaching was developed in 1966 and has evolved over time. The version used to develop the instrument for this study was revised in 2013 and was a better fit for this study than CLASS. The instrument consists of four domains with a number of subdomains within each. Domain 1 is Planning and Preparation. It consists of several subdomains that range from the teacher’s knowledge of content and current pedagogy, knowledge of individual student abilities and backgrounds, ability to set outcomes, knowledge of resources, and the ability to plan and assess instruction. Domain 2 considers different aspects of the classroom environment. Within this domain the teacher’s ability to create and maintain a respectful environment, emphasize the importance of learning, manage classroom procedures
and behavior as well as how the organize physical space. Domain 3 evaluates instruction. Within this domain as with the others, multiple aspects of instruction are observed including the teacher’s ability to communicate with students, level of questioning used by the teacher, student engagement, assessment, and teacher responsibility. The final domain is Domain 4, Professional Responsibility. In domain 4 the teacher’s ability to reflect on learning, maintain records, communicate with families, participate in professional learning, and demonstrate professionalism are assessed. Taken together this tool has the ability to paint a thorough picture of a teacher’s ability and for that reason, it is used by many school districts. For the purpose of this study, only the subdomains that directly relate to teacher expectations will be recorded, Domain 1C – Setting Instructional Outcomes and Domain 2B – Establishing a Culture for Learning. These subdomains will be downloaded into a single form online using the Google Forms application. Observers will be assigned a unique log-in and will complete all observations within this tool. The data will then be analyzed.

Data Collection

Rater data included for this study was collected from two raters who used the Danielson Framework for Teaching and Learning to observe teachers simultaneously on two separate occasions. Teachers also completed surveys based on questions contained in the CABI. A minimum of two observations were collected. Upon completion of the data collection sessions the data were exported from Google Forms into and Excel spreadsheet and organized for analysis using SPSS software.

Data Analysis

Scores from raters were collected in Google Forms. The rubric consisted of two questions and each question could be scored as either ineffective, developing, accomplished, or exemplary.
Each category was assigned a score 1-4 and the sum of those scores makes up the data to be analyzed. A similar process was used for the survey questions. Each teacher completed a survey sent to them from a link on Google Forms. The survey contained five questions. Just like the CABI that the survey questions were modeled after, a higher score meant a teacher had higher expectations and a lower represented lower expectations. Each answer was assigned a score with the sum of these scores comprising the final teacher score. This process was repeated two weeks later and these sets of scores were analyzed. The study used a fully crossed two-facet design for the observation data where the object of measurement was teachers (t), and the facets were raters (r), and occasions (o). The survey data was analyzed using a single facet design where the object of measurement was teachers (t), and the facet was occasions (o).

Following estimation of the terms in the G-study, decision studies were conducted to determine the ideal facet conditions between raters and occasions to reduce the amount of error. Additional analysis was completed to determine whether or not observations and surveys constitute parallel forms.

Summary

This study was designed to identify sources of variance in the most commonly used methods of evaluating teacher expectations. Two G-studies, a two-facet fully crossed and a single-facet design was used. Two raters were selected to conduct a total of 62 observations over a two week time period. During this same time, the teachers also completed a survey based on the CABI which was designed to assess teacher expectations. Rater and survey data was analyzed and will be discussed thoroughly in chapters four and five.
Chapter 4

Results

The purpose of this study was to identify sources of variance in the data produced by the most common measures of teacher expectations. The questions guiding these generalizability and decision studies were:

1. How reliable are the data from the most common published measures of teacher expectations?

2. What sources of variance affect reliability across the most common published measures of teacher expectations?

3. Do observed measures and self-reported measures of teacher expectations constitute parallel forms of measurement?

The study used rater data collected from two observations in the spring of 2018. The two raters observed 31 elementary classrooms ranging from kindergarten to fifth grade. The observed teachers also completed surveys designed to assess expectations.

Generalizability and decision studies were completed to analyze sources of variance and determine facet conditions for reliability. The following tables detail the results of the generalizability study. The data set that was included for study analysis contains the complete set of 31 teachers (t) two occasions (o) and two raters (r). The second set of data also contains the complete set 31 teachers (t) and two occasions (o). The results of these analyses are organized by type (observations and surveys). The results include ANOVA tables, G-study results, and D study scenarios all summarized by their relevance to the research questions for this study.
The first question guiding this study was related to the reliability of the most common published measures of teacher expectations. A review of the literature established that the two most common methods of measuring teacher expectations were observation and teacher self-report of expectations. Separate G study analyses was completed for each measure. Descriptive statistics are presented in tables 4.1 and 4.2.

| Table 4.1 |
| Descriptive Statistics (Rater X Occasion) for observations using Danielson Framework |

<table>
<thead>
<tr>
<th></th>
<th>Occasion 1</th>
<th></th>
<th>Occasion 2</th>
<th></th>
<th>Total</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 1</td>
<td>6.32</td>
<td>1.58</td>
<td>6.16</td>
<td>1.34</td>
<td>6.24</td>
<td>1.46</td>
<td>62</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 2</td>
<td>5.68</td>
<td>1.38</td>
<td>5.65</td>
<td>1.47</td>
<td>5.66</td>
<td>1.41</td>
<td>62</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6.24</td>
<td>1.46</td>
<td>5.66</td>
<td>1.41</td>
<td>5.95</td>
<td>1.46</td>
<td>124</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Table 4.2 |
| Descriptive Statistics for Self-report surveys |

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>SD</th>
<th>N</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Occasion 1</td>
<td>6.00</td>
<td>1.50</td>
<td>62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occasion 2</td>
<td>5.90</td>
<td>1.42</td>
<td>62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5.95</td>
<td>1.46</td>
<td>124</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Teacher Observation

The two facet fully crossed study where the unit of measure was teachers and the facets were occasion and rater yielded the following results. The main effect variables accounted 17.9% of the error. The variance component for teacher\( (\sigma_T^2) \) shows the amount of variance due to individual differences among teachers and accounts for 11.3% of the overall variance. Occasion \( (\sigma_O^2) \) accounted for 0% of the total variance indicating little if any difference between teacher behavior due to time. The last main effect, rater\( (\sigma_R^2) \), accounted for 6.6% of the total variance indicating that some variance was due to differences in how raters scored individual teachers.

The G-coefficient, which is analogous to the reliability coefficient in Classical Test theory was \( (E\rho^2) = .25 \).

