Title
Discrepancies Between Students' and Teachers' Ratings of Instructional Practice: A Way to Measure Classroom Intuneness and Evaluate Teaching Quality

Author
Dockterman, Daniel Milo

Publication Date
2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Discrepancies Between Students' and Teachers' Ratings of Instructional Practice: A Way to Measure Classroom Intuneness and Evaluate Teaching Quality

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Education

by

Daniel Milo Dockterman

2017

ABSTRACT OF THE DISSERTATION

Discrepancies Between Students' and Teachers' Ratings of Instructional Practice:
A Way to Measure Classroom Intuneness and Evaluate Teaching Quality

by

Daniel Milo Dockterman

Doctor of Philosophy in Education

University of California, Los Angeles, 2017

Professor Noreen M. Webb, Co-Chair

Professor Michael H. Seltzer, Co-Chair

Student surveys have gained prominence in recent years as a way to give students a voice in their learning process, and teacher self-reports have always been an effective instrument for revealing the planning, intentions, and expectations behind a given lesson. Though student and teacher surveys are widely used, extant research in education has primarily treated these ratings as separate sources of evidence. Little research in education has directly compared student and teacher reports or examined the potential predictive quality of teacher–student perceptual discrepancy. However, inconsistencies or "discrepancies" in perceptions across constructs has a rich history in organizational psychology and psychopathology.

Using similarly-worded student and teacher survey items from the Quality Assessment in

Science (QAS) Surveys, this dissertation explores whether the degree and direction of perceptional congruence may be revealing of instructional practice and teaching quality. Two principal research topics were investigated:

1) What are the various methods one can use to measure discrepancy between student and teacher ratings within classrooms?
2) How do these different methods of examining discrepancy between student and teacher ratings perform for different purposes?

The first research question was investigated within classrooms, principally by computing "unstandardized differences in means" measuring perceptual discrepancy between students and teachers. These unstandardized differences in means were used in conjunction with other scoring measures and plots of student and teacher item responses to investigate whether perceptual discrepancy is greater for some classes and some instructional practices. The second research question was investigated by comparing how the discrepancy rankings of classrooms and instructional practices changed depending on the employed method of discrepancy.

Teaching is complex and, consequently, as many measures as possible should be used to capture its multidimensionality. Considering the views of students in tandem with the views of their teachers may allow teaching quality to be examined through a wider lens. That is, perceptual discrepancies may act as a barometer for the synchronous relationship or "intuneness" between students and teachers. The results of this dissertation suggest analyzing student and teacher perceptions together can help researchers better understand and differentiate quality practice, thereby providing constructive feedback to teachers.

The dissertation of Daniel Milo Dockterman is approved.

James Stigler

Mark Hansen

Noreen M. Webb, Committee Co-Chair

Michael H. Seltzer, Committee Co-Chair

University of California, Los Angeles

2017

For my parents.

Table of Contents

1. Introduction 1
1.1 Importance of Teaching on Student Learning 1
1.2 Teacher Credentials Are Inadequate 1
1.3 New Measures of Teacher Practice 2
1.3.1 Value-Added Models 3
1.3.2 Classroom Observations 4
1.3.3 Portfolios 4
1.3.4 Student Surveys 5
1.3.5 Teacher Surveys 5
1.4 Using Student and Teacher Survey Responses in Combination 5
1.5 Research Goals and Questions 6
1.6 Chapter Overview 8
2. Surveys of Instructional Practice 9
2.1 Student Surveys 9
2.1.1 History of Use in Higher Education 9
2.1.2 Increasing Prominence of Student Surveys in Teacher Evaluation Frameworks 11
2.1.3 Students Are Natural Observers of the Classroom 13
2.1.4 Practical Advantages of Student Surveys as a Measure of Teaching Quality 14
2.1.5 Student Ratings Are Reliable 15
2.1.6 Student Ratings Are Valid 16
2.1.7 Unit of Analysis 18
2.2 Teacher Surveys 20
2.3 Chapter Summary 23
3. Discrepancy Research in the Social Sciences 24
3.1 Discrepancy in Psychology 24
3.1.1 Supervisor–Subordinate Congruence in Organizational Psychology 24
3.1.2 Parent–Child Discrepancy in Psychopathology 25
3.2 Discrepancy in Education 26
3.2.1 Item and Domain Comparisons Between Educational Stakeholders 27
3.2.2 Investigating Stakeholder Discrepancy Using Difference Scores 30

3.2.3 Section Summary 32
3.3 Discrepancy Methodologies 32
3.3.1 Limitations of Correlations 33
3.3.2 Methodological Shortcomings of Difference Scores 33
3.3.3 Polynomial Regression 36
3.4 Chapter Summary 36
4. Description of the Data Source 38
4.1 QAS Student and Teacher Matching Items 38
4.2 QAS Items Are Climate Variables 40
5. Is Discrepancy Greater for Some Classes than for Others? 41
5.1 Ranking Classes Using Teacher–Student and Teacher–Classroom Discrepancy Scores 43
5.2 Ranking Classes Using Unstandardized Differences in Means 47
5.2.1 Computing the Unstandardized Difference in Means 47
5.2.2 Computing the Standard Error of the Unstandardized Difference in Means 52
5.2.3 Ranking Classes by Statistically Significant Unstandardized Differences in Means 53
5.2.4 Section Summary 55
5.3 Ranking Classes by Composite Unstandardized Absolute Differences in Means 55
5.3.1 Exploratory Analysis of Discrepancy Across Classes Using Box Plots 56
5.3.2 Computing Composite Unstandardized Absolute Difference in Means 58
5.3.3 Ranking Classes by Composite Unstandardized Absolute Differences in Means 62
5.3.4 Section Summary 62
5.4 Chapter Summary 63
6. Is Class Discrepancy Dependent on the Instructional Practice Examined? 65
6.1 Dependency of Discrepancy Across Instructional Practice Items 65
6.1.1 Class Rankings Across All Items 66
6.1.2 Class Rankings Across Items #1 and #14 68
6.1.3 Section Summary 71
6.2 Dependency of Discrepancy Across Instructional Domains 71
6.2.1 Class Rankings Across All Instructional Practice Domains 72
6.2.2 Class Rankings Across Experimental Work and Assessment Instructional Practice Domains 74
6.2.3 Section Summary 78

6.3 Examining Discrepancy in the Most and Least Discrepant Classes 78
6.4 Chapter Summary 85
7. Is Discrepancy Greater for Some Instructional Practices than for Others? 86
7.1 Comparing Instructional Practices Using Proportional Scoring Methods 86
7.2 Comparing Instructional Practices Using Unstandardized Differences in Means 90
7.3 Comparing Instructional Practices Using Item Composite Unstandardized Differences in Means 92
7.3.1 Exploratory Analysis of Discrepancy Across Items Using Box Plots 93
7.3.2 Computing Item Composite Unstandardized Differences in Means 94
7.3.3 Ranking Instructional Practices by Item Composite Unstandardized Differences in Means 96
7.4 Examining Discrepancy for the Most Discrepant Instructional Practices 98
7.5 Chapter Summary 102
8. Discussion 104
8.1 Summary of Findings 104
8.2 Meaningfulness of "Intuneness" as a Construct 105
8.3 Practical Significance of Classroom Discrepancy Research 108
8.3.1 The Case for a Multiple Measures Approach Based on Feasibility 108
8.3.2 Discrepancy Can Reveal a More Complete Picture of Teacher Performance 111
8.3.3 Discrepancy Can Help to Further Differentiate Teacher Practice 112
8.3.4 Discrepancy Can Provide Teachers with Feedback for an Untapped Construct 114
8.4 Limitations 116
8.5 Future Directions 118
8.5.1 Predictiveness of Discrepancy 118
8.5.2 Source of Discrepancy 119
8.5.3 Qualitative Follow-Up 120
8.5.4 Methodological Refinements 121
8.6 Conclusion 123
Appendix: Additional Tables and Figures 126
References 144

List of Figures

Figure 5.1 Student and class composite absolute discrepancy scores, by class……44

Figure 5.2 Classroom counts of significant unstandardized differences in means……....54

Figure 5.3 Classroom box plots of unstandardized differences in means and unstandardized absolute differences in means……..…….....…..57

Figure 5.4 Class composite unstandardized absolute differences in means…….…....59

Figure 6.1 Unstandardized differences in means by classroom, Item #1…….…….....70

Figure 6.2 Unstandardized differences in means by classroom, Item #14…..……..…….....71

Figure 6.3 Unstandardized differences in means and unstandardized absolute differences in means, by classroom, Items #15–#19 in Experimental Work domain.....…..75

Figure 6.4 Unstandardized differences in means and unstandardized absolute differences in means, by classroom, Items #20–#23 in Assessment domain.....…..77

Figure 6.5 Class discrepancy plot: Classroom #21…….....…..80

Figure 6.6 Class discrepancy plot: Classroom #26.……...……...…..81

Figure 6.7 Class discrepancy plot: Classroom #17.……...……...…..83

Figure 6.8 Class discrepancy plot: Classroom #15.……...……...…..84

Figure 7.1 Item counts of significant unstandardized differences in means……..…….....91

Figure 7.2 Item box plots of unstandardized differences in means……...…….....93

Figure 7.3 Item composite unstandardized differences in means…….…….....96

Figure 7.4 Item discrepancy plot: Item #1……..…….....…..100

Figure 7.5 Item discrepancy plot: Item #4……..…….....…..101

Figure A.1 Unstandardized differences in means and unstandardized absolute differences in means, by classroom, Items #2, #6, and #7 in Individual Work domain.....132

Figure A.2 Unstandardized differences in means and unstandardized absolute differences in means, by classroom, Items #3, #8, and #9 in Interactive Work domain.....133

Figure A.3 Unstandardized differences in means and unstandardized absolute differences in means, by classroom, Items #4 and #5 in Discussion domain…...….……....134

Figure A.4 Unstandardized differences in means and unstandardized absolute differences in means, by classroom, Items #10–#13 in Reports and Projects domain….....135

Figure A.5 Forest plot of unstandardized differences in means by item, Class #21……136

Figure A.6 Forest plot of unstandardized differences in means by item, Class #26……137

Figure A.7 Forest plot of unstandardized differences in means by item, Class #17……138

Figure A.8 Forest plot of unstandardized differences in means by item, Class #15……139

Figure A.9 Classes with similar unstandardized differences in means on Items #1 and #4 yet different underlying patterns in student response variation.....143

List of Tables

Table 1.1 Comparison of prominent student classroom perception surveys……..…...12

Table 4.1 QAS matching student and teacher survey items…...……...39

Table 5.1 Chapter organizer: research questions by measure of discrepancy……..…….42

Table 5.2 Sample calculation of composite absolute discrepancy scores.....45

Table 6.1 Classroom unstandardized absolute difference in means rankings for each item.67

Table 6.2 Classroom unstandardized absolute difference in means rankings for each QAS domain……...…….....73

Table 7.1 Proportional and mean comparison of student and teacher ratings, by item...…..89

Table A.1 Student ratings on QAS survey, descriptive statistics…….126

Table A.2 Teacher ratings on QAS pre-survey, descriptive statistics ……127

Table A.3 Teacher ratings on QAS post-survey, descriptive statistics ……128

Table A.4 Student variance within and between classes, by QAS item……...129

Table A.5 Unstandardized differences in means, by QAS item.……..…130

Table A.6 Confirmatory factor analysis of student ratings of QAS items.……...131

Table A.7 Proportion of students within a classroom who rated each QAS item lower than their teacher …….……...……..….....140

Table A.8 Proportion of students within a classroom who rated each QAS item equivalently as their teacher ……...…….……...……..….....141

Table A.9 Proportion of students within a classroom who rated each QAS item higher than their teacher …….……...……..….....142

Acknowledgements

I gratefully acknowledge the guidance of my advisors and committee co-chairs, Dr. Michael Seltzer and Dr. Noreen Webb, who provided editorial guidance and methodological support throughout the dissertation process. I am also thankful to the other members of my committee, Dr. Mark Hansen and Dr. James Stigler, for their mentorship and feedback.

In addition, I would like to thank Professor José Felipe Martínez for providing access to the data used in this dissertation and for all he taught me about the field of teacher practice and about thinking critically.

I am very grateful for the support from past UCLA colleagues: Alejandra Priede Schubert, Glory Tobiason, Jon Schweig, Scott Monroe, and Larry Thomas. I am also especially thankful to my SRM cohort: Megan Kuhfeld, Jane Li, Kevin Schaaf, Jenn Ho, Jason Tsui, and Liz Perez. I am truly fortunate to have shared my time in graduate school with all of you.

Finally, I want to thank my parents and my uncles, David and Michael Dockterman, for their constant encouragement and optimism. Without my family's unwavering support, this would not have been possible.

Vita

Education

2006 Bachelors of Arts, University of Virginia, Psychology and Economics.

2012 Masters of Arts, University of California, Los Angeles, Education.

Work

2007 Mathematics Teacher, Shell Bank Junior High School, Brooklyn, NY.

2008–2011 Research Assistant, Pew Research Center, Washington, DC.

2011–2017 Graduate Student Researcher, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), UCLA

Publications

Schaaf, K. & Dockterman, D. (2014). VAM in Greek, English, and implication: Explanations of different models and their effects on aggregate and individual teacher outcomes. *InterActions*: *UCLA Journal of Education and Information Studies*, 10(1). Retrieved from http://escholarship.org/uc/item/5kq9j901.

Nava, I., Park, J., Kawasaki, J., Dockterman, D., Quartz, K., J.F. Martinez & Schweig, J.D. (under review). Measuring teaching quality of secondary mathematics and science residents: A classroom observation framework and pilot generalizability study. *Journal of Teacher Education*.

Wang, J. & Dockterman, D. (under review). A literature review of the effect of magnet schools on student achievement via a quasi-experimental lens. *Journal of School Choice*.

Dockterman, D. Novice and mentor teacher experiences in an urban teacher residency program: A final report of IMPACT survey findings from 2010–2014. *Xpress Working Papers*. Retrieved from https://ucla.app.box.com/s/daed0v3ggmrmf9hwa99i.

CHAPTER 1

1. Introduction

1.1 Importance of Teaching on Student Learning

Since the turn of the century, teacher evaluation has become an increasingly prominent educational policy issue in the United States. A chief reason for this is the body of evidence indicating wide variation in student learning experiences and achievement by classroom (Aaronson, Barrow, & Sander, 2007; Hanushek & Rivkin, 2006; Rockoff, 2004). Indeed, although factors outside the classroom—and typically beyond the scope of educational policy—contribute greatly to student academic performance (e.g., family background, community experiences, peer effects, aptitudes for schooling), teaching quality has consistently been identified as the most influential school-based factor on student learning gains (McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Haertel, 2013). Variation in teaching quality results in different levels of achievement gains, which are associated with significant long-term benefits including higher college attendance rates, higher salaries, better savings habits, and lower teen pregnancy rates (Chetty, Friedman, & Rockoff, 2011). As noted by Sanders, Wright, and Horn (1997), "more can be done to improve education by improving the effectiveness of teachers than by any other single factor" (p. 3).

1.2 Teacher Credentials Are Inadequate

In response to mounting research revealing the important effect teachers have on student learning, and aided by monetary incentives provided by the U.S. Department of Education's 4.35 billion dollar Race to the Top initiative, states and local school districts have been rapidly developing new teacher evaluation systems. As Cochran-Smith (2010) comments, "as far as education goes, we live in an age of accountability" (p. xiii). This recent push for performance-based standards is due in part to research showing that traditional measures of teacher effectiveness (based on qualifications) inadequately differentiate teaching quality. Several

studies have found that teachers' degree of formal educational attainment, certification status, and years of experience are only minimally related to student achievement (see e.g., Rivkin, Hanushek, & Kain, 2005; Kane, Rockoff, & Staiger, 2008; Clotfelter, Ladd, & Vigdor, 2007). Additionally, intermittent on-the-job evaluations of teachers performed by an administrator have also failed to reveal meaningful variation in teachers' ability to produce student achievement gains (Toch & Rothman, 2008). As a result, teacher evaluation systems have not sufficiently differentiated teachers more effective at raising student achievement from those less successful. Due to the shortcomings of these evaluation systems, salary and high-stakes employment decisions have been based, instead, on signals or correlates of teaching quality, which are largely unrelated to student achievement (Harris, 2009).

The inability of evaluation systems based on teacher qualifications (educational attainment, certification status, and seniority) to differentiate between levels of teaching performance has been coined "the Widget Effect" (Weisberg, Sexton, Mulhern, & Keeling, 2009). Under these traditional evaluation systems, nearly all teachers were viewed as being of similar quality, and were simply rated as satisfactory. That is to say, old evaluation systems were "undifferentiated, unhelpful, and inconsequential" (Teacher Evaluation 2.0, 2010, p. 1). As a result, exceptional teaching went unrecognized and many struggling teachers did not receive professional development, which could have helped them to improve. Despite the lack of evaluation systems to effectively capture variation in teaching performance, it is nonetheless clear that individual teachers are not of interchangeable quality. Over the past decade, this realization has led to a call for new performance measures related to student learning growth to assess teachers. This has resulted in what Mihaly, McCaffrey, Staiger, and Lockwood (2013) called, "a shift in focus from teacher qualifications to teacher effectiveness" (p. 3).

1.3 New Measures of Teacher Practice

Several measures of teaching quality or effectiveness have been developed in recent years with the specific aim of capturing individual teacher-level differences. The belief is that these

measures of classroom practice will help practitioners and policymakers unlock the "*black box* of the classroom" (Correnti & Martinez, 2012, p. 51) and better comprehend the complex relationship between quality instruction and student learning. Measures of teacher practice can be subdivided into two categories: student achievement and professional teacher practice (Partee, 2012). Broad consensus exists among researchers about the need for a multiple-measure approach in order to capture a multitude of student achievement, perception, and attitudinal outcomes (Baker et al., 2010; Braun, Chudowsky, & Koenig, 2010; Kane, Kerr, & Pianta, 2014).

1.3.1 Value-Added Models

A prominent and contentious measure of teacher practice is value-added analysis of student achievement data. Value-added models are a type of longitudinal growth modeling techniques that seek to measure a teacher's performance by comparing his or her students' growth on standardized test scores to the average growth of other students (Ballou, Sanders, & Wright, 2004). To do this, these models isolate the contribution (or "value added") that each teacher provides to students in a given year by controlling for the effects of background characteristics not under his or her control, such as prior student performance and socioeconomic status. In short, this modeling process allows researchers to investigate questions such as "What proportion of the observed variance can be attributed to a school or teacher?" and "How effective is an individual school or teacher at producing [student test score] gains?" (Doran & Lockwood, 2006, p. 206).

Although value-added models are increasingly used for summative evaluation to hold teachers accountable for how much their students are achieving, their predictive strength is limited and far from a panacea for measuring teaching quality. Value-added models inherently suffer from a host of serious confounds, including school and district influences (Baker et al., 2010), the ubiquitousness of non-ignorable missing data (McCaffrey & Lockwood, 2011), and the non-random assignment of students to classrooms (Rothstein, 2010). Furthermore, it is important to remember that value-added models use standardized test scores as their primary

data—scores that are only a proxy for student learning (Kennedy, 1999) and only a relevant measure in grades and academic subjects in which they are administered.

1.3.2 Classroom Observations

Classroom observations are another useful "first-level approximation" (Kennedy, 1999, p. 346) of student learning, and are often considered the "gold standard" method of collecting teaching practice data (Rowan & Correnti, 2009, p. 121). Classroom observations can be conducted across grade levels and academic subjects to provide rich information about classroom behavior and teachers' professional practices. However, classroom observations are expensive and labor intensive due to the time necessary for training and calibrating raters to specific rubrics such as CLASS (Pianta & Hamre, 2009) or Framework for Teaching (Danielson, 1996). Furthermore, raters must spend considerable time inside the classroom observing teachers; typically more than one rater at a time observes a given class period, and ratings must be conducted over multiple occasions in order to increase their reliability. Additionally, daily observations capture only a small portion of an entire curriculum and thus tend to be limited in scope (Goe, Bell, & Little, 2008).

1.3.3 Portfolios

Analyses of student work (e.g., assignments, homework, artifacts, and projects) and teacher lesson plans collected in portfolios are another alternative measure of teaching quality. In these portfolios teachers compile rich information about classroom practice, which is then scored across domains on a rubric by trained raters. Portfolios can serve as a "window into actual classroom practice" without incurring the full rater training costs inherent to classroom observations (Goe et al., 2008, p. 28). However, like classroom observations, the accuracy of raters' scoring is paramount to maintaining the validity of portfolios. Thus, sufficient training and calibration with the rubric, as well as a background in the subject matter, is important for raters. Although they are less popular than other measures, the reliability of portfolios is similar to that of classroom observations (Martinez, Borko, & Stecher, 2012). The collection process can

be useful because it facilitates teacher self-evaluation and serves as a form of professional development. However, portfolios require considerable resources to develop and score as well as a significant time commitment from teachers. Like classroom observations, the feasibility of this measure hinges on whether its formative assessment value outweighs the substantial time and monetary costs incurred in its execution.

1.3.4 Student Surveys

Student surveys are increasingly recognized as a source of teaching quality evidence, and have the singular advantage of being the only measure to reveal the perceptions of students. Student surveys are cost-effective relative to other evaluation methods, and their use has become more popular in recent years as a complementary measure to value-added models and classroom observations. Researchers have found that student surveys provide information about instructional quality that can be both reliable (Raudenbush, 2013; Peterson, Wahlquist, & Bone, 2000) and valid (Kane & Staiger, 2012; Kyriakides, 2005). A complete discussion of the historical use and methodological characteristics of student surveys is presented in Section 2.1.

1.3.5 Teacher Surveys

Like other teacher self-report measures (e.g., logs and interviews), teacher surveys have the advantage of capturing the direct attitudes and beliefs of the teacher. Whereas classroom observations examine how a lesson unfolds in practice and portfolios uncover how the students respond to and comprehend the lesson, teacher self-reports can reveal how the lesson should have proceeded in theory. However, though teacher surveys clearly provide a valuable pedagogical perspective, the validity of this measure remains a concern due to social desirability effects (Goe et al., 2008). Section 2.2 provides greater background on the methodological qualities of teacher surveys.

1.4 Using Student and Teacher Survey Responses in Combination

Once collected, these various measures of instructional practice are either reported separately or combined into a weighted composite measure of a teacher's overall quality

(Cantrell & Kane, 2013). Likewise, student and teacher surveys have traditionally been considered two wholly separate measures of teacher efficacy. Yet a classroom is a dynamic environment with constant teacher–student interaction, and so considering the perceptions of students in conjunction with those of their teachers may provide a means of examining the complex construct of teaching quality through a wider lens. Moreover, given that student survey responses exhibit substantial within-class variation and teacher survey responses are susceptible to social desirability biases, utilizing each of these measures together may reveal a fuller and more complete picture of what goes on inside the classroom. Both student and teacher perceptions are independently important to understanding teaching quality, but what if an exploration of the degree of alignment between them reveals additional information of practical importance? In other words, could the whole of information acquired from analyzing the alignment and discrepancy of student and teacher perceptions be greater than the information obtained from analyzing each survey measure separately?

1.5 Research Goals and Questions

The goal of this dissertation is to assess the value of considering student and teacher surveys in conjunction by examining the congruence (or lack of congruence, herein called "discrepancy") between student and teacher survey responses. The utility of comparing student and teacher perceptions is investigated using various scoring and graphical methods in order to gain a richer understanding of teacher practice and "intuneness" with their students. The study of perceptual alignment has a long tradition in clinical psychology, which has primarily considered alignment between parent and child, as well as in organizational psychology, which has focused on alignment between supervisor and subordinate. Greater perceptual congruence between these parties has been associated with a range of positive outcomes, whereas greater discrepancy has been shown to exhibit opposite effects.

Few studies in education have directly compared student and teacher reports or explored the role of discrepancy in teachers' and students' characterization of instructional practices. The

purpose of the present inquiry is to use various methods to comprehensively explore whether the degree and direction of perceptual congruence may be revealing of instructional practice and teaching quality. The value of this research is proof of concept. By using various methods to measure discrepancy, one can determine if comparing student and teacher classroom perceptions is worthy of additional empirical study, and whether congruence has formative value in teacher evaluative frameworks.

To study these research questions, student and teacher parallel items (i.e., items worded similarly and asked on the same scale) were utilized from the Quality Assessment in Science Surveys (QAS). The QAS surveys were administered to eighth grade students and their teachers in science classrooms. The matching items on the QAS surveys pertain to instructional practices (e.g., frequency of working on worksheets, frequency of conducting experiments in class). It is the tangible nature of these practices that makes them ideal for studying perceptual congruence.

The research objective of this dissertation can be divided into two components. First, various methods—including proportional scoring, computations of unstandardized differences in means, and plots of item responses—are used to explore whether perceptual discrepancy between students and teachers is greater for some classes and some instructional practices. Additionally, this dissertation examines whether the degree of perceptual difference across classes is dependent on the particular instructional practices. Investigating these questions of congruence is complex, as the considerable variation in student responses within classrooms presents intrinsic methodological challenges. Whereas other types of congruence research (e.g., supervisor–subordinate, parent–child, and principal–teacher) involve relatively straightforward, one-to-one comparisons, teachers' perceptions must be compared with the perceptions of their entire classroom of students (i.e., one perception with that of many).

Secondarily, this dissertation explores how these different methods of examining the discrepancy between student and teacher ratings perform for difference purposes. This question was investigated by comparing how the discrepancy rankings of classrooms and instructional

practices change depending on the method used to measure perceptual discord. Modest differences observed in class and instructional practice rankings could be explained by whether a method accounted for student and teacher response variation as well as whether a method measured only magnitudinal differences (and not directional discrepancy).

1.6 Chapter Overview

This dissertation is organized as follows: Chapter 2 provides an overview of student and teacher surveys of instructional practice with a particular focus on the validity and reliability of these teaching quality instruments. Chapter 3 reviews discrepancy literature in the social sciences and compares two prominent methodologies for measuring discrepancy: difference score computations and polynomial regression. Chapter 4 introduces the QAS matching student and teacher surveys. Chapter 5 examines whether discrepancy is greater for some classrooms than for others. Chapter 6 investigates whether discrepancy between classes is dependent on the instructional practice. Chapter 7 examines whether discrepancy is greater for some instructional practices than for others. In total, Chapters 5–7 employ six methods to address these research questions: proportional scoring, discrepancy scores, unstandardized differences in means, class composite unstandardized absolute differences in means, item composite unstandardized differences in means, and discrepancy plots. Finally, Chapter 8 summarizes the findings, provides strengths and limitations of discrepancy research, explains its practical significance, and outlines directions for future research.

CHAPTER 2

2. Surveys of Instructional Practice

This chapter reviews the literature on and discusses the important methodological issues of student and teacher surveys of instructional practice. Though student surveys have primarily been used in higher education to give a voice to students, this measure has gained prominence in recent years as a cost-effective alternative component in teacher evaluation frameworks. Teacher self-reports, though less widely used, have significant formative assessment value by revealing teacher intentions and expectations. Particular attention is given to the validity and reliability of these measures.

2.1 Student Surveys

2.1.1 History of Use in Higher Education

Collecting student perceptual information through surveys of their learning environments can be useful for providing feedback to teachers about the strengths and weaknesses of their instructional practices. Historically, student surveys have been used predominantly in higher education. Consequently, most published research on this instrument reflects this reality. Indeed, a meta-analysis conducted by Kyriakides (2001) on student ratings revealed that the overwhelming majority of empirical investigations were concerned with their use in higher education as opposed to the K–12 setting. These studies have shown student surveys to be both a reliable and valid measure of instructional quality at the university level. A review of the literature by Aleamoni (1999) showed substantial year-to-year correlations in student ratings of instructors (0.87 to 0.89) and Murray (1997) concluded that student ratings of instructors were highly stable across "items, groups of raters, time periods and courses" (p. 8). Similarly, Marsh and Hocevar (1991) found undergraduate and graduate student ratings of faculty to be relatively consistent over a 13-year period, even after accounting for years of teaching experience. The authors concluded "teaching effectiveness as perceived by students is stable" (p. 312).

Though evaluating the validity of student ratings is less straightforward, survey items and subscales have been judged by expert researchers to measure important aspects of instruction (Costin, Greenough, & Menges, 1971). Marsh (1984) concluded that the primary factor affecting student ratings was the characteristics of the instructors, not of the courses. Providing further validity evidence, student ratings have been shown to be moderately to highly correlated with other indicators of instructor quality, such as instructor self-ratings, colleague ratings, alumni ratings, and other student learning measures (Aleamoni, 1999).

The K–12 literature examining student ratings is scarcer than the higher education literature because most states and districts have only recently started administering student surveys. However, extant research suggests that K–12 students are fully capable of accurately assessing their teachers if survey items are low-inference. For example, in measuring a teacher's lesson planning, the item, "My teacher reviews what we have just learned before the lesson is over" is more concrete and therefore lower in inference than the item, "My teachers presents well-designed lessons", which is inherently vague. Administering a survey composed of low-inference items, Worrell and Kuterbach (2001) found the teacher ratings of high-achieving high school students were as reliable and valid as the teacher ratings of college students. Providing further evidence of the validity of student surveys, Peterson and colleagues (2000) found that high school students could identify whether the learning environment in the classroom was more teacher- or student-centered. The authors believed students did not consider the surveys of their teachers to be popularity contest, and could adequately distinguish between likeability and the ability to enable their learning. This finding is significant because halo effects—systematic biases caused by generalizing the perception of a person's specific trait to an overall evaluation of their personality (Thorndike, 1920)—can be a legitimate concern with surveys. Nonetheless, Peterson et al. (2000) concluded that K–12 students responded to survey items with "reason, intent, and consistent values" (p. 148).

2.1.2 Increasing Prominence of Student Surveys in Teacher Evaluation Frameworks

As research has regularly found student surveys to be methodologically sound, these instruments have become an increasingly adopted source of evidence in teacher evaluation. Tripod is the predominant student survey currently being administered. Used in the Bill and Melinda Gates Foundation's 45 million dollar Measures of Effective Teaching (MET) Project, the Tripod survey consists of seven classroom-level domains (care, clarify, control, challenge, captivate, confer, and consolidate) believed to be related to quality teaching. Between the fall of 2011 and spring of 2015, Tripod surveys were used by 106 partners (at the school, district, and state level) across 29 states. This included 9,579 schools, 148,920 teachers, and 226,953 classrooms, with a total of 4,580,314 student surveys completed (Tripod email correspondence, September 4, 2015). Analysis by Schweig (2014b) also found that of the states and local districts citing a specific survey instrument as part of their teacher evaluation systems, 75% were using the Tripod survey.

Though states and districts have overwhelmingly elected to use Tripod, there are several other student survey instruments available (see Table 1.1). The STeP survey, a component of My Student Survey of teacher practice, asks students about their teachers' classroom roles: presenter ("ability to present information and structure lessons"), manager ("ability to manage a classroom and foster productivity"), counselor ("awareness of student need and teacher–student relations"), coach ("providing feedback and challenging students"), motivator ("engaging and investing students in learning"), and content expert ("knowledge of subject and encouraging student thinking") (My Student Survey, 2016). Another prominent measure, the YouthTruth Feedback for Teachers survey, gathers student feedback across six classroom-level domains—relevance, classroom culture, academic rigor and expectations, student engagement, instructional methods, and personal relationships (YouthTruth Student Survey: Design and Methodology, 2016)—whereas the iKnow My Class survey is designed to measure students' perceptions across eight domains of their teachers' instructional practice: engagement, relevance, relationships, class efficacy, cooperative learning environment, critical thinking, positive pedagogy, and discipline

Table 1.1: *Comparison of Prominent Student Classroom Perception Surveys*

Instrument	Domains	Response Scale	Example Items
Tripod Student Survey • Developed by Ronald Ferguson at Harvard University • Versions for grades K–2, 3– 5, and 6–12	1. Care 2. Control 3. Clarify 4. Challenge 5. Captivate 6. Confer 7. Consolidate	1. No, never/Totally untrue 2. Mostly not/ Mostly untrue 3. Maybe, sometimes/ Somewhat 4. Mostly yes/Mostly true 5. Yes, always/Totally True	**Care:** My teacher really tries to understand how students feel about things. **Clarify:** My teacher knows when the class understands, and when we do not. **Control:** Our class stays busy and doesn't waste time. **Challenge:** In this class, my teacher accepts nothing less than our full effort. **Captivate:** My teacher makes lessons interesting. **Confer:** My teacher gives us time to explain our ideas. **Consolidate:** My teacher takes the time to summarize what we learn each day.
My Student Survey: STeP Teacher Practice Survey • Developed by Ryan Balch at Vanderbilt University • Versions for grades 4–5 and 6–12	1. Counselor 2. Manager 3. Coach 4. Motivator 5. Content Expert 6. Presenter	1. Never 2. Sometimes 3. Often 4. Almost always 5. Every time	**Counselor:** If I do not understand something in class, my teacher explains it in a different way to help me understand. **Manager:** We are learning or working during the entire class period. **Coach:** I have to work hard to do well in this class. **Motivator:** My teacher has us apply what we are learning to real-life situations. **Content Expert:** After asking us questions, my teacher lets us think for a few seconds before we have to answer. **Presenter:** My teacher reviews what we have just learned before the lesson is over.
YouthTruth: Feedback for Teachers Survey • Developed by the Center for Effective Philanthropy • Versions for grades 6–8 and 9–12	1. Relevance 2. Classroom Culture 3. Academic Rigor & Expectations 4. Student Engagement 5. Instructional Methods 6. Personal Relationships	1. Strongly disagree 2. Somewhat disagree 3. Neither agree nor disagree 4. Somewhat agree 5. Strongly agree	**Relevance:** How much do you think your teacher cares about you? **Classroom Culture:** How much is student behavior under control in this class? **Academic Rigor & Expectations:** When the work gets difficult, how hard does your teacher expect you to try? **Student Engagement:** How often do you enjoy coming to this class? **Instructional Methods:** How often does your teacher ask students to explain more about answers they give? **Personal Relationships:** The teacher presents lessons in ways I understand.
Quaglia School Voice: iKnow My Class Survey • Developed by Russell Quaglia at the Quaglia Institute for Student Aspirations • Version for grades 3–5 and 6–12	1. Engagement 2. Relevance 3. Relationships 4. Class Efficacy 5. Cooperative Learning Environment 6. Critical Thinking 7. Positive Pedagogy 8. Discipline Problems	1. Strongly disagree 2. Disagree 3. Undecided 4. Agree 5. Strongly agree	**Engagement:** I put forth my best effort in class. **Relevance:** I understand how I can apply what I am learning in my everyday life. **Relationships:** The teacher cares if I am absent from class. **Class Efficacy:** I am comfortable being myself in this class. **Cooperative Learning Environment:** The teacher encourages students to work together **Critical Thinking:** I explore issues, events, and problems from different perspectives **Positive Pedagogy:** The teacher presents lessons in ways I understand **Discipline Problems:** Discipline is a problem in this class.

problems (iKnow My Class Survey, 2016). Therefore, instructional dimensions of support, engagement, awareness, organization, rigor, and discipline are all common across student survey measures, but with different item wording and response scales. In general, student ratings are weighted to account for 5% or 10% of a teacher's overall evaluation, though variation in this assigned weight exists across states and districts.

Proponents cite several key reasons for including student surveys in teacher evaluation frameworks. These include: 1) students are natural observers of the classroom; 2) student surveys hold several practical advantages over other measures of instructional practice (e.g., classroom observations and value-added models); 3) students can reliably discriminate between teacher abilities; and 4) student ratings of teachers are correlated with student achievement. The following sections describe each of these arguments in greater detail.

2.1.3 Students Are Natural Observers of the Classroom

Student surveys have the singular advantage of revealing the perspective of the consumers and direct recipients of the educational practices in question. As Follman (1992) argues, "students are clients," and therefore, they "should play a meaningful role in the evaluation of their teacher" (p. 169). Administering surveys gives students a voice and a sense of classroom democratization. As one New York City teacher remarked of her students, "They're the ones that are in the room. As many walkthroughs [by administrators] as you have, the students are the ones who see it all" (Decker, 2012, para. 13). Furthermore, as stated by the MET Project (2012), "no one has a bigger stake in teaching effectiveness than students. Nor are there any better experts on how teaching is experienced by its intended beneficiaries" (p. 2). Given that students' classroom experiences are inextricably tied to their teachers, it seems sensible to provide them with some agency in the learning process.

Additionally, students are of course the evaluators who have the greatest contact with teachers (Goe et al., 2008). Many students spend more of their waking hours in a classroom than at home (Pianta & Hamre, 2009). Over the course of a school year, students spend hundreds of

hours with their teachers—far more time than any trained classroom observer spends with any one teacher. Consequently, no number of observations can replace the breadth of direct knowledge gleaned by students about the daily classroom practices and routines they experience (Stronge & Ostrander, 1997). Students' perceptions provide first-hand accounts of the overall classroom environments and rapport developed with teachers (Aleamoni, 1981). Moreover, students' ratings benefit from several instances of comparison due to their significant experience in other classrooms observing current and former teachers (Peterson, 1987). Put simply, no measure of instructional practice can substitute the depth of information revealed by studying student perceptions.

2.1.4 Practical Advantages of Student Surveys as a Measure of Teaching Quality

Beyond providing unique and extensive evidence about teacher instruction, there are additional strengths inherent to student surveys that make them an attractive alternative or complement to other measures of teacher evaluation. Student surveys are easy to administer, cost-effective, and non-intrusive. They allow for observation of a large number of classrooms across a broad range of practices. Relative to trained rater observations or portfolio measures (because student surveys do not require rater training and calibration), they have the potential to provide instructional practice information more inexpensively and efficiently. Student surveys can also help assess teachers in elective courses or in non-tested grades where value-added measures are often unavailable (Balch, 2012). And, whereas value-added measures are primarily administered for summative purposes to evaluate job performance, student survey responses can provide valuable formative information to teachers for targeted professional development. As a National Teacher of the Year explained, collecting student feedback can be particularly vital to providing first-hand information about practice:

> "They are the experts about what goes on in the class. Even if I intended it to come out one way, if that's not how they perceive it, that's not reality. Certainly students also bear responsibility for that reality, but their perception is our reality. So my

intentions are not as important as their expertise…There are some things that I can't do better if they don't tell me" (*MET Project Q & A*, 2012, p. 1).

2.1.5 Student Ratings Are Reliable

In order for student surveys to effectively measure teacher instructional quality, aggregate ratings of teachers must be stable across academic years and between class periods (i.e., between different groups of students in separate classes). And yet, there is widespread (if unsupported) concern that students lack the capacity to be consistent in their ratings. Indeed, Aleamoni (1999) cites the belief that "students cannot make consistent ratings about the instructor and instruction because of their immaturity, lack of experience, and capriciousness" (p. 1) as a prevalent myth about student surveys. However, evidence indicates students can reliably distinguish between the practices of different teachers. Recent research of student perceptions from the Tripod survey found stability in student aggregate ratings during a school year. Corrected for measurement error, correlations in the seven classroom-level domains of Tripod ranged from 0.70 to 0.85 between months. In addition, a composite of the Tripod indices was found to be a fairly stable indicator of the overall learning environment, as 78% of classrooms shifted by no more than one decile between months (Ferguson, 2010). Because of Tripod's stability over time, MET Project's culminating report recommended assigning a weight to student surveys equivalent to that of value-added models and teacher observations in order to create the most reliable composite metric (Cantrell & Kane, 2013).

In additional to being stable, student aggregate ratings are also internally consistent. Peterson and colleagues (2000) found the internal consistency of student ratings to be similar across grade levels, and Follman (1995) found teacher ratings of elementary school students were as internally consistent as the teacher ratings of high school students. Likewise, Raudenbush (2013) found teacher-level aggregate ratings collected with Tripod were internally consistent, with reliabilities ranging from 0.74 to 0.81 across survey domains. Still, classroom-aggregated student ratings, while internally consistent and stable, can be highly inter-correlated. Correlations across Tripod domains ranged from 0.56 to 0.95 (Raudenbush, 2013). This is

because high internal consistency is often the result of halo effects, as students frequently cannot or choose not to differentiate between different dimensions of teaching.

2.1.6 Student Ratings Are Valid

Sax (1997) defines validity as the degree to which measurements can be utilized in making decisions and providing explanations applicable to a given purpose. For student perceptions of teacher instruction to be a worthwhile measurement tool, ratings must capture some aspects of teaching quality, yet there is no all-encompassing definition or criterion of this construct. Thus, as Benton & Cashin (2012) state, "the best one can do is to try various [measurement] approaches, collecting data that either support or contest the conclusion that student ratings reflect effective teaching" (p. 3). In this case, the validity of student surveys can be evaluated by its content (i.e., whether the tool is actually measuring teaching quality) and by its ability to predict other accepted measures of teaching quality, such as classroom observation ratings or, more commonly, teacher value-added estimates.

Student survey responses have consistently been shown to correlate with student achievement. Conducting research in Cyprus, Kyriakides (2005) found student surveys to be strongly associated with increases in student achievement in both mathematics and Greek language. Similarly, Wilkerson, Manatt, Rogers, and Maughan (2000) found student survey responses to be significantly correlated to student achievement in reading (0.75), whereas teacher self-ratings, principal ratings, and principal summative evaluations were not. Research has also shown student perceptions of teachers to be related to non-testing teaching quality outcomes. For instance, piloting the now widely used My Student Survey, Balch (2012) found that higher survey ratings correlated with student academic engagement and self-efficacy.

Analyses of student responses from the Tripod survey are also moderately predictive of student outcomes. Raudenbush (2013) found that student perceptions could explain 7.8% of the variability in student learning gains in a different year, while each of the seven dimensions of Tripod independently yielded significant predictions of teachers' value-added scores, with the

dimensions' control and challenge showing the highest correlation. Furthermore, the MET Project found that student perceptions of their teacher predicted learning gains in different class periods taught by the same teacher.

Interestingly, these low correlations between Tripod and teacher value-added estimates also suggest that student surveys capture different components of quality teaching than those measured by standardized tests. Significant correlations between student ratings and teacher value-added scores do not necessarily prove the validity of this measure. It is easy to imagine that students doing well may be more likely to provide higher ratings under the assumption their teacher is doing a good job, while struggling students may be more likely to fault their teacher regardless of his or her instructional quality. Thus, the significant correlations of student survey ratings with standardized tests and other outcomes provide supporting—though not incontrovertible—evidence that students can validly evaluate their teachers.

Assigning some portion of teacher assessment to student perceptions is, understandably, an uncomfortable proposition. It is natural to question whether students have the maturity and overall awareness or their classroom environment to rate their teachers' behaviors, such as degree of support, instructional engagement, lesson rigor, and discipline. Indeed, these reservations are perhaps why student surveys have only recently been adopted in non-higher education settings. However, studies have consistently shown that students can validly evaluate their teachers' practices when the survey items are low-inference. Comparing student ratings to the behavioral ratings teachers routinely make of students, Worrell and Kuterbach (2001) state why this finding should not be wholly unexpected. "It is perhaps not surprising that students can also provide accurate ratings of teacher behavior as students spend as much time observing their teachers as their teachers spend observing them" (p. 245). Though collecting student perceptions of teachers may appear to be an exercise fraught with bias, the consistently positive associations found between student reports and other accepted measures of teaching quality—including student achievement gains—provide evidence to support its inclusion in teacher evaluative

frameworks.

2.1.7 Unit of Analysis

Despite the advantages and untapped potential of student surveys as a measurement tool, researchers are still learning a great deal about the psychometric properties of these instruments. In particular, the unit of analysis with student surveys can be a conceptual concern (Schweig, 2014b). Because students are grouped (or nested) within classrooms, survey responses can be used to analyze two different phenomena: responses can be examined separately to study the individual perceptions of students, or responses can be aggregated to measure the shared instructional environment (Lüdtke, Robitzsch, Trautwein, & Kunter, 2009). This aggregation process presents researchers with unconventional methodological challenges. For instance, suppose the student responses in a classroom are averaged across items or groups of items. Do these aggregated means now represent the "central tendency" of a property of students (i.e., individual-level perceptions), an "intrinsic property" of the classroom or instructor's practice (i.e., group-level characteristics), or a measure of something relating to students and the classroom in tandem (Sirotnik, 1980, p. 261)? Though questions about unit of analysis are subtle, understanding how students choose to interpret items greatly influences what the survey ultimately measures.

If the unit of analysis of student responses is the classroom—as is assumed by student survey measures administered for teacher evaluation—a key question presents itself: are differences in student responses substantively meaningful, or simply a representation of measurement error? Marsh et al. (2012) explains this question as the contrast between "context" and "climate". Context variables are individual in nature and, as such, measuring the differences between students is worthwhile. For example, student background or demographic items such as gender or socioeconomic status are classified as context variables, as are items asking students to rate their personal classroom experience or one-on-on interaction with their teacher (e.g., "My teacher gives me individualized feedback when I need help"). Aggregating the student responses

in a classroom on a context variable basis provides information about the proportional characteristic of the classroom (e.g., percent of students qualifying for free and reduced lunch or percent of students who receive individualized feedback from their teacher). Thus, student ratings on aggregated context variables "represent the central tendency of a distribution of measures" as stated by Sirotnik (1980, p. 261).

By contrast, climate variables reference the classroom or teacher, and it is assumed every student experiences this classroom environment identically (or nearly identically). For instance, the items "In this class we work hard every day" or "Our teacher encourages us to ask questions when we don't understand" are climate variables, as the subject item is clearly defined as the whole class or teacher. Notice that a simple structural change of the referent of an item from the class to the student (e.g., "In this class *I* work hard every day") can alter the item type from climate to context. However, given that the referent of climate variables is of the group, the shared variance among item responses represents a class of students' common conceptions of their classroom environment or the teacher's instructional practice. In this way, the student responses on climate variables can be thought of as "fundamentally exchangeable" (Schweig, 2014b, p. 23). Therefore, the variance between students within the same classrooms can be ascribed to sampling error, and the variance in aggregate ratings between classrooms can be assumed to represent an "intrinsic property" of the classroom, as described by Sirotnik (1980, p. 261).

Survey items asking students about the instructional practices of their teachers—such as those on the QAS survey—are climate variables. It is assumed that students in the same classroom experience their teacher's instructional practices (e.g., Item #1 *listen to lectures or instruction directed by the teacher*) with the same frequency. Likewise, items on Tripod (e.g., "My teacher takes the time to summarize what we learn each day") are climate variables. For these items, student ratings are exchangeable, as the researcher is typically only concerned with the meaning of the aggregated ratings within a classroom, and variation between student ratings

is attributed to sampling error.

It is important to note that the distinction between climate and context variables is only theoretical. Even student responses to climate variable items specifically designed to measure individual teacher practice still exhibit considerably more variance within classrooms than between classrooms. For example, analysis of the Tripod items (classified as climate variables) found intraclass correlation estimates (ICC's) ranging from 0.06 to 0.26 with a median item ICC of 0.12 (Schweig, 2014a). Likewise, ICC's for QAS classroom practice items range from 0.09 to 0.46 with a median item ICC of 0.16 (see Appendix Table A.4). ICC's of instructional practice items are affected by the amount of variation between classrooms in teacher practice and the degree of agreement between students on the practice within classrooms. As such, the consistently small ICC's for student survey items indicate that either (1) classrooms of students perceive their teachers' practices similarly (relatively low between-classroom variation), (2) students in the same classroom often judge their teacher differently or struggle to recognize when a practice is being performed (relatively high within-classroom variation), or (3) a combination of these forces. Thus, a student's classroom experiences—even when one classroom environment is presumed—can be perceived very differently from those of his or her peers. Being cognizant of the inherent noisiness of student survey responses and their multilevel structure is crucial to better understanding how student perceptions of their teachers compare within and across classrooms. For a more detailed discussion of unit of analysis in the context of student survey research, see Schweig (2014b).

2.2 Teacher Surveys

Teacher surveys have traditionally been a popular, cost-effective instrument for exploring teacher views of classroom practice. Like other teacher self-report measures (e.g., logs and interviews), teacher surveys have the advantage of capturing the direct attitudes and beliefs of a teacher—information that is uniquely valuable. Whereas classroom observations examine how a lesson unfolds in practice and classroom artifacts uncover how students respond to and

comprehend the lesson, teacher self-reports reveal the planning, intention, and expectations behind a given lesson (i.e., how the lesson should have proceeded in theory). As a science teacher in New York City succinctly explained, "Student feedback is important but it's also limited. They don't get to see the behind-the-scenes work" (Decker, 2012, para. 15). As a result, because teachers themselves are the only people fully knowledgeable about their own capabilities and of the daily classroom context, teacher surveys can produce insight that a principal, outside observer, or student may not recognize (Goe et al., 2008). Teacher self-reports can also facilitate pedagogical growth by providing a healthy space for teacher reflection (i.e., what worked, what did not work, and why).

Research has shown the reliability and validity of teacher surveys to be mixed. Burstein and colleagues (1995) investigated the instructional practices of math teachers using self-report surveys in conjunction with analysis of classroom artifacts (i.e., samples of student work and teacher lesson plans). The researchers found the reliability for the lessons covered and the instructional strategies implemented was satisfactory between teachers' self-reports on the surveys and analysis of artifacts. However, consistency in terms of teacher goals between the measures was considerably lower, to the point that validity was compromised. As Burstein and colleagues explain, "Instructional goals cannot be validly measured through national surveys of teachers. The data are inconsistent not only with artifact data but also with teachers' own self-reports on other survey items such as those describing their exam formats" (p. xiii). The inconsistencies of the survey instrument were difficult to interpret, but the researchers believed many teachers had confounded the frequency of performing instructional practices with the practices' importance.

Mayer's (1999) examination of the validity of Algebra teachers' self-reports on instructional practices found similarly conflicting results. Comparing teachers' reported time spent on specific instructional practices with observational measures of their time spent, the author found the measures to be highly correlated (0.85)—lending credence to the validity of the

survey—yet the survey responses were systematically inflated. Specifically, teachers reported using professional math teaching standards more frequently than was recorded during the observations. The degree of reliability of teacher self-reports was also mixed. Individual teacher responses of instructional practices were not particularly reliable (e.g., frequency of the teacher using manipulative objects; frequency of teacher-led whole-group discussion). However, test-retest reliability of the composite of teaching practices (0.69) suggests the measure was fairly stable in the aggregate. In sum, when true variation existed in teaching styles, the survey could detect the extent to which a teacher used a composite of instructional practices relative to other teachers. However, the survey failed to accurately provide the amount of time teachers spent implementing specific practices.

Still, perhaps the biggest validity concern regarding teacher surveys is inseparable from the measure itself. As with all self-report measures, teacher surveys suffer from social desirability effects—the tendency to project a favorable image to others (Moorman & Podsakoff, 1992). Social desirability is a common phenomenon in psychological research, and can take the form of overreporting favorable behaviors (e.g., levels of patriotism or acts of charity) or underreporting undesirable behaviors (e.g., frequency of drug usage or feelings of racial intolerance). For teachers, social desirability bias includes the conscious or unconscious misrepresentation of performing certain higher quality classroom practices. For instance, some teachers might overreport performing "better" practices such as active learning or differentiated instruction thought to be favored by the researchers. Ensuring the confidentiality of teacher responses can help control socially desirable responding. Nonetheless, the methodological limitations of teacher surveys suggest that, although this measure provides a valuable pedagogical perspective and is certainly a constructive element in any teacher evaluation framework, it should be utilized exclusively for formative assessment purposes without stakes attached (Goe et al., 2008).

2.3 Chapter Summary

Student and teacher surveys both offer distinct value as measures of instructional practice. Through the use of student surveys, reliable perceptual information can be collected from the natural observers of the classroom and direct recipients of the learning practices. Conversely, teacher surveys allow researchers to "peer under the hood" and learn about instructional planning and how a given lesson should have proceeded in theory. As Ferguson (2010) notes, "No one knows more about what happens in classrooms than the students and teachers who inhabit them" (p. 1). Though the classroom learning environment is determined by the interplay between teachers and students, little research has specifically compared student and teacher survey reports to determine whether perceptual alignment and discord may reveal a richer conceptual picture of the classroom environment and instructional practice. Building on psychological research, the next chapter reviews the literature and common methodologies used to measure perceptual discrepancy and its effect on outcome measures.

CHAPTER 3

3. Discrepancy Research in the Social Sciences

This chapter provides an overview of research in the social sciences that has investigated the degree of perceptual discrepancy or congruence as well as the various methodologies researchers have employed to measure this agreement. Specifically, the first half of the chapter reviews the literature of discrepancy in the fields of psychology and education. Modern congruence research was born in organizational psychology, but has only been studied in education using conventional methodologies—about which reservations exist as to whether these approaches examine the real questions of interest. The second half of the chapter details the methodological problems inherent to discrepancy or difference scores, and explains why polynomial regression is a more informative and revealing technique for examining nuance between two predictor variables and an outcome measure.

3.1 Discrepancy in Psychology

3.1.1 Supervisor–Subordinate Congruence in Organizational Psychology

The comparison of perceptions and attitudes of affiliated or related people has a rich history in the psychological sciences. In organizational behavior research, perceptions across constructs are often compared between workers in an organization (i.e., supervisors, peers, and subordinates) with the goal of measuring compatibility (Laurie, 1966; Turban, 1988). Alignment is typically measured using surveys between the attitudes of the employee and those of the supervisor or organization in regards to culture, structure, mission, norms, or support. These types of studies examining perceptual agreement or person-environment fit (i.e., match, similarity, continuity, convergence, and commonality) are commonly referred to as "value congruence research" in the organizational psychology literature (Edwards, 1994; Bao, Dolan, & Tzafrir, 2012).

Once measured, congruence between workers is typically operationalized as a predictor

of outcomes. In organizational behavior research these outcomes are most often relevant to the employee or organization, with congruence found to be associated with a variety of positive consequences. Greater congruence between employees and their supervisor or organization has been linked to job choice intentions (Cable & Judge, 1996), higher job satisfaction (Mount & Muchinsky, 1978; Michalos, 1986), greater employee well-being (Assouline & Meir, 1987; Edwards, 1991), greater work engagement (Larson, Norman, Hughes, & Avey, 2013), greater commitment to the organization (Lauver & Kristof-Brown, 2001), and higher sales performance (Ahearne, Haumann, Kraus, & Wieseke, 2013). Thus, perceptual agreement across constructs for stakeholders can indicate transparent communication and clearly defined organizational goals and guidelines.

3.1.2 Parent–Child Discrepancy in Psychopathology

Perceptions are also frequently compared in the field of psychopathology in clinical assessments of children. Psychologists often collect reports from multiple informants in a child's life (e.g., the child, parents, teachers, peers, and clinicians) and inconsistencies invariably arise across these reports (Achenbach, 2006; De Los Reyes, 2011). In the past, researchers have ignored incongruence in informant responses by treating discrepancies as measurement error; however, more recently they have explored exactly what it means when informants disagree, and have begun to study whether discrepancies are potentially useful for improving the assessment of children. As Laird and Weems (2011) note, this belief that discrepancy in perceptions between affiliated or related people may be informative is both "conceptually and intuitively appealing":

> When parents report that they know everything that goes on in their child's life and the child reports that the parents know very little, the discrepancy in perspectives seems, at least, to suggest that something is amiss in the family, and it seems reasonable to expect that the discrepancy has implications for the child's behavior. Likewise, when parents report that a child has many behavior problems and teachers report few behavior problems, the discrepancy suggests that children may be better behaved at school than at home or that parents and teachers may be attending to different behaviors or have different standards for acceptable behavior (p. 388).

Thus, rather than considering discrepancies to be "methodological nuisances" (p. 2) to be accounted or controlled for in the model, some psychological researchers have begun actively treating discrepancy as an informative byproduct worthy of study (De Los Reyes, 2011). Renk (2005) notes that though disagreement may be the result of informant error, it may also stem from an informant's "lack of access to certain types of behavior", "denial of the behavior of interest" or "distortion of information" (p. 459). In the psychopathology literature, investigating these perceptual differences (i.e., dissimilarities, disagreements, divergences, dissonances, incongruities) across reports is commonly referred to as "informant discrepancy research". For the purposes of this paper, the terms "discrepancy" and "congruence" are used interchangeably—though with opposite meanings—to describe perceptual fit and agreement.

Like research in organizational psychology, studies in psychopathology have attempted to isolate discrepancy between two informants with extended contact. Research has found discrepancy to be negatively related to child behavioral outcomes across several domains. Greater parent–child discrepancy has been shown to predict riskier teen driving (Beck, Hartos, & Simons-Morton, 2006), greater number of therapy visits (Brookman-Frazee, Haine, Gabayan, & Garland, 2008), greater instances of child delinquency (De Los Reyes, Goodman, Kliewer, & Reid-Quiñones, 2010), greater likelihood of future drug use (Ferdinand, van der Ende, & Verhulst, 2006), and lower parental involvement in the child's therapy (Israel, Thomsen, Langeveld, & Stormark, 2007). In addition, Guion, Mrug, and Windle (2009) found that discrepancy predicted child internalization of problems and lower social competence. They surmised discrepancy was a symptom of family dysfunction resulting from a breakdown in family communication. Taken together, it seems as though when two informants know each other well by spending considerable time together, divergently held views of their environment may negatively impact their interactions and functioning (De Los Reyes & Kazdin, 2006).

3.2 Discrepancy in Education

Surveys are frequently used in educational research to gather valuable perceptual

information from students, teachers, principals, and parents. Most commonly, perceptual discrepancy has been studied at the school level between principals and teachers or at the classroom level between teachers and students. Some researchers have also studied the perceptions of more than two educational stakeholders in order to triangulate the ratings and reduce stakeholder biases. These discrepancy studies have uncovered differences in student, teacher, parent, and principal attitudes and beliefs about their environment, and have found negative associations between greater discrepancy and desirable outcomes. Still, the body of research on discrepancy in education is less robust than that found in psychology. Some educational research has been focused on perceptual differences between stakeholders (i.e., do students and teachers rate matching items higher or lower). Other research in education has examined the predictive or predicated quality of discrepancy by computing difference scores between stakeholders' responses and then inserting these scores as either predictor or outcome variables in regression equations. However, educational research has not yet employed polynomial regression, a more advanced and flexible method for measuring the relationship between predictor variables and an outcome of interest.

3.2.1 Item and Domain Comparisons Between Educational Stakeholders

Several studies in education have measured discrepancy by broadly comparing the mean scores of students and teachers on matching questionnaires. Desimone, Smith, and Frisvold (2010) investigated discrepancy in terms of mean differences between student and teacher survey reports. Using National Assessment of Education Progress (NAEP) data from 2000, the researchers directly compared responses of middle school students and their teachers on similarly worded math instruction items—discussion, partner work, mathematical writing, and use of computers, measurement instruments, textbooks, and calculators. The researchers found small yet significant differences across the items with less subjective, tangible instructional practices, such as frequency of calculator use and frequency of textbook use, exhibiting the smallest differences. Lin, Lee, and Tsai (2014) also compared student and teacher responses by examining Taiwanese high school students' and their teachers' comprehension of learning and

evaluation in science across dimensions. Calculating mean differences, the authors found the conceptions of teachers and students did not always align. Whereas students generally viewed science learning at a more superficial level (memorization, practice, and assessment), teachers perceived science learning with greater depth. Specifically, teachers tended to understand scientific learning as implementing science to practical situations, constructing reasonable knowledge structures, and interpreting the natural world in a novel way. Finally, Stone (1997) examined discrepancies between the self-perceptions of learning-disabled high school students' and those of their special education teachers across various cognitive, social, and behavioral domains. The author found students' self-ratings were generally higher relative to the ratings given by their teachers. Thus, by comparing stakeholder responses, the author learned students with learning disabilities either underestimate the difficulty of the cognitive skills asked of them or overestimate their own capabilities, perhaps as a self-protective psychological defense mechanism.

Still other studies in education have uncovered discrepancies between students and teachers using more qualitative and open-ended measures. Montgomery and Baker (2007) examined perceptual congruence in the written feedback teachers provided to students in an English as a Second Language (ESL) program. Comparing student perceptions with teacher self-evaluation of the feedback, the researchers discovered, perhaps counterintuitively, that students perceived receiving more feedback than teachers felt they were providing. Virtanen and Lindblom-Ylanne (2010) also utilized open-ended questionnaires to compare college students' and faculty members' comprehension of teaching and learning in Biology. The researchers found a "substantial gap" between teachers' and students' conceptions of teaching and learning. Teachers viewed teaching and learning in the context of "scholarship of science" and saw their responsibility as one of guiding students to think independently and problem-solve. By contrast, students' conceptions of teaching were knowledge- and teacher- centered rather than student-centered, and teaching was regarded as a "task to make learning possible". These divergent results of student and teacher conceptions of learning were similar to those found by Lin and

colleagues (2014). Discrepancy between students' and teachers' conceptions may signal that when a teacher's practices are misaligned with his or her students' learning strategies, student learning may be negatively impacted.

Other research in education comparing parallel survey responses has helped to reveal areas of accord and discord in principal and teacher perceptions and beliefs. Using data from the Five-State Study, Desimone (2006) compared the mean responses of fourth and seventh grade teachers with those of their principals on survey items about their policy environment. The author found principals and teachers shared similar attitudes on the deleterious effects of student barriers (low socio-economic status, learning disabilities, limited English proficiency, and high mobility) and outdated resources (textbooks and technology) on adequately implementing state and district math content standards. However, teachers were more critical than principals of the content standards themselves. Whereas teachers believed the rigor of the standards was inappropriate, principals felt the standards increased the depth and consistency of math instruction. This insight reveals that teachers and principals can often agree on the obstacles of student learning and problems facing their schools, but differ in their opinions regarding solutions of reform. Likewise, Bingham, Haubrich and White (1993) found principals and teachers often disagree on issues of student discipline, with principals portraying a more positive picture of their school than teachers. These results are in accordance with those from a MetLife survey examining leadership in public schools (Markow & Scheer, 2003). The report found principals, as compared with teachers, were more "pleased with the current state of affairs", more likely to describe their school as "friendly, caring and safe", and more likely to view the principal–teacher relationships as "open, collaborative, and mutually respectful". Thus, by comparing teacher and principal responses across similarly worded items, researchers have learned that school-based perceptions can be greatly influenced by the occupational role of the respondent.

3.2.2 Investigating Stakeholder Discrepancy Using Difference Scores

Another method of measuring respondent perceptual congruence or fit commonly employed in educational research is computing difference scores to matching questionnaire items. Difference scores can be calculated by taking an algebraic (raw) difference (X-Y), absolute difference ($|X$-$Y|$), or squared difference ($(X$-$Y)^2$) of respondents' answers. Algebraic difference scores are typically utilized when the researcher is interested in determining relative or directional difference, while absolute or squared difference scores are typically employed to determine the overall magnitude of divergence. A perceived benefit of difference scores is that they can be used to predict outcomes measures in a process–product manner.

Miller and Davis' 1992 study was one of the first to apply this difference score approach to the study of perceptual congruence. Specifically, the researchers compared students' actual performance with students', teachers', and parents' predictions of performance across a range of cognitive tasks (vocabulary, math, and memory), as well as items asking about the child's preferences (interests, hobbies, activities, and school subjects) and personality (behaviors and traits). In addition to correlating students' actual performance, self-rated preferences, and self-rated personality traits with student, teacher, and parent predictions, "discrepancy scores" were computed by calculating the absolute difference between predicted performance and actual performance. In general, teachers were found to be as accurate as parents in judging students' cognitive abilities but were less accurate than parents in predicting their interests and personality traits. Students' self-predictions were less often accurate than those of the adults across all the tasks.

In a similar vein, Patel and Stevens (2010) used difference scores to examine the relationship between both parent–teacher and parent–student discrepancies on instances of parental involvement. The researchers computed "discrepancy factors" of students' scholastic abilities by calculating the absolute differences between parents' perceptions and teachers' grade reports and between students' perceptions and teachers' grade reports. Unlike Miller and Davis

(1992), these scores were then correlated with the outcomes volunteerism and home learning activities. Overall, as discrepancies increased—between parents and teachers, or between parents and students—parents reported lower involvement and teachers facilitated fewer programs for parental involvement.

More recently, Glueck (2013) calculated difference scores to investigate the influence of discrepancy in parent and teacher expectations on student outcomes. Using data from the Educational Longitudinal Study (ELS), the author developed "congruence scores" by computing absolute differences between parents' and teachers' expectations of how far the student would go in school. These congruence scores were then used as independent variables in various multilevel models. The author found that higher levels of parent–teacher congruence significantly predicted students' current and future math achievement, but not students' educational attainment (i.e., highest level of education attempted, ranging from "some high school" to "enrolled in a 4-year college").

Other studies in education have placed difference scores on the left side of the regression equation by treating this variable as the outcomes measure of interest. Houseman (2007) examined congruence between principal and teacher perceptions of school leadership behavior. The author computed algebraic (raw) difference scores by subtracting the teachers' responses from the principals' matching responses across items measuring principal leadership behavior. Principal–teacher discrepancy scores were then grouped by leadership dimension and used as outcome variables in multiple regression analysis. Results showed that principals who followed teacher evaluation procedures and modeled ideal leader behavior exhibited lower discrepancy in perceptions with their teacher counterparts. Desimone and colleagues (2010) employed a similar design using difference scores as the outcome measure. Specifically, the researchers investigated whether student and teacher background characteristics predicted the degree of teacher–student absolute discrepancy on similarly worded items of math instructional practices. The researchers were interested in whether higher achieving students and classrooms held perceptions more

closely aligned with their teachers, and whether more experienced and educated teachers held perceptions more closely aligned with their students. Results from multilevel modeling showed that the survey responses of female students, students with higher levels of parental education, higher achieving students, and students residing in advanced classes or classes with higher average math achievement more often agreed with their teacher's perceptions. Conversely, teacher experience and attainment of advanced math or education degrees were not significant predictors of greater teacher–student perceptual congruence. Thus, student and classroom factors—rather than teacher background characteristics—influenced the frequency of perceptual alignment between students' and teachers' ratings of instructional practice.

3.2.3 Section Summary

These studies demonstrate that there is no consensus method for measuring the degree and influence of perceptual differences between stakeholders in education. Rather, several research designs have been used to investigate discrepancy between students, teachers, principals, and parents, including calculating mean differences and computing absolute or algebraic difference score variables. Often, difference scores have been regressed or correlated with external criteria such as student achievement measures. However, no studies in education have utilized relatively newer polynomial regression methods, a more informative and methodologically flexible procedure for measuring discrepancy.

3.3 Discrepancy Methodologies

Though it is clear that greater discrepancy in perceptions or attitudes of respondents (e.g., between supervisor and subordinate, parent and child, principal and teacher, or teacher and students) is related to negative outcomes, the methods researchers have chosen to capture and measure discrepancy have varied. Two common and rudimentary methods researchers have employed are (1) computing correlations between component survey measures and (2) collapsing component survey measures into a single index by computing a difference score. Although these methods can be useful for measuring discrepancy or congruence in a broad sense, correlations

provide only supplementary information and difference score indices are methodologically limited.

3.3.1 Limitations of Correlations

A correlation coefficient is widely accepted as a way to examine reliability and agreement (e.g., test–retest reliability). Though this measure has been employed to measure congruence in responses (Derlin & Schneider, 1994; Kunter & Baumert, 2006; Desimone et al., 2010), correlations between items or factor scores is in actuality a measurement of *association*, not agreement (Jakobsson & Westergren, 2005; Glueck & Reschly, 2014). This distinction is evident when one respondent provides similar but consistently higher ratings than another respondent. In this scenario, the correlation in respondents' ratings is high, yet their overall agreement is low. As such, while there is certainly value in reporting correlations as a measure of relational strength, correlations can be misleading as a measure of respondent congruence.

3.3.2 Methodological Shortcomings of Difference Scores

As shown in Section 3.2.2, difference scores can be calculated algebraically, absolutely, or by computing a squared difference. Some researchers have advocated for standardizing ratings before calculating a difference score—by converting each person's ratings into a z-score relative to the rest of the same person's ratings in the sample (De Los Reyes & Kazdin, 2004). Other researchers have developed heuristics for comparing the difference scores of individuals across a sample. For instance, some have suggested that an individual's difference score be categorized as "divergent" if the score is greater than its associated measurement error (Brekelmans & Wubbels, 1991; den Brok, Bergen, & Brekelmans, 2006). A series of absolute difference scores has also been collapsed into a "profile similarity index" (Cronbach & Gleser, 1953), most commonly by summing absolute differences, summing squared differences, or summing squared differences and then taking the square root. Still, these variant methods of computing, applying, and interpreting difference scores only sidestep the methodological limitations of the measure.

Though difference indices have been widely used, perhaps emanating from their

"seductive face validity" (Johns, 1981), these techniques suffer from a host of methodological shortcoming (Cronbach, 1958; Edwards, 1994). First, while difference scores are not uniformly unreliable (Rogosa & Willett, 1983; Rogosa, Brandt, & Zimowski, 1982), they are often less reliable than the alternative metric of their component measures taken together (Johns, 1981). This is because, as Edwards (2001) notes, when *X* and *Y* are positively correlated (as is typical in discrepancy research), the reliability of the algebraic difference score between *X* and *Y* is often less than the reliability of *X* or *Y* computed separately. Only in situations when predictors are precisely uncorrelated will the reliability of a computed difference be equivalent to the average reliability of its component measures.

Second, difference scores can be difficult to accurately interpret as they combine conceptually distinct constructs into a single index, thereby confounding the effects of each of the component measures on the outcome (Edwards, 2001). For instance, it is appealing to think of an algebraic difference score as representing an equal contribution of each opposing component measure. However, this is only true when the variances of the component measures are equal. In practice, it is of course unlikely that component measures will exhibit equivalent variances. Thus, by computing difference scores one is unable to separate the relative contribution of each component measure in the index on the outcome measure. As Cronbach and Furby (1970) note, "There is little reason to believe and much empirical reason to disbelieve the contention that some arbitrarily weighted function of two variables will properly define a construct" (p. 79). Rather, a "considerable burden of proof" (Cronbach & Furby, 1970, p. 79) is necessary for an index to be valid. Yet, because difference scores confound the effects of each predictor measure that comprises the index, this requirement is not met. As such, it is not sufficient to claim difference scores can capture some latent congruence construct between stakeholders (e.g., Guion, Mrug, & Windle, 2009). Instead, as Edwards (2001) contends, congruence should be viewed as the distance or proximity between two constructs (e.g., between students' perceptions and their teacher's perceptions on a given instructional domain).

Third, difference scores constrain the relationship between the component measures and outcome. As Edwards (2002) states, "constraints should not be simply imposed on the data, but instead should be viewed as hypotheses that, if confirmed, lend support to the conceptual model upon which the difference score is based" (p. 33). For instance, consider the following example originally presented by Edwards (1994) using a regression equation of an algebraic difference score as a predictor of outcomes:

$$Z = b_0 + b_1(X–Y) + e \quad (3.1)$$

In Equation 3.1, *X* and *Y* are the predictor variables and *Z* is the outcome. If one distributes b_1 the equation expands to:

$$Z = b_0 + b_1X – b_1Y + e \quad (3.1)$$

Consider now a basic equation that uses components *X* and *Y* to separately predict *Z*:

$$Z = b_0 + b_1X + b_2Y + e \quad (3.2)$$

As can be seen by comparing Equation 3.1 with Equation 3.2, utilizing an algebraic difference score as a predictor variable is akin to constraining the coefficients of *X* and *Y* in Equation 3.2 to exhibit equivalent magnitudes but opposite signs ($-b_1 = b_2$). In other words, Equation 3.1 is identical to testing the hypothesis that *X* is positively related to the outcome while *Y* is negatively related. Squared difference scores and absolute difference scores similarly impose unproven constraints on the relationship between component measures and outcome.

Finally, difference scores can misrepresent and oversimplify the relationship between the component measures and the outcome by reducing an inherently three-dimensional relationship into two dimensions. Using a difference score predictor variable disregards the absolute levels of the components. That is to say, by assuming a linear relationship in two dimensions, difference scores cannot distinguish between the effects of predictors having equivalently low versus equivalently high ratings. It is only by considering the relationship in three dimensions that one can capture the complexity of the interaction between predictors and outcome. For a more

detailed discussion of the methodological shortfalls of difference scores, see Johns (1981), Edwards (1994), and Edwards (2001).

3.3.3 Polynomial Regression

Polynomial regression is a nonlinear approach to modeling discrepancy consisting of two predictor variables, two higher-order terms, and an interaction term (Edwards, 1994). Unlike difference scores, polynomial regression does not collapse the effects of the component measures but retains each predictor variable's independent effects (Shanock, Baran, Gentry, Pattison, & Heggestad, 2010). As explained by Edwards (2002), polynomial regression treats perceptual congruence "not as a single score, but instead as the correspondence between the component measures" (p. 360). As a result, this approach can be utilized to generate response surfaces conceptualizing the joint relationship between predictor variables and outcome in three dimensions. This makes polynomial regression ideal for investigating distinct patterns in perceptual differences.

The general form of the polynomial regression equation modeling the effects of perceptual congruence between two predictors on an outcome is shown below:

$$Z = b_0 + b_1X + b_2Y + b_3X^2 + b_4XY + b_5Y^2 + \text{e} \qquad (3.3)$$

where Z is the outcome variable (e.g., student achievement), X is Predictor 1 (e.g., student ratings on an item), and Y is Predictor 2 (e.g., teacher ratings on an item). Thus, in this design the outcome variable is regressed on each of the predictor variables (X and Y), the interaction between the predictor variables (XY), and the squared terms for each of the predictor variables (X^2 and Y^2).

3.4 Chapter Summary

As Laird and Weems (2011) note, computing difference scores can be an informative method of identifying broad disagreements between respondents. However, interpreting difference score coefficients can result in spurious conclusions, such as the appealing but

misguided belief that a difference score represents a fleeting discrepancy or congruence construct. Additionally, difference scores "may 'steal' variance from their components [and] carry this variance less reliably" (Johns, 1981, p. 454). Furthermore, difference scores inherently discard information by reducing a relationship between two configurations into a less parsimonious third variable (Cronbach & Gleser, 1953). Therefore, congruence is best conceptualized in reference to its component measures since a difference score "does not itself represent a construct but instead refers to the proximity between two constructs" (Edwards, 2001, p. 280).

Polynomial regression offers a more informative alternative for measuring congruence by examining the influences of both the degree and direction of discrepancy between respondents. While beyond the scope of this dissertation, this flexible, more modern procedure is now widely used in organizational psychology, and more recently has replaced difference scores in psychopathology as the consensus method for measuring congruence. Unlike difference scores, polynomial regression allows for component ratings to be taken jointly, and thus their interactive influences on an outcome of interest can be analyzed with nuance in three dimensions (Edwards, 2001). However, this approach to investigating congruence has not been widely adopted in the field of education. The next chapter presents the QAS student and teacher surveys used in the discrepancy analyses.

CHAPTER 4

4. Description of the Data Source

This chapter describes the dataset used to measure discrepancy in student and teacher perceptions: Quality Assessment in Science (QAS) surveys. Unlike other measures, which survey only students, survey only teachers, or survey students and teachers separately, the QAS survey is unique in that it was composed of parallel survey forms for students and teachers (i.e., matching items worded similarly and asked on the same scale). Thus, the design of this survey allowed for a direct, apples-to-apples comparison of students' and their science teachers' perceptions and attitudes, making the investigation of discrepancy substantively meaningful.

Students and teachers were recruited from a pilot study of a portfolio measure of classroom assessment practice during the 2009–2010 academic year (see Martinez, Borko, & Stecher, 2012). The sample of 645 students was enrolled in 39 eighth grade science classrooms from 37 public middle schools in 12 districts across California. Of these students, 74% were minority, 54% female, 20% English language learners, and 43% on free or reduced lunch. No demographic information was collected on the participating teachers.

4.1 QAS Student and Teacher Matching Items

Information about classroom practices came from surveys of student and teacher perceptions of learning and instructional activities in their eighth grade science classrooms. Teachers were administered the QAS survey twice, once when they completed their first portfolio measure (pre-survey) and then again when they completed their second portfolio (post-survey). The amount of time between teacher survey administrations averaged around three months, with variation across teachers ranging from one to six months. Students were administered the QAS survey only once, concurrent with their teachers' post-survey. While the surveys were originally developed to explore the domains of classroom assessment practice represented in the National Research Council assessment framework (NRC, 2001), there were 23

common items across the surveys pertaining more broadly to daily classroom practices (see Table 4.1). The focus of the QAS survey items on classroom practices—rather than the specific behavior and instruction of teachers—uniquely separated the measure from other prominent student surveys instruments, as outlined in Table 1.1.

Table 4.1: *QAS Matching Student and Teacher Survey Items*

Student stem: "How often do you do each of the following activities in your eighth grade science class?" Teacher stem: "How often do students engage in each of the following activities in your classroom?"	*Never*	*Less than once a month*	*Once or twice a month*	*Once a week or more*	*Multiple times per week*	*Every day*
1. Listen to lectures or instruction directed by the teacher						
2. Read a science textbook						
3. Read science articles in magazines or newspapers						
4. Discuss science topics with other students in small groups						
5. Discuss science topics with other students as a class						
6. Work on worksheets						
7. Work on homework tasks during class time						
8. Watch science videos, movies, TV shows, etc.						
9. Use science-related software or internet resources						
10. Work on science projects individually						
11. Work on science projects in pairs or groups						
12. Work on written science reports						
13. Present oral science reports						
14. Watch teacher demonstrate experiment or investigation						
15. Use lab instruments or materials						
16. Plan/design experiments or investigations						
17. Conduct experiments or investigations						
18. Analyze data / relationships between variables						
19. Write reports about labs or experiments						
20. Review materials to prepare for a test or quiz						
21. Develop or practice test-taking skills						
22. Take tests with questions where you choose the answer						
23. Take tests with questions where you write out the answer						

The matching items on the student and teacher QAS surveys were worded similarly to capture the same dimensions of classroom practice. The sentence stem for the student items was, "How often do you do each of the following activities in your eighth grade science class?" The stem for teacher items was, "How often do students engage in each of the following activities in your classroom?" For example, the first QAS survey item asked students how often they listened to lectures or instruction during class, and asked teachers how often students in their classrooms listened to lectures or instruction. All 23 matching QAS items were scored on a 6-point Likert scale indicating the frequency or degree with which students or teachers experienced a classroom practice. The response categories were 0 = *never*, 1 = *less than once a month*, 2 = *once or twice a month*, 3 = *once a week or more*, 4 = *multiple times per week*, 5 = *every day*.

4.2 QAS Items Are Climate Variables

Because the QAS student survey investigates perceptions of tangible classroom practice frequency, these matching items can be classified as climate variables. Theoretically then, variation in student responses within classrooms should reflect rater measurement error. For instance, Item #1 *listen to lectures or instruction directed by your teacher* should not elicit discrepant answers due to students' differing classroom perceptions. Rather, one would expect students in the same classroom to similarly experience the given practice. Consequently, the QAS items are ideal for examining student variation within classrooms and response pattern differences between classrooms.

CHAPTER 5

5. Is Discrepancy Greater for Some Classes than for Others?

The main objective of this research was to investigate the perceptual congruence between students and teachers on instructional practices. The study of this question presents particular methodological challenges due to the variation of student survey responses within classrooms. Whereas supervisor–subordinate, parent–child, and principal–teacher responses all involve relatively straightforward one-to-one comparisons, the comparison of a teacher's perception is with his or her classroom of students—a contrast of one person's perception versus the perceptions of many. Using matching items on the QAS surveys, this dissertation employs six methods to measure the discrepancy between student and teacher survey ratings of instructional practice frequency: proportional scoring, discrepancy scores, unstandardized differences in means, class composite unstandardized absolute differences in means, item composite unstandardized differences in means, and discrepancy plots. It is important to note that not all methods of examining discrepancy are appropriate for answering each proposed research question (see Table 5.1). Additionally, in order to achieve a greater representativeness of teacher perceptions and decrease measurement error, all of these methods compare the student and class item rating to the *average* teacher item rating across pre-survey and post-survey occasions.

Using a combination of these six methods of measuring discrepancy, three main research questions are investigated in Chapters 5–7 as well as the influences of these methods in answering these questions:

1. Is teacher–classroom discrepancy greater for some teacher–classroom pairs than for others (Chapter 5)?
2. Is discrepancy between classes dependent on the instructional practice examined (Chapter 6)?
3. Is teacher–classroom discrepancy greater for some classroom practices than for others (Chapter 7)?

Table 5.1: *Research Questions by Measure of Discrepancy*

Measure of Discrepancy	RQ1 (Chapter 5): Is discrepancy greater for some classes than for others and does the measure of discrepancy influence which classes are considered more discrepant?	RQ2 (Chapter 6): Is discrepancy between classes dependent on the instructional practice examined?	RQ3 (Chapter 7): Is discrepancy greater for some instructional practices than for others and does the measure of discrepancy influence which practices produce more discrepancy?
Proportional scoring			X
Student and class composite absolute discrepancy scores	X		
Unstandardized differences in means	X	X	X
Class composite unstandardized absolute differences in means	X	X	
Item composite unstandardized differences in means			X
Discrepancy plots		X	X

Note. Chapter 6 employs "Unstandardized differences in means" and the magnitude of this discrepancy measure, "unstandardized absolute differences in means".

This chapter examines perceptual differences across classes using three methods of measuring discrepancy (other measures of discrepancy in Table 5.1 are explained in tandem with their use in subsequent chapters). First, composite absolute discrepancy scores are calculated for students and classes. Second, unstandardized differences in means are examined between each of the 39 teacher–classroom pairs on each of the 23 QAS items. The unstandardized difference in means is the building block of composite unstandardized differences in means across classrooms and items. Third, classroom composite (aggregate) unstandardized absolute differences in means are compared across each of the 39 teacher–classroom pairs. Results from these variegated methods are triangulated to address the main research questions of this chapter:

1a. Is teacher–classroom discrepancy greater (in magnitude or direction) for some teacher–classroom pairs than others?
1b. Does the measure of discrepancy used change the ranking of classrooms on discrepancy?

Example discrepancy analyses using the student and teacher ratings in Class #16 on Item #16 *plan/design experiments or investigations* are provided throughout the chapter.