Table 4.3
Estimated Variance Components in Fully Crossed 2 Facet Design for observation data Elementary Classrooms (Teachers x Raters x Occasion)

<table>
<thead>
<tr>
<th>Source</th>
<th>Variance Component</th>
<th>Estimate</th>
<th>Percent of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teachers (p)</td>
<td>( \sigma_p^2 )</td>
<td>.2525</td>
<td>11.3%</td>
</tr>
<tr>
<td>Rater (r)</td>
<td>( \sigma_r^2 )</td>
<td>.1468</td>
<td>6.6%</td>
</tr>
<tr>
<td>Occasion (t)</td>
<td>( \sigma_t^2 )</td>
<td>-.0151*</td>
<td>0%</td>
</tr>
<tr>
<td>teacher X rater</td>
<td>( \sigma_{pr}^2 )</td>
<td>.6115</td>
<td>27.3%</td>
</tr>
<tr>
<td>teacher X occasion</td>
<td>( \sigma_{pt}^2 )</td>
<td>.5475</td>
<td>24.5%</td>
</tr>
<tr>
<td>rater X occasion</td>
<td>( \sigma_{rt}^2 )</td>
<td>-.0177*</td>
<td>0%</td>
</tr>
<tr>
<td>p x r x t, e</td>
<td>( \sigma_{prt,e}^2 )</td>
<td>.679</td>
<td>30.3%</td>
</tr>
</tbody>
</table>

**Notes.** Relative error variance \( (\sigma_d^2) = .75 \)
Generalizability coefficient \( (E\rho^2) = .25 \)
Absolute error \( (\sigma_{A}^2) = .82 \)
Index of Dependability \( (\phi) = .23 \)
* negative estimates set to zero for calculations
When just examining the main effects one might assume that the variability is close to where it should be with most of the variance attributable to individual difference among teachers, however, the interaction effects paint a different picture. The teacher by occasion ($\sigma^2_{TO}$) interaction accounts for 24.5% of the total variance which was interesting considering that the main effect of occasion was negligible. The teacher by rater ($\sigma^2_{TR}$) interaction accounts for 27.3% of the overall variance while occasion by rater ($\sigma^2_{OR}$) accounted for 0% of the variance. These interaction effects are substantial and were considered when designing the D studies.

Although 69.7% of error could be identified, another 30.3% of the variance was confounded within the error term, meaning that it was a combination of all of the facets and unknown error. It is important to note that in table 4.3, the occasion facet and interaction between occasion and rater facet were negative. Brennan (2001), explained issues with variance component estimation, specifically how one should handle negative estimates that arise due to sampling error. Brennan argued that negative estimates should be set to zero, however, researchers should use the negative estimates for other components. This method was used in this study.

According to Brennan (2001), the absolute error variance refers to the difference between the observed and universe score. The absolute error variance was calculated at .8227. Relative error refers to the error found when using a person’s deviation score as an estimate of the universe deviation score (Brennan, 2001). For this data, the relative error was calculated as .7493. According to Shavelson and Webb (1991), the G coefficient shows how accurate the generalization is from the observed score to the universe score and is analogous to a reliability coefficient in classical test theory. The G coefficient was .2520. Index of dependability is similar to a G coefficient for absolute decision. The Index of Dependability is reported as .2348.
Self-report

The results of the single facet G study on teacher self-report of expectations is found in table 4.4. The overwhelming majority of the variance is found among teachers ($\sigma_T^2$) at 97.5%. Very small amounts of error variance was reported with the occasion facet ($\sigma_O^2$) at 0% and confounded error ($\sigma_{TO}^2$) at 2.5%. The g-coefficient was calculated at .9875. The index of dependability was .9873. Relative and absolute error was reported as .2950 and .2983 respectively.

<table>
<thead>
<tr>
<th>Source</th>
<th>Variance Component</th>
<th>Estimate</th>
<th>Percent of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teachers (p)</td>
<td>$\sigma^2(p)$</td>
<td>23.22</td>
<td>97.5%</td>
</tr>
<tr>
<td>Occasion (o)</td>
<td>$\sigma^2(o)$</td>
<td>.0065</td>
<td>0%</td>
</tr>
<tr>
<td>Teacher X Occasion (po)</td>
<td>$\sigma^2(po)$</td>
<td>.590</td>
<td>2.5%</td>
</tr>
</tbody>
</table>

Notes. Relative error variance ($\sigma_\delta^2$) = .2950
Generalizability coefficient (E$\rho^2$) = .9875
Absolute error ($\sigma_\Delta^2$) = .2983
Index of Dependability ($\phi$) = .9873

The second question guiding this study was concerned with which sources of variance affected reliability across the most common published measures of teacher expectations. The data collected to address this question is also found in tables 4.3 and 4.4. The sources of variance that most affected reliability for teacher observations, were the interaction effects between teacher and occasion with more than 24.5% of the variation and between teacher and rater which
accounted for 27.3% of the variation. The largest source of error was from the confounded error at 30.3%.

The self-report data from the single facet study showed that the source of variance most effecting reliability was from the teacher with 97.5%. Only a small portion of variance was found in occasion (0%) and interaction (2.5%).

Next, a D study was conducted. According to Shavelson and Webb (1991), D studies are used to design procedures that reduce error. To accomplish this task the data from G studies are used.

The results of this D study provide data to address research question number two. D studies were conducted to assess whether or not reliability of teacher observation could be increased. Each facet was systematically manipulated while the others were held constant. This method ensured that each facet’s contribution to generalizability and reliability could be studied. According to Shavelson and Webb (1991), “increasing the number of raters or occasions would reduce error and consequently, increase the level of generalizability” (p. 102). As shown in table 4.5 both number of occasions and raters were increased and reliability ranged from .1208 at its lowest and .4936 at its highest with five raters and five occasions. An attempt was also made to assess if it were possible to increase reliability to acceptable ranges with only two raters. A d-study calculation was completed with two raters and ten occasions. The results were still low at .0547 for teacher by occasion and .3058 for teacher by rater. The estimated generalizability coefficient for five raters and five occasions was .4936.

After a review of the G-coefficients for teacher self-report, it was decided that additional d-studies were not needed considering that the coefficient was so high at .9875.
The last question guiding this study sought to discover if teacher observation measures and self-reported measures of teacher expectations constituted parallel forms. To answer this question, the data was correlated using Pearson’s $R$ and supported by the work of Crocker and Algina (1986). The two concepts guiding this process were the Coefficient of Equivalence and the Alternative Forms Method. The Coefficient of Equivalence endeavors to approximate comparable measures by administering two forms of a test on one occasion to the same

Table 4.5

<table>
<thead>
<tr>
<th>Source</th>
<th>$\sigma^2$</th>
<th>G-Study</th>
<th>D-Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$Occasion(n_0)$</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>$Rater(n_r)$</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Teacher</td>
<td>$\sigma_T^2$</td>
<td>.2525</td>
<td></td>
</tr>
<tr>
<td>Occasion</td>
<td>$\sigma_O^2$</td>
<td>-.0151</td>
<td></td>
</tr>
<tr>
<td>Raters</td>
<td>$\sigma_R^2$</td>
<td>.1468</td>
<td></td>
</tr>
<tr>
<td>teacher X occasion</td>
<td>$\sigma_{TO}^2$</td>
<td>.2737</td>
<td>.5475</td>
</tr>
<tr>
<td>teacher X rater</td>
<td>$\sigma_{TR}^2$</td>
<td>.3058</td>
<td>.6115</td>
</tr>
<tr>
<td>occasion X rater</td>
<td>$\sigma_{OR}^2$</td>
<td>-.0177</td>
<td></td>
</tr>
<tr>
<td>$Ei$</td>
<td>$\sigma_{tor,ei}^2$</td>
<td>.1698</td>
<td>.6790</td>
</tr>
<tr>
<td>Relative</td>
<td>$\sigma^2(\delta)$</td>
<td>.7439</td>
<td>1.838</td>
</tr>
<tr>
<td>Absolute</td>
<td>$\sigma^2\Delta$</td>
<td>.8107</td>
<td>1.9520</td>
</tr>
<tr>
<td>Generalizability</td>
<td>$E\hat{p}^2$</td>
<td>.2520</td>
<td>.1208</td>
</tr>
<tr>
<td>Dependability</td>
<td>$\Phi$</td>
<td>.2345</td>
<td>.1145</td>
</tr>
</tbody>
</table>