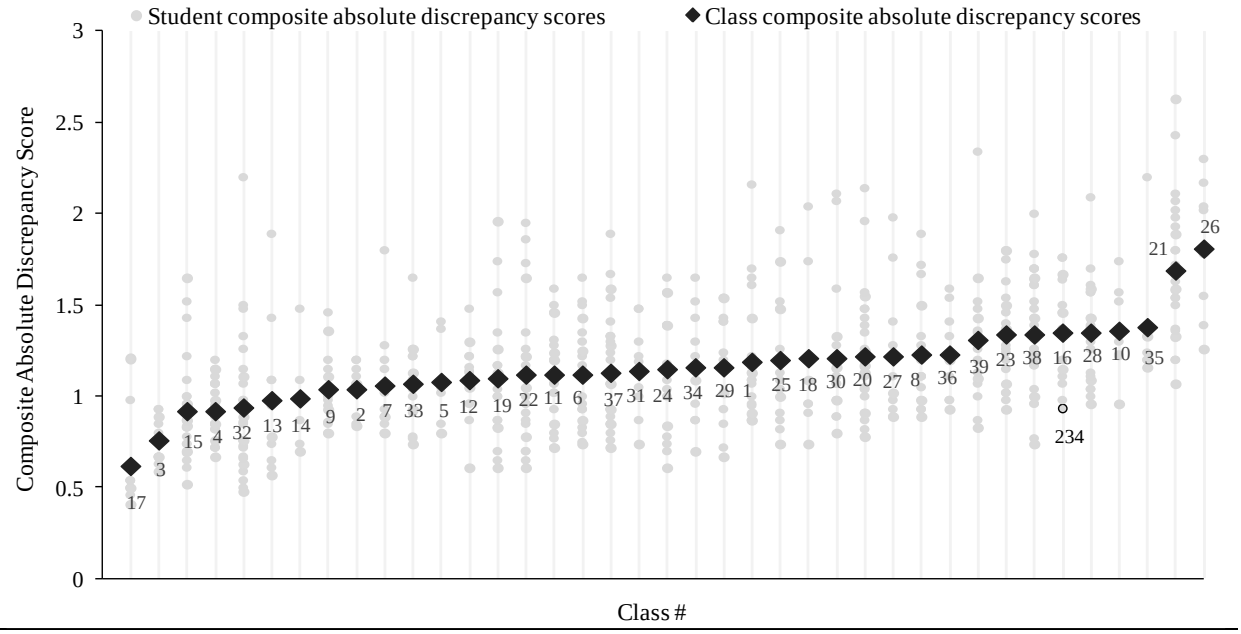
5.1 Ranking Classes Using Teacher–Student and Teacher–Classroom Discrepancy Scores

Figure 5.1 provides the ranking of classrooms using student and class composite absolute discrepancy scores. Absolute discrepancy scores are a simple measure of the magnitude of perceptual difference within classrooms found by computing the distance (i.e., absolute value) between a student's and teacher's average ratings on a given item. While directional information is discarded using this procedure, the advantage of computing absolute differences is that discrepancy can be aggregated across items without positive and negative item values cancelling out. In this dissertation the average absolute difference between a student's and teacher's ratings across all 23 matching QAS items is referred to as the "student composite absolute discrepancy score" and the average absolute difference between a classroom of student's and teacher's ratings across all 23 matching QAS items is referred to as the "class composite absolute discrepancy score". Using scores for a student's discrepancy with his or her teacher and scores for a teacher's discrepancy with his or her classroom provides a straightforward means to identifying the students and teachers most discrepant in absolute terms.

Consider a student (Student #234) in Class #16 who rated Item #16 a 3, and whose teacher (Teacher #16) rated Item #16 a 1 on the pre-survey and a 2 on the post-survey for an average item rating of 1.5. Student #234's absolute discrepancy score on Item #16 was, then, 1.5. Of the other 16 students in Class #16, four rated Item #16 a 0 (producing absolute discrepancy scores of 1.5), one student rated Item #16 a 1 (producing an absolute discrepancy score of 0.5), four students rated Item #16 a 2 (producing absolute discrepancy scores of 0.5), five students rated Item #16 a 3 (producing absolute discrepancy scores of 1.5), and two students rated Item #16 a 4 (producing absolute discrepancy scores of 2.5). Thus, the *average* absolute discrepancy score of the 17 students' responses (Students #234–#250) in Class #16 on Item #16 was 1.32.

This value is Class #16's absolute discrepancy score on Item #16.

Figure 5.1: *Student and Class Composite Absolute Discrepancy Scores, by Class*



Note. Classes sorted by class composite absolute discrepancy score.

Aggregating absolute discrepancy scores across items creates the composite indices. Class composite absolute discrepancy scores (black diamonds in Figure 5.1) were found by averaging the class absolute discrepancy scores across the 23 QAS items. For example, the composite absolute discrepancy score for Class #16 was 1.36 (right-most box in Table 5.2). This ranked Class #16 34th out of the 39 classes in terms of its magnitude of perceptual alignment—this is why the black diamond of Class #16 is sixth from the right in the figure.

Vertically plotted around each class composite absolute discrepancy score are grey circles corresponding to each of the 645 students' absolute discrepancy composite scores with their teacher. For example, Table 5.2 shows that Student #234's absolute discrepancy composite score computed across the 23 QAS items was 0.93 (second box to the right in Table 5.2). This ranked Student #234 in the 71st percentile of all students and first among the 17 students in Class #16 in terms of the magnitude of his or her perceptual alignment—this is why the grey circle of

Table 5.2: *Calculation of Composite Absolute Discrepancy Scores Between Student and Teacher Reports of Instructional Practice*

		Student Ratings (pre-survey)				**Teacher Average Rating (pre-survey and post-survey)**				**Absolute Discrepancy Score**				**Student Composite Absolute Discrepancy Score**	**Class Composite Absolute Discrepancy Score**
Teacher ID	Student ID	Q1	Q2 …	Q16…	Q23	Q1	Q2 …	Q16…	Q23	Q1	Q2 …	Q16…	Q23		
#1	#1	5	4	4	4					1	2.5	1	1	1.61	
#1	#2	4	3	2	1					0	1.5	1	2	0.87	
#1	#3	4	2	2	2	4	1.5	3	3	0	0.5	1	1	0.87	1.20
#1	…	…	…	…	…					…	…	…	…	…	
#1	#23	5	5	4	3					1	3.5	1	0	2.09	
#2	#24	4	4	0	2					0	0	3	0	0.85	
#2	#25	5	3	2	3					1	1	1	1	1.20	
#2	#26	5	3	2	3	4	4	3	2	1	1	1	1	1.15	1.05
#2	…	…	…	…	…					…	…	…	…	…	
#2	#32	4	3	3	2					0	1	0	0	0.89	
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
#16	#234	0	1	3	4					1	0.5	1.5	3	0.93	
#16	#235	4	1	4	2					3	0.5	2.5	1	1.24	
#16	#236	4	3	3	4					3	1.5	1.5	3	1.37	
#16	…	…	…	…	…	1	1.5	1.5	1	…	…	…	…	…	1.36
#16	#250	2	1	3	3					1	0.5	1.5	2	1.76	
Class #16		3.18	1.00	2.06	2.47					2.18	0.50	1.32	1.47	---	
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
#39	#630	3	0	3	1					0	0.5	1.5	0.5	1.13	
#39	#631	5	0	3	0					2	0.5	1.5	1.5	1.48	
#39	#632	3	0	0	1	3	0.5	1.5	1.5	0	0.5	1.5	0.5	1.43	1.33
…	…	…	…	…	…					…	…	…	…	…	
#39	#645	5	3	2	2					2	2.5	0.5	0.5	1.04	

Student #234 is the lowest vertically plotted point of the class.

Figure 5.1 shows that class composite absolute discrepancy scores roughly followed a trend line associated with a higher order polynomial function when ranked from lowest to highest. Teacher–Classroom Pairs #17 and #3 had the lowest class composite absolute discrepancy scores (0.62 and 0.76) while Teacher–Classroom Pairs #26 and #21 had the highest composite discrepancy scores (1.81 and 1.69). Therefore, the class composite absolute discrepancy score of the most discrepant classroom was nearly three times the score of the least discrepant classroom. This difference was a little more than one point on the six-point QAS response scale. The remaining 35 teacher–classroom pairs had discrepancy scores ranging from 0.92 to 1.38, a difference of a little less than half a point on the QAS response scale

Students in the most discrepant classrooms were also much more discrepant in absolute terms with their teachers than students in the least discrepant classrooms. The composite absolute discrepancy score of only one student in the least discrepant classes (#17 and #3) was greater than the least discrepant student in the most discrepant classes (#26 and #21). Thus, for both students and teachers in the least and most discrepant teacher–classroom pairings, clear differences were evident in degree of absolute discrepancy. However, it is important to note that the degree to which differences in the highest and lowest discrepancy classroom are statistically significant or substantively meaningful is more difficult to determine using this method. This is an undisputed shortcoming of composite absolute discrepancy scores. Clearly, the magnitude of perceptual difference between students in Class #26 and Teacher #26 appears greater than the magnitude of perceptual difference between students in Class #17 and Teacher #17. Yet, questions as to how much greater and whether these differences are statistically consequential cannot be adequately answered using composite absolute discrepancy scores. The next section details the use of unstandardized differences in means, an alternative approach to measuring classroom perceptual differences.

5.2 Ranking Classes Using Unstandardized Differences in Means

Another method of measuring perceptual misalignment within classrooms is that of computing an unstandardized difference between a classroom of student's average rating and the teacher's average rating on an item. Whereas student composite absolute discrepancy scores measure the average student's deviation from his or her teacher on a given item and class composite absolute discrepancy scores measure the average teacher deviation from all his or her students on a given item, unstandardized differences in means measure the difference between the aggregate response of all the students (i.e., the class perception) and the average teacher response on a given item. Thus, unstandardized differences in means do not consider variation in student responses (i.e., how discrepant students are from each other) in assessing class perceptual difference. Instead, uncertainty is reflected in standard error estimates, which can be used to construct confidence intervals around an unstandardized difference in means parameter to show the plausible values of a discrepancy. While the width of this surrounding confidence interval is affected by the degree to which students deviate from their mean response (and the degree to which a teacher deviates from his or her mean response), and on the size of the class, these factors do not affect the magnitude of the difference between the student's mean response and the teacher's mean response. Thus, difference scores and unstandardized differences in means offer two conceptually distinct approaches to measuring classroom discrepancy.

5.2.1 Computing the Unstandardized Difference in Means

An unstandardized mean difference $\overline{X}_{\text{Dif\_Item}(i)\text{\_Class}(j)}$ between student and teacher responses on a single item (i) in a given classroom (j) can be calculated using the formula below:

$$\overline{X}_{\text{Dif\_Item}(i)\text{\_Class}(j)} = \overline{X}_{\text{Item}(i)\text{\_Class}(j)} - \overline{X}_{\text{Item}(i)\text{\_Teacher}(j)} \quad (5.1)$$

with $\overline{X}_{\text{Item}(i)\text{\_Class}(j)}$ the average rating for all students on item i in class j on one occasion (post-survey) and $\overline{X}_{\text{Item}(i)\text{\_Teacher}(j)}$ the average rating on item i for teacher j on two occasions (pre-survey and post-survey).

The average rating of all 17 students in Class #16 on Item #16 was 2.06 and the average ratings of Teacher #16 across the pre-survey and post-survey was 1.5. This means on average students in Class #16 perceived that they participated in experimental design about once or twice a month while Teacher #16 felt his or her students conducted this instructional practice slightly less frequently. The unstandardized mean difference ($\overline{X}_{Dif\_Item16\_Class16}$) between student and teacher responses on Item #16 in Class #16 is computed as follows:

$$\overline{X}_{Dif\_Item16\_Class16} = \overline{X}_{Item16\_Class16} - \overline{X}_{Item16\_Teacher16} = 2.06 - 1.50 = 0.56$$

Note that this value is different from Class #16's absolute discrepancy score on Item #16 of 1.32. This is because some students in this class rated Item #16 lower than their teacher's average rating and some students rated the item higher. However, had all students in Class #16 rated Item #16 higher than Teacher #16's average ratings, these two values would have been equivalent. Indeed, the unstandardized difference in means will always be less than or equal to the absolute discrepancy score for a given class. All 897 unstandardized differences in means can be found in Appendix Table A.5.

Measures of central tendency. In the example provided above, mean is used as the measure of item central tendency. However, using this parameter to compute an unstandardized difference is by no means a settled choice. To begin with, it is not clear what measure of central tendency (mean, median, or mode) best describes student classroom responses on a given item. The mean is the most typically used measure of central tendency, but this metric is more susceptible than the median to an outlier student response and more sensitive when the response distribution is heavily skewed. Computing the median may also be preferable to the mean when the data is ordinal, as was the case with the QAS survey items. The mode is even less sensitive than the median to a heavily skewed response distribution and may be the most appropriate measure if the distribution of an item is bimodal. Nonetheless, the six-point scales of the QAS items ameliorated concerns about the effects of outlier ratings. As a result, the current analyses use mean as the sole measure of central tendency.

Measures of spread. It is also not fully clear what measure of spread best captures the variation in the student responses within classrooms and teacher responses across occasions. While not needed for computing unstandardized differences in means, a measure of spread is necessary in determining the standard error of this parameter. Standard deviation is the most commonly used statistical metric to describe variation. However, the mean absolute deviation around the mean is also a valid measure of spread that is robust to outliers and more easily understood (Pham-Gia & Hung, 2001; Leys, Ley, Klein, Bernard, & Licata, 2013). Additionally, median absolute deviation around the median (MAD) is a frequently used measure of variation. However, this method was inappropriate in the current analyses as 22% of the within-classroom student responses on the QAS items had a MAD = 0. This occurred when a majority of students in a classroom rated an item equivalently. Therefore, the current analyses employed standard deviation (i.e., variance) exclusively as the measure of spread.

Two different methods were used to calculate the variance for teachers across survey occasions and the variance for students within classrooms. Each of these metrics capably separates true student and teacher response variation from error. Beginning first with teachers, the variance across the QAS pre-survey and post-survey on each instructional practice was pooled (or aggregated) as teacher survey responses were collected on multiple timepoints. Each pooled variance represented the overall spread in teacher responses for a given item under the assumption that, while average teacher ratings on an item may have been different, the rating variance for each teacher was the same. Had teachers in the sample answered an item an unequal number of times, greater weight could have been assigned to the variances from the responses of those teachers who completed the survey item on more occasions.

The average amount of the time between teacher QAS survey administrations was three months. While it is certainly possible (and even likely) that teachers made incremental adjustments to their instructional practices over the course of the school year, teacher perceptions were largely similar in each survey administration. Nearly half (47%) of ratings were in exact

agreement, 89% of ratings were within one scale unit, and 97% of ratings were within two scale units across pre-survey and post-survey. Furthermore, patterns of agreement with student ratings were similar across teacher survey administrations. Specifically, 27% of teacher responses on the pre-survey were in exact agreement with student responses and 68% of teacher responses were within one scale unit; on the post-survey 28% of teacher responses were in exact agreement with student responses and, again, 68% of teacher responses were within one scale unit. Moreover, the same instructional practices were rated highest and lowest by teachers compared to their students with no item rated higher or lower by more than 10% of students across administrations. Put simply, teachers were no more discrepant with their students overall and displayed largely similar discrepancy patterns on the individual instructional practices. As such, even though teacher perceptions were of course not identical across administrations, the pre-survey teacher ratings were included to double the available teacher data for discrepancy analyses, which was particularly important given the unreliability of single teacher item ratings.

In total, 23 separate classroom-pooled teacher variances were computed, one for each QAS item. The pooled variance for teachers on Item #16 was 0.43. This ranked Item #16 as the 11th most variable instructional practice for teachers of the 23 total QAS items. The instructional practice rated most variably by teachers was Item #11 *work on science projects in pairs or groups*, $s^2_{\text{Teacher\_Item11}} = 1.12$. The instructional practice rated least variably by teachers was Item #22 *take tests or quizzes where you choose the answer*, $s^2_{\text{Teacher\_Item22}} = 0.41$.

In the current analysis, 38 teachers completed the QAS survey twice and one teacher (Teacher #22) completed only the post-survey. Thus, the variance across survey occasions for Teacher #22 was zero. The pooled sample variance ($s^2_{\text{Teacher\_Item}(i)}$) of teacher responses for a given item (*i*) was simply the average of the variances across survey occasions for each of the 38 teachers ratings (because weights are assigned by the number of occasions - 1, the zero variance estimates of Teacher #22 always cancel out in pooling variances across teachers).

By contrast, the metric of spread used to capture students' response variation was not a

pooled estimate but was simply the variance of the survey responses for a classroom of students (j) on a given item (i), $s^2_{\text{Student\_Item}(i)\_\text{Class}(j)}$. This computation resulted in a distinct item variance for each classroom of students, as opposed to teachers who all shared a single, universal pooled variance for each instructional practice item. In total, 897 unique student response variances were computed across the QAS sample, one variance in each of the 39 classrooms on each of the 23 items. For example, the variance on Item #16 for students in Class #16 was 1.93. This value is the variability of the 17 student ratings in Class #16 on Item #16 under the assumption that the variances in ratings on this item are distinct across the QAS classrooms. For context, the range in variances for students on Item #16 (across the 39 classes) was [0.36, 2.37] and the range of variances for students in Class #16 (across the 23 items) was [0.76, 3.35].

Class-pooled and class-unique measures of spread offer different interpretations of item response variation. Pooling the variances of teacher responses on a given item discards response spread information within-classrooms. Therefore, by choosing to use this metric one assumes the spread in the ratings of teachers over time is interchangeable. That is, the responses provided by a teacher on a given item do not exhibit a meaningful spread pattern nor are more variable than the responses provided by a teacher in any other classroom. Instead, by using this computational approach, one presupposes variation in teacher responses results solely from inherent item differences in spread (i.e., some items elicit greater teacher response variation over time than others). In this sense, the particular variation across occasions is irrelevant as it is assumed each instructional practice has a true, underlying teacher response variation. Importantly, this variation is not assumed equivalent to variation of the corresponding class of students on the instructional practice.

Conversely, computing the distinct variance of student responses for a given item within a classroom posits that the grouping unit can uniquely influence the variation in responses. Just as item means are unique from classroom to classroom, item variances are also treated as distinct from one classroom to another. That is, students in Class #16 may exhibit legitimately greater

disagreement on Item #16 than, for example, the students in Class #17—with perhaps the opposite response pattern true for a different item. In other words, it is assumed response variability in instructional practices is different across classrooms of students. Some classrooms of students simply disagree more often than other classrooms of students on the whole, or disagree to a greater or lesser degree than other classrooms depending on the practice.

5.2.2 Computing the Standard Error of the Unstandardized Difference in Means

After computing an unstandardized difference in means describing the direction and magnitude of perceptual differences between students and teacher on a single item, a standard error estimate can be found to evaluate the statistical significance of this parameter. The standard error ($SE_{\text{Dif\_Item}(i)\text{\_Class}(j)}$) of the unstandardized difference in means between student and teacher responses on item *i* in classroom *j* can be calculated using the formula below:

$$SE_{\text{Dif\_Item}(i)\text{\_Class}(j)} = \sqrt{\frac{s^2_{\text{Teacher}(j)\text{\_Item }(i)}}{n_{\text{Teacher}(j)}} + \frac{s^2_{\text{Student\_Item}(i)\text{\_Class}(j)}}{n_{\text{Class}(j)}}} \tag{5.2}$$

where $s^2_{\text{Teacher}(j)\text{\_Item}(i)}$ is the pooled sample variance of teacher responses on item *i*, $n_{\text{Teacher}(j)}$ is number of survey occasions teacher *j* completed item *i* (QAS pre-survey and post-survey), $s^2_{\text{Student\_Item}(i)\text{\_Class}(j)}$ is the unique sample variance of student responses in class *j* on item *i*, and $n_{\text{Class}(j)}$ is the number of students in class *j* who completed item *i*.

For example, the standard error ($SE_{\text{Dif\_Item16\_Class16}}$) of the unstandardized difference in means in Class #16 on Item #16 is computed as follows:

$$SE_{\text{Dif\_Item16\_Class16}} = \sqrt{\frac{0.43}{2} + \frac{1.93}{17}} = 0.58$$

The standard error can then be used to construct the lower and upper bounds of an approximate 95% confidence interval quantifying the margin of error around the unstandardized difference in means parameter. Setting $\alpha = 0.05$ and basing the critical value (t_{critical}) on a *t* distribution with

the degrees of freedom (*df*) equal to $n_{\text{Class}(j)}$-1, the lower and upper bounds of a 95% confidence interval around the unstandardized difference in means estimate in Class #16 on Item #16 is shown below:

Lower limit $= \overline{\text{X}}_{\text{Dif\_Item16\_Class16}} - (t_{\text{critical }(df\_16)} * SE_{\text{Dif\_Item16\_Class16}}$

$= 0.56 - (2.13 * 0.58) = -0.68$

Upper limit $= \overline{\text{X}}_{\text{Dif\_Item16\_Class16}} + (t_{\text{critical }(df\_16)} * SE_{\text{Dif\_Item16\_Class16}}$

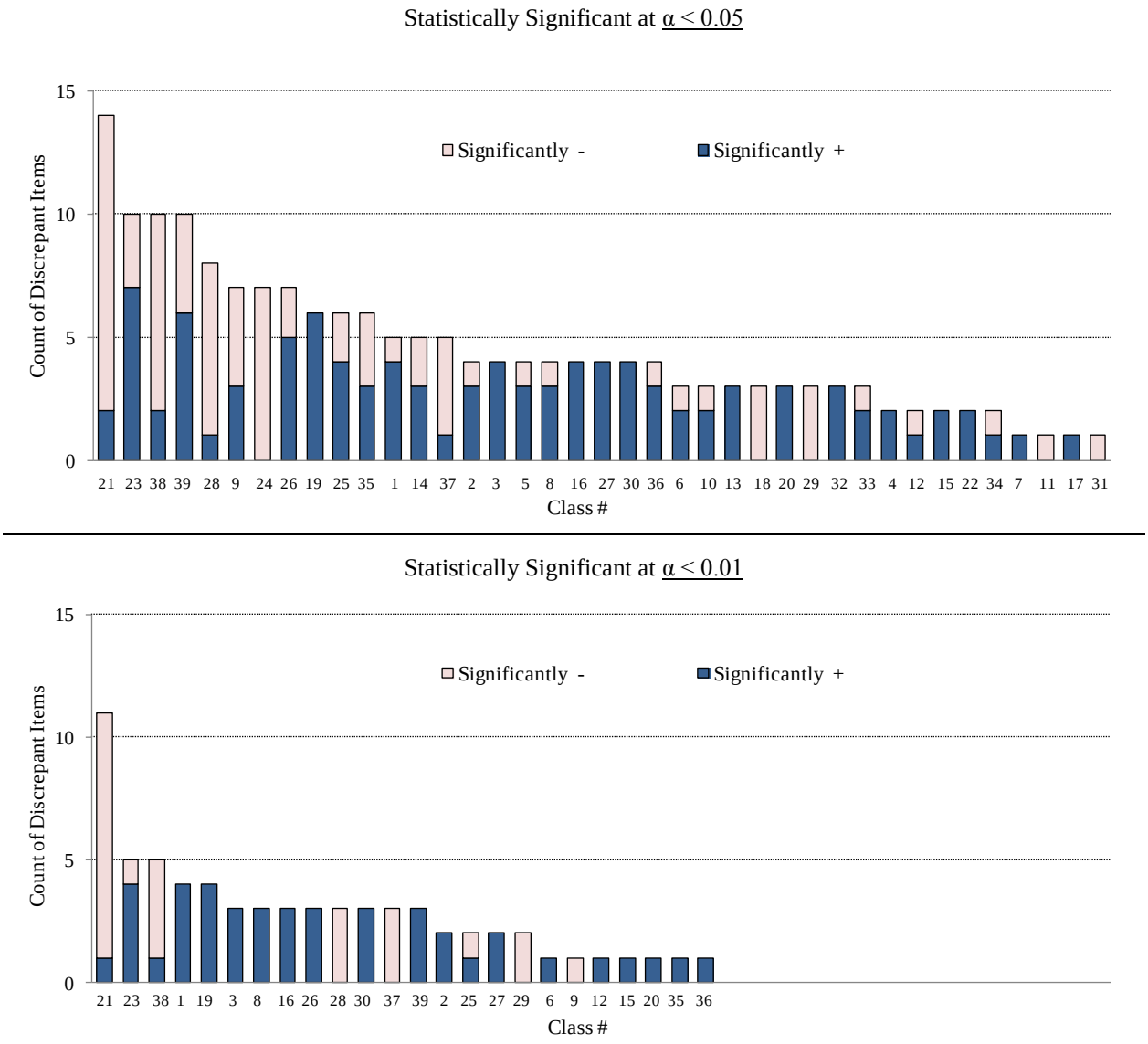
$= 0.56 + (2.13 * 0.58) = 1.80$

Thus, the unstandardized difference in means in Class #16 on Item #16 between students and their teacher is 0.56 with a confidence interval of [-0.68, 1.80]. This positive mean difference suggests students in Class #16 perceived the instructional practice *plan/design experiments or investigations* as more frequently occurring in the classroom than their teacher. However, as this interval overlaps with zero, one cannot make this conclusion with statistical confidence. Indeed, the fact that the lower bound of the confidence interval is negative means there is a non-negligible probability Teacher #16 actually perceives this practice as occurring *more* frequently than the students in Class #16, not less. Had Teacher #16 been surveyed on more occasions, or had more students comprised Class #16, a greater degree of statistical confidence could be ascribed to this unstandardized difference in means measurement.

5.2.3 Ranking Classes by Statistically Significant Unstandardized Differences in Means

To explore class differences in discrepancy on an item-by-item level, the statistical significance (evaluated at $\alpha < 0.05$ and $\alpha < 0.01$) for all 897 unstandardized differences in means were tallied across teacher–classroom pairs. Results showed some classes accounted for a disproportionate number of the total instances of significant unstandardized mean difference estimates. As evident in Figure 5.2, the five most discrepant classes tallied 52 out of 176 item-level discrepancies (30%) evaluated at $\alpha < 0.05$ and 29 out of the 68 item-level discrepancies (43%) evaluated at $\alpha < 0.01$.

Figure 5.2: *Classroom Counts of Statistically Significant Unstandardized Differences in Means*



Note. In total, 176 unstandardized differences in means were found to be statistically significant at α < 0.05 (76 significantly negative and 100 significantly positive) and 68 unstandardized differences in means were found to be statistically significant at α < 0.01 (25 significantly negative and 43 significantly positive).

The two most discrepant classes in terms of composite absolute discrepancy scores—Classes #21 and #26—accounted for 21 out of the 176 item-level discrepancies (12%) evaluated at α < 0.05. Class #26 tallied 7 significant instances of discrepancy (five positive and two negative) across the 23 items while Class #21 particularly stood apart from the other classes tallying 12 significant negative instances of discrepancy and two significant positive instances.

Moreover, 10 of Class #21's 12 negative instances of discrepancy were also significant at $\alpha < 0.01$. This was 40% (10 out of 25) of the total number of negatively significant unstandardized differences in means evaluated at $\alpha < 0.01$ observed in the entire QAS sample. Clearly, Teacher #21 either consistently rated practice frequencies higher, students in Class #21 consistently rated practice frequencies lower, or there existed a combination of these forces. By comparison, more than a third of the other teacher–classroom pairs (13 out of 38) did not record a single negatively significant unstandardized difference in means on any of the items. These two classes were also discrepant across a range of practices. Either Class #21 or Class #26 was discrepant on 17 of the 23 total QAS items at $\alpha < 0.05$ while 17 of the other 37 classes (45%) were discrepant on three or fewer QAS practices.

5.2.4 Section Summary

In this section classes were ranked by their total number of statistically significant unstandardized differences in means on the 23 QAS instructional practices. The unstandardized difference in means is a disaggregate measure of the discrepancy between a class average student rating and an average teacher rating on a given QAS item. Class #21 emerged as consistently discrepant at both $\alpha < 0.05$ and $\alpha < 0.01$ statistical significance levels while Class #26 tallied fewer statistically significant item-level unstandardized differences in means as its smaller class size produced wider confidence intervals. Overall, the students and teachers in these two classes were discrepant across a range of instructional practices while, by contrast, nine other classes recorded two or fewer item discrepancies. This suggests perceptual difference was greater for some classes than for others. The next section further examines this question by comparing the magnitude of class-aggregated unstandardized absolute differences in means across items.

5.3 Ranking Classes by Composite Unstandardized Absolute Differences in Means

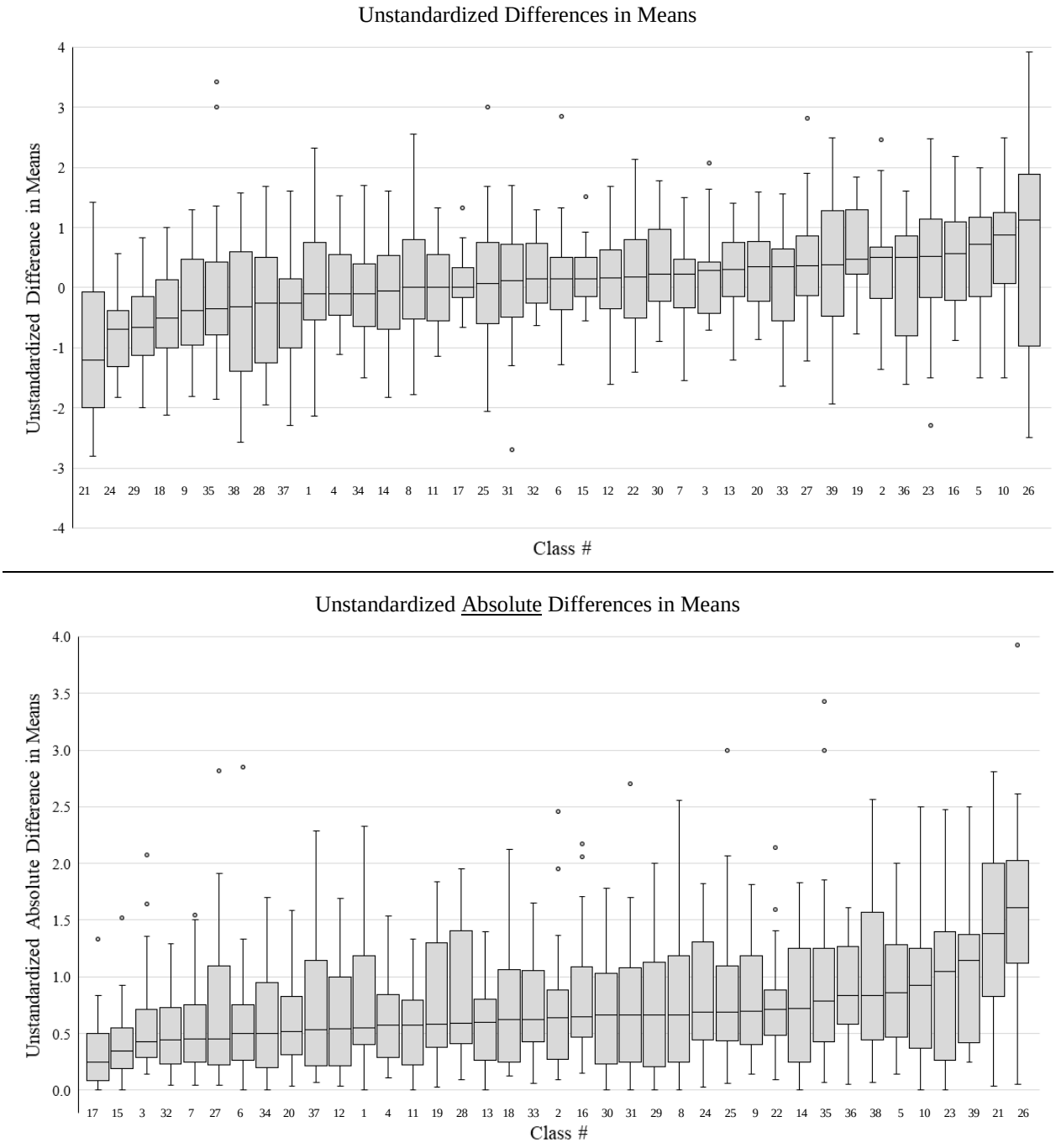
Class composite unstandardized absolute differences in means are a more computationally complex measure of magnitude of perceptual difference within classrooms,

found by aggregating the absolute values of all 23 item-level unstandardized difference in means estimates between students and teacher for a single classroom. Like composite absolute discrepancy scores in Section 5.1, directional information is sacrificed using this measure so that discrepancy can be appropriately summed across items (aggregating differences in means across items without first computing absolute values produces an inconsequential composite measure, as positive and negative difference estimates are canceled out). Broadly speaking, discrepancy results corroborated the class differences found from tallying instances of item-level statistically significant unstandardized differences in means across classrooms.

5.3.1 Exploratory Analysis of Discrepancy Across Classes Using Box Plots

Before computing class composite unstandardized absolute differences in means, it can useful to construct box plots of the 23 item unstandardized differences in means (top) and 23 item unstandardized *absolute* differences in means (bottom) for each class as shown in Figure 5.3. This exploratory, less formal approach revealed considerable differences in the direction and degree of discrepancy across classes. The two most discrepant teacher–classroom pairs in terms of composite absolute discrepancy scores—Classes #26 and #21— also had the highest and lowest median unstandardized difference in means, respectively. Indeed, the interquartile range of unstandardized differences in means for Class #26 was the widest of all the teacher–classroom pairs, as was its overall range, which extended from -2.50 to 3.93. By contrast, the median unstandardized difference in means for two of the least discrepant teacher–classroom pairs, Classes #17 and #15, were both around zero and their interquartile ranges were noticeably more compact. Box plots of the item unstandardized absolute differences in means corroborated that clear differences in the magnitude of discrepancy exist by class. The most discrepant teacher–classroom pairs, Classes #26 and #21, had the largest median unstandardized difference in means and wide interquartile ranges; the least discrepant teacher–classroom pairs, Classes #17 and #15, had the smallest median unstandardized difference in means and compact interquartile ranges.

Figure 5.3: *Classroom Box Plots of Unstandardized Differences in Means and Unstandardized Absolute Differences in Means*

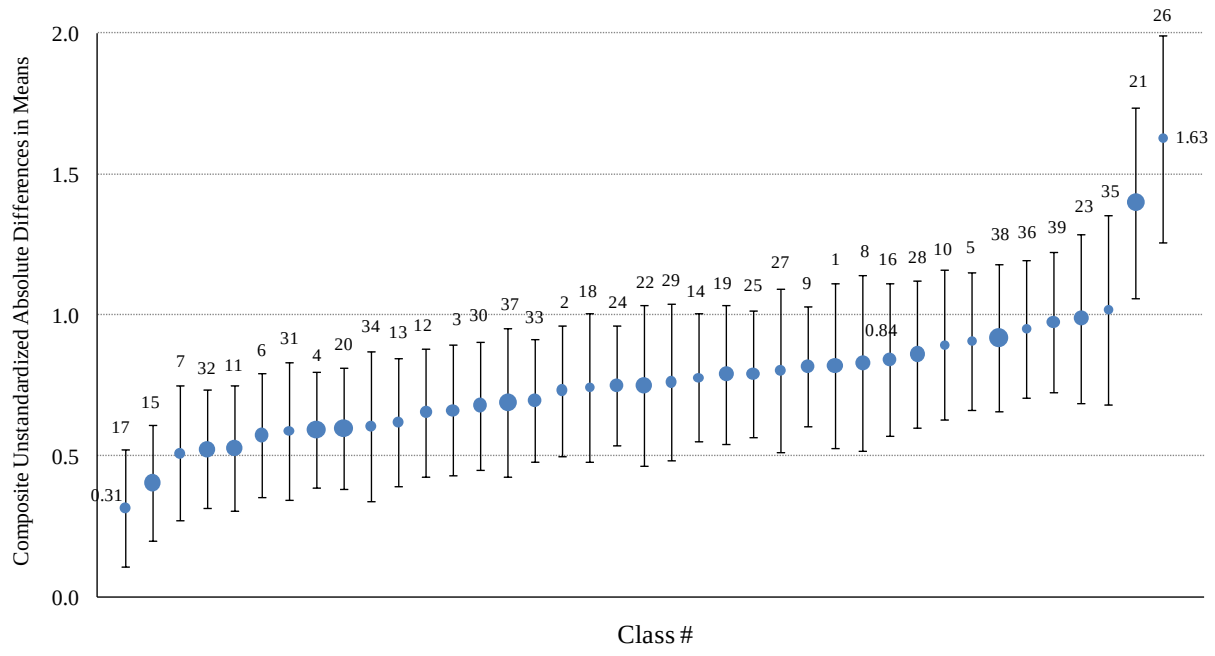


Note. Classrooms in top box plot sorted by median unstandardized difference in means; classrooms in bottom box plot sorted by median unstandardized absolute difference in means.

5.3.2 Computing Composite Unstandardized Absolute Difference in Means

Class composite unstandardized absolute differences in means are shown in Figure 5.4. For example, the composite unstandardized absolute difference in means for Class #16 was 0.84. This ranked it 29th out of the 39 teacher–classroom pairings in terms of magnitude of perceptual alignment. The composite unstandardized absolute difference in means for Class #16 was calculated by (1) using Equation 5.1 to find the unstandardized differences in means for each of the 23 QAS items between the students and teacher in Class #16, (2) taking the absolute value for each of the 23 mean difference estimates, and (3) computing a weighted average absolute difference in means estimate across items. To compute this weighted average, each item absolute difference in means estimate for Class #16 was inversely weighted by its error and parameter variances. Specifically, the weight of each item for Class #16 was found by computing the inverse or reciprocal of the sum of the item's unique squared standard error estimate from Equation 5.2 and an estimate of the parameter variance for Class #16 (tau or the symbol T) found by specifying an unconditional ANOVA model using the statistical program HLM 6 (Raudenbush, Bryk, & Congdon, 2004). This resulting weighted average absolute difference in means for a given class is referred to as the class composite unstandardized absolute difference in means.

Figure 5.4*: Class Composite Unstandardized Absolute Differences in Means*



Note. Larger bubbles indicate classrooms with more students. Classrooms sorted by composite unstandardized absolute difference in means.

The process of aggregating or synthesizing unstandardized differences in means across items for a given class is similar to that of combining standardized differences in means (i.e., effect sizes) across studies in a basic meta-analysis (Raudenbush & Bryk, 2002). The series of computational steps outlined in the previous paragraph is detailed for Class #16.

Recall that the unstandardized difference in means for Class #16 on Item #16 found in Section 5.2.1 using Equation 5.1 was $\overline{X}_{Dif\_Item16\_Class16} = 0.56$ (shown below)

$$\overline{X}_{Dif\_Item16\_Class16} = \overline{X}_{Item16\_Class16} - \overline{X}_{Item16\_Teacher16} = 2.06 - 1.50 = 0.56$$

and standard error of this estimate found in Section 5.2.2 using Equation 5.2 was $SE_{Dif\_Item16\_Class16} = 0.58$. This estimate is a conservative approximation of the standard error of the absolute value of the unstandardized difference in means for Class #16 on Item #16.

$$SE_{Dif\_Item16\_Class16} = \sqrt{\frac{0.43}{2} + \frac{1.93}{17}} = 0.58$$

The standard error of this estimate can be used to compute the error variance (*V*) of Item #16 for Class #16:

$V_{\text{Dif\_Item16\_Class16}} = SE^2_{\text{Dif\_Item16\_Class16}} = 0.58^2 = 0.34$

This process was repeated to compute unstandardized difference in means, associated standard errors of these estimates, and error variances for Class #16 on each of the other 22 QAS items.

Next, the parameter variance for Class #16 across the 23 items (*T*) was found by specifying an unconditional Level-2 model (or between-item model) in HLM. There was no predictor variable in this model. Instead, the unstandardized absolute differences in means for Class #16 were predicted using known class variances. More specifically, the Level-1 model provides a summary of the data for Class #16 on each item:

$\overline{X}_{\text{Dif\_Item16\_Class16}} = \mu_{\text{Dif\_Item16\_Class16}} + e_i,\ e_i \sim N(0, SE_i^2)$

where e_i represents the deviation of the estimate of the difference in means for item *i* from the true difference and the error variance is the squared standard error (*SE*). Here $i = 1, \ldots, 23$. Note that the formula for the *SE* appears in Equation 5.2.

The Level-2 model is as follows:

$\mu_{\text{Dif\_Class16}} = \gamma_0 + U_i,\ U_i \sim N(0, T)$

where U_i is a random effect, and *T* is the random effects variance component or parameter variance. The amount of variance in the true absolute differences in means across the 23 items for Class #16 was found to be $T_{\text{Dif\_Class16}} = 0.12$. Therefore, the composite unstandardized absolute difference in means for Class #16 ($Comp_{\text{Dif\_Class16}}$) can be obtained from the HLM model itself (γ_0) or computed longhand using item error variances (*V*'s) and parameter variance (*T*) as follows:

$$Comp_{\text{Dif\_Class16}} = \frac{\sum_{i=1}^{23}\left[\bar{X}_{\text{Dif\_Class16}}(i) * (1/(V_{\text{Dif\_Class16}}(i) + T_{\text{Dif\_Class16}}))\right]}{\sum_{i=1}^{23}\left[1/(V_{\text{Dif\_Class16}}(i) + T_{\text{Dif\_Class16}})\right]}$$

$$= \frac{\sum_{i=1}^{23}\left[\bar{X}_{\text{Dif\_Class16}}(i) * (1/(V_{\text{Dif\_Class16}}(i) + 0.12))\right]}{\sum_{i=1}^{23}\left[1/(V_{\text{Dif\_Class16}}(i) + 0.12))\right]} = \frac{44.13}{52.37} = 0.84$$

As shown in Figure 5.4, confidence intervals can be built around class composite unstandardized absolute difference in means estimates. To compute these confidence intervals, the standard error of each class composite unstandardized absolute difference in means was found by taking the square root of the inverse of the sum of the inverse total variance estimates (error variance + parameter variance) across the 23 items for the given class. Note that the sum of the inverse of the total variance estimates, 52.37, is the denominator in the equation (above) used to find $Comp_{\text{Dif\_Class16}}$. These standard errors provide a reasonably good estimate of the level of uncertainty in each class composite unstandardized absolute difference in means. The standard error for Class #16 ($SE_{\text{Dif\_Class16}}$) is shown below:

$$SE_{\text{Dif\_Class16}} = \sqrt{\frac{1}{\sum_{i=1}^{23}\left[1/(V_{\text{Dif\_Class16}}(i) + T_{\text{Dif\_Class16}})\right]}} = \sqrt{\frac{1}{\sum_{i=1}^{23}\left[1/(V_{\text{Dif\_Class16}}(i) + 0.12)\right]}} = \sqrt{\frac{1}{52.37}} = 0.14$$

Setting $\alpha < 0.05$ and basing the critical value (t_{critical}) on a *t* distribution with the degrees of freedom (*df*) equal to $n_{\text{Total Items}}$ - 1, the standard errors were used to construct approximate 95% confidence intervals quantifying the margin of error around each class composite unstandardized difference in means. The computation of the lower and upper bounds of the 95% confidence interval around the composite unstandardized absolute difference in means for Class #16 is shown below:

Lower confidence interval limit = $\bar{X}_{\text{Dif\_Class16}} - (t_{\text{critical (}df\_22\text{)}} * SE_{\text{Dif\_Class16}})$

$$= 0.84 - (2.07 * 0.14) = 0.55$$

Upper confidence interval limit = $\bar{X}_{\text{Dif\_Class16}} + (t_{\text{critical (}df\_22\text{)}} * SE_{\text{Dif\_Class16}})$

$$= 0.84 + (2.07 * 0.14) = 1.13$$

As can be seen in Figure 5.4, the confidence interval for Class #16 ranged from 0.55 to 1.13. Thus, one can be 95% confident that the true composite unstandardized absolute difference in means for Class #16 resides within this interval. Note that the confidence interval of the mean difference measurement overlapped with all but two of the 38 other QAS teacher–classroom pairs: the lower limit of the confidence interval did not overlap with Class #17 (the least discrepant class) while the upper limit did not overlap with Class #26 (the most discrepant class). The composite unstandardized absolute difference in means for Class #16 was not statistically different from the differences in means of the remaining 36 classes.

5.3.3 Ranking Classes by Composite Unstandardized Absolute Differences in Means

Figure 5.4 shows the considerable variation between classrooms in the magnitudes of the composite unstandardized absolute differences in means. Indeed, the composite difference in means for the highest class, Class #26, was 5.2 times the lowest unstandardized difference in means of Class #17 ($\overline{X}_{Dif\_Class26}$ = 1.63 versus $\overline{X}_{Dif\_Class17}$ = 0.31). Furthermore, the lower limit of the 95% confidence interval of the difference in means estimate for Class #26 did not overlap with the upper limits of the 95% confidence intervals of the estimates for 35 other classes. Likewise, the lower limit of the 95% confidence interval of the difference in means estimate for Class #21—the most discrepant class in terms of item counts of statistically significant unstandardized differences in means—did not overlap with the upper limits of the confidence intervals of the estimates of 25 other classes. Thus, in absolute terms, students on the whole in Classes #26 and #21 held more discrepant perceptions from their teacher than students overall in the majority of the other classes.

5.3.4 Section Summary

This section ranked classes by composite unstandardized absolute differences in means, a measure of magnitude of teacher–classroom perceptual difference aggregated across all 23 QAS items. Class rankings on composite absolute difference in means—as well as descriptive class box plots—revealed conclusions broadly similar to item counts of statistically significant

unstandardized differences in means. Class #21 was the most discrepant class in the top and bottom histograms in Figure 5.2 and the second most discrepant class in Figure 5.4. Similarly, the least discrepant class in terms of composite unstandardized absolute differences in means—Class #17—was tied for the fewest statistically significantly discrepant items. The 95% confidence intervals of the composite absolute difference in means estimates for the most and least discrepant classes also did not overlap. Thus, comparisons of composite unstandardized absolute differences in means across classes further supported greater perceptual difference between students and their teacher in some classrooms than in others.

5.4 Chapter Summary

The six methods of measuring discrepancy employed in this dissertation are not universally applicable or without limitations, but best utilized complementarily. For this reason, the analyses in this chapter employed a multiple measures approach—discrepancy scores, item-level unstandardized differences in means, and classroom composite unstandardized absolute differences in means—to compare the perceptual discrepancy between students and teachers in the 39 QAS classrooms.

Perceptual differences were more pronounced between some teacher–classroom pairs than between others. For example, Class #21 and Class #26 had the largest class composite absolute discrepancy scores, the greatest count of statistically significant item-level unstandardized differences in means evaluated at $\alpha < 0.05$, and the largest class composite unstandardized absolute differences in means. Conversely, Class #17 had the smallest class composite absolute discrepancy score, fewest number of statistically significant item unstandardized differences in means, and the smallest composite unstandardized absolute differences in means. Thus, regardless of the approach employed, definite differences were observed between the most and least discordant classrooms in their degree (and subsequent ranking) of discrepancy.

While there were ranges in perceptual differences for the most and least discrepant

classes, discrepancy rankings for classes in the middle of the distribution were more fluid depending on the discrepancy approach utilized. For example, Class #3 was ranked as the second least discrepant class using composite absolute discrepancy scores, yet the thirteenth least discrepant class using composite unstandardized absolute differences in means. This ranking difference was due to the fact that composite discrepancy scores do not explicitly account for student and teacher response variation, while composite unstandardized absolute differences in means *do* account for variation. However, the spread estimates in Class #3 were based on a sample of only seven student ratings and thus were quite noisy. The rankings of Classes #5, #9, and #20 also changed considerably depending on whether rankings were based on composite absolute discrepancy scores or composite unstandardized absolute differences in means. Yet, these ranking differences were less substantively meaningful as differences were accentuated by the clustering of composite estimates for classes in the middle of the distribution. Indeed, raw differences across discrepancy method were markedly less pronounced. The next chapter takes a closer look at these class discrepancy differences by exploring the degree to which perceptual differences are dependent on the particular instructional practices examined.

CHAPTER 6

6. Is Class Discrepancy Dependent on the Instructional Practice Examined?

Measuring perceptual differences using composite unstandardized absolute differences in means, the previous chapter found that Classes #21 and #26 had the highest discrepancy estimates without considering particular instructional practices and Class #17 had the lowest discrepancy estimates. Ranking classes by the magnitude of their unstandardized differences in means, this chapter investigates whether this finding remains the same or changes when particular QAS practices are considered. Student and class composite absolute discrepancy scores are not used in these analyses as these measures aggregate differences across all items, and thus are not applicable to investigating the influence of particular QAS practices. The main research question of this chapter is the following:

2. Are discrepant teacher–classroom pairs dependent on the instructional practice examined?

Additionally, in Section 6.3 of this chapter discrepancy plots are used to closely examine perceptual differences in the two most and least discrepant teacher–classroom pairs.

6.1 Dependency of Discrepancy Across Instructional Practice Items

One approach to exploring the impact of particular QAS items on discrepancy is to replicate the analyses in Section 5.3 for *each* of the 23 QAS instructional practice items. This section begins by broadly examining the influence of instructional practices on class discrepancy by ranking all 39 classes on all 23 QAS items by the magnitudes of their unstandardized differences in means. Then, the unstandardized differences in means are examined in greater depth for two select instructional practices: Item #1 *listen to lectures or instruction directed by the teacher* and Item #14 *watch teacher demonstrate experiment or investigation*. Overall,

perceptual differences for the most and least discrepant classes were not especially reliant on specific instructional practice items.

6.1.1 Class Rankings Across All Items

Defining discrepancy in absolute terms on a single instructional practice, Table 6.1 displays the classroom rankings using unstandardized differences in means for each QAS item—from 1 (most discrepant) to 39 (least discrepant)—with QAS classrooms sorted by their composite unstandardized absolute differences in means. Table 6.1 provides an informative, bird's-eye view of differences in class rankings by instructional practice. For example, for the most discrepant classes (Class #26 and Class #21), it does not appear that perceptual differences were particularly practice-dependent. Class #26 was one of the ten most discrepant classrooms for 17 of the 23 items, while Class #21 was one of the ten most discrepant classrooms for 15 of the 23 items (see Table 6.1).

Examining the least discrepant classes in Table 6.1 (#17 and #15) also revealed perceptual differences were generally not especially practice-dependent. In contrast to the two most discrepant classrooms, Classes #17 and #15 tallied far fewer top-ten discrepancy rankings across the items. Even combining the instances, these two classrooms only produced one top-ten discrepancy ranking. This single instance occurred for Class #17, which actually ranked as the *most* discrepant overall classroom in terms of magnitude on Item #13 *present oral science reports*. Thus, there were some items for which the most and least discrepant classrooms did not follow the broader trends.

Table 6.1: *Classroom Unstandardized Absolute Difference in Means Rankings for Each Item*

		Individual Work					Interactive Work			Discussion		Reports & Projects				Experimental Work					Assessments				Class
ID	N	#1	#14	#2	#6	#7	#3	#8	#9	#4	#5	#10	#11	#12	#13	#15	#16	#17	#18	#19	#20	#21	#22	#23	Comp
#17	12	34	20	23	32	25	39	25	33	36	32	20	19	36	1	38	37	39	20	38	32	39	26	26	0.31
#15	26	22	22	38	38	29	32	30	38	35	18	24	31	18	30	36	25	27	31	29	13	32	19	14	0.41
#7	11	32	19	10	4	23	33	32	28	20	21	25	9	38	29	31	18	37	21	23	30	18	35	28	0.51
#32	25	13	11	16	35	27	17	19	16	34	37	37	26	16	34	37	27	30	37	25	19	16	23	33	0.52
#11	21	18	33	38	14	19	14	22	32	37	20	19	11	17	16	9	39	10	39	21	23	33	30	12	0.53
#6	18	39	16	28	21	3	37	26	25	24	9	27	28	26	3	28	30	12	35	27	27	6	22	38	0.57
#31	9	37	27	9	18	37	24	27	4	23	16	16	2	21	31	38	10	6	8	7	33	27	38	6	0.59
#4	28	7	33	29	11	20	20	17	27	18	19	36	30	35	37	13	16	31	27	22	24	12	18	25	0.59
#20	29	23	24	25	6	9	12	36	15	39	11	34	8	24	19	10	28	33	30	32	28	22	20	10	0.60
#34	10	29	12	36	39	27	11	7	10	22	29	39	15	22	22	26	14	38	33	37	22	35	10	2	0.60
#13	10	17	6	27	8	26	26	14	39	27	12	38	25	30	38	3	20	35	32	15	39	24	23	19	0.62
#12	13	16	23	35	29	38	16	3	11	29	35	3	12	34	12	33	7	15	23	12	37	8	33	36	0.65
#3	14	5	10	26	24	32	31	31	34	26	28	14	36	15	26	24	22	20	16	34	8	31	8	30	0.66
#30	18	21	4	33	36	7	2	8	19	33	7	11	14	23	39	17	11	19	14	33	35	37	36	18	0.68
#37	28	36	37	29	11	17	9	35	30	4	15	6	7	7	33	30	36	31	17	20	3	2	28	32	0.69
#33	17	28	38	14	10	10	18	20	7	11	10	18	23	8	14	19	4	22	9	10	31	34	34	29	0.70
#2	11	20	35	4	22	23	36	11	36	2	39	27	35	19	8	14	8	21	28	18	4	17	27	31	0.73
#18	8	14	31	18	34	30	5	12	35	3	23	5	24	33	34	15	31	8	4	35	25	30	31	17	0.74
#24	17	33	28	20	26	12	25	23	13	9	27	4	6	3	2	18	5	14	13	6	29	38	25	37	0.75
#22	22	12	21	15	31	22	29	15	36	16	29	14	27	4	5	11	29	23	18	3	26	20	13	20	0.75
#29	12	34	18	37	15	34	19	39	9	15	36	7	19	6	7	2	31	9	2	14	38	26	37	23	0.76
#14	9	27	8	11	28	35	6	16	6	7	5	35	33	9	27	8	33	23	38	38	9	9	9	34	0.78
#19	19	31	17	32	3	5	38	4	5	31	22	23	29	28	23	25	26	26	19	11	12	11	6	21	0.79
#25	16	24	15	17	30	1	21	9	18	5	23	2	38	12	13	22	38	11	36	28	16	14	21	22	0.79
#27	11	3	36	13	2	16	8	5	8	38	33	33	32	29	9	27	35	36	34	23	5	19	17	35	0.80
#9	15	9	25	2	27	21	27	9	21	10	8	16	10	31	24	5	16	16	24	26	21	25	28	24	0.82
#1	23	26	39	1	1	4	15	34	17	28	38	26	21	20	21	16	24	25	29	36	6	15	3	8	0.82
#8	18	18	29	12	37	18	30	29	20	8	17	13	13	39	36	12	33	18	22	31	1	3	1	4	0.83
#16	17	4	25	20	16	15	22	24	24	25	26	10	18	25	20	34	23	17	26	13	7	7	1	5	0.84
#28	22	25	13	5	33	14	1	18	23	14	3	27	22	32	16	7	21	1	3	30	36	29	39	3	0.86
#10	8	11	31	6	7	39	28	33	29	21	23	12	3	10	4	32	12	29	4	8	14	4	14	7	0.89
#5	7	14	5	3	11	11	13	2	21	30	34	32	36	5	32	35	8	28	10	9	20	23	11	9	0.91
#38	31	38	9	24	20	31	34	20	1	12	1	22	17	1	6	29	3	34	11	5	10	13	16	16	0.92
#36	9	10	29	20	17	8	23	13	12	13	6	27	39	11	11	6	19	12	12	17	15	28	5	11	0.95
#39	16	8	1	29	25	13	7	6	26	17	4	27	5	26	16	22	15	7	24	19	11	5	12	27	0.98
#23	20	6	2	34	9	33	35	27	3	32	13	9	15	14	15	20	1	4	6	2	18	36	15	38	0.99
#35	7	2	7	7	19	1	10	37	31	19	29	21	34	13	27	4	2	5	15	16	34	21	32	13	1.02
#21	26	30	3	19	23	36	4	38	14	1	1	1	1	2	10	1	6	3	1	4	17	10	7	15	1.40
#26	7	1	14	8	4	6	3	1	2	6	14	8	4	37	24	21	13	2	6	1	2	1	4	1	1.63

Note. ID is the Class ID #; N is the class size; Class Comp is the Class composite unstandardized absolute differences in means; column headings (#1–#23) correspond to QAS items (see Table 4.1); items grouped according to confirmatory factor analysis of QAS student items in Appendix Table A.6; numbers highlighted in darker **Red** indicate greater absolute difference in means.

It is also worth noting that the most discrepant classrooms were not discrepant on the same items with certainty. For example, in terms of unstandardized absolute differences in means, Class #26 was ranked as the most discrepant class on Item #8 *watch science videos, movies, TV shows, etc* and Class #21 was ranked as second *least* discrepant class. Conversely, Class #21 was ranked as the second most discrepant class on Item #12 *work on written science reports* while Class #26 was ranked as the third *least* discrepant class. Therefore, even though Classes #21 and #26 were clearly the two most discrepant classes in the QAS sample, there were item-level exceptions to these aggregate discrepancy trends.

Further analyses are based on factor analyses (see Appendix A.6), which revealed six instructional practice domains: Individual Work (Items #2, #6, and #7), Interactive Work (Items #3, #8, and #9), Discussion (Items #4 and #5), Reports and Projects (Items #10–#13), Experimental Work (Items #15–#19), and Assessments (Items #20–#23). Additionally, two singleton items were identified that did not load on these common factors: Item #1 *listen to lectures or instruction directed by the teacher* and Item #14 *watch teacher demonstrate experiment or investigation*. The two items were distinct from the other QAS instructional practices as they describe passive learning in the classroom from observing teacher pedagogy. For simplicity of interpretation, class discrepancy rankings on the two singleton items are discussed first, followed by a discussion of class discrepancy rankings by the items clustered in the six instructional practice domains.

6.1.2 Class Rankings Across Items #1 and #14

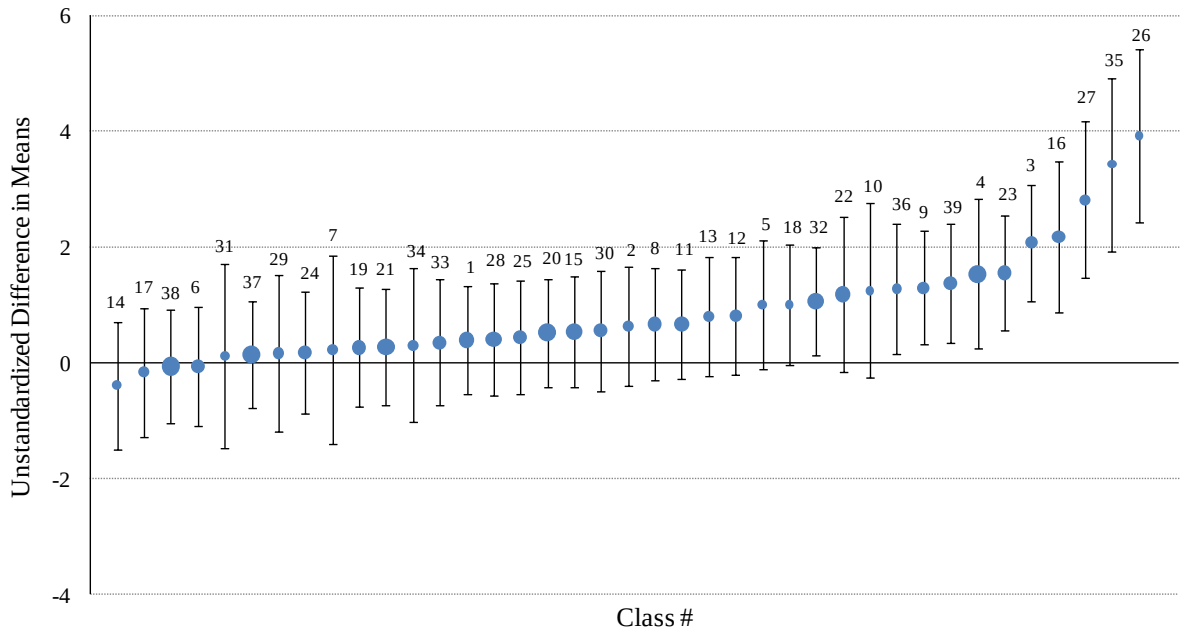
It can also be useful to examine the impact of discrepancy for a few choice practices in greater depth. Therefore, the unstandardized differences in means were ranked across the classrooms on Item #1 *listen to lectures or instruction directed by the teacher* and Item #14 *watch teacher demonstrate experiment or investigation*. These two items were investigated separately because they are the only instructional practices not easily categorized using factor analytic procedures. For the remainder of the QAS items, discrepancy patterns across classrooms

are examined as clusters of items that define a particular instructional domain. These analyses are presented in the subsequent section.

It is important to note that, unlike class discrepancy rankings in 6.1.1, in this section the absolute value of the unstandardized differences in means for Items #1 and #14 was *not* computed. This was because when a sampling distribution of the unstandardized difference in means spans positive and negative values—as occurs when the true difference is small—the mean of the sampling distribution of the unstandardized *absolute* differences in means will be positively skewed, and thus greater than the mean of the sampling distribution of the unstandardized differences in means, which is normally distributed. Additionally, this positive skewness will result in smaller standard error estimates of the sampling distribution of the unstandardized absolute differences in means than the standard errors of the sampling distribution of the unstandardized differences in means. Simply put, the sampling distribution of the unstandardized absolute difference in means is not as well understood. Fortunately, the positive, modest bias of the sampling distribution of the unstandardized absolute difference in means is greatest in situations with the least discrepancy (i.e., when the true difference in means is close to zero). Nonetheless, because of this bias, only unstandardized differences in means across the classes are plotted in Figures 6.1 and 6.2.

The unstandardized differences in means across the classrooms on Item #1 *listen to lectures or instruction directed by the teacher* is shown in Figure 6.1. In total, 11 of the 39 classrooms had differences in means different from zero—all statistically significantly positive. Class #26 ranked as the most discrepant class while the unstandardized difference in means rankings of Class #21 were considerably lower (29th overall). Thus, of the two most discrepant classrooms overall, Class #26 had the highest unstandardized differences in means on one of the most discordant items (as will be shown in Chapter 7), while the mean difference for Class #21 was ranked in the middle of the distribution.

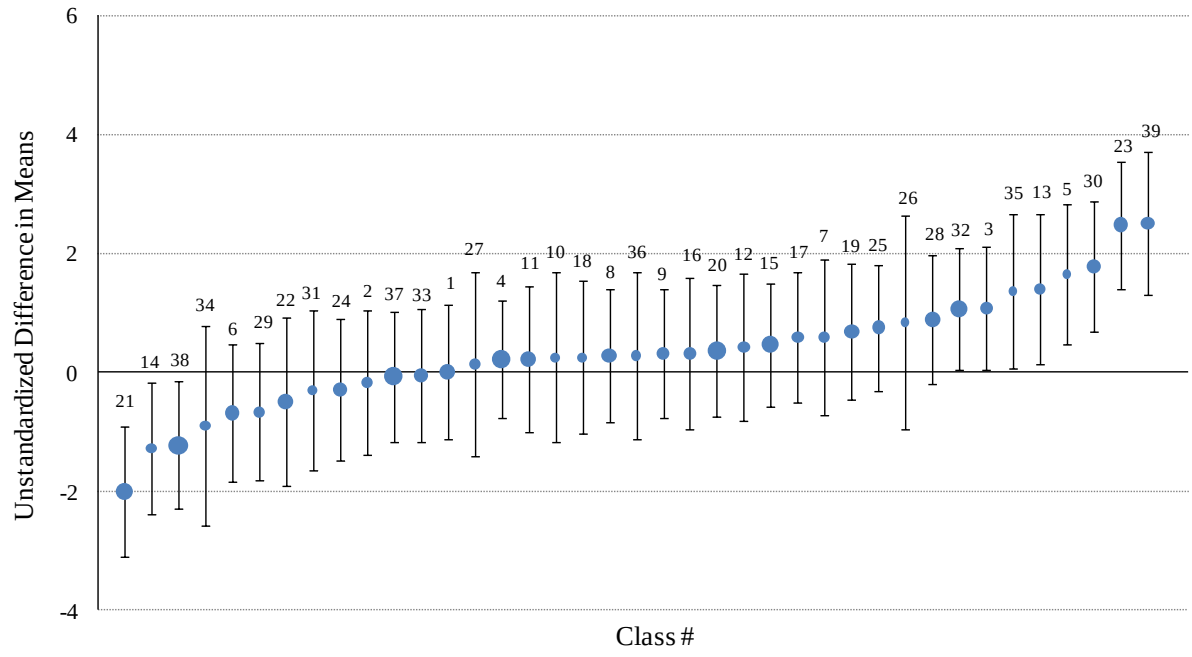
Figure 6.1: *Unstandardized Differences in Means, by Classroom, Item #1 Listen to Lectures or Instruction Directed by the Teacher*



Note. Larger bubbles indicate classrooms with more students. Classrooms sorted by unstandardized difference in means on Item #1.

Discrepancy across classrooms on Item #14 *watch teacher demonstrate experiment or investigation* yielded a different ranking pattern than that found on Item #1. In total, 10 of the 39 classrooms had differences in means statistically different from zero—seven statistically significantly positive and three statistically significantly negative. In terms of the most and least discrepant classrooms on Item #14, Class #21 ranked as the most negatively discrepant classroom on this item. The other most discrepant teacher–classroom pair (Class #26) as well as the least discrepant teacher–classroom pairs (Classes #15 and #17) ranked in the middle of the discrepancy distribution on this item (see Figure 6.2).

Figure 6.2: *Unstandardized Differences in Means, by Classroom, Item #14 Watch Teacher Demonstrate Experiment or Investigation*



Note. Larger bubbles indicate classrooms with more students. Classrooms sorted by unstandardized difference in means on Item #14.

6.1.3 Section Summary

This section investigated the influence of instructional practice items on class unstandardized differences in means and class unstandardized absolute differences in means. Ranking all 39 classes on all 23 QAS items by the magnitudes of their unstandardized differences in means revealed, broadly speaking, that perceptual differences for the most and least discrepant classes were not especially dependent on specific instructional practice items. For example, either Class #21 or #26 was one of the ten most discrepant classrooms for all 23 of the instructional practices. The next section expands upon these analyses by examining whether particular instructional practice domains affected perceptual differences in discrepant classes.

6.2 Dependency of Discrepancy Across Instructional Domains

Another approach to exploring the influence of particular QAS items on discrepancy is to replicate the analyses in Section 5.3 for clusters of items that define a particular instructional domain. Using exploratory and then confirmatory factor analyses (see Appendix Table A.6), six

instructional practice domains were identified in the QAS surveys: Individual Work (Items #2, #6, and #7), Interactive Work (Items #3, #8, and #9), Discussion (Items #4 and #5), Reports and Projects (Items #10–#13), Experimental Work (Items #15–#19), and Assessments (Items #20–#23). Magnitudes of unstandardized differences in means were aggregated across the items in each domain in a similar process to that performed in creating composite unstandardized absolute differences in means—which aggregate mean differences across all 23 QAS items. Classrooms were then ranked on these six domains from 1 (most discrepant) to 39 (least discrepant) by the magnitude of their unstandardized differences in means. Results are presented in Table 6.2.

6.2.1 Class Rankings Across All Instructional Practice Domains

Examining class discrepancy rankings by instructional domain revealed clear patterns for the most and least discrepant classes. Class #21 was the most discrepant class across three of the six instructional practice domains (Discussion, Reports and Projects, and Experimental Work) in terms of the magnitude of its unstandardized differences in means. Likewise, Class #26 was one of the ten most discrepant classes in all six instructional practice domains. In a similar pattern, the classrooms with the lowest unstandardized absolute differences in means (Classes #15 and #17) were often ranked as the least discrepant classes in all (or nearly all) of the domains. For example, Class #15 was ranked as the least discrepant class in Individual Work and the second least discrepant class in Interactive Work; Class #17 was ranked as the least discrepant class in Experimental Work and Assessments. Consequently, the most and least discrepant classes were often discrepant on different domains. This aligns with item-level discrepancy patterns, which showed that perceptual differences were evident across many of the items. By contrast, for classes in the middle of the distribution of composite unstandardized absolute differences in means, there was less of a discernible interaction pattern between these classes and the instructional domains on which they were discordant. These results suggest discrepancy was driven as much by the actual teacher–classroom pair as the practice itself.

Table 6.2: *Classroom Unstandardized Absolute Difference in Means Rankings for QAS Domains*

ID	N	Individual Work	Interactive Work	Discussion	Reports and Projects	Experimental Work	Assessments	Composite
#17	12	33	39	38	7	39	36	0.31
#15	26	39	38	32	35	37	20	0.41
#7	11	5	36	21	28	34	33	0.51
#32	25	28	18	37	37	38	22	0.52
#11	21	29	27	34	14	26	28	0.53
#6	18	15	34	16	17	29	25	0.57
#31	9	17	20	18	13	16	35	0.59
#4	28	24	22	17	39	25	21	0.59
#20	29	11	25	29	22	31	24	0.60
#34	10	38	8	26	30	35	15	0.60
#13	10	16	30	20	38	18	30	0.62
#12	13	37	4	33	8	19	31	0.65
#3	14	34	37	30	27	22	12	0.66
#30	18	30	3	25	31	17	39	0.68
#37	28	21	31	4	12	33	13	0.69
#33	17	10	16	10	15	9	38	0.70
#2	11	13	19	13	20	20	19	0.73
#18	8	35	11	14	16	10	27	0.74
#24	17	22	23	15	2	7	32	0.75
#22	22	26	29	19	6	15	17	0.75
#29	12	36	24	22	5	2	37	0.76
#14	9	25	9	3	36	32	9	0.78
#19	19	8	6	31	33	24	7	0.79
#25	16	12	13	8	10	28	18	0.79
#27	11	4	2	39	26	36	16	0.80
#9	15	7	15	7	21	13	26	0.82
#1	23	1	26	36	25	30	4	0.82
#8	18	20	32	11	32	27	1	0.83
#16	17	19	28	28	19	23	3	0.84
#28	22	14	10	5	29	6	34	0.86
#10	8	9	35	23	3	11	5	0.89
#5	7	6	5	35	34	21	11	0.91
#38	31	32	14	2	4	12	14	0.92
#36	9	18	12	6	23	8	8	0.95
#39	16	27	7	12	18	14	10	0.98
#23	20	23	21	27	11	4	23	0.99
#35	7	3	33	24	24	5	29	1.02
#21	26	31	17	1	1	1	6	1.40
#26	7	2	1	9	9	3	2	1.63

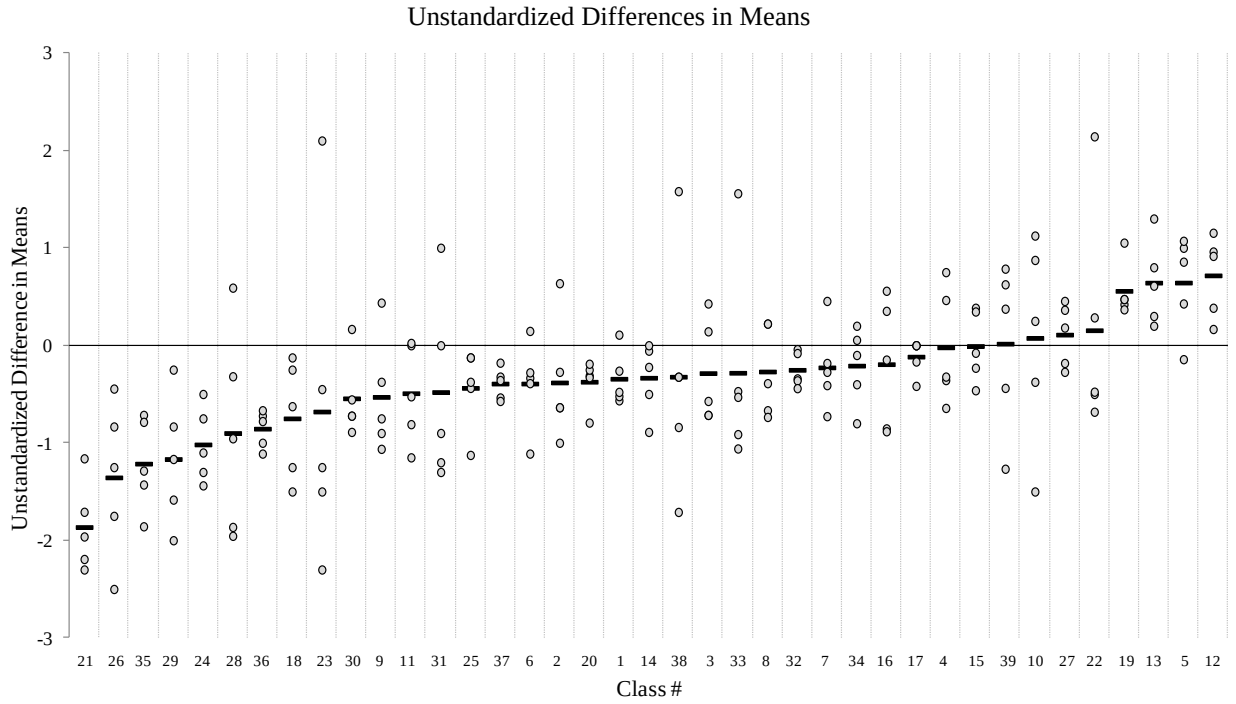
Note. ID is the Class ID #; N is the class size; Composite is the composite unstandardized absolute difference in means; domains from confirmatory factor analysis of QAS student items in Appendix Table A.6. Each domain comprised of the following items: Individual Work (Items #2, #6, and #7), Interactive Work (Items #3, #8, and #9), Discussion (Items #4 and #5), Reports and Projects (Items #10–#13), Experimental Work (Items #15–#19), and Assessments (Items #20–#23). Numbers highlighted in darker **Red** indicate greater absolute differences in means; rows sorted from lowest to highest composite unstandardized absolute difference in means.

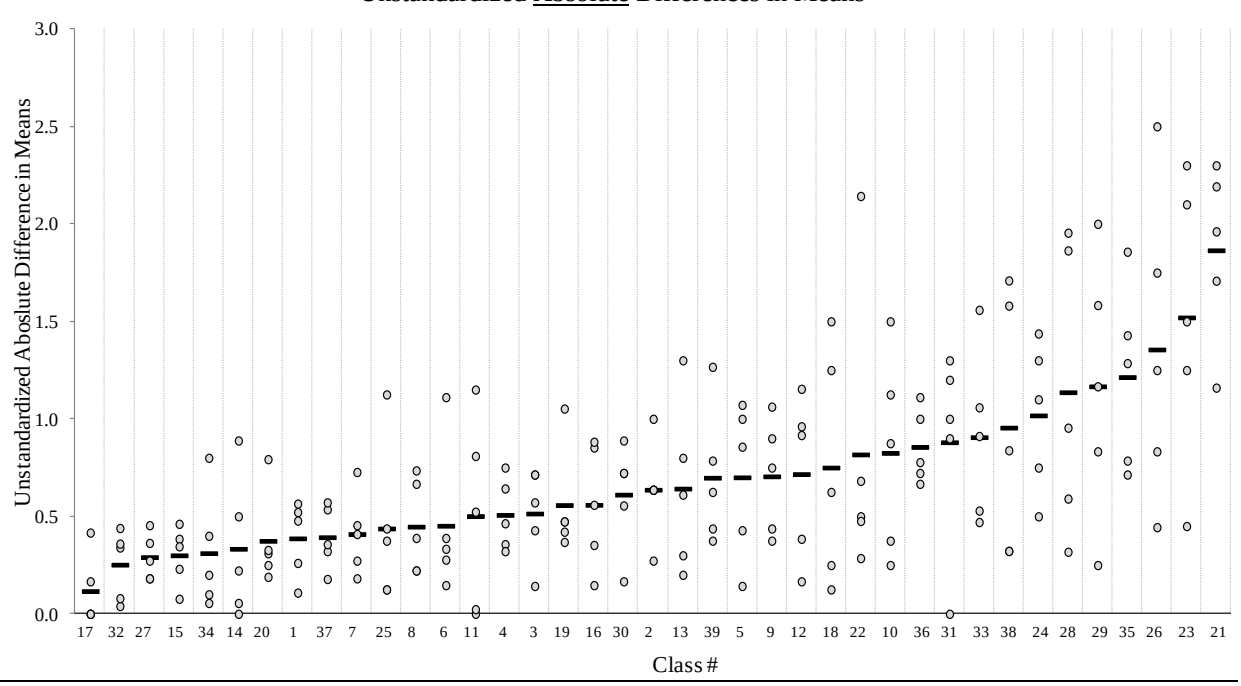
6.2.2 Class Rankings Across Experimental Work and Assessment Instructional Practice Domains

It is also useful to examine the impact of discrepancy for a few select domains in greater depth. Therefore, the unstandardized differences in means and unstandardized absolute differences in means for the items in the Experimental Work and Assessment instructional domains are plotted in Figure 6.3 and Figure 6.4, respectively. Plotting the unstandardized differences in means reveals directional discrepancy for items comprising a given domain, whereas plots of unstandardized absolute differences in means illustrate magnitudinal differences (i.e., the absolute deviation of a class's average ratings from its corresponding teacher's average rating). Experimental Work and Assessment domains were chosen because of their relatively high factor loadings (see Appendix Table A.6). Plots of the unstandardized differences in means and unstandardized absolute differences in means for the items comprising the other four domains of instructional practice can be found in Appendix Figures A.1–A.4.

The plots in Figure 6.3 display the unstandardized differences in means (top) and unstandardized absolute differences in means (bottom) across the classrooms for Items #15–#19 comprising the Experimental Work domain. Discrepancy patterns on this domain were a microcosm of class composite unstandardized absolute differences in means (see Section 5.3). The two most discrepant teacher–classroom pairs, Class #21 and Class #26, had the lowest (most negative) average unstandardized difference in means on Experimental Work practices. Thus, Teacher #21 and Teacher #26 consistently rated Experimental Work items as occurring more frequently than their students' average ratings. In terms of magnitude of perceptual difference, Class #21 and Class #26 also had the highest and third-highest average unstandardized absolute differences in means on the Experimental Work domain. Likewise, the least discrepant teacher–classroom pairs using class composite unstandardized absolute differences in means had low average unstandardized absolute difference in means on Experimental Work items. Classes #17 and #15 ranked as the least discrepant and the fourth least discrepant teacher–classroom pairs in terms of magnitude.

Figure 6.3: *Unstandardized Differences in Means and Unstandardized Absolute Differences in Means, by Classroom, Items #15–#19 in Experimental Work Domain*

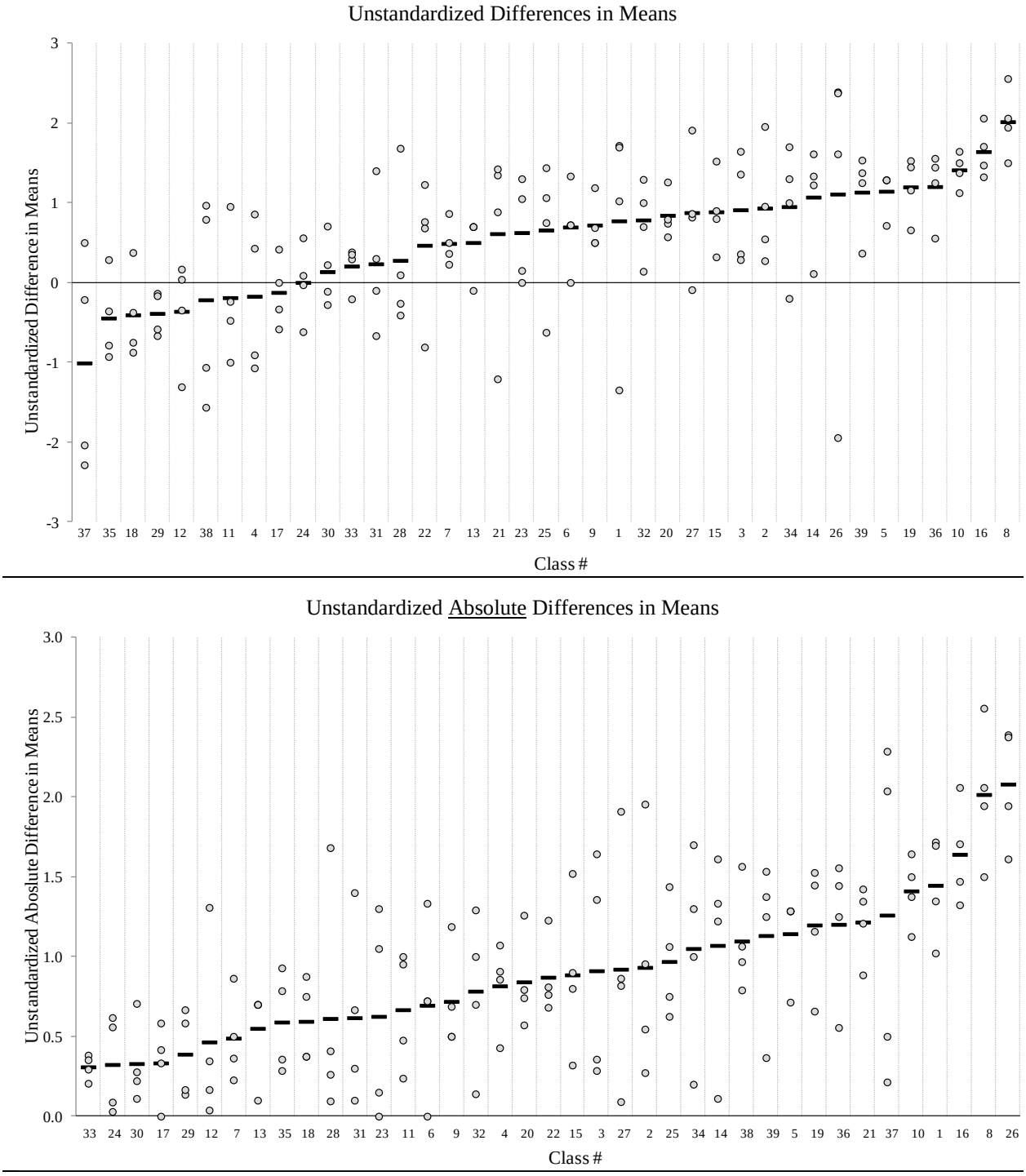




Note. Classrooms in top plot sorted by average unstandardized difference in means on Experimental Work domain (represented by the dashes); classrooms in bottom plot sorted by average unstandardized absolute difference in means on Experimental Work domain (represented by the dashes).