44
examinees. The Alternate forms method involves the creation of two versions of a test that are given to the same group of participants. The forms are to be given within a very short time period. The study procedures met these requirements. Both the survey and observation instruments were constructed to test the same content, teacher expectations and were given within a two-week period to the same teachers. The average score of observed raters across all occasions was correlated to the average score of self-report measures. No correlation was found between the results of the self-report and teacher observation scores, $r = 0.03$.

**Summary**

Overall, results from both the generalizability and decision studies indicate that reliability is low for teacher observation and high for teacher self-report of expectations. Results of the D-study reveal increasing the number of raters and occasions to increase the generalizability coefficients to acceptable values (.80 or greater) were unsuccessful. Also, attempts to correlate the measures were unsuccessful suggesting that the two (teacher observation and teacher self-report) are not parallel forms.

The following chapter includes additional discussion of the results, implications for practitioners, and recommendations for future studies.

**Chapter 5: Summary, Recommendations, and Conclusion**

**Introduction**

The purpose of this study was to examine the reliability of the most commonly used published measures of teacher expectations. The foundation on which this study was built consisted of three pillars of research, the seminal work of Rosenthal and Jacobsen (1968), a teacher expectancy critique by Hoge (1984), and Visible Learning (2008).
When stripped down to their most basic arguments, the seminal study as well as subsequent studies on teacher expectations make three claims. The first assertion is that teachers form expectations for students (Rosenthal & Jacobsen, 1968; Brophy & Good, 1970; Dusek & O’Connell, 1973; O’Connell, Dusek & Wheeler, 1974; Rist, 1970). The second claim follows that once these expectations are formed they lead teachers to treat students differently (Braun, 1976; Brophy & Good, 1970; Rothbart, Dalfen, & Barrett, 1971, Good & Nichols, 2001). Finally, these differences in treatment can affect student achievement (Brophy & Good, 1970; Jussim & Eccles, 1992; Alvidrez & Weinstein, 1999; Goldenberg, 1992, Madon, Jussim, & Eccles, 1997). Although these studies looked at different populations, age groups, genders, and settings with slight variations, the evidence overwhelmingly supports the existence of this phenomenon.

Hoge (1984) wrote a critique that looked at a number of issues concerning the teacher expectancy phenomenon. One of his critiques examined what he saw as issues of validity and reliability of the methods used to measure teacher expectations. He implored future researchers to strengthen these areas. A more recent review of the literature on teacher expectations published over the last five years revealed that only nine of the 38 studies mentioned reliability at all, confirming that this issue was still mostly unaddressed more than three decades later and an obvious gap in the literature.

The last piece of the foundation of this study came from Hattie’s work and was included to address the work of practitioners today. In his text, Visible Learning, Hattie synthesized multiple meta-analyses on a number of educational topics or methods and ranked them according to effect size. He then suggested that resources be allocated to those topics, methods, and concepts with the largest effect size. Hattie calculated an effect size of \( d = .43 \) for teacher
expectation placing it in the highly effective category. Although, Hattie’s work has been critiqued for a number of reasons, the criticism that aligns with this study was his failure to evaluate the quality of the meta-analyses that he chose to synthesize (Snook, O’Neill, Clark, O’Neill & Openshaw, 2009). Some of these study characteristics include sample size, methodology, and the validity and reliability. Simply put, Rosenthal and Jacobsen introduced a phenomenon of teacher expectation, it was critiqued by Hoge and others for its lack of reliability and measurement, little was done to remedy these issues and now with the help of Hattie it is considered a must do strategy amongst practitioners. This study was not an attempt to disprove the theory, but rather to evaluate and possibly improve the reliability of measures used to assess the phenomenon.

The study sought to answer the following questions:

1. How reliable are the data from the most common published measures of teacher expectations?

2. What sources of variance affect reliability across the most common published measures of teacher expectations?

3. Do observed measures and self-reported measures of teacher expectations constitute parallel forms of measurement?

**Summary of Findings**

To answer the questions generalizability theory was used to identify sources of variance amongst the most commonly used published measures of teacher expectations with a goal of possibly increasing reliability. Two G-studies were conducted. One was a 2-facet fully crossed G-study where the unit of measurement was teacher, the facets were occasion and rater. The
second was a single-facet G-study where the unit of measure was again teacher and the facet was occasion. In addition to the G-studies, the data produced from the 2-facet fully crossed G-study on teacher observation was used to create a D-study with the purpose of increasing the reliability of the data by manipulating the occasion and rater facets. The reliability was extremely high for teacher self-report so it was decided that a D-study for this data was unnecessary. The final statistical analysis was an attempt to correlate the two measures to assess whether or not they constituted parallel forms. An in-depth discussion of these results follows.

The first question, How reliable are the most common published measures of teacher expectations? was answered through the results of the separate G-studies and more specifically the generalizability coefficient. The generalizability coefficient is comparable to the reliability coefficient in Classical Test Theory. According to Webb, Shavelson, & Haertel (2006), if little variability exists in observed scores across different conditions, then it is assumed that the observed score must be close to the person’s true score. There should be a high correlation between true and observed scores. However, if there are large variances in a persons observed score across conditions then the true and observed scores are not close, substantial error exists and the correlation will be low. According to Webb, Shavelson, and Haertel (2006), coefficients are considered sufficiently reliable when they are greater than or equal to 0.80. If substantial consequences accompany the decisions, a higher value is preferred. Table 4.3 gives the estimated error variances and generalizability coefficients for the two facet fully crossed study where the unit of measure was teachers and the facets were occasion and rater. The generalizability coefficient ($E\rho^2$) = .25 far below the acceptable value .80 indicating that the data from teacher observations is unreliable.
Data from the d-study of teacher observation data went further by systematically manipulating both the occasion and rater facets in an attempt to improve reliability. Although reliability was increased slightly, (see table 4.5) even with five raters and five occasions reliability did not reach acceptable ranges ($\hat{E}_p^2 = .4936$). Further, even when the number of occasions was increased to 10 which is equivalent to one observation every month of a typical school year with two raters the generalizability coefficient is less reliable with $\hat{E}_p^2 = .3903$.

A separate G-study was conducted to assess the reliability of the data from teacher self-reports of expectations and the data is found in table 4.4. The generalizability coefficient was extremely high ($E_\rho^2 = .9875$). The generalizability coefficient was so high that it would be considered sufficient for serious decision-making. Due to the high reliability of this measure a d-study was not conducted.

The second question, *what sources of variance affect reliability across the most common published measures of teacher expectations?* was answered using both the results from the G-studies. The main effects accounted for 17.9% of the total variance. The interaction effects, accounted for 51.8% of the overall variance and 30.3% of the variance was confounded in the error term. The observation data indicated that the sources of variance most effecting reliability were the interaction effects at 51.8% of the overall variance and the unexplained error with 30.3% of the overall variance. For the self-report survey data, the individual teacher differences had the biggest effect accounting for 97.5% of the overall variance.

When examining teacher self-report the findings suggest that there are serious threats to the validity of the measure. According to Hoskin (2012), there are several reasons why self-report surveys might not be entirely valid. One reason is participant honesty. The research is
based on the participants telling the truth. The more controversial the questions the more likely that a participant might not tell the truth. Teacher expectations fall into this category because teachers are expected to have high expectations for students. Teacher are unlikely to admit that they have low expectations. One must also consider the participant’s ability to be introspective. Even if participants believe they are answering honestly, some people lack the ability to see themselves like others do. Also, the participants ability to understand the questions can contribute to the lack of validity. This is of particular concern when the survey is meant to measure abstract concepts like teacher expectations. It is impossible to know if everyone who answered the question interpreted the questions in the same way. Finally, participant interpretation of the rating scales is problematic. Research suggests that some people are known as extreme responders. They answer at either end of the scale while others prefer to stay in the middle. This tendency might naturally produce differences that have little if anything to do with what the purpose of the survey.

The final question, *Do observed measures and self-reported measures of teacher expectations constitute parallel forms?* was answered by attempting to correlate the two measures to assess whether or not they constituted parallel forms. This procedure was suggested from Crocker and Algina (1986). No correlation was found between the measures and it was determined that they do not constitute parallel forms.

**Interpretation of Findings**

Similar to Hoge (1984), the results of this study suggest that issues of reliability exists in the most common published methods of teacher expectations. Further, this research suggests that teacher observations are unreliable (*E p^2 = .25*) and self-report surveys are highly reliable
\(Ep^2 = .98\) although there are serious threats to the validity of this measure. Last these two measures do not constitute parallel forms \((p=.031)\). Several reasons might explain these findings.

First, the lack of a clear definition for the teacher expectancy construct could be one explanation of these results (Hoge, 1984). Even in the seminal study, Rosenthal and Jacobsen (1968) fail to give a definition of teacher expectations although one can extract their intended meaning from their hypothesis which states, “within a given classroom those children for whom the teacher expected greater intellectual growth would show greater growth” \((p. 61)\). Later, Brophy and Good (1983) go on to operationalize the phenomenon when they align specific classroom behaviors to the teachers level of expectation. An analysis of the direction and variety in which the subsequent studies go, indicate more work might need to occur on the conceptualization of teacher expectations. For example, when one examines teacher expectations should the study focus on general academic ability (Rosenthal & Jacobsen, 1968) or specific academic areas (Rubie-Davies, Peterson, & Flint, 2015)? Should the studies include present levels of performance or only future prediction? What is the connection between current levels of performance and the teacher’s assertion of future performance studies? This lack of definition makes instrument design and consistency across instruments very difficult.

Another possible reason for these results is the research community’s failure to acknowledge the differences in the types of questions used on tools designed to measure expectations, and the tendency of researchers to treat the items/questions the same without evidence that they were measuring the same thing (Hoge, 1984). Hoge (1984) examined the types of questions and tasks included in the research and used what he called a “crude categorization” to organize each type. These categories were described as either unidimensional or multidimensional measures. The literature today still shows variability in the types of
questions used to measure the phenomenon (see appendix A) and these results support the assumption that this is problematic. The vastly different results between the reliability of the two instruments and the lack of correlation between the two instruments indicates that they are indeed measuring something different. This questions not only the reliability of the instruments, but the validity as well.

Again, the goal of this study was not to challenge the existence of teacher expectations or the work of John Hattie. Hattie used current research to suggest that practitioners should, “be prepared to be surprised” (p. 124). By this, he meant look for evidence to disprove low expectations and encourage teachers to examine their current practices to make sure all students are receiving a challenging curriculum. No one could argue the practicality of this advice.

**Limitations**

The current study was designed to assess the reliability of the most common published measures of teacher expectations. The instruments used and the procedures might limit the empirical findings from this study. According to Hill, Charalambos, and Kraft (2011) reliability cannot be linked to an instrument alone and instead requires the combination of rater training, certification, and specific scoring rubrics. These results suggest an overwhelming lack of reliability for teacher observation however, they might be skewed by the lack of such a comprehensive system described above, specifically extensive rater training. This study was designed to mirror the current practices used by the district studied. It is believed that these results then reflect those that might be found in this District but might not translate in other areas where more training is provided.
Recommendations for Future Research

Future studies should continue to focus on reliability, validity, and measurement of the teacher expectation phenomenon. The tendency thus far has been for researchers to build on the current research in this area, but this study indicates that time and resources might be better used on further conceptualization and definition of the concept. The ability of research to clearly conceptualize teacher expectancy and develop tools to reliably measure the construct would be invaluable not only for researchers but for practitioners as well. The amount of time and financial resources spent each year observing teachers, at least for the purposes of identifying expectations might be better spent elsewhere.

Conclusion

This study highlights the issues related to measuring teacher expectations. Expectancy theory is one of the most recognized, accepted, and marketed theories in education. Most, if not all of the leading educational researchers incorporate pieces of this theory into their work, and this lays the foundation for what happens at the state level where laws are written, district level where policy is written, university level where curriculum is decided for teacher preparation programs, as well as in the classrooms. With 50 years of research studies supporting teacher expectation theory, it would be tough to find someone to dispute it. That is what makes the results of this study so important. These results suggest that unlike other theories, which are based on measurable concepts, we lack the ability to measure the teacher expectation concept on which expectancy theory is built.