Discrepancy across classrooms on Items #20–#23 comprising the Assessment domain highlights many of the same most and least discrepant classrooms (see Figure 6.4). For instance, one of the two most discrepant teacher–classroom pairs, Class #26, had the highest unstandardized absolute difference in means while Class #21 was the seventh most discrepant teacher–classroom pair in terms of magnitude. Interestingly, the unstandardized difference in means for Classes #26 and #21 masked this magnitudinal discrepancy, as these classes were ranked more in the middle of the distribution using this metric. This was because each of these teacher–classroom pairs had both positive and negative unstandardized difference in means on individual assessment items. As a result, the average unstandardized difference in means on the Assessment domain for these classes was somewhat muted. In terms of the least discrepant teacher–classroom pairs, Class #17 had the fourth lowest ranked unstandardized absolute difference in means. Finally, the average unstandardized difference in means and average unstandardized absolute difference in means for Class #15 on Assessment practices did not align with its composite ranking, as this class ranked in the middle of the distribution on each of these metrics.

Figure 6.4: *Unstandardized Differences in Means and Unstandardized Absolute Differences in Means, by Classroom, Items #20–#23 in Assessment Domain*



Note. Classrooms in top plot sorted by average unstandardized difference in means on Assessment domain (represented by the dashes); classrooms in bottom plot sorted by average unstandardized absolute difference in means on Assessment domain (represented by the dashes).

6.2.3 Section Summary

This section investigated the influence of instructional practice domains on class unstandardized difference in means and class unstandardized absolute difference in means. Ranking all 39 classes on six domains—Individual Work, Interactive Work, Discussion, Reports and Projects, Experimental Work, and Assessments—domain-specific magnitudes of their unstandardized differences in means showed that perceptual differences for the most discrepant classes were not dependent on a single dimension of instructional practice. Rather, students and their teacher in the most discrepant classes were perceptually different across multiple, often non-overlapping, domains—a finding similar to that of discrepancy patterns on particular items. These results suggest that discrepancy was driven as much by the students and teachers in the class as by the instructional practice itself.

6.3 Examining Discrepancy in the Most and Least Discrepant Classes

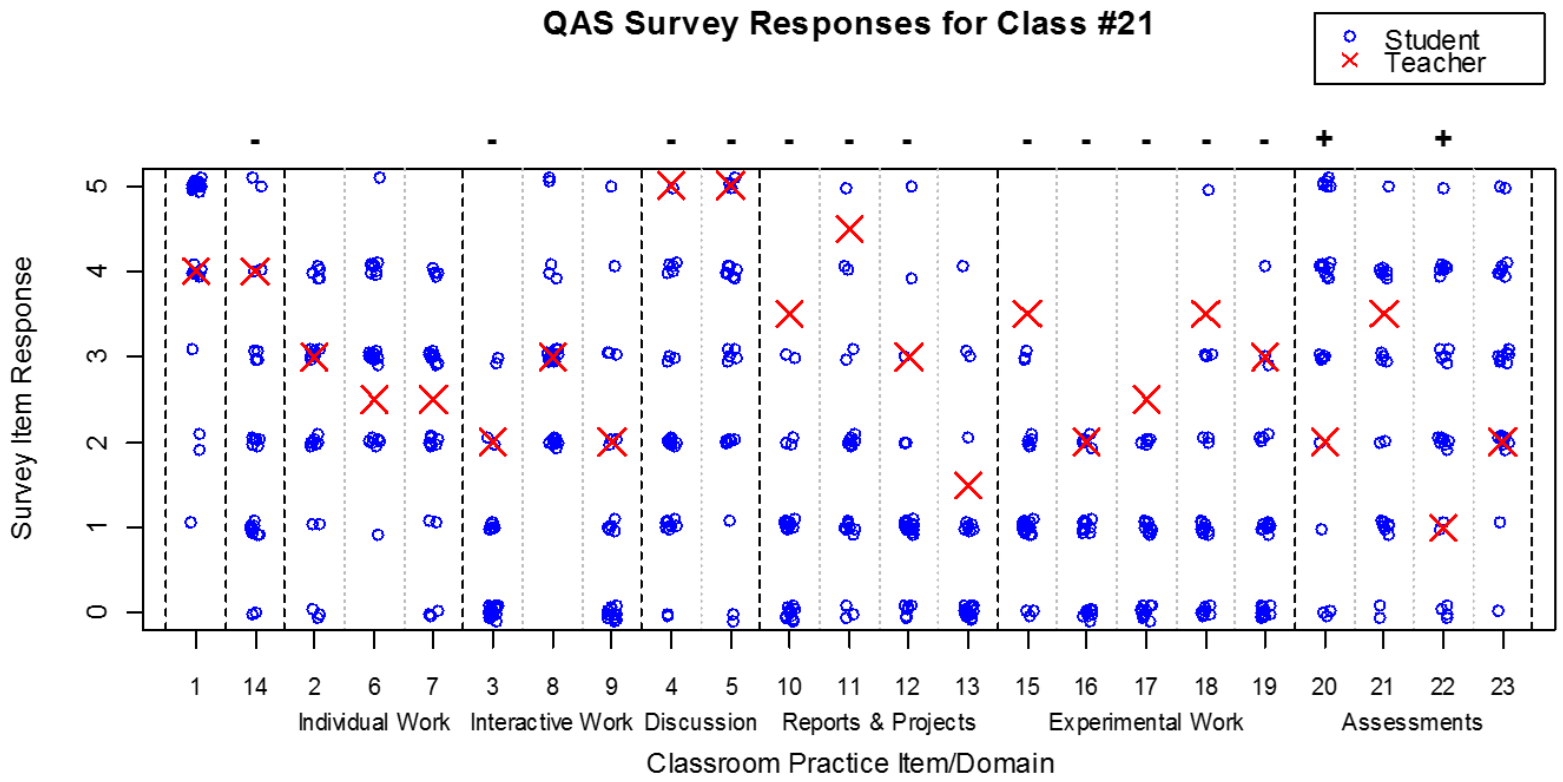
Plots are useful as a supplementary tool for exploring teacher–classroom discrepancy, as they allow a researcher to peer inside a given classroom to examine how teacher and student responses align for particular items and across instructional domains. In each "classroom discrepancy plot" presented in Figures 6.5–6.8, blue circles represent student responses in the given class and red X's represent the teacher's average response with each vertical column corresponding to a QAS item grouped by classroom practice dimension. The y-axis is a scale of the individual student and teacher item responses (ranging from 0 "never" to 5 "every day"). Printed across the top of these plots are the symbols "–" and "+" indicating whether the corresponding unstandardized difference in means is negatively or positively statistically significant.

Figures 6.5 and 6.6 plot the student and teacher item ratings in the two most discrepant classes in terms of composite unstandardized absolute differences in means: Class #21 and Class #26. Even though these classes had similar composite unstandardized absolute differences in means, their patterns of discrepancy by instructional domains and items were markedly different.

Perceptual discord in Class #21 was especially prevalent for items categorized in three instructional practice domains: Discussion, Reports and Projects, and Experimental Work. In the Discussion domain, Teacher #21's average rating for both Item #4 *discuss science topics with other students in small groups* and Item #5 *discuss science topics with other students as a class* was a 5, yet students' most common rating was a 2. Likewise, Teacher #21's average ratings on items comprising the Reports and Projects and Experimental Work instructional dimensions ranged from 1.5 to 4.5, yet the most common student ratings for all these items was either a 0 or 1. The plot in Figure 6.5 also reveals that all 26 students in Class #21 perceived they performed the Item #17 *conduct experiments or investigations* only *once or twice a month* or less frequently. As a result, even though Teacher #21's average rating of this item seemed reasonable, it actually fell outside the bounds of the clustered student responses.

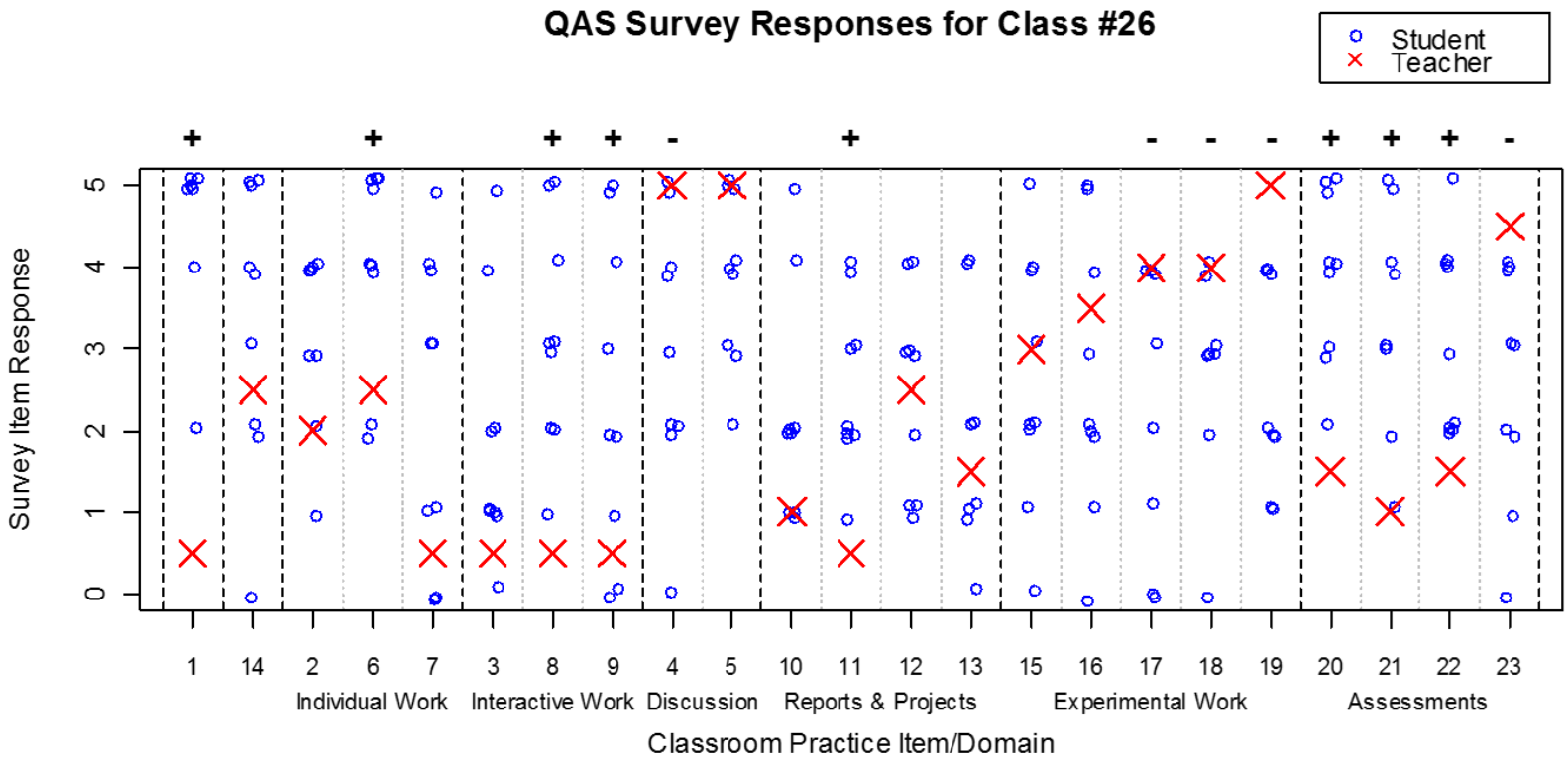
The plot of Class #26 in Figure 6.6 reveals a noticeably different pattern of perceptual discord, as discrepancy was most prevalent for items in the Interactive Work and Assessment domains. For instance, Teacher #26 rated Item #22 *take tests or quizzes with questions where you choose the answer* as occurring less frequently than all of his or her students, yet rated Item #23 *take tests or quizzes with questions where you write out the answer* as occurring *more* frequently than all of his or her students. Like Class #21, Class #26 was also negatively discrepant on Item #17 *conduct experiments or investigations*, yet student ratings of this instructional practice were much more variable. Also of note was the discrepancy pattern for Class #26 on Item #1 *listen to lectures or instruction directed by the teacher*. Whereas the majority of students felt Teacher #26 lectured *every day*, Teacher #26 believed he or she performed this practice *less than once a month* to *never*. Finally, there is Item #11 *work on science projects in pairs or groups*, for which both Class #21 and Class #26 were discrepant. Had the unstandardized absolute differences in means across Item #11 been profiled (instead of Items #1 and #14), estimates of discrepancy for Class #21 and #26 would have appeared similar. However, the plots revealed that the discrepancy in Class #21 was directionally negative while the discrepancy in Class #26 was directionally positive, a consequential finding that the previous methods could not show.

Figure 6.5: *Classroom Discrepancy Plot Comparing Student and Teacher Ratings Within Classroom #21 Across QAS Items*



Note. Scale is 0 = *Never*, 1 = *Less than once a month*, 2 = *Once or twice a month*, 3 = *Once a week or more*, 4 = *Multiple times per week*, 5 = *Every day*. Blue circles represent student responses and red X's represent teacher average responses. Each vertical column corresponds to the class' responses on the item. Statistical significance of unstandardized difference in means for each item printed across the top with the symbol "–" indicating negative significance and the symbol "+" indicating positive significance. Classroom Practice Item/Domains according to confirmatory factor analysis of student items shown in Appendix Table A.6.

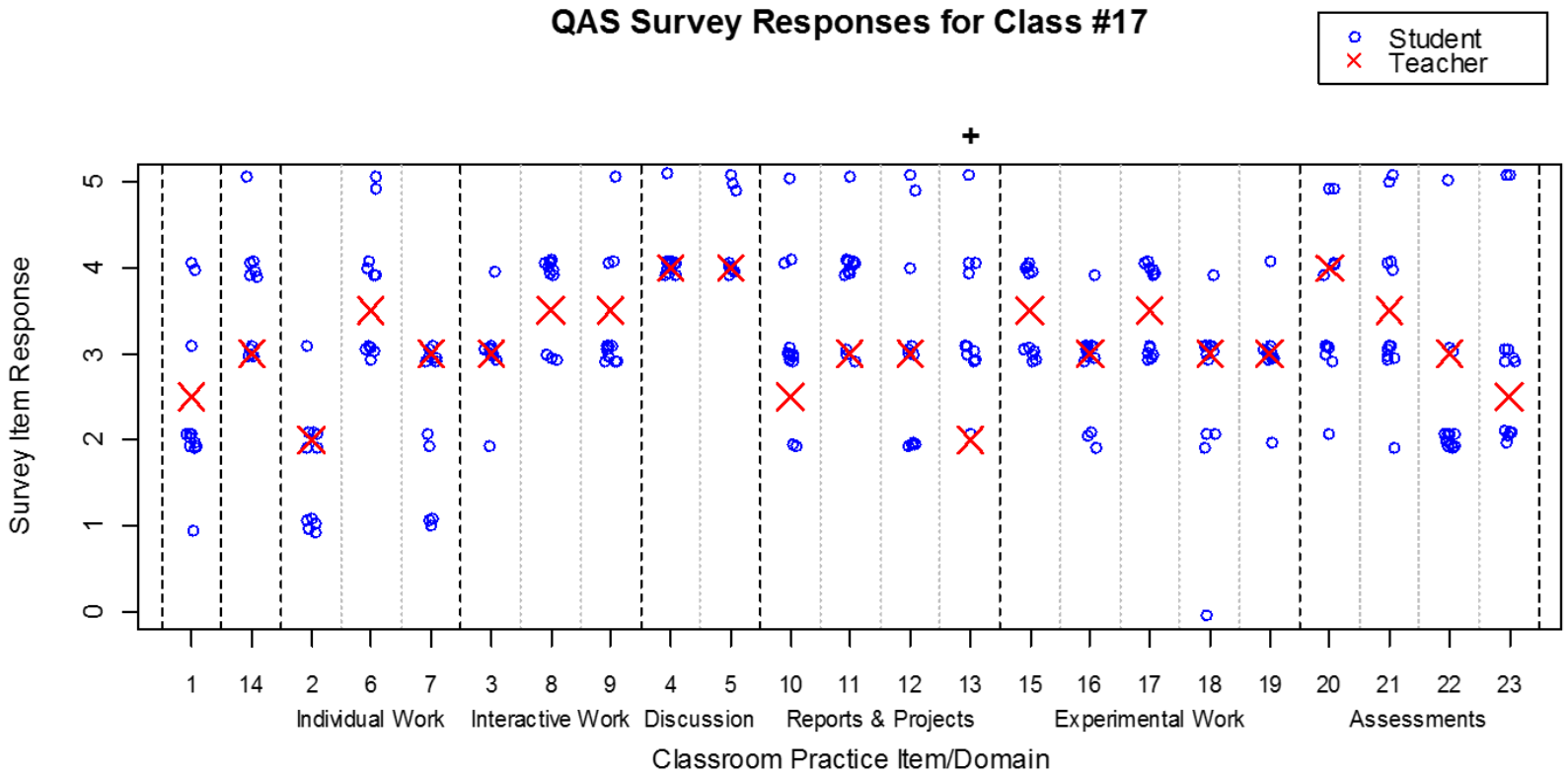
Figure 6.6: *Classroom Discrepancy Plot Comparing Student and Teacher Ratings Within Classroom #26 Across QAS Items*



Note. Scale is 0 = *Never*, 1 = *Less than once a month*, 2 = *Once or twice a month*, 3 = *Once a week or more*, 4 = *Multiple times per week*, 5 = *Every day*. Blue circles represent student responses and red X's represent teacher average responses. Each vertical column corresponds to the class' responses on the item. Statistical significance of unstandardized difference in means for each item printed across the top with the symbol "–" indicating negative significance and the symbol "+" indicating positive significance. Classroom Practice Item/Domains according to confirmatory factor analysis of student items shown in Appendix Table A.6.

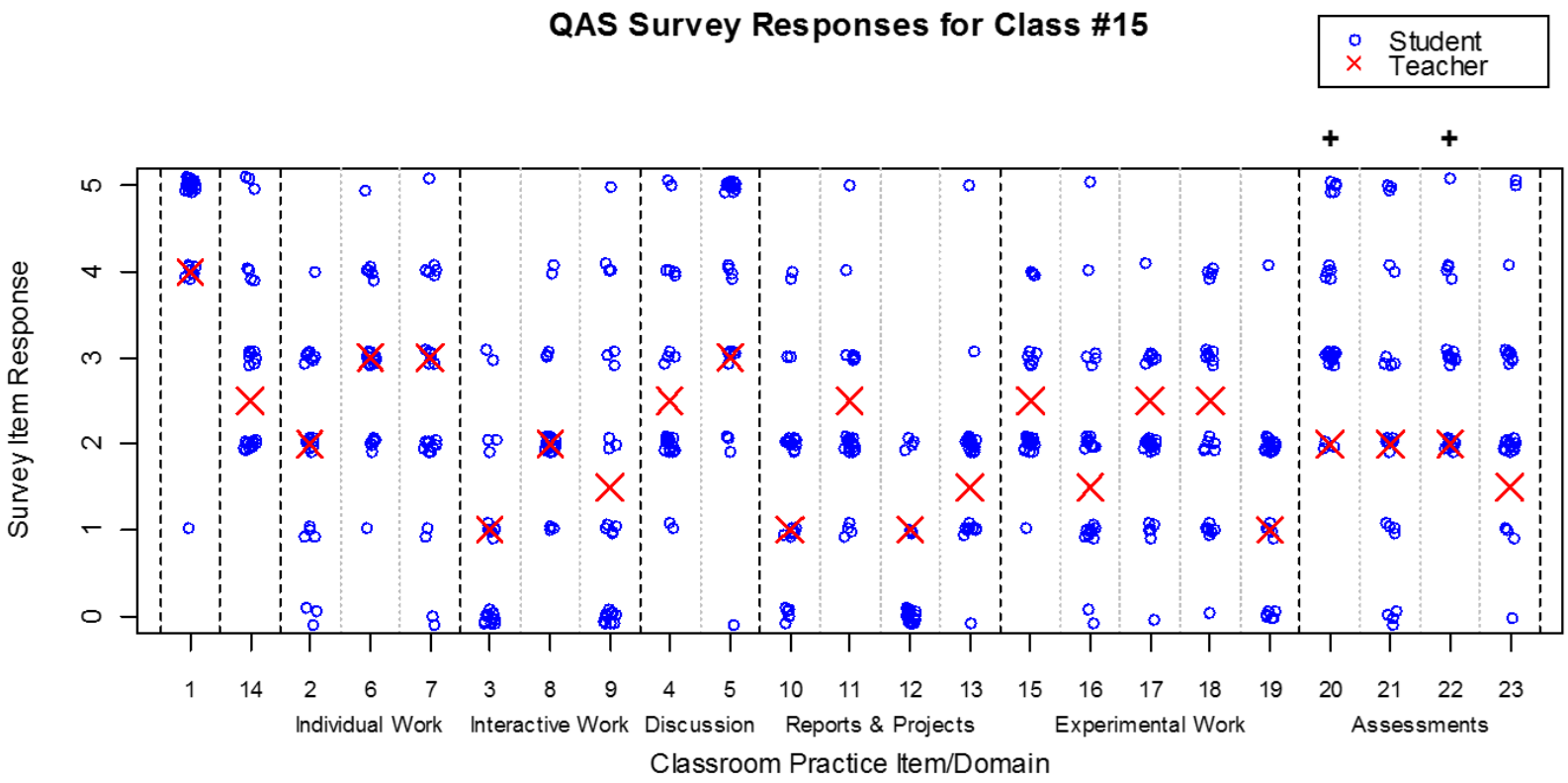
Figures 6.7 and 6.8 plot the student and teacher item ratings in the two least discrepant classes: Class #17 and Class #15. Like the most discrepant classes, Class #17 and Class #15 had similar composite unstandardized absolute differences in means yet noticeably different patterns of discrepancy across instructional domains and items. In particular, the student responses were strikingly less variable in Class #17, given that the range of responses on 14 of the 23 items spanned across only 3 scale ratings. This was especially evident for items in the Experimental Work domain. With the exception of a single student response, every student rating on Items #15–19 was within one scale unit of the Teacher #17's average rating. Conversely, while Teacher #15's average rating also fell within his or her student ratings, the spread in student ratings was much greater. Indeed, Class #15's student responses on Item #16 *plan/design experiments or investigations* ranged from 0 to 5. Thus, even though the unstandardized absolute differences in means suggested discrepancy for Classes #17 and #15 were quite similar in the Experimental Work domain (see Figure 6.3) the plots reveal clear differences in rating patterns. Matching forest plots, showing 95% confidence intervals of unstandardized differences in means across items for the most discrepant teacher–classroom pairs (Class #21 and Class #26) and least discrepant teacher–classroom pairs (Class #17 and Class #15), can be found in Appendix Figures A.5–A.8.

Figure 6.7: *Classroom Discrepancy Plot Comparing Student and Teacher Ratings Within Classroom #17 Across QAS Items*



Note. Scale is 0 = *Never*, 1 = *Less than once a month*, 2 = *Once or twice a month*, 3 = *Once a week or more*, 4 = *Multiple times per week*, 5 = *Every day*. Blue circles represent student responses and red X's represent teacher average responses. Each vertical column corresponds to the class' responses on the item. Statistical significance of unstandardized difference in means for each item printed across the top with the symbol "–" indicating negative significance and the symbol "+" indicating positive significance. Classroom Practice Item/Domains according to confirmatory factor analysis of student items shown in Appendix Table A.6.

Figure 6.8: *Classroom Discrepancy Plot Comparing Student and Teacher Ratings Within Classroom #15 Across QAS Items*



Note. Scale is 0 = *Never*, 1 = *Less than once a month*, 2 = *Once or twice a month*, 3 = *Once a week or more*, 4 = *Multiple times per week*, 5 = *Every day*. Blue circles represent student responses and red X's represent teacher average responses. Each vertical column corresponds to the class' responses on the item. Statistical significance of unstandardized difference in means for each item printed across the top with the symbol "–" indicating negative significance and the symbol "+" indicating positive significance. Classroom Practice Item/Domains according to confirmatory factor analysis of student items shown in Appendix Table A.6.

6.4 Chapter Summary

Discrepancy in the most discordant teacher–classroom pairs did not substantively depend on particular instructional practices—though it is difficult to definitively answer this question with only 23 instructional practices and relatively few instances of discrepancy. Students and teachers in the most discrepant classes according to composite absolute difference in means were also consistently ranked as highly discrepant across the full range of QAS items and in multiple instructional domains. This suggests the magnitude of perceptual difference was influenced as much by the particular teacher–classroom pairing as by the instructional practice.

This chapter also used discrepancy plots to examine student and teacher item response patterns in the two most and least discrepant classes. Discrepancy plots are transparent in that they reveal several ways in which the same average discrepancy plays out—both in magnitude and in direction. Because of this, these plots can reveal a deeper level of response pattern detail than can be ascertained by solely examining unstandardized difference in means or composite absolute unstandardized difference in means. Thus, discrepancy plots have formative assessment value, by telling an easily understood story of the clear differences in perceptions that may exist between a teacher and his or her class of students. The next chapter pivots in focus by exploring whether perceptual discrepancies between students and teachers are more common on some instructional practice items than on others.

CHAPTER 7

7. Is Discrepancy Greater for Some Instructional Practices than for Others?

This chapter examines how the degree of perceptual difference between students and their teachers varies by QAS instructional practice item. Four methods are used to measure discrepancy across practices: (1) proportional scoring, (2) unstandardized differences in means, (3) item composite unstandardized differences in means, and (4) discrepancy plots. Additionally, particular focus is given to whether the approach utilized for measuring perceptual difference influences item discrepancy rankings. The main research questions of this chapter are as follows:

3a. Is discrepancy greater for some instructional practices than for others?
3b. Does the measure of discrepancy influence which practices produce more discrepancy?

Like in Chapter 5, the example discrepancy analyses provided throughout use the student and teacher ratings on Item #16 *plan/design experiments or investigations* in Class #16.

7.1 Comparing Instructional Practices Using Proportional Scoring Methods

A basic technique for determining discrepancy within classrooms is computing the percent of student responses on a given item in a classroom that are rated above, rated below, and are equivalent to their teacher. For example, there were 17 students in Class #16 who answered Item #16 *plan/design experiments or investigations*. Eight students (47%) rated the practice as occurring more frequently than their teacher rated, five students (29%) rated the practice as occurring less frequently than their teacher rated, and four students (24%) rated the practice as occurring with a frequency equivalent to what their teacher reported. Appendix Tables A.7–A.9 display these proportional values, comparing students' item ratings with those of their teachers on the QAS post-survey.

Using this method, the degree of perceptual alignment between students and teachers was found to vary considerably across QAS classroom practice items. Overall, 27% of student responses were in exact agreement with teacher responses from the first survey administration and 28% of student responses were in exact agreement with teacher responses from the second administration. In both survey administrations, the two instructional practices rated highest by students compared to their teacher's perceptions were Item #1 *listen to lectures or instruction directed by the teacher* and Item #20 *review materials to prepare for a test or quiz*. Likewise, in both survey administrations Item #4 *discuss science topics with other students in small groups* and Item #17 *conduct experiments or investigations* were the two practices rated lowest by students compared to their teacher's perceptions. Broadly speaking, agreement patterns between student and teacher ratings by instructional practice item were similar in each survey administration.

Table 7.1 compares student ratings with their teacher's *average* ratings across the two survey administrations. Each row of the table displays the percent and resulting means of the QAS classroom practice item for which student ratings are greater than (by at least one scale unit), less than (by at least one scale unit), and in agreement with their teacher ratings (equivalent or within 0.5 scale units). The practice Item #1 *listen to lectures or instruction directed by the teacher* was most often rated higher by students than by their teachers. Sixty percent of students rated this practice at least one scale unit higher than their teacher's average ratings on the item across the surveys. Given that ratings are influenced by recall, this practice may have been rated comparatively higher by students because to some it *felt* as if it occurred every day when in actuality it only occurred a couple times a week.

By contrast, other instructional practices items involving peer interaction and exploratory learning may have felt to some students as if they were not performed frequently enough. This may help explain why about half of teachers' average ratings on Item #4 *discuss science topics with other students in small groups* (49%) and Item #17 *conduct experiments or investigations*

(48%) were at least one scale unit higher than their students' ratings on these practices. Lastly, Item #13 *present oral science reports* had the highest teacher–student agreement within half a scale unit (occurring 53% of the time) while Item #20 *review materials to prepare for a test or quiz* had the lowest teacher–student agreement (occurring only 29% of the time). Given these wide disparities in student and teacher ratings across items, the results in Table 7.1 convey that discrepancy appears to be greater for some instructional practices than for others.

As an investigative tool, calculating proportions can be informative, yet this method is somewhat limited in its utility. First of all, proportions do not account for the magnitude of a discrepancy. For example, of the eight students in Class #16 who rated Item #16 as occurring more frequently than Teacher #16, six students were discrepant by one scale unit and two students were discrepant by two scale units. Proportional calculations also do not take into consideration the number of students per classroom. This shortcoming is important because a perceptual difference in a classroom with more students should logically be given greater credence than a perceptual difference in a classroom with fewer students. Finally, using proportions to examine perceptual differences in individual classes, item-by-item, is not especially efficient, as this method requires considering each of these three percentages in tandem in order to gain an understanding of the discrepancy. Because of these limitations, the subsequent sections use unstandardized differences in means to measure perceptual difference between student and teachers.

Table 7.1: *Percentage and Corresponding Means of Student Ratings Greater than, Less than, and in Agreement with Teacher Ratings, by QAS Instructional Practice Item*

Classroom Practice Items	Percentage			Mean Rating					
				Student			Teacher		
	S < T	**S = T**	**S > T**	**S < T**	**S = T**	**S > T**	**S < T**	**S = T**	**S > T**
#1 Listen to lectures or instruction directed by the teacher	8%[a]	32%	60%	2.04[b]	3.93	4.78	3.78[c]	3.87	3.30
#2 Read a science textbook	26%	42%	32%	1.40	2.61	3.83	3.00	2.66	2.40
#3 Read science articles in magazines or newspapers	35%	44%	21%	0.41	1.39	2.90	1.87	1.41	1.08
#4 Discuss science topics with other students in small groups	49%	37%	14%	1.68	3.43	4.07	3.91	3.53	2.42
#5 Discuss science topics with other students as a class	41%	38%	21%	2.13	3.93	4.43	4.18	3.94	2.77
#6 Work on worksheets	11%	42%	47%	1.96	3.28	4.35	3.31	3.20	2.82
#7 Work on homework tasks during class time	28%	32%	40%	1.50	2.53	3.79	3.39	2.53	1.73
#8 Watch science videos, movies, TV shows, etc.	17%	43%	39%	1.46	2.57	3.67	2.86	2.50	2.05
#9 Use science-related software or internet resources	30%	39%	31%	0.91	2.10	3.73	2.73	2.14	1.79
#10 Work on science projects individually	31%	39%	30%	0.73	1.83	3.06	2.36	1.85	1.28
#11 Work on science projects in pairs or groups	34%	35%	31%	1.49	2.32	3.60	3.40	2.33	1.75
#12 Work on written science reports	39%	35%	26%	0.73	1.82	3.01	2.38	1.84	1.41
#13 Present oral science reports	20%	53%	26%	0.25	0.90	2.75	1.66	0.94	1.05
#14 Watch teacher demonstrate experiment or investigation	24%	38%	38%	1.55	2.75	4.01	3.31	2.72	2.24
#15 Use lab instruments or materials	40%	45%	14%	1.78	3.17	4.25	3.38	3.21	2.97
#16 Plan/design experiments or investigations	33%	36%	31%	0.89	2.14	3.37	2.42	2.10	1.61
#17 Conduct experiments or investigations	48%	39%	12%	1.57	3.02	4.14	3.25	3.09	2.83
#18 Analyze data / relationships between variables	44%	37%	19%	1.56	2.99	3.96	3.25	2.99	2.56
#19 Write reports about labs or experiments	37%	36%	27%	0.92	2.51	3.51	2.76	2.53	1.88
#20 Review materials to prepare for a test or quiz	17%	29%	54%	1.72	3.04	4.01	3.75	3.00	2.14
#21 Develop or practice test-taking skills	30%	31%	39%	1.20	2.42	3.78	3.10	2.41	1.96
#22 Take tests or quizzes with questions where you choose the answer	10%	39%	51%	1.30	2.18	3.61	2.67	2.13	1.84
#23 Take tests or quizzes with questions where you write out the answer	20%	41%	39%	0.94	2.04	3.48	2.39	2.04	1.70

Note. Scale is 0 = *Never*, 1 = *Less than once a month*, 2 = *Once or twice a month*, 3 = *Once a week or more*, 4 = *Multiple times per week*, 5 = *Every day*. S = Student, T = Teacher. S < T includes matching responses where the average student rating is less than the average teacher rating by at least one scale unit; S = T includes matching responses where the average student rating is equivalent to the average teacher rating or within 0.5 scale units; S > T includes matching responses where the average student rating is greater than the average teacher ratings by at least one scale unit. Mean student and teacher ratings from condition S = T are often not equivalent as matching average student and teacher ratings can be within 0.5 scale units.

[a] For example, 8% is the percentage of the 645 students in the QAS sample who rated Item #1 at least one scale unit lower than the average rating of their teacher across the two survey occasions.

[b] For example, 2.04 is the average student rating for those 8% of students who rated Item #1 at least one scale unit lower than the average rating of their teacher.

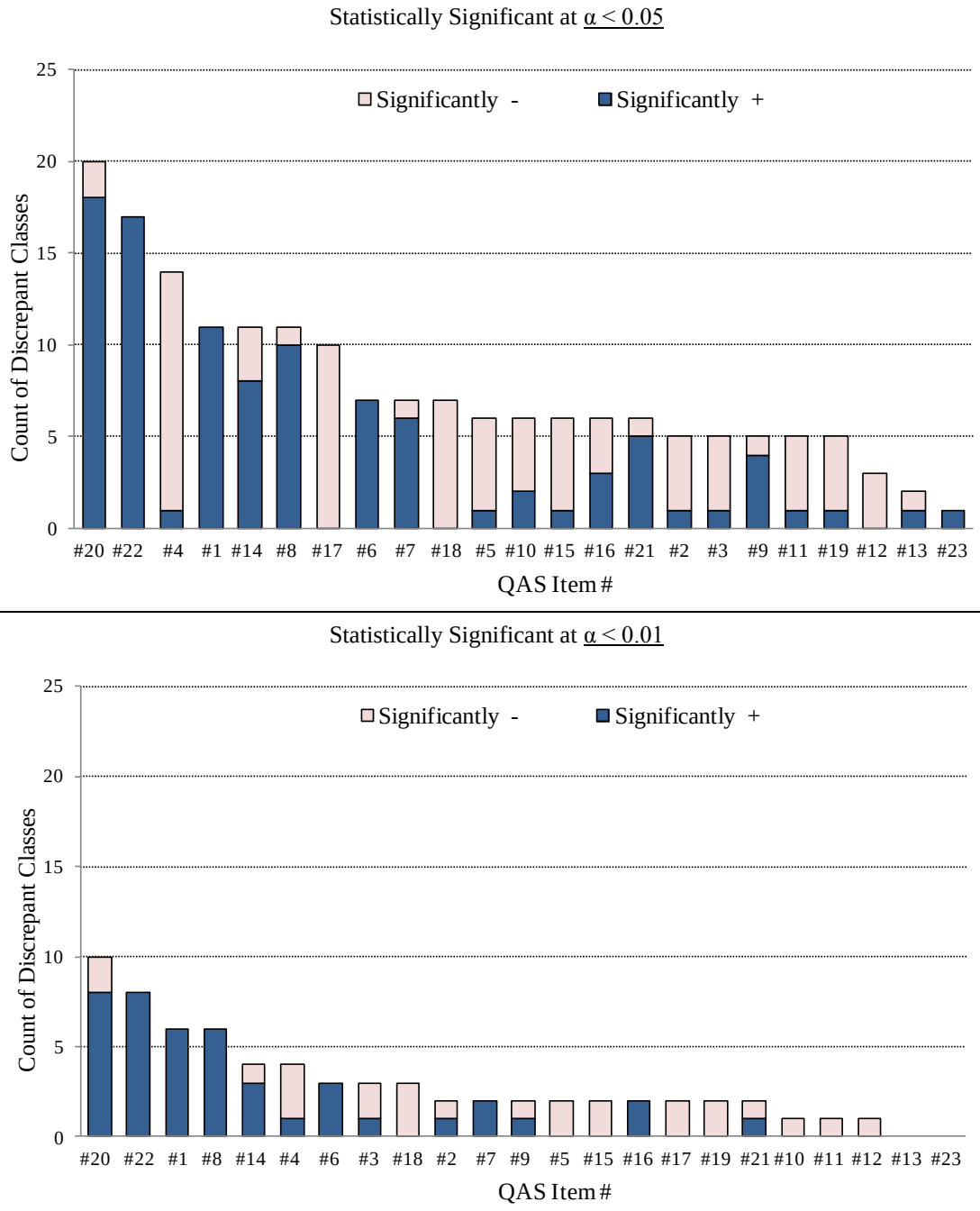
[c] For example, 3.78 is the average teacher rating corresponding to those 8% of students who rated Item #1 at least one scale unit lower than the average rating of their teacher.

7.2 Comparing Instructional Practices Using Unstandardized Differences in Means

Another method of exploring differences in student and teacher perceptions across instructional practices is by comparing the number of discrepant teacher–classroom pairs on a given item. As detailed in Section 5.2.1, an unstandardized difference in means describing the discrepancy between student and teacher responses on a single QAS item in a given classroom can be computed using Equation 5.1. In total, 897 item-level unstandardized differences in means were previously computed in Chapter 5 (one for each of the 23 items in each of the 39 classrooms). Additionally, the statistical significance of these mean differences—evaluated at $\alpha < 0.05$ and $\alpha < 0.01$ using standard error estimates from Equation 5.2—was tallied across classes to explore differences in discrepancy by teacher–classroom pairs. This section tallies statistically significant unstandardized differences in means across items to explore instructional practice differences in discrepancy.

In total, 20% of the total 897 unstandardized differences in means were found to be statistically significant, indicating a class perceptual difference between students and their teacher. Class counts of positive and negative statistical significance were clustered differently depending on the instructional practice. Two assessment instructional practices accounted for more than a third (35 out of 100) of the total mean differences found to be positively significant (item rated higher by the class of students). As shown in Figure 7.1, 18 classes were positively discrepant at $\alpha < 0.05$ on Item #20 *review materials to prepare for a test or quiz* and 17 classes were positively discrepant on Item #22 *take tests or quizzes with questions where you choose the answer*. Eleven and ten classes, respectively, were also positively discrepant at $\alpha < 0.05$ on Item #1 *listen to lectures or instruction directed by the teacher* and Item #8 *watch science videos, movies, TV shows, etc.* In addition, these four instructional practices accounted for 28 of the 43 (65%) instances of classes positively discrepant at $\alpha < 0.01$. For 13 instructional practices there were two or fewer classes positively discrepant at $\alpha < 0.05$ while for 10 instructional practices not a single class was found to be discrepant when evaluated at $\alpha < 0.01$.

Figure 7.1: *Item Counts of Statistically Significant Unstandardized Differences in Means*



Note. In total, 176 unstandardized differences in means were found to be statistically significant at α <0.05 (76 significantly negative and 100 significantly positive) and 68 unstandardized differences in means were found to be statistically significant at α <0.01 (25 significantly negative and 43 significantly positive).

Class counts of the 85 negatively significant (evaluated at $\alpha < 0.05$) unstandardized differences in means (item rated higher by the teacher) were similarly dispersed across items. Classes were disproportionately more likely to be negatively discrepant on two instructional practices—Item #4 *discuss science topics with other students in small groups* and Item #17 *conduct experiments or investigations*. In total, these two practices accounted for three-in-ten (30%) of the total mean differences found to be negatively significant at $\alpha < 0.05$. Conversely, there were nine instructional practices for which two or fewer classes were negatively discrepant. Class counts of negatively significant unstandardized differences in means evaluated at $\alpha < 0.01$ were further dispersed across items, as no more than three classes were negatively discrepant on any one instructional practice. Finally, only two classes were discrepant on Item #13 *present oral science reports* and only one class was discrepant on Item #23 *take tests or quizzes with questions where you write out the answer*. This suggests some instructional practices were more likely to elicit perceptual differences between students and their teachers than other practices.