The two most common published measures of teacher expectations, teacher observation and teacher self-report were examined in this study. Based on the results, teacher observation
was found to be unreliable. Further, there was no feasible amount of raters or occasions that brought the generalizability coefficient to the acceptable level for decision making of .80. When examining teacher self-report the findings suggest that there are serious threats to the validity of the measure. According to Hoskin (2012), participant honesty, ability to be introspective, understanding of the questions, and interpretation of the rating scales all contribute this issue. Finally, because some studies use observation as a check of self-report the two measures were assessed to see if they constituted parallel forms. The two measures of teacher expectations were not found to be parallel measures and therefore are not measuring the same thing.
Appendix A

Matrix of Current Research, Methodology, and Reliability

<table>
<thead>
<tr>
<th>Citation</th>
<th>Method</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adumitroaie, E., &amp; Dafinoiu, I. (2013). Measuring Teachers' Expectations towards Students with Migrant Parents. <em>Annals of All Cuza University. Psychology Series</em>, 22(2), 87.</td>
<td>Researchers designed a scenario describing a student who is preparing to be admitted to high school. The scenarios were exactly the same with one exception. Half of the scenarios did not contain any reference to the student’s family and for the other half it was mentioned that student’s parents worked abroad. Using the scenario teacher expectations were evaluated using a questionnaire.</td>
<td>No reliability coefficient</td>
</tr>
<tr>
<td>Akifyeva, R. N., &amp; Alieva, A. (2016). The Influence of Ethnicity on Teacher Expectations and Teacher Perceptions of Student Warmth and Competence.</td>
<td>Surveys</td>
<td>No reliability coefficient</td>
</tr>
<tr>
<td>Boonen, T., Van Damme, J., &amp; Onghena, P. (2014). Teacher effects on student achievement in first grade: which aspects matter most?. <em>School Effectiveness and School Improvement</em>, 25(1), 126-152.</td>
<td>A teacher questionnaire and a school team questionnaire</td>
<td>No mention of reliability</td>
</tr>
<tr>
<td>Boser, U., Wilhelm, M., &amp; Hanna, R. (2014). The Power of the Pygmalion Effect: Teachers’ Expectations Strongly Predict College Completion. <em>Center for American Progress</em>.</td>
<td>Self-report survey in which teachers were asked to predict “how far in school … you expect this student to get,” including high school, college, and beyond.</td>
<td>No mention of reliability</td>
</tr>
<tr>
<td>Source</td>
<td>Study Description</td>
<td>Methodology</td>
</tr>
<tr>
<td>--------</td>
<td>-------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Carder, R. J. (2015). The influences of teachers regulative discourse, specifically teachers expectations of learner achievement on teachers' pedagogic practice in teaching Grade 6 Natural Science in the Western Cape: two case studies.</td>
<td>Qualitative study that used interviews, seating charts, and lesson plans etc. as a data collection method.</td>
<td>No reliability coefficient</td>
</tr>
<tr>
<td>Dabach, D. B., Suárez-Orozco, C., Hernandez, S. J., &amp; Brooks, M. D. (2017). Future Perfect?: Teachers’ Expectations and Explanations of their Latino Immigrant Students’ Postsecondary Futures. <em>Journal of Latinos and Education</em>, 1-15.</td>
<td>Researchers conducted interviews and their analysis was based on a series of questions in the interview protocol concerning teachers’ perceptions of their students’ postsecondary futures. We asked teachers to project what percentage of their immigrant and nonimmigrant students would attend college, whether immigrant students were aware of college information (and if not, why not), and what immigrant youth should know to go to college. We also asked: For kids who are not going to college, do you have any ideas about the kinds of things they are likely to be doing after they leave high school?</td>
<td>No reliability coefficient</td>
</tr>
<tr>
<td>Fischbach, A., Baudson, T. G., Preckel, F., Martin, R., &amp; Brunner, M. (2013). Do teacher judgments of student intelligence predict life outcomes?. <em>Learning and Individual Differences</em>, 27, 109-119.</td>
<td>Students filled out a background questionnaire and completed a comprehensive intelligence test. Teachers completed a questionnaire in which they reported student grades and judged their intelligence. TJs of student intelligence at age 12 were assessed in the middle of the term by the following question: “How would you rate this child's intelligence?” Responses were given on a 5-point rating scale (1 = very low, 2 = low, 3 = average, 4 = high, 5 = very high).</td>
<td>No reliability coefficient</td>
</tr>
<tr>
<td>Author(s)</td>
<td>Method</td>
<td>Reliability Coefficient</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>Garrett, L., Rubie-Davies, C., Alansari, M., Peterson, E., &amp; Flint, A. (2015).</td>
<td>Teachers were asked to compile a class list and estimate what level they expected each of their students to achieve in reading by the end of the current school year, relative to the New Zealand curriculum levels for their particular group of students. That is, teachers were provided with guidance as to what constituted below, average and above average levels in relation to the curriculum levels for each year group.</td>
<td>No reliability coefficient</td>
</tr>
<tr>
<td>Glock, S., &amp; Krolak-Schwerdt, S. (2013).</td>
<td>Teachers were provided student information that either confirmed or disconfirmed a stereotype. Expectations were then examined using a questionairre</td>
<td>No reliability coefficient</td>
</tr>
<tr>
<td>Hansen, K. (2016).</td>
<td>Teachers were asked to rate the child in terms of attractiveness. Teachers could rate children as either attractive, not attractive, undernourished, slovenly or dirty. Teachers were also asked a number of questions related to their perception of the child’s academic ability (a binary response question of whether the child showed any outstanding ability) and their behaviour (whether they were delinquent, rebellious, aggressive or showed other bad behaviours). Teachers were further asked to rate each child’s ability on a scale from 1 to 5 with 1 being exceptional, 2 above average, 3 average, 4 below average and 5 very limited in different areas of</td>
<td>No reliability coefficient</td>
</tr>
<tr>
<td>Study</td>
<td>Description</td>
<td>Methodology</td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Irizarry, Y. (2015).</td>
<td>Selling students short: Racial differences in teachers’ evaluations of high, average, and low performing students. Social science research, 52, 522-538.</td>
<td>Teachers were asked to provide an overall rating of their student’s language and literacy skills compared to other students of the same grade by selecting one of five answer options ranging from far below average to far above average.</td>
</tr>
<tr>
<td>Kaiser, J., Südkamp, A., &amp; Möller, J. (2017).</td>
<td>The effects of student characteristics on teachers’ judgment accuracy: Disentangling ethnicity, minority status, and achievement. Journal of Educational Psychology, 109(6), 871.</td>
<td>Participants were asked to rate the percentage of correct answers given by each student on a scale from 0%–100%</td>
</tr>
<tr>
<td>Kim, H. S. (2015).</td>
<td>Foregone opportunities: unveiling teacher expectancy effects in kindergarten using counterfactual predictions. Social Psychology of Education, 18(2), 273-296.</td>
<td>Self-report survey where each item was rated on a scale from one (“Not Yet”: child has not yet demonstrated skill, knowledge, or behavior) to five (“Proficient”: child demonstrates skill, knowledge, or behavior competently and consistently). There was also the choice of “not applicable,” which was appropriate when a relevant item had not been introduced into the classroom.</td>
</tr>
<tr>
<td>Langan, K. (2015).</td>
<td>Kill the Stigma! Teacher Expectancy in the Information Literacy Classroom. ACRL 2015.</td>
<td>Participants completed a questionnaire with 28 questions loosely connected to the “Star Teacher Pre-Screening” test</td>
</tr>
<tr>
<td>Malone, L. (2015).</td>
<td>My existence didn’t make no difference to them: Perceptions of teacher expectations among African-American students and their families.</td>
<td>Quantitative study using NCES data. Unique because each student was rated by two teachers. This allowed researchers to look at differing expectations. Teachers were asked about the probability that a student would complete college.</td>
</tr>
<tr>
<td>Metzger, A. N. (2016).</td>
<td>The Influence of the ADHD Label on Teacher’s Expectations of Academic Achievement (Doctoral dissertation, University of California, Merced).</td>
<td>Teachers rate their students based on their own perceptions of how students are performing in these subjects. These students received a rating of far below average, below average, average, above average, or far above average in comparison to children of the same grade for their mathematical, science and reading skills.</td>
</tr>
<tr>
<td>Papageorge, N. W., Gershenson, S., &amp; Kang, K.</td>
<td>Qualitative study that used interviews as a data collection method</td>
<td></td>
</tr>
</tbody>
</table>
## MEASURING TEACHER EXPECTATIONS

<table>
<thead>
<tr>
<th>Source</th>
<th>Methodology and Findings</th>
<th>Reliability/Validity Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peterson, E. R., Rubie-Davies, C., Osborne, D., &amp; Sibley, C. (2016). Teacher Expectations Matter.</td>
<td>Teachers completed both a self-report and a modified Implicit Association Task designed to assess ethnic stereotypes associated with academic achievement and failure. A Likert scale was used and Teachers were asked to list the students in their class and, without referring to school records, indicate the level they expected each student to reach by the end of the academic year (relative to national curriculum levels). Teachers were provided with information about the average national achievement of students at various year levels, but not individual student achievement data.</td>
<td>No reliability coefficient</td>
</tr>
<tr>
<td>Ready, D. D., &amp; Chu, E. M. (2015). Sociodemographic inequality in early literacy development: The role of teacher perceptual accuracy. Early Education and Development, 26(7), 970-987.</td>
<td>Teacher misestimation measures were calculated using two achievement tests and teacher self-report</td>
<td>Reliability coefficient was reported for the ARS data $\alpha = .87$, but for the purpose of this study that rating was then subtracted from actual student performance and then a new scale was created for which no reliability coefficient was reported.</td>
</tr>
<tr>
<td>Regalla, M. (2013). Teacher Expectations and Students from Low Socioeconomic Background: A Perspective from Costa Rica. Online Submission.</td>
<td>Surveys</td>
<td>No reliability coefficient</td>
</tr>
<tr>
<td>Riley, T. (2015). “I Know I'm Generalizing but…”: How Teachers’ Perceptions Influence ESL Learner Placement. TESOL Quarterly, 49(4), 659-680.</td>
<td>Specifically, the study examines 21 teachers’ responses to and decisions regarding fictional student record cards. Teachers were asked to participate in an interview in which they read about fictional students and sorted them into categories. They were then invited to explain their thinking.</td>
<td>No reliability coefficient</td>
</tr>
<tr>
<td>Ross, L. (2015). Pygmalion or plekhanov in the classroom: The subtle role</td>
<td>Used national KEYS data which consisted of teachers completing a self administered survey. Three questions were used to assess teacher expectation. On average, what is the performance level of all students in your</td>
<td>No reliability coefficient</td>
</tr>
</tbody>
</table>
| of social class in teacher perceptions. | TARGET CLASS?  
2. On average, what is the performance of racial and ethnic minority students in your TARGET CLASS?  
3. On average, what is the performance of Caucasian, not of Hispanic origin, students in your TARGET CLASS? Teachers could respond using primarily high achieving; |  |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sedova, K., &amp; Salamounova, Z. (2016). Teacher expectancies, teacher behaviour and students' participation in classroom discourse. <em>The Journal of Educational Enquiry, 15</em>(1).</td>
<td>Videotaped observations from literacy lessons at a Czech lower secondary school were analyzed. Each teacher was to assess students in their classroom by answering the following question: “This student is predisposed to study at a university” with a number ranging from 1 (I completely agree) to 10 (I completely disagree).</td>
<td>No reliability coefficient</td>
</tr>
</tbody>
</table>
Language Literacy Subscale $\alpha = .94$ |
<p>| St Amant, T. (2017). The Effect of Teacher Mindset on Low-Track Students. | Surveys, observations, and interviews | No reliability coefficient |
| Tedeschi, D. M. (2016). Factors Influencing a Teacher's Decision to Make | Online self-report instrument consisting of two sections. The first asked teachers to respond to a scenario and determine how likely they would be to refer the student | No reliability coefficient |</p>
<table>
<thead>
<tr>
<th>Source</th>
<th>Method</th>
<th>Reliability</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thys, S., &amp; Van Houtte, M. (2016).</td>
<td>Teachers’ expectations were measured using a self-report survey that consisted of three items. These items measured the expectations for future school progress. The three items are: “I expect most of my pupils will perform well in their future school career”, “I expect most of my pupils to perform well in secondary education”, and “I think most students will find their way in life”.</td>
<td>$\alpha = .0804$</td>
<td></td>
</tr>
<tr>
<td>Timmermans, A. C., Boer, H., &amp; Werf, M. P. (2016).</td>
<td>Teacher expectations were operationalized using the track recommendations given by teachers at the end of primary education. The track recommendation was considered an informed expectation of the teacher indicating which track is the most optimal for a student given the student’s potential. The recommendation was measured using a teacher questionnaire in which teachers could choose the most optimal tracks in secondary education for each student.</td>
<td>No reliability coefficient</td>
<td></td>
</tr>
<tr>
<td>Timmermans, A. C., Kuyper, H., &amp; Werf, G. (2015).</td>
<td>The teacher’s recommendations at the end of primary education were measured using a questionnaire. For each student, the teachers could mark one or more of eight categories that corresponded to the tracks in secondary education (1 = practical training; 2 = basic track with additional support; 3 = basic track; 4 = middle track; 5 = combined track; 6 = theoretical track; 7 = higher general secondary education; and 8 = pre-university education).</td>
<td>No reliability coefficient</td>
<td></td>
</tr>
<tr>
<td>Van Houtte, M., &amp; Demanet, J. (2016).</td>
<td>Self-report survey was used to measure expectations. The ‘Teachable Pupil Survey’ developed by Kornblau (1982) was comprised of 31 items, including ‘school-adjusted behaviors’ (e.g., ‘concentrates well’ and ‘enjoys school work’), ‘cognitive-motivational behaviors’ (e.g., ‘intelligent’ and ‘curious, inquisitive’), and ‘personal-social behaviors’ (e.g., ‘calm’ and ‘confident’).</td>
<td>$\alpha = 0.94$</td>
<td></td>
</tr>
</tbody>
</table>
Appendix B

Self-Report Survey Tool

Thank you for your voluntary participation in this research study. By completing and submitting this web-based survey, you are giving your consent for the researcher to include and use your responses in this study. Your participation is strictly voluntary, and you may choose not to participate without fear of negative consequences. Responses will be treated confidentially and no identifiable information will be disclosed or published.