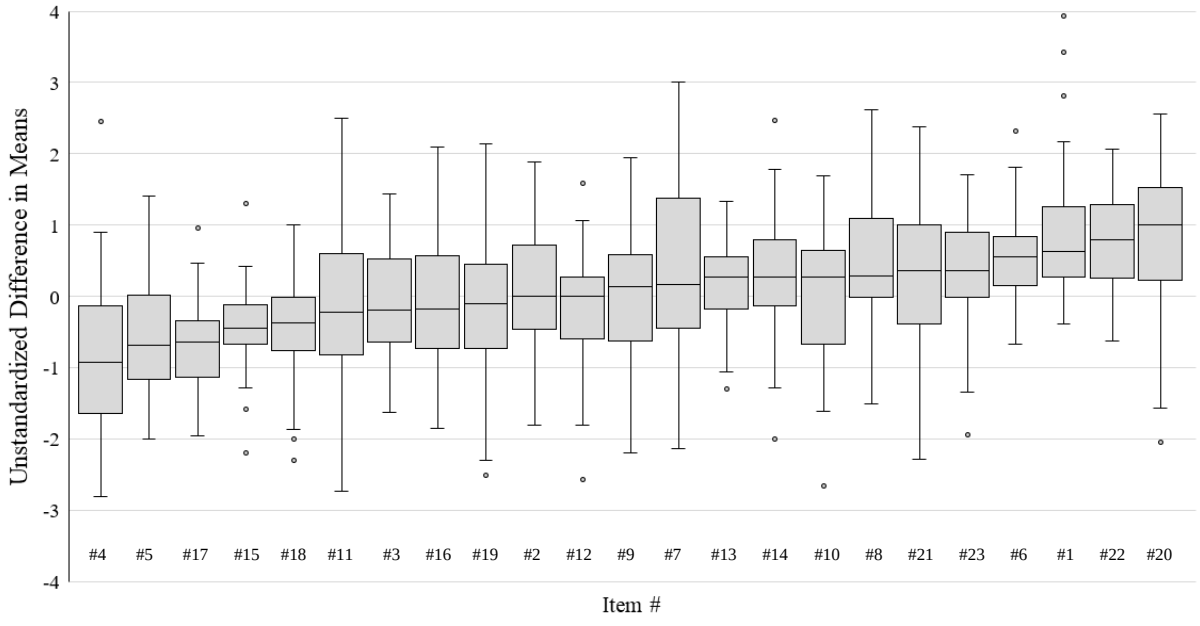
7.3 Comparing Instructional Practices Using Item Composite Unstandardized Differences in Means

Like class composite unstandardized absolute differences in means (Section 5.3), item composite unstandardized differences in means are a measure of perceptual difference within classrooms found by combining unstandardized difference in mean estimates between students and their teacher in a classroom. However, instead of aggregating discrepancy across the 23 QAS items for single classroom—as was done in Chapter 5—difference in means are aggregated across the 39 classrooms on a single item. Importantly, because the unstandardized differences in means being combined are all of the same instructional practice, the absolute value was not computed for each difference in means before aggregation. Thus, item composite unstandardized differences in means are a measure of both directional and magnitudinal perceptual difference for a given instructional practice. Overall, composite discrepancy results predominantly aligned with counts of instances of statistically significant unstandardized differences in means across items as clear differences in discrepancy were observed between instructional practices.

7.3.1 Exploratory Analysis of Discrepancy Across Items Using Box Plots

Before computing item composite unstandardized differences in means, it can be helpful to construct box plots of the 39 class unstandardized differences in means for each item. As shown in Figure 7.2, this exploratory procedure revealed noticeable differences in the direction and spread of class discrepancy depending on the instructional practice. Item #4 *discuss science topics with other students in small groups* and Item #5 *discuss science topics with other students as a class* had the lowest (most negative) median class unstandardized difference in means. Conversely, two assessment practices, Item #20 *review materials to prepare for a test or quiz* and Item #22 *take tests or quizzes with questions where you choose the answer* had the highest median class unstandardized difference in means. There were also noticeable differences in the variability of class discrepancy across instructional practices. The interquartile range was widest for Item #7 *work on homework tasks during class time,* which spanned about two scale points, and was most compact for Item #15 *use lab instruments or materials,* which spanned only about half a scale point.

Figure 7.2: *Item Box Plots of Unstandardized Differences in Means*



Note. Items sorted by median unstandardized difference in means.

7.3.2 Computing Item Composite Unstandardized Differences in Means

The mean difference for Item #16 is used as a computational example of item composite unstandardized difference in means. Specifically, this estimate was calculated by (1) using Equation 5.1 to find the unstandardized differences in means for each of the 39 QAS classes between the students and teacher on Item #16 and (2) computing a weighted average difference in means estimate across classes. To compute this weighted average, each class difference in means estimate for Item #16 was inversely weighted by its error and parameter variances. Specifically, the weight of each class for Item #16 was found by computing the inverse or reciprocal of the sum of the class' unique error variance (*V*) estimate and the parameter variance estimate across all the classes for Item #16 (*T*) from specifying an unconditional ANOVA model using the statistical program HLM 6 (Raudenbush, Bryk, & Congdon, 2004). This resulting weighted average mean difference for a given item is referred to as the item composite unstandardized difference in means.

Like in Section 5.3.2, the computational steps outlined in the previous paragraph are detailed for Item #16 in Class #16. Note that the unstandardized difference in means for Item #16 in Class #16 ($\overline{X}_{\text{Dif\_Item16\_Class16}} = 0.56$) and the standard error of this estimate ($SE_{\text{Dif\_Item16\_Class16}} = 0.58$) were originally found in Sections 5.2.1 and 5.2.2 using Equations 5.1 and 5.2, respectively. The error variance (*V*) was also found previously in Section 5.3.2 for Item #16 in Class #16 by squaring the standard error ($V_{\text{Dif\_Item16\_Class16}} = SE^2_{\text{Dif\_Item16\_Class16}} = 0.58^2 = 0.34$). This process was repeated to compute unstandardized difference in means, associated standard errors of these estimates, and error variances for Item #16 on each of the other 38 QAS classrooms. Then, the parameter variance for Item #16 (*T*) was found by specifying an unconditional model in HLM 6 predicting the unstandardized difference in means across classes for Item #16 using known item variances, with each item's squared standard error estimate for Class #16 originally computed from Equation 5.2. The amount of variance in the true differences in means across the 39 classes for Item #16 was found to be $T_{\text{Dif\_Item16}} = 0.43$. The item unstandardized difference in means for Item #16 ($Comp_{\text{Dif\_Item16}}$) can be obtained

from HLM model itself (γ_0) or can be computed longhand using class error variances (V's) and parameter variance (T) as follows:

$$Comp_{\text{Dif\_Item16}} = \frac{\sum_{c=1}^{39}[\bar{X}_{\text{Dif\_Item16}}(c) * (1/(V_{\text{Dif\_Item16}}(c) + T_{\text{Dif\_Item16}}))]}{\sum_{c=1}^{39}[1/(V_{\text{Dif\_Item16}}(c) + T_{\text{Dif\_Item16}})]}$$

$$= \frac{\sum_{c=1}^{39}[\bar{X}_{\text{Dif\_Item16}}(c) * (1/(V_{\text{Dif\_Item16}}(c) + 0.43))]}{\sum_{c=1}^{39}[1/(V_{\text{Dif\_Item16}}(c) + 0.43)]} = \frac{-1.74}{52.53} = -0.03$$

Like in Figure 5.4, 95% confidence intervals were constructed around each item composite unstandardized difference in means to quantify the margin of error. To compute these confidence intervals, the standard error of each of the 23 item composite unstandardized differences in means was found by taking the square root of the inverse of the sum of the inverse total variance estimates (error variance + parameter variance) across the 39 classes for each item. Note that the sum of the inverse of the total variance estimates, 52.53, is the denominator in the equation (above) used to find $Comp_{\text{Dif\_Item16}}$. The standard error of the composite unstandardized difference in means for Item #16 ($SE_{\text{Dif\_Item16}}$) is computed below:

$$SE_{\text{Dif\_Item16}} = \sqrt{\frac{1}{\sum_{c=1}^{39}[1/(V_{\text{Dif\_Item16}}(c) + T_{\text{Dif\_Item16}})]}} = \sqrt{\frac{1}{\sum_{c=1}^{39}[1/(V_{\text{Dif\_Item16}}(c) + 0.43)]}} = \sqrt{\frac{1}{52.53}} = 0.14$$

The standard error was then used to construct the lower and upper bounds of a 95% confidence interval around the item unstandardized difference in means parameter by setting $\alpha < 0.05$ and basing the critical value (t_{critical}) on a t distribution with the degrees of freedom (df) equal to $n_{\text{Total Classes}}$ - 1. The lower and upper bounds of the 95% confidence interval around the composite unstandardized difference in means for Item #16 is computed below:

Lower confidence interval limit = $\bar{X}_{\text{Dif\_Item16}} - (t_{\text{critical }(df\_38} * SE_{\text{Dif\_Item16}})$

$$= -0.03 - (2.02 * 0.14) = -0.31$$

Upper confidence interval limit = $\bar{X}_{\text{Dif\_Item16}} + (t_{\text{critical }(df\_38} * SE_{\text{Dif\_Item16}})$

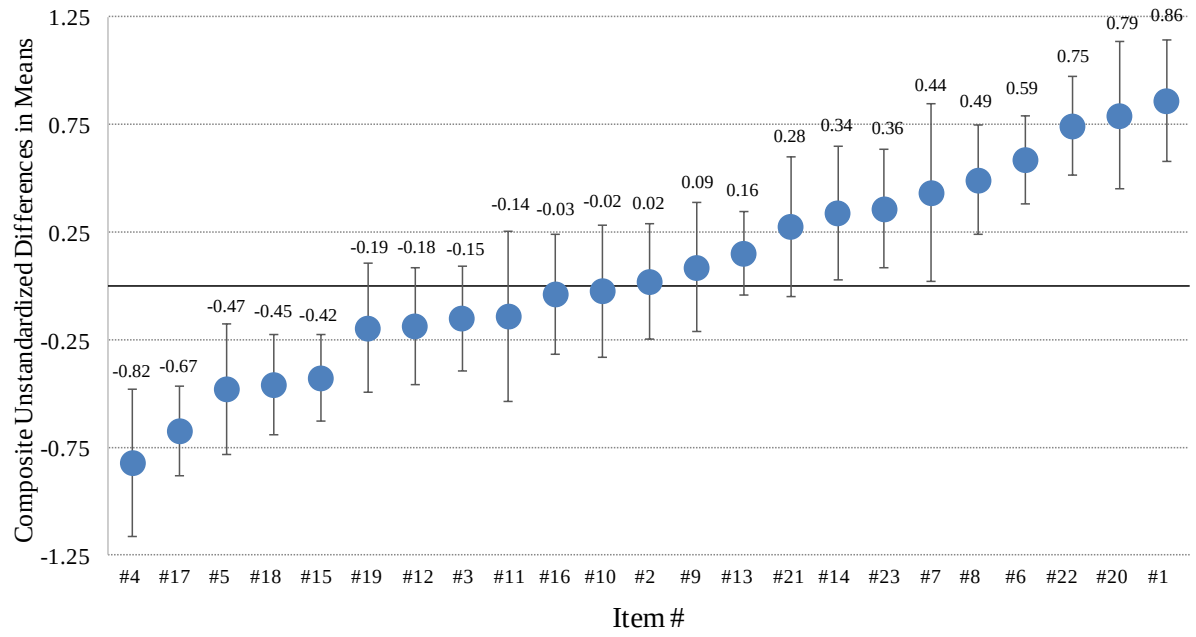
$$= -0.03 + (2.02 * 0.14) = 0.25$$

The item composite unstandardized difference in means for Item #16 was $\overline{X}_{Dif\_Item16}$ = -0.03, with an associated 95% confidence interval ranging from [-0.31, 0.25]. Thus, one can be 95% confident that the true composite unstandardized difference in means for Item #16 resides within this interval. Note that the confidence interval of the mean difference measurement overlaps with zero. Thus, there is no statistical evidence that students or their teachers perceived that they *plan/design experiments or investigations* more frequently during class.

7.3.3 Ranking Instructional Practices by Item Composite Unstandardized Differences in Means

Figure 7.3 provides the ranking of instructional practices using item composite unstandardized differences in means. As evident from the figure, three broad categories emerged: practices with statistically significant negative discrepancy (signifying an item rated higher by teachers); practices with statistically significant positive discrepancy (signifying an item rated higher by students); and practices with non-significant difference in means estimates, signifying a discrepancy not statistically different from zero like that found for Item #16.

Figure 7.3: *Item Composite Unstandardized Differences in Means*



Note. Items sorted from lowest to highest item composite unstandardized difference in means.

With confidence bands entirely below zero, five items were found to be negatively discrepant, meaning teachers tended to rate these practices as occurring more frequently than their students. Interestingly, all five of these items belonged to only two instructional practice domains (see Appendix Table A.6). In the Discussion domain was Item #4 *discuss science topics with other students in small groups* ($\overline{X}_{Dif\_Item4}$ = -0.82) and Item #5 *discuss science topics with other students as a class* ($\overline{X}_{Dif\_Item5}$ = -0.47). In the Experimental Work domain was Item #17 *conduct experiments or investigations* ($\overline{X}_{Dif\_Item17}$ = -0.67), Item #18 *analyze data / relationships between variables* ($\overline{X}_{Dif\_Item18}$ = -0.45), and Item #15 *use lab instruments or materials* ($\overline{X}_{Dif\_Item15}$ = -0.42). A common characteristic of these instructional practices is their active learning and participatory nature.

Conversely, with confidence bands entirely above zero, eight items were found to be positively discrepant, meaning each of these instructional practices was perceived by students as occurring more frequently than their teachers perceived that it occurred. Three of these items pertained to assessment instructional practices: Item #20 *review materials to prepare for a test or quiz* ($\overline{X}_{Dif\_Item20}$ = 0.79), Item #22 *take tests or quizzes with questions where you choose the answer* ($\overline{X}_{Dif\_Item22}$ = 0.75), and Item #23 *take tests or quizzes with questions where you write out the answer* ($\overline{X}_{Dif\_Item23}$ = 0.36). The other five positively discrepant items were passive learning instructional practices. For two items, this passive learning related to observing teacher pedagogy: Item #1 *listen to lectures or instruction directed by the teacher* ($\overline{X}_{Dif\_Item1}$ = 0.86) and Item #14 *watch teacher demonstrate experiment or investigation* ($\overline{X}_{Dif\_Item14}$ = 0.34). The other three items described passive learning in terms of independent work or watching videos: Item #6 *work on worksheets* ($\overline{X}_{Dif\_Item6}$ = 0.59), Item #7 *work on homework tasks during class time* ($\overline{X}_{Dif\_Item7}$ = 0.44), and Item #8 *watch science videos, movies, TV shows, etc.* ($\overline{X}_{Dif\_Item8}$ = 0.49).

Finally, the confidence bands of ten items overlapped with zero, indicating no difference in the frequency with which these instructional practices were perceived by students and their teachers. These practices were scattered among five of the six instructional domains. Of note, all

four instructional practices in the Reports and Projects domain were found to be non-significantly discrepant. This included Item #10 *work on science projects individually* ($\overline{X}_{Dif\_Item10}$ = -0.02), Item #11 *work on science projects in pairs or groups* ($\overline{X}_{Dif\_Item11}$ = -0.14), Item #12 *work on written science reports* ($\overline{X}_{Dif\_Item12}$ = -0.18), and Item #13 *present oral science reports* ($\overline{X}_{Dif\_Item13}$ = 0.16).

7.4 Examining Discrepancy for the Most Discrepant Instructional Practices

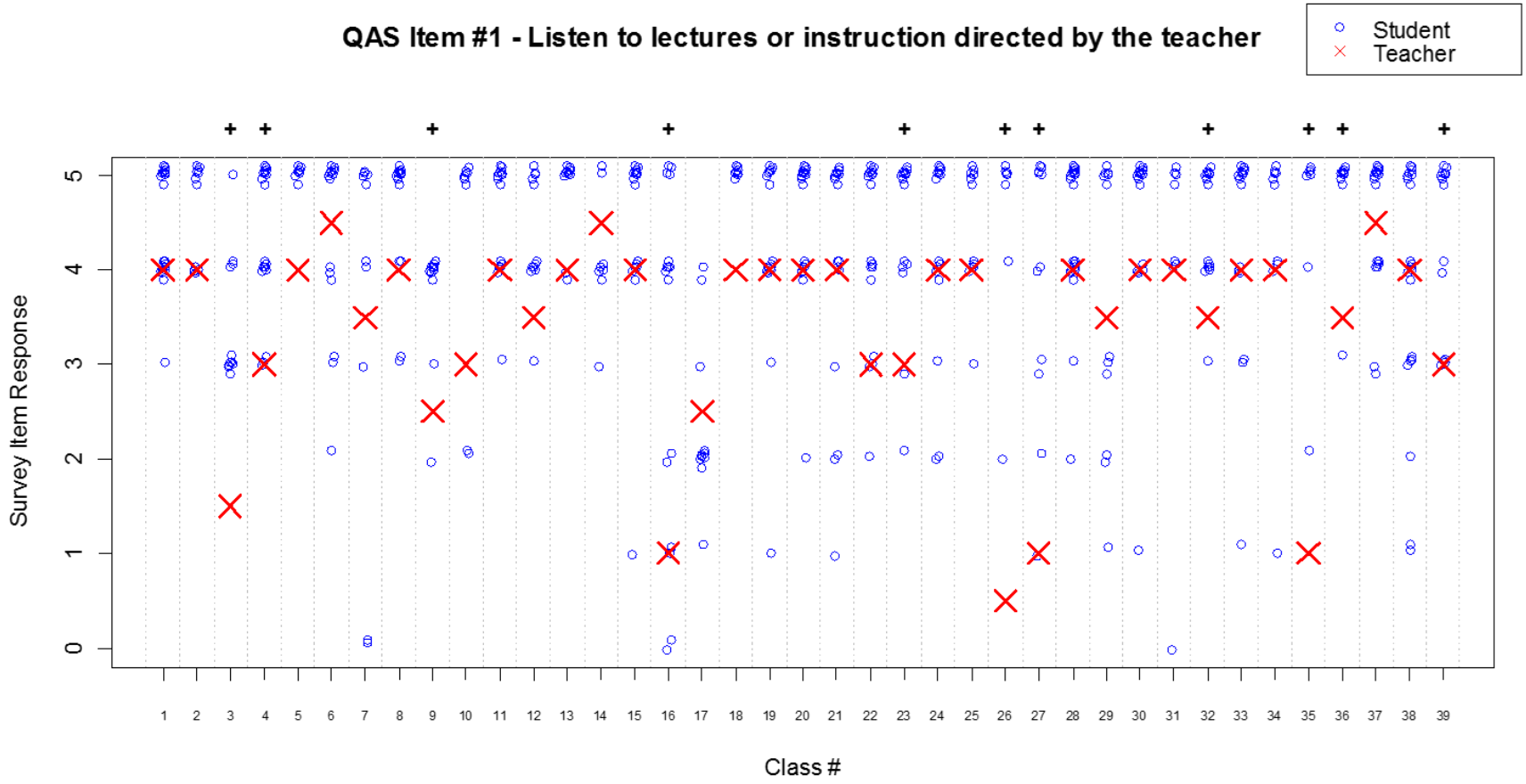
All three of the methods presented in this chapter—proportional scoring, unstandardized differences in means, and item composite unstandardized differences in means—can often mask wide differences in underlying student response variation within classrooms (see e.g., Appendix Figure A.9). Thus, plotting student and teacher ratings across classrooms in "item discrepancy plots" can be a useful tool for highlighting response pattern differences by instructional practice. Similar to class discrepancy plots presented in Chapter 5 (Figures 6.5–6.8), item discrepancy plots compare student and average teacher responses, yet on a single instructional practice across all classes. In these plots blue circles represent student responses and red X's represent teacher average responses with vertical columns now corresponding to each of the 39 QAS classes.

For example, the discrepancy plot of Item #1 *listen to lectures or instruction directed by the teacher* displays classroom differences on one of the most positively discrepant instructional practices (see Figure 7.4). The plot makes evident that discrepancy on this practice was driven by the very high percentage of students who rated this item a 5 compared to the average teacher ratings, which were always lower than 5. Interestingly, relative to the other instructional practices, teacher ratings were often quite high, yet the item was positively discrepant because so many students also rated this practice a 5.

The discrepancy plot of Item #4 *discuss science topics with other students in small groups* reveals a contrasting pattern of perceptual difference between students and their teacher within QAS classrooms (Figure 7.5). While this item also had a high degree of discrepancy, in this case students gave lower ratings than their teachers on average. This is plainly visible from

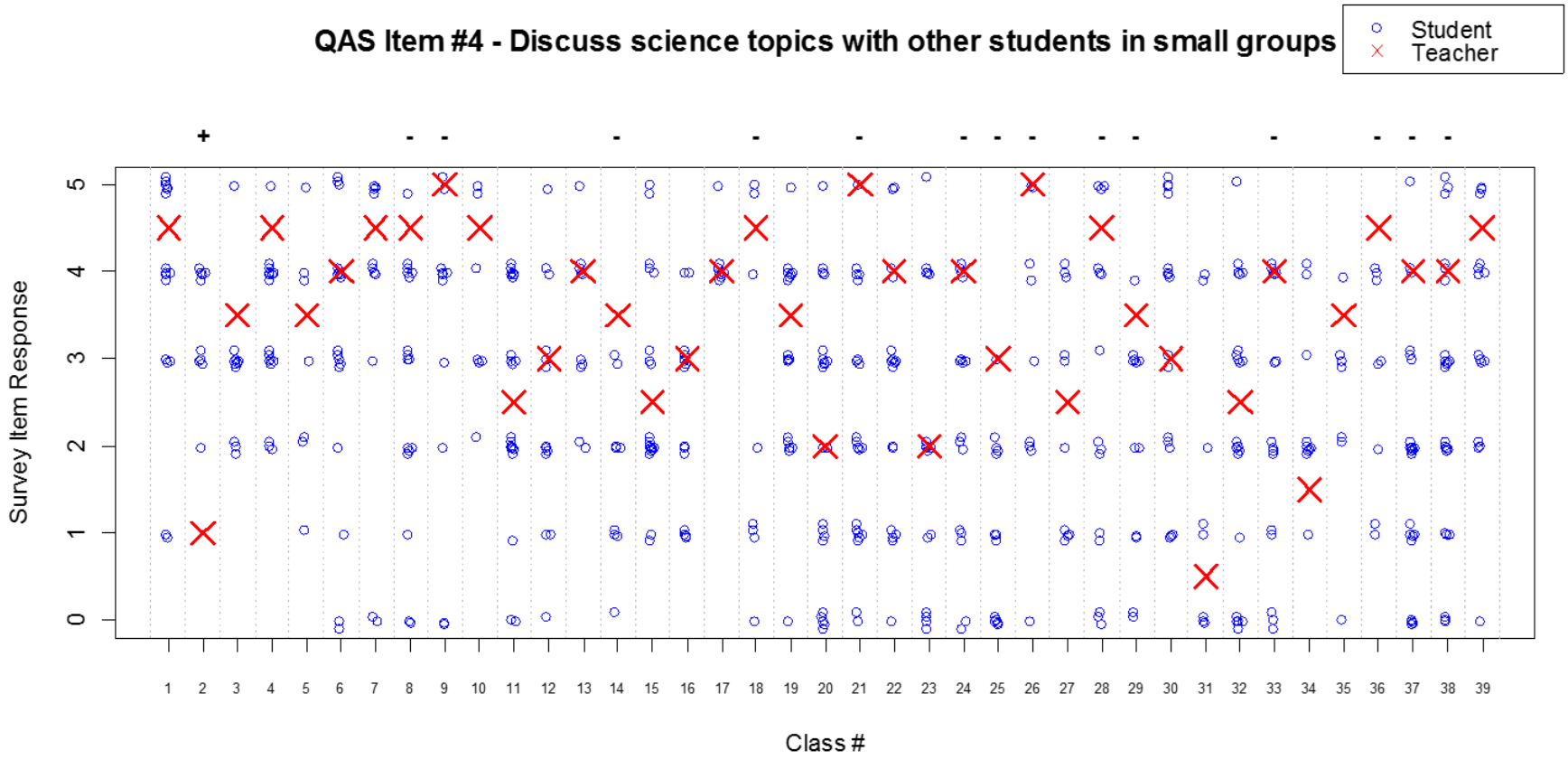
the discrepancy plot. Unlike Item #1, far fewer students rated this practice a 5, and indeed student rating for several classes even ranged from 0 to 5. In sum, the discrepancy plots of Item #1 and Item #4 are helpful in revealing these contrasting patterns in the spread of student responses and in understanding perceptual difference, both in terms of direction and magnitude.

Figure 7.4: *Item Discrepancy Plot Comparing Student and Teacher Ratings Within Classrooms on QAS Item #1, Listen to Lectures or Instruction Directed by the Teacher*



Note. Scale is 0 = *Never*, 1 = *Less than once a month*, 2 = *Once or twice a month*, 3 = *Once a week or more*, 4 = *Multiple times per week*, 5 = *Every day*. Blue circles represent student responses and red X's represent teacher average responses. Each vertical column corresponds to a classroom. Statistical significance of unstandardized difference in means for each teacher–classroom pair printed across the top with the symbol "–" indicating negative significance and the symbol "+" indicating positive significance.

Figure 7.5: *Item Discrepancy Plot Comparing Student and Teacher Ratings Within Classrooms on QAS Item #4, Discuss Science Topics with Other Students in Small Groups*



Note. Scale is 0 = *Never*, 1 = *Less than once a month*, 2 = *Once or twice a month*, 3 = *Once a week or more*, 4 = *Multiple times per week*, 5 = *Every day*. Blue circles represent student responses and red X's represent teacher average responses. Each vertical column corresponds to a classroom. Statistical significance of unstandardized difference in means for each teacher–classroom pair printed across the top with the symbol "–" indicating negative significance and the symbol "+" indicating positive significance.

7.5 Chapter Summary

This section utilized four methods—proportional scoring, unstandardized differences in means, item composite unstandardized differences in means, and discrepancy plots—to investigate the influence of classroom practice on the perceptual discrepancy between students and their teachers. Clear differences by direction and magnitude in student and teacher perceptions within classrooms were found by instructional practice. More "traditional" instructional practices were rated as occurring more frequently by students. Though discrepancy across instructional practices was likely driven by a combination of factors, this finding may have been partly influenced by students' level of engagement in performing these "traditional" practices. Simply put, students may have rated higher the practices they enjoyed doing and given lower ratings to the practices they found cognitively demanding or intellectually unstimulating.

By contrast, more active learning instructional practices were rated as occurring more frequently by teachers. Desimone et al. (2010) found math teachers significantly reported practices associated with the National Council of Teachers of Mathematics (NCTM) with greater frequency, such as student discussion and group work. It seems plausible, then, that teachers in the QAS sample may have reported small group discussion and experimentation as occurring more frequently given these instructional practices are more "reform-oriented". This result provides evidence that teachers may be susceptible to social desirability effects even when they understand the surveys are administered for low stakes purposes.

As in Chapter 5, the method employed only marginally impacted differences in discrepancy rankings, given that trends across instructional practices were broadly similar between item-level and composite unstandardized differences in means. Instructional practices most often rated higher by students (Item #1 *listen to lectures or instruction directed by the teacher*, Item #20 *review materials to prepare for a test or quiz*, and #22 *take tests or quizzes with questions where you choose the answer*) were also the practices with the most individual classes positively discrepant and highest composite unstandardized differences in means

estimates. Likewise, instructional practices rated lower by students (Item #4 *discuss science topics with other students in small groups* and Item #17 *conduct experiments or investigations*) were the practices with the most individual classes' negatively discrepant and lowest composite unstandardized differences in means estimates.

This alignment in the rankings of QAS instructional practice across proportional scoring, unstandardized differences in means, and item composite unstandardized differences in means was partly due to the fact that discrepancy between students and teachers in individual classes was very often directionally similar for particular instructional practices. Indeed, there was no instructional practice highly discrepant in terms of magnitude that did not also exhibit a distinctive directional trend in class discrepancy. This highlights an important limitation of item composite unstandardized differences in means in that the measure may underestimate discrepancy on practices for which students and teachers have a propensity to hold absolute perceptual differences but divergent directional patterns across classes. Nonetheless, this was less of a concern in the current analyses given discrepancy on the QAS survey tended to be systematically directional depending on the instructional practice.

Finally, rankings of QAS instructional practice were broadly similar across discrepancy methods because the magnitudes in mean differences for individual discrepant classes were generally comparable across instructional practices. The lone exception to this trend in dispersion was Item #1 *listen to lectures or instruction directed by the teacher*, as a disproportionate number of the very largest positive estimates of unstandardized differences in means occurred for this instructional practice. As a result, this instructional practice ranked highest in discrepancy using proportional methods and composite unstandardized difference in means, yet ranked somewhat lower in terms of its count of individual discrepant classes. It is unclear why some instructional practices may elicit a small number of class discrepancy estimates of great magnitude while other practices elicit a greater overall count of discrepant classes yet with estimates lower in magnitude.

CHAPTER 8

8. Discussion

8.1 Summary of Findings

This dissertation employed six methods to measure discordance between student and teacher perceptions of classroom practice frequency: proportional scoring, discrepancy scores, unstandardized difference in means, class composite unstandardized absolute difference in means, item composite unstandardized difference in means, and discrepancy plots. Discrepancy was greater between some teacher–classroom pairs (Class #21 and Class #26) than between others (Class #17 and Class #15), and these differences did not substantively depend on particular instructional practices. Instead, for the teacher–classroom pairs ranked most and least discrepant, perceptual difference was as much a function of the students and teacher as the instructional practice itself. This was true regardless of the approach employed to measure discrepancy. By contrast, discrepancy rankings for classes in the middle of the distribution were more fluid depending on the discrepancy method utilized.

This dissertation also found that the directional difference between student and teacher perceptions was influenced by the instructional practice. Instructional practices related to passive learning (listening to the teacher, assessment, and individual work) were rated as occurring more frequently by students. This finding may have been influenced by students' level of engagement while performing these "traditional" practices. Students may have rated higher the practices they preferred doing and given lower ratings to the practices they found cognitively demanding or intellectually unstimulating. Conversely, more active learning or "reform-oriented" instructional practices were rated as occurring more frequently by teachers, such as student discussion and lab work. This suggests teachers may be susceptible to social desirability effects even when surveys are administered in a low stakes environment. Finally, like class discrepancy rankings, the method employed only marginally impacted differences in discrepancy across instructional

practices. This was because discrepancy for particular instructional practices in individual classes was often directionally similar and magnitudes in discrepancy were generally comparable across instructional practices. Such a result indicates discrepancy may stem from two influences: universal perceptual difference between students and teachers due to the instructional practice itself and class-specific perceptual difference (deviation) between students and their teacher in a given class.

8.2 Meaningfulness of "Intuneness" as a Construct

A classroom functions similarly to any hierarchical organization (e.g., corporation, government, religious organization, etc.) that constantly strives to achieve common group objectives. In the context of the classroom, teachers must continually work with their students to achieve student learning and growth. However, for objectives to be realized, a high degree of perceptual alignment or "intuneness" between stakeholders can be a vital ingredient for success. If parties are not on the same page, student learning may be hindered.

"Intuneness" is defined as "a state in which people agree with or understand one another" (*Merriam-Webster Dictionary*, 2017). While this definition provides a helpful starting point, intuneness seems to take on different connotations depending on the field of study in which it is considered. In clinical psychology, intuneness is defined in terms of empathy. Empathic understanding facilitates a kind of "intuneness," an awareness of interconnectedness, which is "in itself healing, confirming, growth-promoting" (Rogers, 1987, p. 182). Other researchers have defined intuneness in terms of communication. According to Schramm (1956), the successful transmission of information "implies a degree of commonness or in-tune-ness [*sic*] between the systems which are communicating" (p. 503). Finally, other fields have foregone using the word altogether. Organizational psychology has utilized the word "congruence" to describe a type of perceptual intuneness (Edwards, 1994; Ostroff et al., 2005; Larson et al., 2013), while relational psychology has most commonly employed the word "discrepancy" to refer to perceptual discord (Israel et al., 2007; De Los Reyes, 2010; Sher-Censor, Parke, & Coltrane, 2010).

While the word "intuneness" is scantly used in educational literature, its usage seems to relate to the degree of "perceptual connectedness" between teacher and students. Silva (2012) embeds the construct of intuneness while describing the challenge of transforming academic knowledge from teacher to students:

> Each individual student is different. Each class of students is different. Each course is different. Each teacher is different. Each teacher–student encounter is different. The good teacher acknowledges and accepts these differences while, at the same time, sees commonalities—a respect for students and for self, an openness to "teachable" moments, an "intuneness" to students' learning and discovery, an "intuneness" to students' joys and anxieties, and a commitment to students' excellence. Yes, I believe that good teaching—that which exemplifies the scholarship of teaching—is both good science and fine art (p. 600).

Thus, Silva (2012) argues that a significant component of high quality pedagogy involves effective teachers who possess an understanding of and relatability to students' learning curiosities and needs, motivations and apprehensions. It is unclear the extent to which this awareness is inherent and personality-driven or developed with teaching experience.

In some ways, Silva's usage of the construct of intuneness may be tangentially related to Shulman's (1987) categories of teacher knowledge. Along with content, pedagogical, and curricular knowledge, Shulman stresses that quality teachers should hold "knowledge of learners and their characteristics" (Shulman, 1987, p. 8). This connectedness between teacher and student is the conduit by which knowledge can be transferred. Shulman (1987) continues, "But the key to distinguishing the knowledge base of teaching lies at the intersection of content and pedagogy, in the capacity of a teacher to transform the content knowledge he or she possesses into forms that are pedagogically powerful and yet adaptive to the variations in ability and background presented by the students" (p. 15). It seems reasonable that an experienced and knowledgeable teacher, highly in tune with students' learning needs, would be able to not only anticipate and address subject-specific challenges, but also understand and relate to students on a humanistic and personal level.

However, intuneness as measured using the matching QAS items may be more a reflection of the degree of teacher–student agreement in classroom perceptibility than of a teacher's skill in personal relatability. That is, intuneness in the current analysis may more critically depend on a combination of teacher awareness and student attention. Strong communication and cooperation may also be critical toward consciousness of the daily instructional practices performed. Furthermore, the role of participants' capacity for recall in explaining discrepancy cannot be discounted. It is also unclear the extent to which social desirability by teachers may influence observed discrepancy. As with discrepancy found in marital survey responses about partner roles, which appear to be partly driven by social desirability effects (Kamo, 2000), it is conceivable that some teachers rated higher the QAS practices they perceived to be associated with "higher quality instruction" and rated lower the QAS practices they perceived to be associated with "lower quality instruction".

A test can be thought of as a form of "communication device" (Cronbach, 1954, p. 267) on learning for students and their teachers. It provides baseline feedback (i.e., a sampling) of students' knowledge or skill in an academic subject after an instructional interval (Schramm, 1956). Perhaps, then, parallel survey items can be viewed similarly, as a type of polling instrument measuring teacher–student perceptual intuneness. While there is little reason to suspect a teacher who excels at raising test scores would necessarily score higher on intuneness metrics, greater mindfulness and attentiveness should positively impact intuneness. This might result from better organization in terms of lesson planning (e.g., by maintaining a weekly log), greater structure of ingrained routines, or perhaps through more explicit and frequent communication with students. Teachers, who explain to students not only why certain curricular topics are studied (i.e., why they are teaching what they are teaching), but also why certain classroom practices are implemented to study those topics, instill in their students greater instructional awareness. Yet, regardless of the relationship between teacher–student perceptual discrepancy and academic outcomes, the construct of intuneness is a meaningful research pursuit in its own right. This is true whether intuneness represents the misalignment in teacher–student

classroom awareness and attention—as was possible using the QAS items—or a potential broader metric of overall connectedness of the teacher–student relationship. The next section explicates the motivation behind further study of classroom perceptual discrepancy.

8.3 Practical Significance of Classroom Discrepancy Research

The practical rationale for studying discrepancy between student and teacher perceptions is its application as a supplementary measure of instructional practice. This section first argues for assessing instructional practice using a multiple measures approach and considers the common selection criteria for a measure's inclusion in such an evaluative framework. The case for continued study of discrepancy in student and teacher perceptions is then defended on the grounds that it can (1) reveal a more complete picture of teacher performance, (2) create more finely differentiated categories for classifying teachers, and (3) provide useful feedback for an untapped construct of teacher practice.

8.3.1 The Case for a Multiple Measures Approach Based on Feasibility

Teaching is complex, and consequently as many measures as is practical should be used (e.g., value-added models of student standardized test scores, classroom observations, portfolios, student and teacher surveys) to capture its intrinsic multidimensionality (Shulman, 1987) and identify potential areas in need of professional development. As Walkington and Marder (2014) comment, "Even the most vigorous advocates of value-added models acknowledge the need for multiple measures of teaching performance, particularly when decisions might lead to financial rewards or dismissal" (p. 235). Employing several evaluative measures also reduces the incentive for teachers to devote a disproportionate amount of class time to test preparation (Booher-Jennings, 2005). Furthermore, teachers who are effective at raising standardized test scores are frequently not as effective at improving nonacademic student outcomes such as social and emotional development (Blazar & Kraft, 2015). Thus, to capture student attitudinal and learning outcomes, and given each measure's inherent advantages and limitations, consensus exists among policymakers (e.g., *Race to the Top Guidance and FAQ*, 2010) and researchers (e.g.,

Kane et al., 2014) that no single measure should be utilized exclusively to evaluate practice. Rather, the contribution of distinct teaching quality evidence by each instrument requires that several measures be used to assess teaching practice through a "confirmation by triangulation strategy" (Popham, 2013, p. 40).

While a multiple measures approach is needed to assess teaching—with, theoretically, the more sources the better—some measures should be precluded from the evaluative framework on either methodological or practical grounds. Peterson (2000) outlines decision criteria for a measure's inclusion for teacher assessment: "For data sources to be acceptable in teacher evaluation systems, they must meet tests of logic, empirical trial, fairness, legality, and cost" (p. 92). The related conditions of fairness (i.e., equality for teachers in different settings, teaching different academic subjects, or in classes with differing characteristics) and legality (i.e., adherence of measures to federal, state, and local regulations) are assumed and not directly addressed in this dissertation. The remaining three stipulations are briefly explained in turn as they pertain to a measure of teacher–student discrepancy.

The first stipulation of Peterson (2000), "tests of logic", refers to the validity of the measure being considered. The tests raise three questions: Are the data "linked to student learning, welfare, or other needs"; "caused by (or the responsibility of) the teacher"; and "of primary importance in consideration of teacher quality?" (p. 93). Essentially, a measure of teacher evaluation must generate data that is related to other accepted measures of teaching quality or to student outcomes deemed important. While beyond the scope of this dissertation, establishing a negative relationship between perceptual differences within classrooms and consequential student achievement and affective outcomes would help to validate discrepancy as a supplementary measure of teaching evaluation.

Second, Peterson (2000) states a measure must undergo empirical trials. By this condition, the author means a measure must be reliable, "in actual use, logistics, and the presence of practical flaws not apparent in logical analysis" (p. 92). While the reliability of discrepancy

measures were not examined directly, proportional scoring methods found overall agreement patterns between student and teacher responses to be similar across teacher survey administrations (percent of student and teacher responses in exact agreement; practices rated highest and lowest by students and teachers). Thus, the degree of discrepancy was consistent across teacher survey administrations. Inter-method reliability was also high, given that the various methods employed to measure discrepancy broadly identified the same items and classrooms as most and least discrepant.

Finally, new measures of teaching quality must be defensible from a monetary perspective. Peterson (2000) cites one-on-one interviews with students as a measure that, while revealing rich information regarding teacher practice, is not feasible due to cost considerations. An advantage of measuring perceptual alignment of students and teachers through surveys is the relative inexpensiveness of the process. Unlike classroom observations and value-added analysis of student standardized test scores, an expenditure analysis conducted by the RAND Corporation found student surveys to be a considerably cheaper teacher evaluative measure to implement. For instance, of the $6.4 million spent on teacher evaluation in Pittsburgh Public Schools from 2009-2012, 47% of total expenditures were allotted for teacher observations and 45% were apportioned for value-added modeling procedures; however, only 8% of total expenditures were spent on the student survey measure. Likewise, the RAND study found that during the same period Memphis City Schools spent five times more on teacher observations than on student surveys (Chambers, De Los Reyes, & O'Neil, 2013). While this study did not investigate the relative quality or effectiveness of the measures as a function of cost, these results suggest student surveys can be created, distributed, and analyzed at a reasonable expense. Teacher surveys are similarly inexpensive relative to other measures as evidenced by many large-scale surveys currently in use to gather instructional information (e.g., National Assessment of Educational Progress; Rowan & Correnti, 2009). Thus, an intuneness metric collected from student and teacher surveys could potentially improve teacher practice at a fraction of the cost of other teaching quality measures.