Name - ______________________ School _______________________

Gender

A. Male
B. Female

Age

A. 21-30
B. 31-38
C. 39-46
D. 47-54
E. Over 55

Ethnicity/Race or ethnicity

A. Black
MEASURING TEACHER EXPECTATIONS

B. White

C. Hispanic

D. Other

Years of Teaching

A. 0-5 years

B. 6-11 years

C. 12-17 years

D. 18-23 years

E. Over 24 years

How many students are in your class? ______

Directions: The questions on this survey focus on teacher expectations of elementary students. Please answer each question as honestly as possible. Please respond by considering how well each statement applies to the students in your classroom. Tally the number of students in your classroom for which each statement applies and record it below.

(1) This child is smart for his/her age. _____

(2) This child catches on quickly, e.g., is quick at learning new games. _____

(3) This child is able to follow directions, remembers what he/she is told. _____

(4) This child understands difficult words. _____

(5) This child learns new skills to cope with new situations or problems. _____
Appendix C

Teacher Observation Tool


Teaching is a purposeful activity; even the most imaginative activities are directed toward certain desired learning. Therefore, establishing instructional outcomes entails identifying exactly what students will be expected to learn; the outcomes describe not what students will do, but what they will learn. The instructional outcomes should reflect important learning and must lend themselves to various forms of assessment through which all students will be able to demonstrate their understanding of the content. Insofar as the outcomes determine the instructional activities, the resources used, their suitability for diverse learners, and the methods of assessment employed, they hold a central place in domain 1.

Learning outcomes may be of a number of different types: factual and procedural knowledge, conceptual understanding, thinking and reasoning skills, and collaborative and communication strategies. In addition, some learning outcomes refer to dispositions; it’s important not only that students learn to read but also, educators hope, that they will like to read. In addition, experienced teachers are able to link their learning outcomes with outcomes both within their discipline and in other disciplines.

The elements of component 1c are:

Value, sequence, and alignment

- Outcomes represent significant learning in the discipline reflecting, where appropriate, the Common Core State Standards.

Clarity

- Outcomes must refer to what students will learn, not what they will do, and must permit viable methods of assessment.

Balance

- Outcomes should reflect different types of learning, such as knowledge, conceptual understanding, and thinking skills.

Suitability for diverse students

- Outcomes must be appropriate for all students in the class.

Indicators include:

- Outcomes of a challenging cognitive level
- Statements of student learning, not student activity
- Outcomes central to the discipline and related to those in other disciplines
- Outcomes permitting assessment of student attainment
- Outcomes differentiated for students of varied ability
<table>
<thead>
<tr>
<th>UNSATISFACTORY • LEVEL 1</th>
<th>BASIC • LEVEL 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>The outcomes represent low expectations for students and lack of rigor, and not all of these outcomes reflect important learning in the discipline. They are stated as student activities, rather than as outcomes for learning. Outcomes reflect only one type of learning and only one discipline or strand and are suitable for only some students.</td>
<td>Outcomes represent moderately high expectations and rigor. Some reflect important learning in the discipline and consist of a combination of outcomes and activities. Outcomes reflect several types of learning, but the teacher has made no effort at coordination or integration. Outcomes, based on global assessments of student learning, are suitable for most of the students in the class.</td>
</tr>
<tr>
<td>• Outcomes lack rigor.</td>
<td>• Outcomes represent a mixture of low expectations and rigor.</td>
</tr>
<tr>
<td>• Outcomes do not represent important learning in the discipline.</td>
<td>• Some outcomes reflect important learning in the discipline.</td>
</tr>
<tr>
<td>• Outcomes are not clear or are stated as activities</td>
<td>• Outcomes are suitable for most of the class.</td>
</tr>
<tr>
<td>• Outcomes are not suitable for many students in the class</td>
<td></td>
</tr>
</tbody>
</table>

A learning outcome for a fourth-grade class is to make a poster illustrating a poem.
- All the outcomes for a ninth-grade history class are based on demonstrating factual knowledge.
- The topic of the social studies unit involves the concept of revolutions, but the teacher expects his students to remember only the important dates of battles.
- Despite the presence of a number of ELL students in the class, the outcomes state that all writing must be grammatically correct.
- None of the science outcomes deals with the students’ reading, understanding, or interpretation of the text.

<table>
<thead>
<tr>
<th>PROFICIENT • LEVEL 3</th>
<th>DISTINGUISHED • LEVEL 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most outcomes represent rigorous and important learning in the discipline and are clear, are written in the form of student learning, and suggest viable methods of assessment. Outcomes reflect several different types of learning and opportunities for coordination, and they are differentiated, in whatever way is needed, for different groups of students</td>
<td>All outcomes represent high-level learning in the discipline. They are clear, are written in the form of student learning, and permit viable methods of assessment. Outcomes reflect several different types of learning and, where appropriate, represent both coordination and integration. Outcomes are differentiated, in whatever way is needed, for individual students.</td>
</tr>
<tr>
<td>• Outcomes represent high expectations and rigor.</td>
<td>• The teacher’s plans reference curricular frameworks or blueprints to ensure accurate sequencing.</td>
</tr>
<tr>
<td>• Outcomes are related to “big ideas” of the discipline.</td>
<td>• The teacher connects outcomes to previous and future learning.</td>
</tr>
<tr>
<td>• Outcomes are written in terms of what students will learn rather than do.</td>
<td></td>
</tr>
</tbody>
</table>
MEASURING TEACHER EXPECTATIONS

<table>
<thead>
<tr>
<th>Outcomes represent a range of types: factual knowledge, conceptual understanding, reasoning, social interaction, management, and communication.</th>
<th>Outcomes are differentiated to encourage individual students to take educational risks.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcomes, differentiated where necessary, are suitable to groups of students in the class.</td>
<td></td>
</tr>
</tbody>
</table>

- One of the learning outcomes is for students to “appreciate the aesthetics of 18th-century English poetry.”
- The outcomes for the history unit include some factual information, as well as a comparison of the perspectives of different groups in the run-up to the Revolutionary War.
- The learning outcomes include students defending their interpretation of the story with citations from the text.
- The teacher encourages his students to set their own goals; he provides them a taxonomy of challenge verbs to help them strive to meet the teacher’s higher expectations of them.
- Students will develop a concept map that links previous learning goals to those they are currently working on.
- Some students identify additional learning.
- The teacher reviews the project expectations and modifies some goals to be in line with students’ IEP objectives.
- One of the outcomes for a social studies unit addresses students analyzing the speech of a political candidate for accuracy and logical consistency.

ESTABLISHING A CULTURE FOR LEARNING

A “culture for learning” refers to the atmosphere in the classroom that reflects the educational importance of the work undertaken by both students and teacher. It describes the norms that govern the interactions among individuals about the activities and assignments, the value of hard work and perseverance, and the general tone of the class. The classroom is characterized by high cognitive energy, by a sense that what is happening there is important, and by a shared belief that it is essential, and rewarding, to get it right. There are high expectations for all students; the classroom is a place where the teacher and students value learning and hard work.

Teachers who are successful in creating a culture for learning know that students are, by their nature, intellectually curious, and that one of the many challenges of teaching is to direct the students’ natural energy toward the content of the curriculum. They also know that students derive great satisfaction, and a sense of genuine power, from mastering challenging content in the same way they experience pride in mastering, for example, a difficult physical skill.

Part of a culture of hard work involves precision in thought and language; teachers whose classrooms display such a culture insist that students use language to express their thoughts clearly. An emphasis on precision reflects the importance placed, by both teacher and students, on the quality of thinking; this emphasis conveys that the classroom is a business-like place.
where important work is being undertaken. The classroom atmosphere may be vibrant, even
joyful, but it is not frivolous.

The elements of component 2b are:

Importance of the content and of learning

In a classroom with a strong culture for learning, teachers convey the educational value of what the students are learning.

Expectations for learning and achievement

In classrooms with robust cultures for learning, all students receive the message that although the work is challenging, they are capable of achieving it if they are prepared to work hard. A manifestation of teachers’ expectations for high student achievement is their insistence on the use of precise language by students.

Student pride in work

When students are convinced of their capabilities, they are willing to devote energy to the task at hand, and they take pride in their accomplishments. This pride is reflected in their interactions with classmates and with the teacher.

Indicators include:

- Belief in the value of what is being learned
- High expectations, supported through both verbal and nonverbal behaviors, for both learning and participation
- Expectation of high-quality work on the part of students
- Expectation and recognition of effort and persistence on the part of students
- High expectations for expression and work products

<table>
<thead>
<tr>
<th>UNSATISFACTORY • LEVEL 1</th>
<th>BASIC • LEVEL 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>The classroom culture is characterized by a lack of teacher or student commitment to learning, and/ or little or no investment of student energy in the task at hand. Hard work and the precise use of language are not expected or valued. Medium to low expectations for student achievement are the norm, with high expectations for learning reserved for only one or two students.</td>
<td>The classroom culture is characterized by little commitment to learning by the teacher or students. The teacher appears to be only “going through the motions,” and students indicate that they are interested in the completion of a task rather than the quality of the work. The teacher conveys that student success is the result of natural ability rather than hard work, and refers only in passing to the precise use of language. High expectations for learning are reserved for those students thought to have a natural aptitude for the subject.</td>
</tr>
</tbody>
</table>
### PROFESSIONAL EXPECTATIONS

<table>
<thead>
<tr>
<th>The teacher conveys that there is little or no purpose for the work, or that the reasons for doing it are due to external factors.</th>
<th>The teacher’s energy for the work is neutral, neither indicating a high level of commitment nor ascribing the need to do the work to external forces.</th>
</tr>
</thead>
<tbody>
<tr>
<td>The teacher conveys to at least some students that the work is too challenging for them.</td>
<td>The teacher conveys high expectations for only some students.</td>
</tr>
<tr>
<td>Students exhibit little or no pride in their work.</td>
<td>Students exhibit a limited commitment to complete the work on their own; many students indicate that they are looking for an “easy path.”</td>
</tr>
<tr>
<td>Students use language incorrectly; the teacher does not correct them.</td>
<td>The teacher’s primary concern appears to be to complete the task at hand.</td>
</tr>
<tr>
<td>The teacher tells students that they’re doing a lesson because it’s in the book or is district-mandated.</td>
<td>The teacher urges, but does not insist, that students use precise language.</td>
</tr>
<tr>
<td>The teacher says to a student, “Why don’t you try this easier problem?”</td>
<td>The teacher says, “Let’s get through this.”</td>
</tr>
<tr>
<td>Students turn in sloppy or incomplete work.</td>
<td>The teacher says, “I think most of you will be able to do this.”</td>
</tr>
<tr>
<td>Many students don’t engage in an assigned task, and yet the teacher ignores their behavior.</td>
<td>Students consult with one another to determine how to fill in a worksheet, without challenging one another’s thinking.</td>
</tr>
<tr>
<td>Students have not completed their homework; the teacher does not respond.</td>
<td>The teacher does not encourage students who are struggling.</td>
</tr>
<tr>
<td>The teacher’s primary concern appears to be to complete the task at hand.</td>
<td>Only some students get right to work after an assignment is given or after entering the room.</td>
</tr>
</tbody>
</table>

### PROFICIENT LEVEL 3

The classroom culture is a place where all value learning; high expectations for both learning and hard work are the norm for most students. Students understand their role as learners and consistently expend effort to learn. Classroom interactions support learning, hard work, and the precise use of language.

- The teacher communicates the importance of the content and the conviction that with hard work all students can master the material.
- The teacher demonstrates a high regard for students’ abilities.
- The teacher conveys an expectation of high levels of student effort.
- Students expend good effort to complete work of high quality.
- The teacher insists on precise use of language by students.

### DISTINGUISHED LEVEL 4

The classroom culture is a cognitively busy place, characterized by a shared belief in the importance of learning. The teacher conveys high expectations for learning for all students and insists on hard work; students assume responsibility for high quality by initiating improvements, making revisions, adding detail, and/or assisting peers in their precise use of language.

- The teacher communicates passion for the subject.
- The teacher conveys the satisfaction that accompanies a deep understanding of complex content.
- Students indicate through their questions and comments a desire to understand the content.
- Students assist their classmates in understanding the content.
- Students take initiative in improving the quality of their work.
<table>
<thead>
<tr>
<th>The teacher says, “This is important; you’ll need to speak grammatical English when you apply for a job.”</th>
</tr>
</thead>
<tbody>
<tr>
<td>The teacher says, “This idea is really important! It’s central to our understanding of history.”</td>
</tr>
<tr>
<td>The teacher says, “Let’s work on this together; it’s hard, but you all will be able to do it well.”</td>
</tr>
<tr>
<td>The teacher hands a paper back to a student, saying, “I know you can do a better job on this.” The student accepts it without complaint.</td>
</tr>
<tr>
<td>Students get to work right away when an assignment is given or after entering the room.</td>
</tr>
<tr>
<td>Students correct one another in their use of language.</td>
</tr>
<tr>
<td>The teacher says, “It’s really fun to find the patterns for factoring polynomials.”</td>
</tr>
<tr>
<td>A student says, “I don’t really understand why it’s better to solve this problem that way.”</td>
</tr>
<tr>
<td>A student asks a classmate to explain a concept or procedure since he didn’t quite follow the teacher’s explanation.</td>
</tr>
<tr>
<td>Students question one another on answers.</td>
</tr>
<tr>
<td>A student asks the teacher for permission to redo a piece of work since she now sees how it could be strengthened.</td>
</tr>
</tbody>
</table>
References


expectations, student sex, and time of year. *Journal of educational Psychology, 72*(3), 378.


stigmatizing information on children’s social interactions. *Journal of Personality and Social
Psychology, 63*(1), 41.


expectations to achievement in the early school years. *Journal of educational psychology, 101*(3), 662.

Hoge, R. D. (1984). The definition and measurement of teacher expectations: Problems and

of Educational Psychology, 76*(5), 777.


Achievement Tests by Content Area Using Teachers’ Ratings as Criteria. *Journal of Educational
Measurement, 22*(3), 177-182.


Rubie-Davies, C. M. (2008). Teacher beliefs and expectations: Relationships with student