If a teacher–classroom discrepancy metric meets Peterson's (2000) standards of validity, reliability, and practically, the utility of including such a measure of classroom perceptual difference in teacher evaluation should be considered. To that end, synthesizing the research on a multiple measures approach to teacher assessment, Schweig (2014b) cites five underlying benefits of collecting evaluative evidence from several sources. A multiple measures evaluative approach should provide (1) "a more complete picture of teacher performance"; (2) "finer, more stable categories for classifying teachers"; (3) "feedback to help improve classroom practice"; (4) "reduced incentives for gaming the system"; and (5) "greater confidence in results among stakeholders" (p. 5). While the efficacy of a discrepancy measure for disincentivizing evaluation system abuse or creating greater confidence in teacher assessment results remains unclear (Reasons 4 and 5), it is not difficult to imagine its usefulness with regard to Reasons 1–3. That is, including discrepancy as a supplementary measure of teaching quality may reveal a more complete picture of teacher performance, create more finely differentiated categories for classifying teachers, and provide worthwhile and distinctive feedback for an untapped construct. The subsequent sections explain each of these justifications in greater depth.

8.3.2 Discrepancy Can Reveal a More Complete Picture of Teacher Performance

Studying discrepancy holds promise for helping researchers better understand the complexities of quality teaching. Currently, a lack of clear consensus on the precise characteristics and practices of an effective teacher has made teaching quality an especially difficult construct to measure (Goe et al., 2008). As noted by Campbell, Kyriakides, Muijs, and Robinson (2004), "Teachers can be effective with some students more than others, with some subjects more than others, in some contexts more than others, and with some aspects of their professional work more than others" (p. 4). For these reasons, it can be difficult to differentiate teaching quality. Moreover, teachers' role within the dynamic environment of the classroom makes it challenging for researchers to come to agreement about "what defines teaching quality and what the corresponding evaluative criteria should be" (Schweig, 2014b, p. 3).

Furthermore, as Lavigne and Good (2013) note, "What it means to be a good teacher varies and encompasses a wide range of dispositions and characters" (p. viii). Hiebert and Grouws (2007) define teaching in terms of its bidirectional quality (i.e., the dynamic relationship between students and teachers around subject matter) and its impact on facilitating student achievement and learning goals. Using this criterion, successful teachers interact and instruct their students by designing and implementing lessons to effectively assist in the progress and achievement of their students' learning goals. Thus, at the heart of discrepancy research is what Hiebert and Grouws (2007) propose is the core question of education: "What is it, exactly, about teaching that influences students' learning?" (p. 371). A measure of discrepancy may help answer these intricate questions by proving to be a worthwhile tool for better understanding teacher–student relationships. Moreover, determining the degree to which the viewpoints of teachers and students align may help researchers better understand the detrimental influence perceptual discord may have on student learning and attitudes. At its most effective, uncovering the meaning and implications of teachers holding discrepant perceptions from their students may allow for a greater comprehension of teacher practice itself.

8.3.3 Discrepancy Can Help to Further Differentiate Teacher Practice

Incorporating a teacher–classroom discrepancy measure into a teacher evaluation framework may also help to more acutely differentiate teaching quality. The existence of the Widget Effect (Weisberg et al., 2009)—the inability of traditional evaluation systems based on teacher qualifications to adequately differentiate teaching performance—is supported by an underlying reality of teaching quality. Value-added analyses of student achievement have revealed that the distribution of teaching quality is not bell-shaped but positively kurtotic (Staiger & Kane, 2014). Likewise, measurements from teacher observational instruments suffer from even greater volatility (Kane & Staiger, 2012). Although the determination of traditional evaluation systems that 99% of teachers are satisfactory cannot be accurate, it does seem that the quality of many teachers (in terms of raising standardized test scores) is broadly indistinguishable for all intents and purposes. That is, the distribution of teaching quality may be

such that there exists a small percentage of struggling teachers, a small percentage of exceptional teachers, and an overwhelming majority (say, 80% to 90%) of teachers who can rightly be classified as satisfactory. For these satisfactory teachers, the standard errors associated with their value-added estimates overlap; thus, teaching quality is much more difficult to differentiate—if substantive differences indeed exist at all (Raudenbush & Jean, 2012).

This seeming imperceptibility of quality for the vast majority of teachers is further justification for employing several feasible measures when designing evaluation systems. Not only is the inclusion of multiple measures useful for capturing a more complete picture of teaching, incorporating additional, sound measures can also be a beneficial technique for more finely distinguishing teaching quality. This is particularly true given that value-added estimates, classroom observation scores, and student ratings are only weakly correlated—an indication that the information these measures yield about teaching quality practice is quite distinct (Chaplin, Gill, Thompkins, & Miller, 2014; Walkington & Marder, 2014). Collecting several measures of teaching quality may result in conflicting signals of quality. Still, the inherent multidimensionality of teaching should not be ignored, but embraced, as it is the reality of the profession. Utilizing a supplementary teacher–classroom discrepancy measure could provide decision-makers with auxiliary evidence, which may corroborate or complicate ratings from other measures, yet nonetheless succeeds in identifying finer differences in practice among teachers.

Necessarily, the inclusion of an additional measure to further differentiate teacher practice requires a precondition that it produces meaningful variation in teaching quality estimates. For example, incorporating a self-assessment measure where 90% of teachers rate themselves as "outstanding"—as was found in the Chilean national teacher evaluation system (Taut & Sun, 2014)—will do little to discriminate practice. Similarly, DePascale (2012) cites teacher professionalism as a measure for which, "Evaluators might find it nearly impossible to distinguish among the vast majority of teachers performing in the middle of the distribution" (p.

9). Therefore, to hold potential value as an instrument to further classify teachers amongst their peers, a teacher–classroom discrepancy measure must both be valid and create score variability for its inclusion in an evaluation system.

Finally, it is important to caveat that a discrepancy measure is clearly exploratory and should only be administered for low-stakes evaluation given its potential limitations. For instance, a survey-taker who commits to answering items at the extremes of the scale opens the door to the possibility of more frequently recording a discrepancy. Thus, it is unclear the degree to which a teacher may be able manipulate the magnitude of his or her overall discrepancy by only choosing middle categories. Furthermore, the connection between perceptual discrepancy and student learning is unknown. It is not necessarily true that teachers should always strive for greater intuneness with their students, as teacher intuneness is naturally not synonymous with teacher quality. Notwithstanding these limitations, by identifying those teachers whose perceptions are most "out of step" with their students, a discrepancy measure could prove potentially useful for administrators as a way to further examine and more finely differentiate teachers' practice.

8.3.4 Discrepancy Can Provide Teachers with Feedback for an Untapped Construct

The goal of formative assessment is to provide information to instructors in order to make responsive changes in teaching and student learning (Boston, 2002). As Gil (1987) notes, feedback "can help teachers not only improve their teaching practices but also change their attitudes toward the act (or art) of teaching" (p. 62). Teacher–classroom discrepancy plots may have value as a supplementary formative assessment tool for teachers by providing easily digestible information regarding their degree of perceptual congruence with their students on specific classroom practices and overall instruction. Specifically, these plots can inform teacher practice by illuminating, in a straightforward manner, instructional practices that fall outside the ideal range of perceptual alignment between teachers and their students. Identifying such discordance can perhaps spark productive teacher–administrator and teacher–student

conversations as well as initiate discussions between colleagues.

As McKeachie (1987) postulates about instructional evaluation more broadly, the process "may be helpful not so much in determining whether or not teaching is less than optimal but in identifying specifically where the problems lie" (p. 4). Teacher–classroom discrepancy plots may help teachers better understand their students' perceptions and attitudes, as well as pinpoint areas of perceptual discord. For instance, comparing teacher self-assessments to student perceptions, Montgomery and Baker (2007) found that students believed they received more feedback than their teachers perceived giving. Similarly, by analyzing matching survey items responses, a teacher can ascertain, for example, whether she engages in test preparation more frequently or provides laboratory work less frequently than her students perceive. After all, inherent in being perceptually out of tune is that, by definition, one is unaware of his or her mismatched attitudes or perceptions. Furthermore, different teachers can implement the same instructional strategy in different ways. For example, an ineffectual teacher can turn a high-level task into a rote exercise by the way he or she implements it. Perhaps then, perceptual discrepancy partly reflects the difference between teacher intent and teacher enactment—and students are evaluators capable of discerning the difference.

In this way, the usage of a discrepancy measure as an identification tool could be an impetus for change by providing information to teachers relevant to improving their pedagogical areas of concern. One could imagine a teacher comparing his or her intuneness with students to that of his or her colleagues', or charting the progress or trajectory of his or her intuneness over time. For example, a discrepant teacher surprised by many of her students' perceptions of instructional practices might make a conscious effort to reduce the frequency with which she lectures or increase the time allotted for experimental work. The teacher could then examine possible trends in her students' ratings in subsequent years to self-evaluate her change in pedagogy as well as responsively adjust her own ratings in hopes of becoming more in tune over time. Thus, measuring teacher–student perceptual congruence may hold professional

developmental promise by giving teachers constructive and actionable feedback to ameliorate a yet to be studied area of their practice.

Finally, it is important to be cognizant that collecting feedback is only effective for improving practice if the amassed information is used in conjunction with other components of the evaluative process. Stevens (1987) posits that the process to change instructional practice consists of four components: evaluation, identification, design, and implementation. Feedback on its own is "unlikely to produce meaningful instructional improvement without the recognition that obtaining evaluative information is only the first step in a larger process" (p. 36). For a measure of teacher–classroom discrepancy to be truly effective the feedback information it elicits must be used to design and implement strategies for instructional improvement which can eventually be properly evaluated. As such, a measure of teacher–classroom discrepancy should be viewed as a supplemental identification instrument to be employed with other measures of teacher practice.

8.4 Limitations

Several limitations merit brief comment in the comparison of student and teacher survey responses. First, a critically important question is whether similarly worded student and teacher items really measure the same constructs. Factor analysis by Kunter and Baumert (2006) has shown parallel student and teacher survey forms may nonetheless still measure different aspects of the learning environment. That is, ratings may not be simply different methodological approaches to evaluating the same characteristics of instruction but rather may reflect "perspective-specific constructs" (p. 234) held by students and teachers. For instance, it is conceivable that students and teachers may have interpreted some of the same matching QAS items differently. This may be particularly true regarding students' and teachers' conceptions of test review activities and assessment practices. For example, whereas students may view all assessments similarly, teachers may be more apt to differentiate between multiple choice tests and short answer assessments given the latter's greater formative value.

A related limitation of the current research pertains to the stated subject of the QAS items. An item asking teachers how often their students work in small groups is very different from an item asking them how often they *think* their students work in small groups. Furthermore, as outlined in Section 2.1.7, it can be difficult to discriminate between items intended to measure individual-level psychological constructs (context variables) and classroom-level organizational constructs (climate variables) (Marsh et al., 2012). In the current analyses, there was an inherent assumption that student variation in responses within classrooms was a reflection of rater measurement error and not substantive differences in students' classroom experiences. That is to say, students within the same classroom were assumed to have a shared environment of teaching. However, even for items designed to explicitly measure class or teacher constructs, it is certainly possible to imagine the existence of classroom microclimates (Seidel, 2006), in which some students in reality *do* experience different types or amounts of instruction. For example, variation in student ratings of an item asking about the frequency with which the teacher gives individualized feedback would reflect true differences if, theoretically, lower achieving students actually did receive more one-on-one feedback during class time. In such a scenario, teacher–student discrepancy would be conflated with a teacher's use of differentiated instruction. The results of this dissertation were limited insofar as it was assumed that the QAS matching classroom practice items reflected similar teacher–student expectations experienced by all students equally. Future research may consider the possibility of student learning subgroups within classrooms, rather than presuppose that students experience all instructional practices similarly—which may not always be the case (Seidel, 2006).

Another limitation pertains to the generalizability of these results, as only one matching student and teacher dataset was employed in the analyses. This shortcoming stems from a dearth of appropriate survey datasets that can be analyzed to explore teacher–student discrepancy. First, there is a lack of publicly available survey data. The most prominent student surveys (Tripod, Student Voice Survey, YouthTruth Student Survey, My Student Survey, and iKnow My Class Survey) are all commercially produced, and items and data are generally not publicly available.

Other surveys simply do not include similarly worded student and teacher items (e.g., Los Angeles Unified School District or New York City Department of Education surveys) or do not link the responses of students and teachers by unique classroom codes (e.g., University of Chicago Consortium on Chicago School Research Surveys). For still other surveys, similarly worded items are administered on non-matching scales (e.g., National Assessment of Educational Progress). Clearly student and teacher surveys have been designed independently in the past, without an eye toward making perceptual comparisons across forms. It is the author's hope that a convincing case for the efficacy of discrepancy research has been made so that decision-makers will consider relatively simple alterations in survey design and data storage to allow for perceptions of the classroom to be appropriately compared between students and teachers.

8.5 Future Directions

The results of this dissertation provide many worthy avenues for future exploration of classroom perceptual discrepancy. These topics include the predictive quality of discrepancy, the source of discrepancy, and various study design and methodological suggestions.

8.5.1 Predictiveness of Discrepancy

First and foremost, a natural extension of measuring classroom perceptual differences is investigating the relationship between discrepancy and student achievement and attitudinal outcomes. A simple means of exploring the predictive impact of discrepancy is the use of regression analyses. For example, unstandardized differences in means, composite absolute unstandardized differences in means, or student and class composite absolute discrepancy scores could be included in single and multilevel regression models to predict various outcomes of interest. Covariate measures could also be included to examine whether background characteristics of students (e.g., socio-economic status, English language ability, prior achievement) as well as those of teachers (e.g., graduate degree attainment, years of teaching experience) influence the strength of these potential relationships. It certainly seems plausible

that a student's linguistic background or a teacher's acquired pedagogical knowledge may influence his or her interpretations and perceptions of classroom practices and thereby impact intuneness.

Additionally, the more sophisticated and flexible method of polynomial regression, as introduced in Section 3.3.3, should be employed to study the potential predictiveness of classroom perceptual discord. Though not widely adopted in the field of education, this procedure allows a researcher to graph the degree to which two associated predictor variables—and extent of their discrepancy—relate to an outcome measure in three-dimensional space (Edwards, 1994; Shanock et al., 2010). Polynomial regression could be used to answer more nuanced research questions of class congruence including how the magnitude of discrepancy, the direction of discrepancy, and the extent of agreement between students and their teachers relate to outcome measures of import. Cross-level polynomial regression—a variant of polynomial regression but within a hierarchical linear modeling framework (Jansen & Kristof-Brown, 2005)—would be a particularly appropriate method of examining the predictiveness of discrepancy on individual items and outcomes. Furthermore, if matching surveys included multiple items of the same instructional dimension, discrepancy could be studied using latent variable modeling procedures. For instance, multilevel latent polynomial regression (MLPM), a variant of polynomial regression, could be used to investigate the impact of perceptual differences on instructional domains and outcomes measures (Zyphur, Zammuto, & Zhang, 2016). A potential lack of predictiveness of intuneness with student achievement or attitudes should not exclude the use of such a measure in formative assessment. Nonetheless, if greater discrepancy between teachers and their students were found to be negatively associated with meaningful outcomes, such findings would help to validate the measure and lend credence to the importance of classroom intuneness research.

8.5.2 Source of Discrepancy

Discrepancy can be the result of a high student rating, low student rating, high teacher

rating, low teacher rating, or some combination of these forces (Kamo, 2000). It is of course impossible to know the "true" frequency with which classroom practices are performed in a given class and whether differences in ratings are due to measurement error or true differences in experience. Nonetheless, future research should consider whether students or their teacher are the greater sources of discrepancy across particular practices. A rudimentary way to examine source could be to explore the extent to which discrepant ratings of students and teachers in some classes diverge (comparatively speaking) from the ratings on the same instructional practices in other classes. For example, one might determine the degree to which, for a given item, the mean student rating in a given class is relatively similar to the mean student ratings in other classes, and then contrast this rank with the extent to which the mean teacher ratings in the given class is considerably different from the mean teacher ratings in other classes. Such comparative procedures may aid in providing evidence towards disentangling the impetuses of discrepancy. To be sure, determining the source of class discord is a thorny and complicated endeavor, but it is a research topic worthy of future pursuit.

8.5.3 Qualitative Follow-Up

While not possible in the current analysis, qualitative follow-up in those classes most and least discrepant would be especially valuable for future discrepancy research. For instance, having students in a classroom rate items separately and then come to a consensus would help reveal the reasoning behind particular instructional practice ratings and allow students to share their class experiences with their peers. Moreover, such a consensus-building process might reveal class microclimates if, perhaps, a teacher was found to be more perceptually aligned with some proportion of students in the class while less in tune with another group of students. In addition, conducting classroom observations would be worthwhile in order to determine whether teachers more perceptually aligned with their students tend to display certain instructional qualities with greater frequency or exhibit a better pulse of the classroom. Finally, interviewing teachers about their divergent item responses with the aid of discrepancy plots would be invaluable to understanding the potential mechanisms of perceptual misalignment. As an

example, Teacher–Classroom Pair #26 was discrepant on Item #6 *work on worksheets*. Was the discrepancy due to true perceptual differences (i.e., Teacher #26 truly perceived students completed worksheets at a less frequent rate than student in the class)? Were semantic differences the source of the discrepancy? Perhaps Teacher #26 held a more encompassing definition of the word "worksheets" than students did, which included "warm-up assignments" and "lesson summary tasks", or utilized different class-specific vocabulary to describe the practice. It seems plausible that lack of intuneness could stem from a mismatch between the lexicon a teacher uses when referring to practices and the vocabulary found on the surveys. Or maybe the discrepancy stemmed from self-serving bias or a misguided attempt to achieve perceptual congruence. Perhaps Teacher #26 *did* believe students worked on worksheets more frequently than originally indicated, but rated the item lower thinking it was the answer researchers were looking for, or that it was the answer the students were most likely to give. There are many possible explanations students and teachers might provide for their response choices. Conducting such think-aloud activities with teachers would help researchers verify the perceptual discord identified using discrepancy plots.

8.5.4 Methodological Refinements

There are inherent limitations to using an existing instrument like the QAS surveys. For this reason, researchers should also consider making several methodological refinements for future studies of perceptual differences between students and teachers. First, matching survey items should be interval in scale, not ordinal like the QAS items. It is important that the magnitude of the difference between response options is consistent so that the degree to which the levels of the predictor variables affect an outcome can be precisely measured. This stipulation is particularly critical in polynomial regression analyses. In addition, matching items should be asked of low inference practices. The results of this study, as well as those of Desimone and colleagues (2010), have shown students are more capable of accurately assessing tangible instructional practices (e.g., *work on worksheets*) than practices that are more inherently vague (e.g., *small group discussion*), as ratings on low-inference practices tend to exhibit less

variation within classrooms.

A future study of perceptual difference might also consider employing matching survey items that de-emphasize students' own classroom experiences. The QAS student surveys framed items in terms of the frequency students in the class engaged in given instructional practices ("How often do you do each of the following activities"). However, in measuring discrepancy a researcher may instead want to ask students more broadly about their shared classroom environment ("How often do students in this class do each of the following activities"). Stressing the collectiveness of the classroom by using de-personalized item stems may influence patterns of discrepancy and help to clarify whether an instructional practice is best classified as climate or context in nature.

In this way, discrepancy research has natural implications for survey development. After all, a prerequisite for discrepancy analysis is that the student and teacher surveys are reasonably reliable and valid. Thus, a natural consequence of measuring intuneness is that by analyzing responses on paired items researchers have the potential to collect convergent types of validity evidence. Discrepancy between student and teacher survey ratings gives researchers an investigative opportunity to determine why the measures diverge. As such, this type of research could help facilitate a better understanding of the functioning of survey items of instructional practice.

Furthermore, future researchers might also consider examining the efficacy of directly asking students and teachers survey items related to the inherent characteristics of classroom intuneness. Though this dissertation utilized teacher–student perceptual discrepancy as an implicit measure of intuneness, the argument can be made that congruence is not an extant construct but an illusion created from component measure differences (Edwards, 1994; Cronbach & Gleser, 1953). For example, triangulating various discrepancy measures—as was done in this dissertation—with student and teacher responses to items describing the characteristics of intuneness (e.g., awareness, engagement, understanding, connectedness, trust) may help

researchers better comprehend perceptual difference. Do teachers rated higher by students in terms of classroom awareness tend to score lower on discrepancy measures? What about teachers who students feel always have their best interests in mind or who self-rate higher on student connectedness and rapport? Furthermore, how does discrepancy relate to other teacher characteristics, such as their values or learning expectations? Like conducting teacher interviews, determining the extent to which ratings on these qualities are associated with discrepancy measures has the benefit of helping researchers better understand why perceptual differences in the classroom occur.

In designing surveys to specifically measure classroom perceptual congruence between students and their teachers, researchers should contemplate several new domains of instructional practice that are ripe for study. These include homework volume, degree of student autonomy, student lesson feedback, classroom rules, classroom learning routines, clarity of teacher communication, and assessment procedures. As outlined in Table 1.1, many of these dimensions of classroom practice are already being studied with student survey measures. Relatedly, future research might also examine intuneness in terms of a teacher's predictiveness of his or her students' achievement. For example, do teachers who better forecast their students' scores on assessment measures exhibit less discrepancy with their students' on matching survey items? Follow-up teacher interviews and classroom observations should be conducted to gain insight into the sources of perceptual disagreement. The QAS surveys are by no means exemplary measures for the study of classroom perceptual discord, as the data was not originally collected with this singular purpose in mind. Yet, the shortcomings of this measure can lead to thoughtful improvements in survey design and implementation for future discrepancy research.

8.6 Conclusion

Some educational economists have suggested (e.g., Hanushek, 2011; Rivkin et al., 2005) that the most efficient way to improve the quality of teachers in the United States is by frequently re-sorting the job pool or "firing to the top"—essentially dismissing the bottom 5% to

8% of teachers annually. But a more realistic, productive, and noncontroversial solution is to improve overall teaching quality—as opposed to working under an inherent teacher quality framework. Improving the preparedness of new teachers graduating from teacher education programs is one obvious way to improve teaching quality (Lavigne & Good, 2013). In addition, Stiggins and Duke (1988) maintain that, "as we pursue excellence in education through the promotion of the professional development of teachers, we cannot overlook the potential contribution of the teacher evaluation process" (p. xi).

Properly utilizing measures that provide valuable formative assessment information to teachers is paramount to improving teaching practice. This means using value-added scores not as irrefutable summative evidence but rather as a flagging mechanism to identify struggling teachers. It also means comparing teachers' classroom observation and students' survey scores across instructional domains to determine areas of individual pedagogical strength and areas in which instructional improvements are necessary. Other formative assessment tools, including teacher logs and portfolios of student work and lesson plans, should be employed to diagnose struggling teachers most in need of improvement and to identify specific pedagogical areas in which the application of professional development would provide the greatest support.

Yet researchers and practitioners should not be limited to measuring teaching quality using only our current set of methodological instruments. This is particularly true given the multidimensionality, not just of teacher practice, but also of student learning in the classroom in the form of attitudinal and behavioral growth (Blazar & Kraft, 2015). Finding and implementing new, financially feasible measures of teaching quality that provide useful and corroborative information would prove fruitful in improving instructional practice. Just as valued-added modeling can aid in recognizing teachers who experience difficulty improving their students' achievement, and as classroom observations can help identify teachers struggling with pedagogy, teacher–classroom unstandardized differences in means and discrepancy plots may help to identify teachers whose perceptions most diverge from their students'. Such metrics of classroom

perception could be included as a supplementary piece of formative evaluative evidence, used in conjunction with other measures, to more fully reveal the complexities of instructional practice and support professional development (Mehrens, 1990). If researchers are serious about comprehensively capturing the multidimensionality of teaching, then *any* measure—including classroom perceptual discrepancy—that is fair, legal, reliable, cost-effective, and valid should be earnestly explored. The findings of this dissertation suggest that comparing student and teacher survey responses holds promise as a method to reveal a richer, more variegated picture of teaching quality and provide teachers with unique diagnostic feedback to self-improve their instructional practice.

Appendix: Additional Tables and Figures

Appendix Table A.1: *QAS Student Responses, Descriptive Statistics*

Classroom Practice Items	n	Never	Less than once a month	Once or twice a month	Once a week or more	Multiple times per week	Every day	Mean	SD
#1 Listen to lectures or instruction directed by the teacher	640	5	13	28	53	186	355	4.29	1.02
#2 Read a science textbook	638	70	68	102	196	155	47	2.69	1.42
#3 Read science articles in magazines or newspapers	632	215	151	127	108	20	11	1.37	1.29
#4 Discuss science topics with other students in small groups	636	68	74	137	139	161	57	2.66	1.46
#5 Discuss science topics with other students as a class	633	40	39	77	146	196	135	3.30	1.42
#6 Work on worksheets	642	2	25	73	167	216	159	3.63	1.11
#7 Work on homework tasks during class time	640	59	84	106	178	139	74	2.74	1.46
#8 Watch science videos, movies, TV shows, etc.	636	22	69	150	203	149	43	2.81	1.21
#9 Use science-related software or internet resources	636	105	101	146	141	100	43	2.25	1.49
#10 Work on science projects individually	639	89	179	191	112	46	22	1.86	1.26
#11 Work on science projects in pairs or groups	641	52	113	184	131	116	45	2.44	1.37
#12 Work on written science reports	631	132	184	146	97	57	15	1.70	1.33
#13 Present oral science reports	632	232	168	122	73	26	11	1.25	1.27
#14 Watch teacher demonstrate experiment or investigation	631	19	61	147	188	143	73	2.94	1.26
#15 Use lab instruments or materials	642	23	69	172	197	135	46	2.76	1.21
#16 Plan/design experiments or investigations	638	90	119	176	157	79	17	2.11	1.31
#17 Conduct experiments or investigations	634	50	95	167	181	118	23	2.46	1.27
#18 Analyze data / relationships between variables	637	51	86	148	202	120	30	2.54	1.29
#19 Write reports about labs or experiments	638	101	108	156	143	99	31	2.19	1.43
#20 Review materials to prepare for a test or quiz	634	20	35	90	172	201	116	3.34	1.26
#21 Develop or practice test-taking skills	635	68	96	113	167	136	55	2.59	1.46
#22 Take tests with questions where you choose the answer	641	16	48	221	173	122	61	2.81	1.19
#23 Take tests with questions where you write out the answer	640	51	93	221	147	100	28	2.37	1.26

Note. Scale is 0 = *Never*, 1 = *Less than once a month*, 2 = *Once or twice a month*, 3 = *Once a week or more*, 4 = *Multiple times per week*, 5 = *Every day*.

Appendix Table A.2: *QAS Pre-Survey Teacher Responses, Descriptive Statistics*

Classroom Practice Items	n	Never	Less than once a month	Once or twice a month	Once a week or more	Multiple times per week	Every day	Mean	SD
#1 Listen to lectures or instruction directed by the teacher	37	1	3	2	7	22	2	3.41	1.14
#2 Read a science textbook	38	0	6	10	14	7	1	2.66	1.05
#3 Read science articles in magazines or newspapers	38	8	7	16	7	0	0	1.58	1.03
#4 Discuss science topics with other students in small groups	38	1	2	3	7	18	7	3.58	1.20
#5 Discuss science topics with other students as a class	38	0	1	2	7	16	12	3.95	0.98
#6 Work on worksheets	38	0	2	7	17	12	0	3.03	0.85
#7 Work on homework tasks during class time	37	6	5	9	9	6	2	2.27	1.47
#8 Watch science videos, movies, TV shows, etc.	37	1	6	13	13	4	0	2.35	0.98
#9 Use science-related software or internet resources	37	3	4	17	8	5	0	2.22	1.08
#10 Work on science projects individually	37	1	13	13	5	4	1	2.03	1.14
#11 Work on science projects in pairs or groups	37	3	6	7	9	9	3	2.65	1.44
#12 Work on written science reports	37	3	9	12	8	4	1	2.11	1.22
#13 Present oral science reports	36	5	18	8	5	0	0	1.36	0.90
#14 Watch teacher demonstrate experiment or investigation	38	0	5	8	18	7	0	2.71	0.93
#15 Use lab instruments or materials	38	0	0	3	19	16	0	3.34	0.63
#16 Plan/design experiments or investigations	38	2	5	15	14	2	0	2.24	0.94
#17 Conduct experiments or investigations	38	0	0	4	18	16	0	3.32	0.66
#18 Analyze data / relationships between variables	38	0	2	8	15	13	0	3.03	0.88
#19 Write reports about labs or experiments	38	0	7	12	12	6	1	2.53	1.06
#20 Review materials to prepare for a test or quiz	38	0	4	18	6	8	2	2.63	1.10
#21 Develop or practice test-taking skills	38	1	7	18	5	6	1	2.29	1.11
#22 Take tests with questions where you choose the answer	38	0	5	21	11	1	0	2.21	0.70
#23 Take tests with questions where you write out the answer	37	2	6	18	8	3	0	2.11	0.97

Note. Scale is 0 = *Never*, 1 = *Less than once a month*, 2 = *Once or twice a month*, 3 = *Once a week or more*, 4 = *Multiple times per week*, 5 = *Every day*.

Appendix Table A.3: *QAS Post-Survey Teacher Responses, Descriptive Statistics*

Classroom Practice Items	n	Never	Less than once a month	Once or twice a month	Once a week or more	Multiple times per week	Every day	Mean	SD
#1 Listen to lectures or instruction directed by the teacher	39	2	1	4	7	23	2	3.38	1.16
#2 Read a science textbook	39	2	6	12	8	11	0	2.51	1.21
#3 Read science articles in magazines or newspapers	38	5	15	11	5	2	0	1.58	1.06
#4 Discuss science topics with other students in small groups	39	1	2	6	7	15	8	3.46	1.27
#5 Discuss science topics with other students as a class	39	0	2	5	7	13	12	3.72	1.19
#6 Work on worksheets	39	1	3	5	15	14	1	3.05	1.07
#7 Work on homework tasks during class time	39	6	7	7	12	6	1	2.21	1.40
#8 Watch science videos, movies, TV shows, etc.	39	1	7	15	10	6	0	2.33	1.03
#9 Use science-related software or internet resources	39	4	5	16	5	9	0	2.26	1.25
#10 Work on science projects individually	39	1	15	15	6	2	0	1.82	0.91
#11 Work on science projects in pairs or groups	39	4	5	9	10	8	3	2.56	1.43
#12 Work on written science reports	39	5	14	10	5	5	0	1.77	1.22
#13 Present oral science reports	39	10	16	13	0	0	0	1.08	0.77
#14 Watch teacher demonstrate experiment or investigation	39	1	2	17	13	6	0	2.54	0.91
#15 Use lab instruments or materials	39	0	2	7	18	11	1	3.05	0.89
#16 Plan/design experiments or investigations	39	3	7	17	7	5	0	2.10	1.10
#17 Conduct experiments or investigations	39	0	2	9	18	9	1	2.95	0.89
#18 Analyze data / relationships between variables	39	0	3	11	11	14	0	2.92	0.98
#19 Write reports about labs or experiments	39	2	7	13	10	6	1	2.36	1.18
#20 Review materials to prepare for a test or quiz	39	0	8	14	11	4	2	2.44	1.10
#21 Develop or practice test-taking skills	39	1	8	13	8	9	0	2.41	1.14
#22 Take tests with questions where you choose the answer	39	0	11	18	8	2	0	2.03	0.84
#23 Take tests with questions where you write out the answer	38	4	8	15	7	3	1	2.00	1.19

Note. Scale is 0 = *Never*, 1 = *Less than once a month*, 2 = *Once or twice a month*, 3 = *Once a week or more*, 4 = *Multiple times per week*, 5 = *Every day*.

Appendix Table A.4: *Student Variance Within and Between Classes, by QAS Classroom Practice Item*

Classroom Practice Item	Within-Class	Between-Class	Total	ICC
#1 Listen to lectures or instruction directed by the teacher	0.89	0.17	1.06	16%
#2 Read a science textbook	1.08	0.92	2.00	46%
#3 Read science articles in magazines or newspapers	1.25	0.44	1.69	26%
#4 Discuss science topics with other students in small groups	1.83	0.30	2.13	14%
#5 Discuss science topics with other students as a class	1.81	0.22	2.03	11%
#6 Work on worksheets	0.89	0.36	1.25	29%
#7 Work on homework tasks during class time	1.85	0.28	2.13	13%
#8 Watch science videos, movies, TV shows, etc.	1.02	0.40	1.42	28%
#9 Use science-related software or internet resources	1.85	0.33	2.18	15%
#10 Work on science projects individually	1.43	0.16	1.59	10%
#11 Work on science projects in pairs or groups	1.49	0.45	1.94	23%
#12 Work on written science reports	1.37	0.39	1.76	22%
#13 Present oral science reports	1.10	0.49	1.59	31%
#14 Watch teacher demonstrate experiment or investigation	1.45	0.14	1.59	9%
#15 Use lab instruments or materials	1.06	0.39	1.45	27%
#16 Plan/design experiments or investigations	1.48	0.22	1.70	13%
#17 Conduct experiments or investigations	1.17	0.37	1.54	24%
#18 Analyze data / relationships between variables	1.38	0.26	1.64	16%
#19 Write reports about labs or experiments	1.55	0.46	2.01	23%
#20 Review materials to prepare for a test or quiz	1.47	0.16	1.63	10%
#21 Develop or practice test-taking skills	1.85	0.30	2.15	14%
#22 Take tests with questions where you choose the answer	1.21	0.21	1.42	15%
#23 Take tests with questions where you write out the answer	1.31	0.29	1.60	18%

Appendix Table A.5: *Unstandardized Differences in Means, by QAS Item*

				Individual Work			Interactive Work			Discussion		Reports & Projects				Experimental Work					Assessments				Class
ID	N	#1	#14	#2	#6	#7	#3	#8	#9	#4	#5	#10	#11	#12	#13	#15	#16	#17	#18	#19	#20	#21	#22	#23	Comp
#1	23	0.39	0.00	1.89	2.33	-2.13	0.70	-0.11	0.82	-0.55	0.14	-0.54	-0.80	-0.48	-0.41	-0.57	-0.52	-0.48	-0.26	0.11	1.72	1.02	1.70	-1.35	0.82
#2	11	0.64	-0.18	-1.36	0.55	0.73	0.14	1.14	-0.09	2.45	-0.10	0.50	-0.18	0.55	0.82	-0.64	-1.00	-0.64	-0.27	0.64	1.95	0.95	0.55	0.27	0.73
#3	14	1.30	0.31	-1.81	0.31	-1.00	-0.38	1.19	-0.57	-1.69	-1.38	-0.70	-1.43	0.14	0.38	-1.06	-0.75	-0.90	-0.38	0.44	1.19	0.69	0.50	0.50	0.82
#4	28	1.54	0.21	-0.25	0.71	1.02	-0.57	0.65	-0.29	-1.11	-0.82	0.25	0.29	-0.11	-0.11	-0.64	0.75	-0.36	-0.32	0.46	-0.91	-1.07	0.86	0.43	0.59
#5	7	1.00	1.64	-1.50	0.71	1.43	-0.71	2.00	0.57	-0.50	-0.36	0.43	0.14	-1.17	-0.17	-0.14	1.00	0.43	0.86	1.07	1.29	0.71	1.29	1.29	0.91
#6	18	-0.06	-0.69	0.28	0.56	2.85	-0.11	0.22	0.50	-0.78	-1.28	0.50	0.29	0.25	-1.06	-0.33	-0.28	-1.11	0.15	-0.39	0.72	1.33	0.72	0.00	0.57
#7	11	0.23	0.59	1.05	1.50	0.73	-0.23	0.14	0.27	-0.95	-0.77	-0.55	-1.55	-0.05	0.27	-0.27	-0.73	-0.18	-0.41	0.45	0.50	0.86	0.23	0.36	0.51
#8	18	0.67	0.28	0.94	-0.06	1.15	-0.33	-0.17	0.61	-1.78	-0.94	-0.72	-1.22	0.00	-0.11	-0.67	0.22	-0.74	-0.39	0.22	2.56	1.94	2.06	1.50	0.83
#9	16	2.07	1.07	0.31	0.43	-0.29	0.29	-0.14	0.14	-0.64	-0.57	-0.71	-0.14	0.71	0.36	0.43	-0.57	-0.71	-0.71	0.14	1.64	0.36	1.36	0.29	0.66
#10	8	1.25	0.25	1.25	1.25	0.00	0.36	0.13	-0.25	-0.93	-0.75	0.75	2.50	0.88	1.00	0.25	0.88	-0.38	-1.50	1.13	1.50	1.64	1.13	1.38	0.89
#11	21	0.67	0.21	0.00	0.67	-1.05	-0.70	0.42	0.17	0.07	0.79	0.67	1.33	-0.57	-0.50	-0.81	0.00	-1.15	0.02	-0.52	0.95	-0.24	-0.48	-1.00	0.53
#12	13	0.81	0.42	-0.12	0.27	0.08	0.65	1.69	-1.04	-0.54	-0.35	-1.62	-1.23	-0.12	0.62	0.17	1.15	0.96	0.38	0.92	0.17	-1.31	-0.35	0.04	0.65
#13	10	0.80	1.40	-0.30	1.20	0.60	-0.40	0.80	0.00	-0.60	-1.20	-0.20	-0.40	0.22	0.10	1.30	0.61	0.30	0.20	0.80	-0.10	0.70	0.70	0.70	0.62
#14	9	-0.39	-1.28	-1.00	-0.28	0.17	-1.17	0.72	1.33	-1.83	-1.50	0.28	-0.22	0.89	0.33	-0.89	-0.22	-0.50	-0.06	0.00	1.61	1.22	1.33	0.11	0.78
#15	26	0.54	0.46	0.00	-0.04	-0.38	-0.27	0.15	-0.04	0.14	0.92	0.58	-0.27	-0.56	0.27	-0.08	0.38	-0.46	-0.23	0.35	1.52	0.32	0.80	0.90	0.41
#16	17	2.18	0.31	-0.50	0.62	1.21	0.56	0.29	0.53	-0.65	0.74	0.97	0.85	-0.26	-0.44	-0.15	0.56	-0.85	0.35	-0.88	1.71	1.32	2.06	1.47	0.84
#17	12	-0.17	0.58	-0.42	0.17	-0.67	0.00	0.25	-0.17	0.08	0.42	0.67	0.83	0.08	1.33	0.00	-0.17	0.00	-0.42	0.00	-0.33	0.00	-0.58	0.42	0.31
#18	8	1.00	0.25	-0.63	0.13	-0.38	-1.25	-1.13	0.13	-2.13	-0.75	-1.50	-0.50	0.13	0.13	-0.63	-0.25	-1.25	-1.50	-0.13	-0.88	-0.38	0.38	-0.75	0.74
#19	19	0.26	0.68	0.21	1.68	1.84	0.03	1.66	1.55	-0.45	-0.76	-0.58	0.29	0.24	-0.39	0.42	0.37	0.47	0.47	1.05	1.53	1.16	1.45	0.66	0.79
#20	29	0.52	0.35	0.31	1.34	1.59	0.74	-0.07	-0.86	0.03	1.21	0.34	1.59	0.29	-0.46	-0.79	-0.31	-0.33	-0.25	-0.19	0.57	0.74	0.79	1.26	0.60
#21	26	0.27	-2.00	-0.56	0.54	-0.12	-1.38	-0.04	-0.88	-2.81	-2.00	-2.65	-2.73	-1.81	-0.77	-2.19	-1.16	-1.71	-2.30	-1.96	1.35	-1.21	1.42	0.88	1.40
#22	22	1.18	-0.50	0.82	0.18	-0.77	0.33	0.77	-0.09	-1.41	-0.50	-0.71	-0.33	1.59	0.95	-0.68	0.29	-0.50	-0.48	2.14	0.76	-0.81	1.23	0.68	0.75
#23	20	1.55	2.47	-0.15	1.15	0.25	-0.18	-0.20	1.90	0.28	1.12	1.05	1.10	0.75	0.53	-0.45	2.10	-1.50	-1.25	-2.30	1.30	0.15	1.05	0.00	0.99
#24	17	0.18	-0.29	-0.50	-0.35	-1.32	-0.47	-0.41	-0.91	-1.71	-0.69	-1.53	-1.82	-1.71	-1.30	-0.50	-1.44	-1.10	-0.75	-1.30	0.56	0.09	-0.62	-0.03	0.75
#25	16	0.44	0.75	0.69	0.19	3.00	-0.56	1.19	-0.69	-2.07	-0.75	1.69	0.06	-0.81	0.56	-0.44	-0.13	-1.13	-0.13	-0.38	1.44	1.06	0.75	-0.63	0.79
#26	9	3.93	0.83	1.13	1.50	1.83	1.39	2.61	1.94	-2.00	-1.11	1.22	2.06	-0.06	0.38	-0.44	-0.83	-1.75	-1.25	-2.50	2.39	2.38	1.61	-1.94	1.63
#27	11	2.82	0.14	0.86	1.82	1.20	-1.00	1.64	-1.23	-0.05	-0.41	0.36	-0.23	0.23	0.77	0.36	0.18	-0.27	-0.18	0.45	1.91	0.82	0.86	-0.09	0.80
#28	22	0.41	0.88	-1.27	0.14	-1.23	-1.62	0.64	0.55	-1.55	-1.60	0.50	-0.59	0.14	0.50	-0.95	0.59	-1.95	-1.86	-0.32	-0.26	-0.41	0.10	1.68	0.86
#29	12	0.17	-0.67	0.08	-0.67	0.17	-0.58	0.00	-1.18	-1.42	-0.33	-1.25	-0.83	-1.08	0.83	-1.58	-0.25	-1.17	-2.00	-0.83	-0.14	-0.67	-0.17	-0.58	0.76
#30	18	0.56	1.78	-0.18	-0.11	1.67	1.44	1.33	0.67	0.24	1.41	0.78	1.17	-0.33	0.00	-0.56	-0.89	-0.72	-0.72	0.17	-0.28	-0.11	0.22	0.71	0.68
#31	11	0.11	-0.30	1.11	0.60	-0.10	0.50	0.20	1.70	0.83	1.06	-0.70	-2.70	0.40	-0.20	0.00	-0.90	-1.30	1.00	-1.20	0.30	-0.67	-0.10	1.40	0.59
#32	25	1.06	1.07	0.76	0.12	-0.50	0.65	0.54	0.86	-0.22	-0.29	0.24	0.38	-0.63	0.13	-0.04	-0.34	-0.36	-0.08	-0.44	1.29	1.00	0.70	0.14	0.52
#33	17	0.35	-0.06	-0.82	0.97	1.56	0.63	0.50	1.31	-1.65	-1.26	0.68	-0.59	1.06	0.56	-0.47	1.56	-0.53	-0.91	-1.06	0.38	-0.21	0.29	0.35	0.70
#34	10	0.30	-0.90	-0.10	0.00	-0.50	-0.80	-1.50	-1.10	0.90	0.50	-0.20	-1.10	0.40	0.40	-0.40	-0.80	0.06	0.20	-0.10	1.00	-0.20	1.30	1.70	0.60
#35	7	3.43	1.36	1.21	0.57	3.00	0.83	0.07	0.21	-1.07	-0.50	-0.64	-0.21	-0.79	-0.33	-1.29	-1.86	-1.43	-0.71	-0.79	0.29	-0.79	-0.36	-0.93	1.02
#36	9	1.28	0.28	0.50	0.61	1.61	0.56	1.06	-1.00	-1.61	-1.44	0.50	0.06	-0.83	0.67	-1.00	-0.72	-1.11	-0.78	-0.67	1.44	0.56	1.56	1.25	0.95
#37	28	0.14	-0.07	-0.25	0.71	-1.18	-0.93	0.11	0.21	-2.07	-1.11	1.29	1.61	-1.07	0.14	-0.32	-0.18	-0.36	-0.54	-0.57	-2.04	-2.29	0.50	-0.21	0.69
#38	31	-0.06	-1.23	-0.39	0.57	-0.29	0.23	-0.50	-2.19	-1.65	-2.00	0.61	1.03	-2.57	0.84	-0.32	1.58	-0.32	-0.84	-1.71	-1.56	-1.06	0.97	0.79	0.92
#39	16	1.38	2.50	0.25	0.40	1.31	-1.14	1.53	0.33	-1.19	-1.56	-0.50	-1.94	0.25	0.50	-0.44	0.79	-1.27	0.38	0.63	1.53	1.38	1.25	0.37	0.98
Item Comp		0.86	0.34	0.02	0.59	0.44	-0.15	0.49	0.09	-0.82	-0.47	-0.02	-0.14	-0.18	0.16	-0.42	-0.03	-0.67	-0.45	-0.19	0.79	0.28	0.75	0.36	

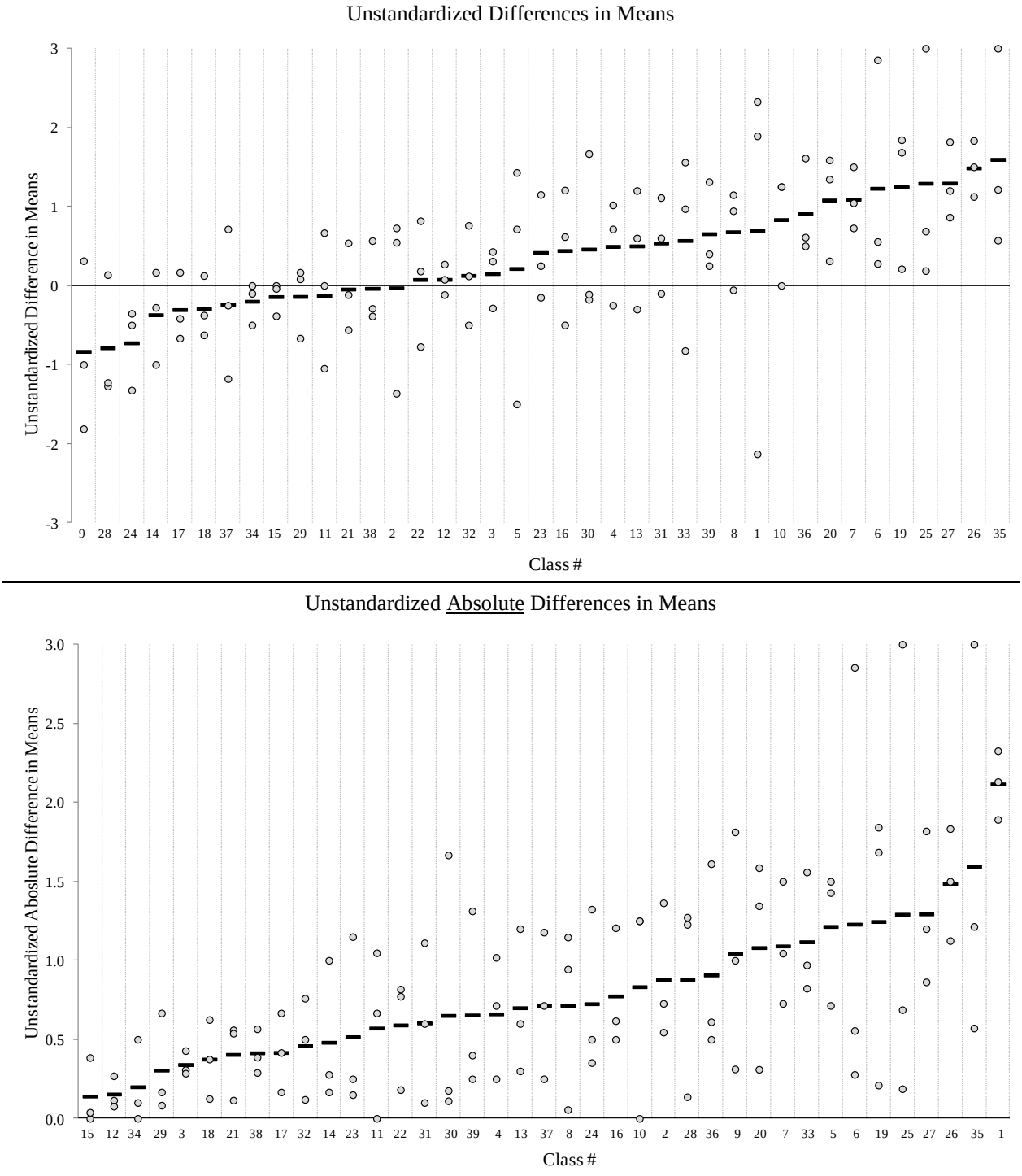
Note. ID is the Class ID #; N is the class size; column headings (#1–#23) correspond to QAS items (see Table 4.1); items grouped according to confirmatory factor analysis of QAS student items in Appendix Table A.6; Item Comp is the item composite unstandardized differences in means; Class Comp is the Class composite unstandardized absolute differences in means; numbers highlighted in **Red** indicate negative unstandardized differences in means; numbers highlighted in **Green** indicate positive unstandardized differences in means.

Appendix Table A.6: *Confirmatory Factor Analysis of QAS Student Ratings of Classroom Practice Items*

QAS Instructional Practice Domains and Items	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Individual Work						
#2 Read a science textbook	0.38					
#6 Work on worksheets	0.60					
#7 Work on homework tasks during class time	0.48					
Interactive Work						
#3 Read science articles in magazines or newspapers		0.58				
#8 Watch science videos, movies, TV shows, etc.		0.50				
#9 Use science-related software or internet resources		0.67				
Discussion						
#4 Discuss science topics with other students in small groups			0.82			
#5 Discuss science topics with other students as a class			0.57			
Reports and Projects						
#10 Work on science projects individually				0.61		
#11 Work on science projects in pairs or groups				0.58		
#12 Work on written science reports				0.72		
#13 Present oral science reports				0.63		
Experimental Work						
#15 Use lab instruments or materials					0.73	
#16 Plan/design experiments or investigations					0.80	
#17 Conduct experiments or investigations					0.78	
#18 Analyze data / relationships between variables					0.72	
#19 Write reports about labs or experiments					0.68	
Assessment						
#20 Review materials to prepare for a test or quiz						0.68
#21 Develop or practice test-taking skills						0.74
#22 Take tests with questions where you choose the answer						0.66
#23 Take tests with questions where you write out the answer						0.61

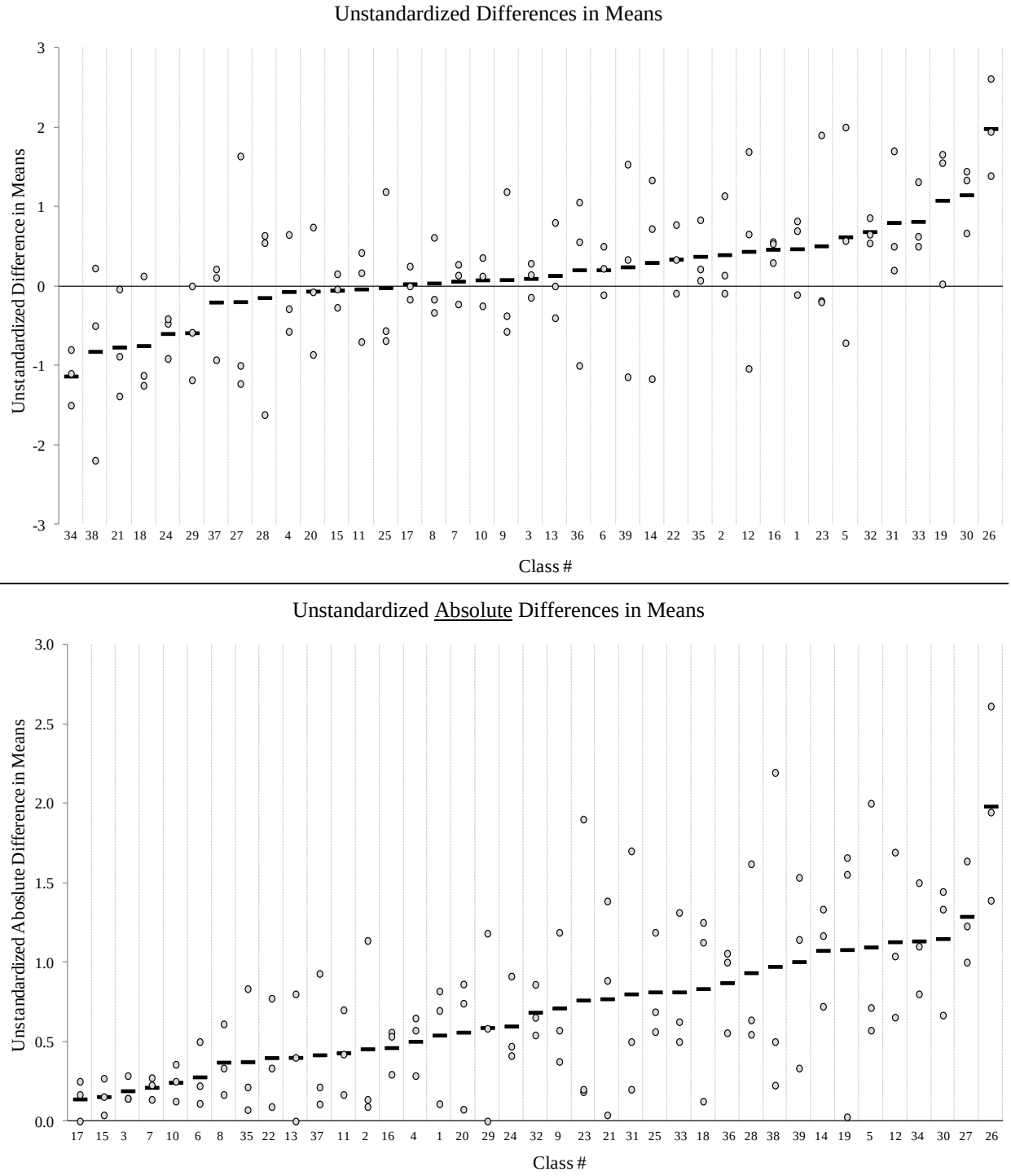
Note. Factor analysis performed using Mplus Version 6 software (Muthén & Muthén, 1998-2010) with STDYX Standardization. $\chi^2 = 592.59(174)$; RMSEA= 0.061; CFI=0.897; TLI=0.875. Item #1 *listen to lectures or instruction directed by the teacher* and Item #14 *watch teacher demonstrate experiment or investigation* omitted from analysis.

Appendix Figure A.1: *Unstandardized Differences in Means and Unstandardized Absolute Differences in Means, by Classroom, Items #2, #6, and #7 in Individual Work Domain*



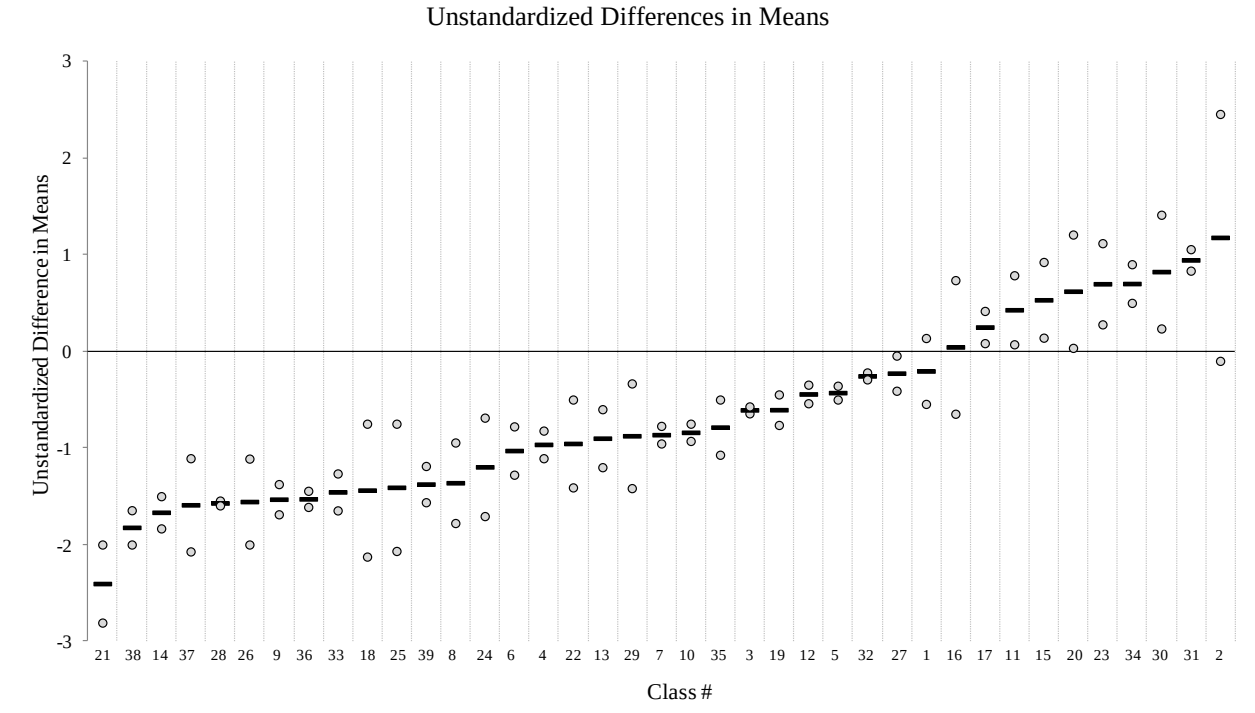
Note. Classrooms in top plot sorted by average unstandardized difference in means on Individual Work domain (represented by the dashes); classrooms in bottom plot sorted by average unstandardized absolute difference in means on Individual Work domain (represented by the dashes).

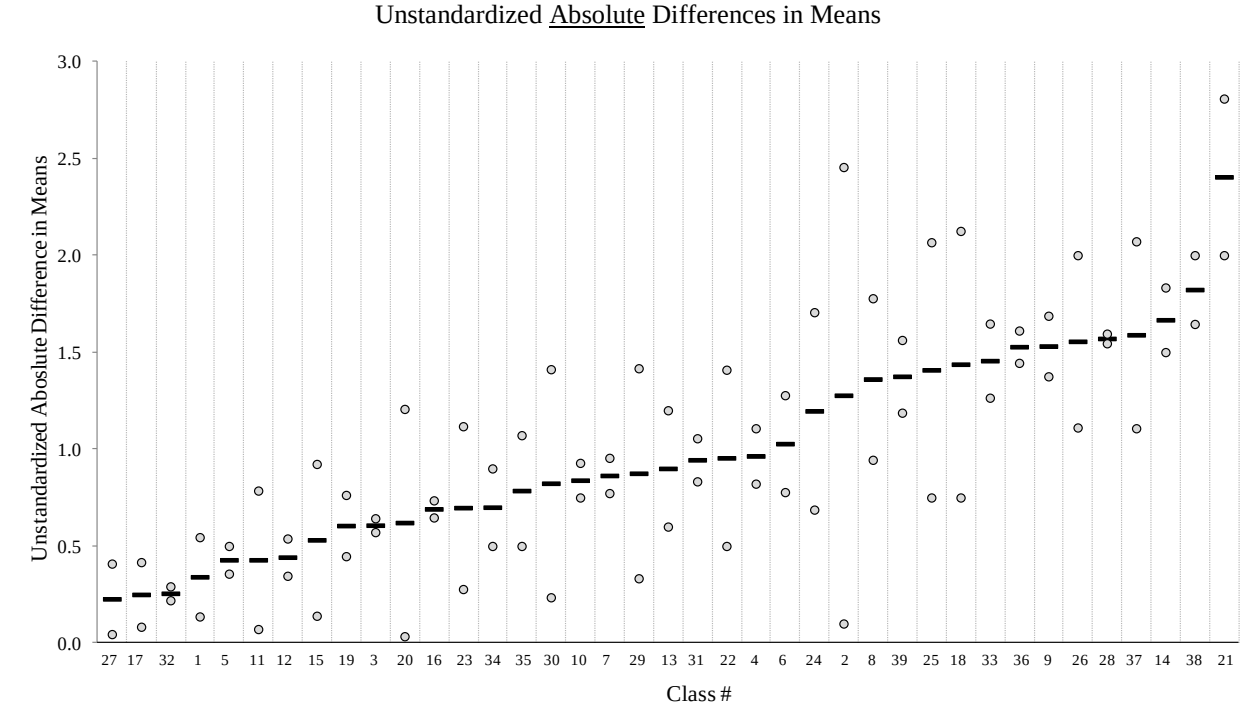
Appendix Figure A.2: *Unstandardized Differences in Means and Unstandardized Absolute Differences in Means, by Classroom, Items #3, #8, and #9 in Interactive Work Domain*



Note. Classrooms in top plot sorted by average unstandardized difference in means on Interactive Work domain (represented by the dashes); classrooms in bottom plot sorted by average unstandardized absolute difference in means on Interactive Work domain (represented by the dashes).

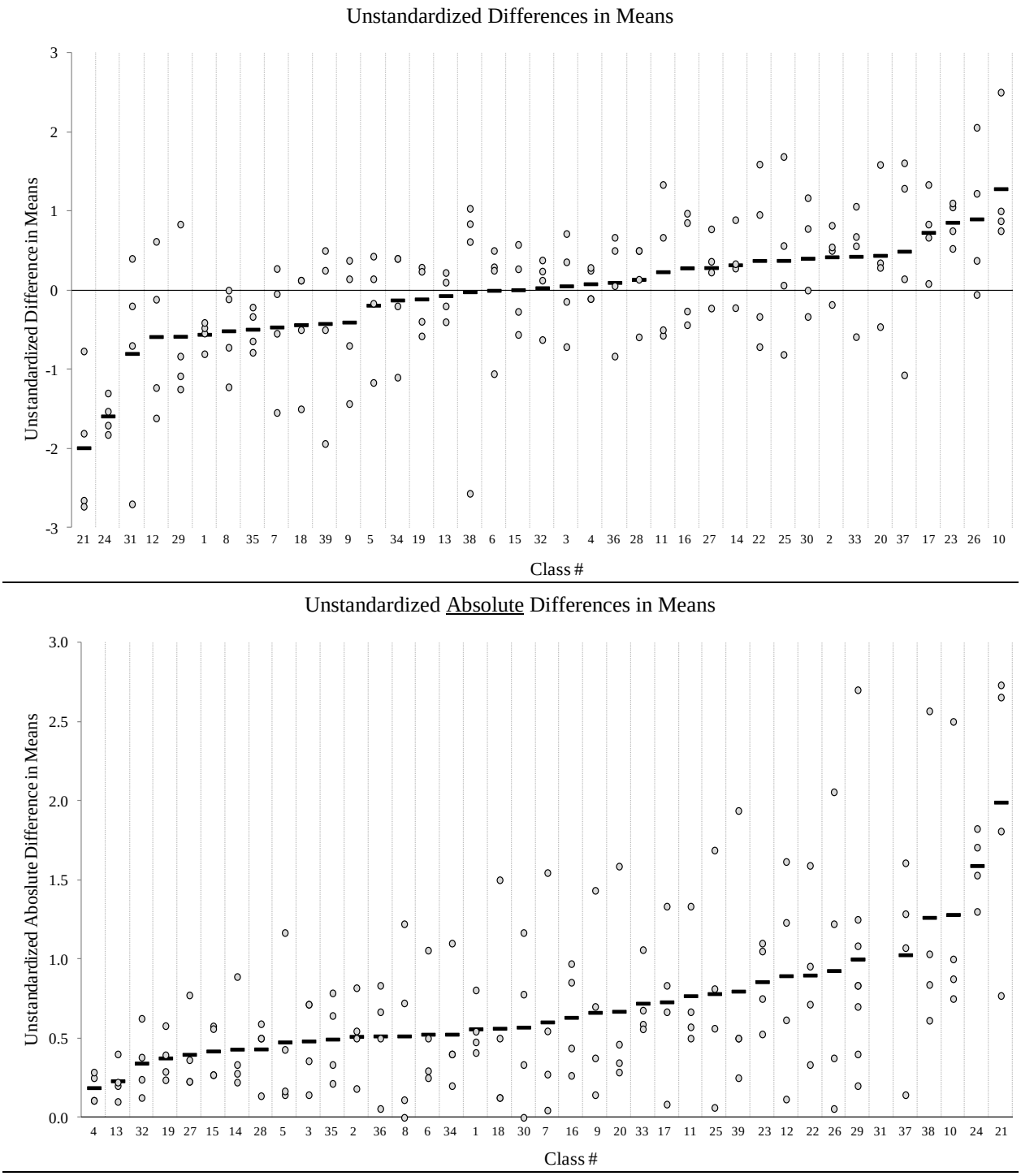
Appendix Figure A.3: *Unstandardized Differences in Means and Unstandardized Absolute Differences in Means, by Classroom, Items #4 and #5 in Discussion Domain*





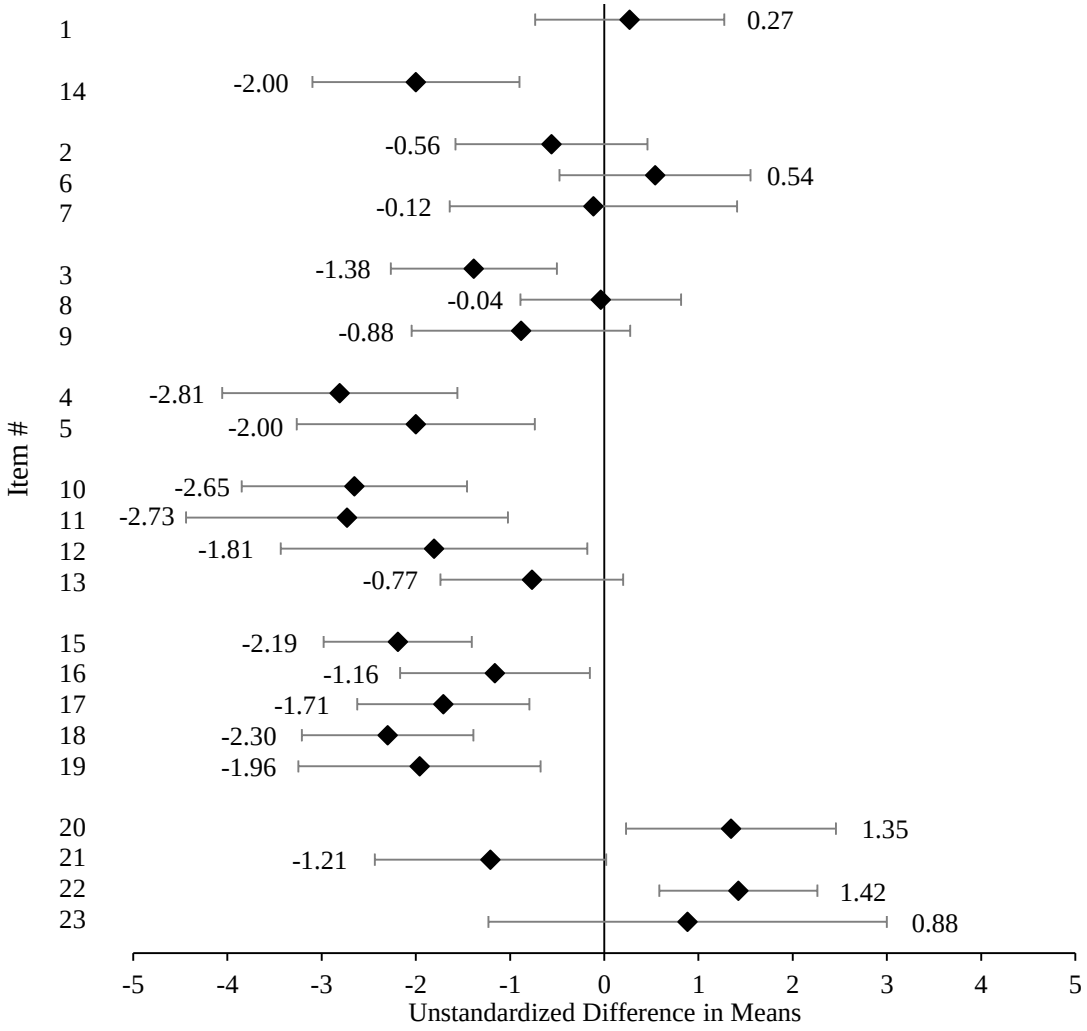
Note. Classrooms in top plot sorted by average unstandardized difference in means on Discussion domain (represented by the dashes); classrooms in bottom plot sorted by average unstandardized absolute difference in means on Discussion domain (represented by the dashes).

Appendix Figure A.4: *Unstandardized Differences in Means and Unstandardized Absolute Differences in Means, by Classroom, Items #10–#13 in Reports and Projects Domain*



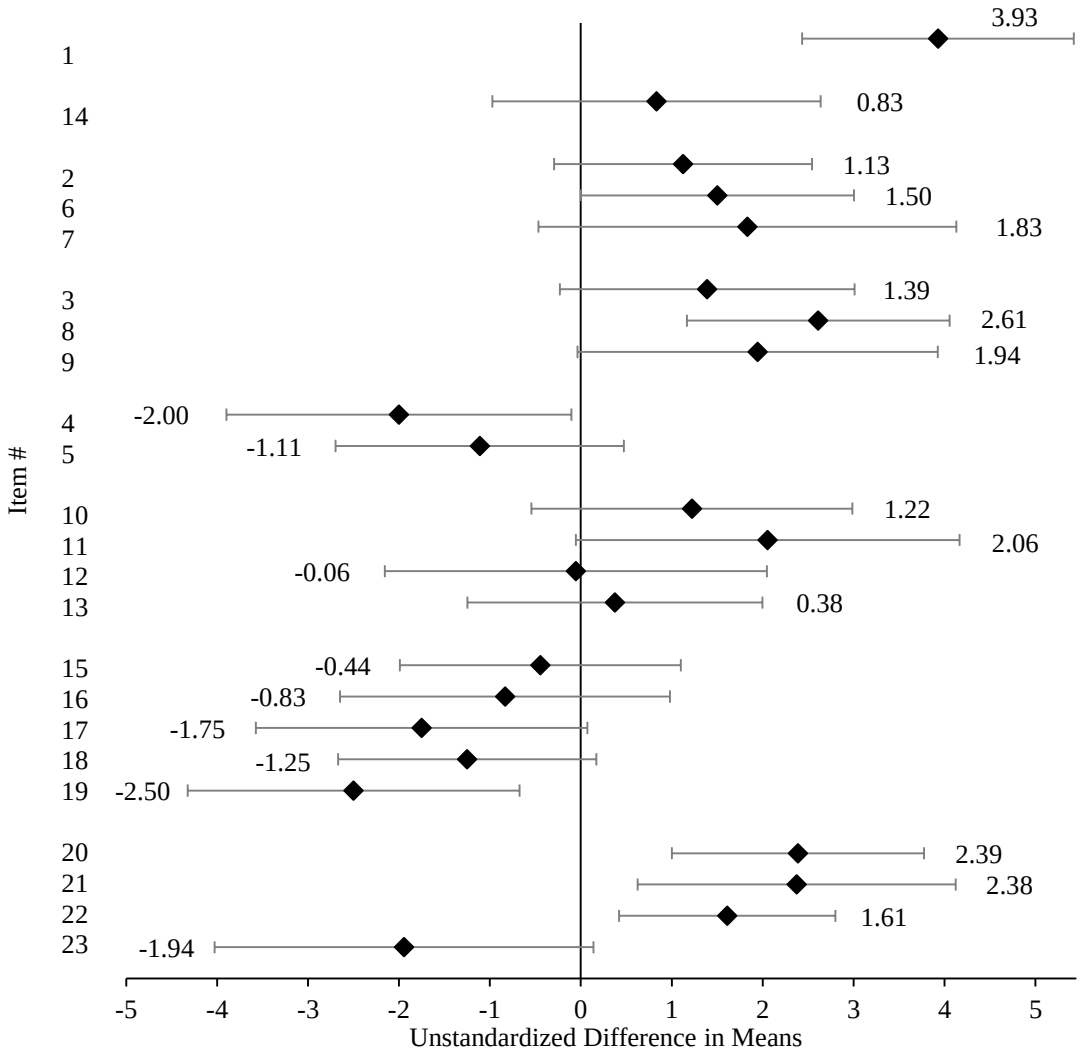
Note. Classrooms in top plot sorted by average unstandardized difference in means on Reports and Projects domain (represented by the dashes); classrooms in bottom plot sorted by average unstandardized absolute difference in means on Reports and Projects domain (represented by the dashes).

Appendix Figure A.5: *Forest Plot of Unstandardized Differences in Means, by Item, Class #21*



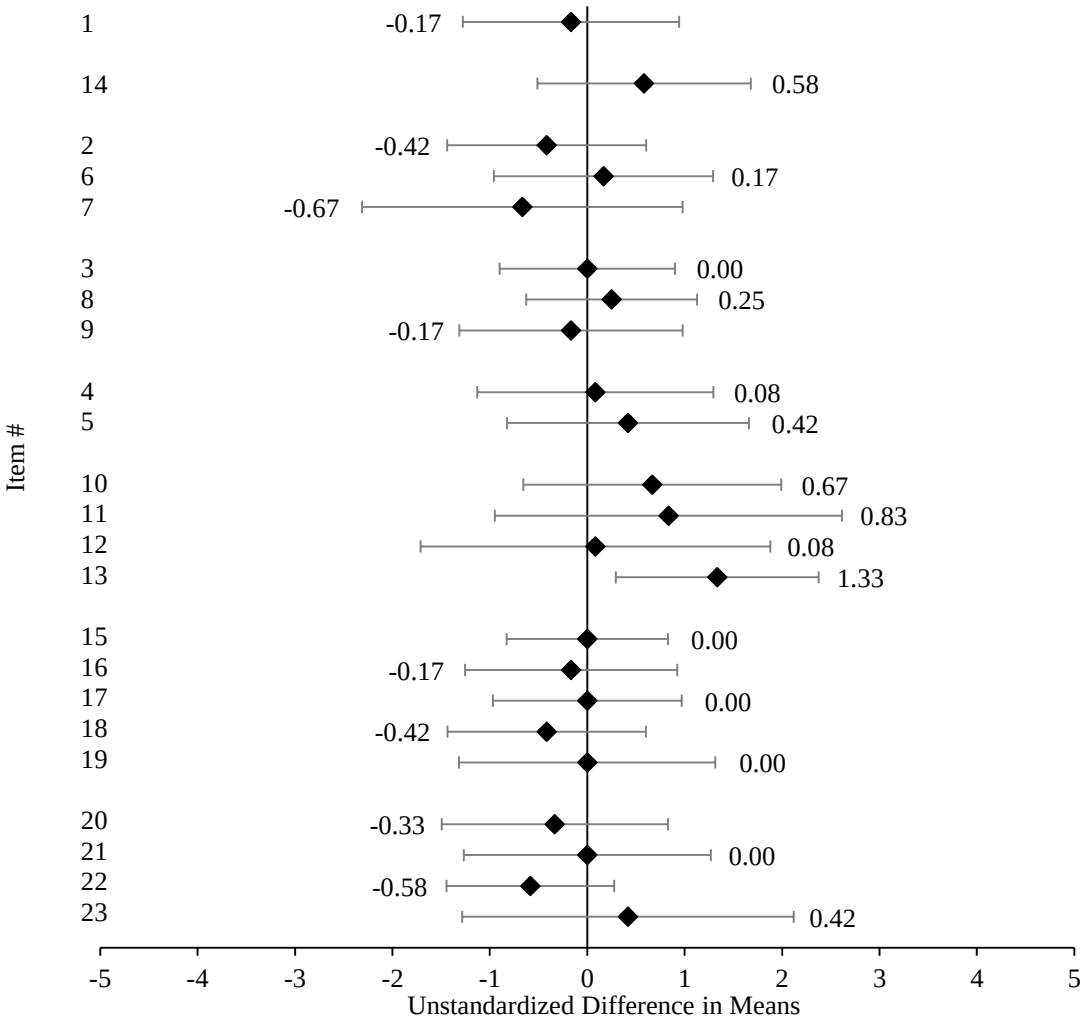
Note. Intervals displayed are the 95% confidence intervals of each unstandardized difference in means estimate.

Appendix Figure A.6: *Forest Plot of Unstandardized Differences in Means, by Item, Class #26*



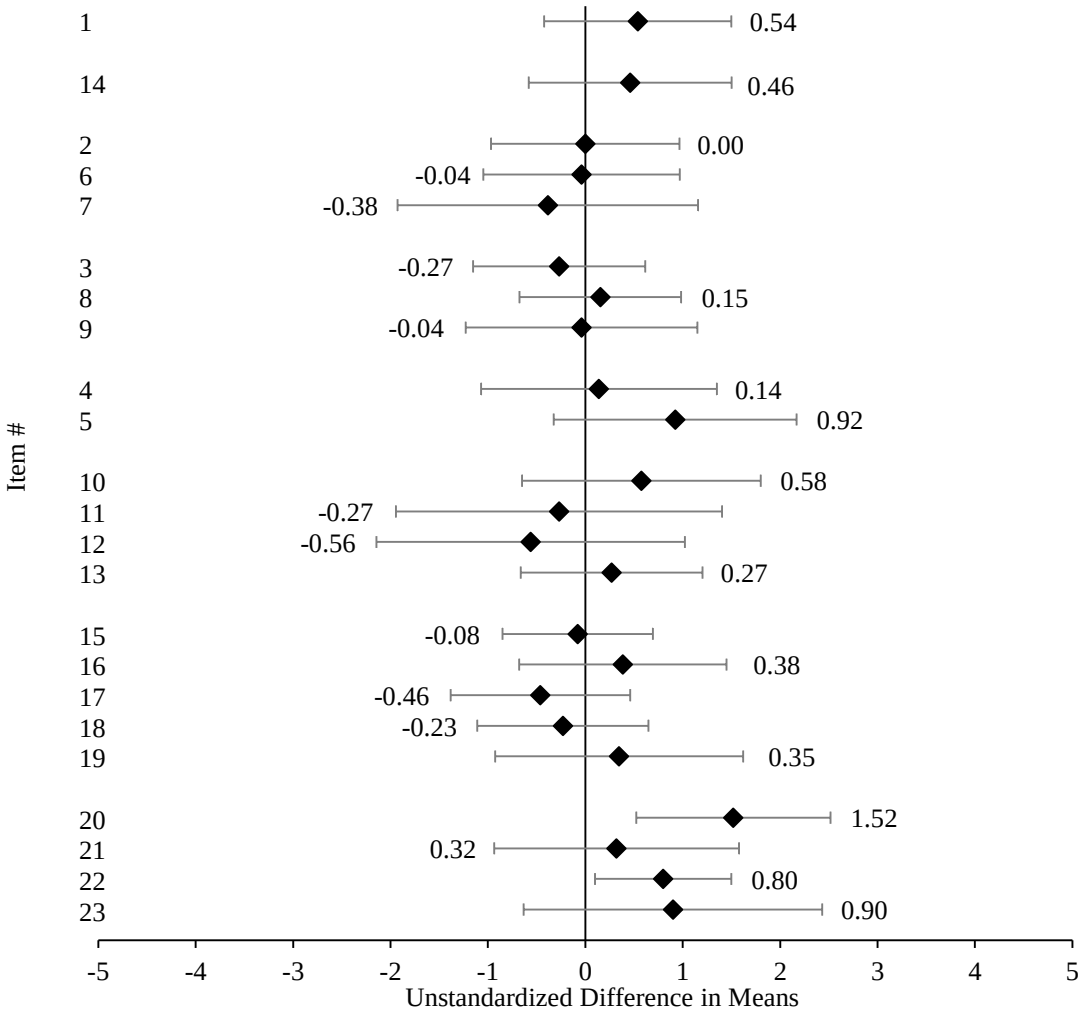
Note. Intervals displayed are the 95% confidence intervals of each unstandardized difference in means estimate.

Appendix Figure A.7: *Forest Plot of Unstandardized Differences in Means, by Item, Class #17*



Note. Intervals displayed are the 95% confidence intervals of each unstandardized difference in means estimate.

Appendix Figure A.8: *Forest Plot of Unstandardized Differences in Means, by Item, Class #15*



Note. Intervals displayed are the 95% confidence intervals of each unstandardized difference in means estimate.

Appendix Table A.7: *Proportion of Students Within a Classroom Who Rated QAS Item as Occurring Less Frequently than Their Teacher's Post-Survey Rating (%)*

		Individual Work					Interactive Work			Discussion		Reports & Projects				Experimental Work					Assessments				
ID	N	#1	#14	#2	#6	#7	#3	#8	#9	#4	#5	#10	#11	#12	#13	#15	#16	#17	#18	#19	#20	#21	#22	#23	**Avg.**
#1	23	4	18	0	0	61	0	13	0	23	23	48	48	67	0	61	57	48	43	48	4	22	0	87	**29**
#2	11	0	36	73	9	18	36	9	73	0	20	9	36	27	0	64	64	55	55	45	0	18	9	18	**29**
#3	14	0	0	0	7	36	7	64	36	29	7	100	21	0	36	0	21	0	29	14	0	7	0	0	**18**
#4	28	0	4	4	0	0	57	0	14	96	50	11	39	71	36	75	18	39	46	29	70	64	0	21	**32**
#5	7	0	0	83	0	0	43	0	14	43	43	14	14	83	50	14	0	14	0	14	0	0	0	0	**19**
#6	18	17	69	17	6	0	0	17	11	50	72	38	18	25	72	44	44	61	29	33	0	17	6	33	**29**
#7	11	18	9	0	0	18	36	9	18	27	18	64	45	18	9	27	55	55	27	18	9	9	0	0	**21**
#8	18	11	17	6	28	0	22	17	6	67	61	11	44	0	17	22	22	29	39	11	0	0	6	0	**19**
#9	16	0	44	88	0	38	13	0	64	88	81	33	87	0	0	69	63	67	56	25	6	25	6	0	**37**
#10	8	25	13	25	13	50	0	38	38	71	63	13	13	25	38	13	13	63	75	0	0	14	0	0	**26**
#11	21	5	33	19	5	48	45	16	38	14	10	19	0	62	67	62	38	65	19	43	5	48	48	67	**34**
#12	13	0	0	15	23	31	38	0	46	23	31	92	77	38	8	0	0	0	31	17	17	77	62	85	**31**
#13	10	0	0	30	0	20	90	30	50	50	60	60	60	33	30	0	0	0	0	0	10	10	10	0	**24**
#14	9	78	100	67	56	22	56	0	0	100	78	44	33	0	11	78	44	22	11	0	0	11	0	22	**36**
#15	26	4	0	27	27	50	54	12	62	8	4	19	15	72	35	4	8	23	31	27	0	32	0	4	**22**
#16	17	12	25	41	6	35	29	35	33	47	24	12	24	71	56	18	29	47	29	71	6	0	0	18	**29**
#17	12	8	0	50	0	42	9	25	75	0	0	17	25	33	0	0	25	0	33	8	50	8	75	42	**23**
#18	8	0	50	63	38	50	75	75	25	75	38	63	25	38	50	38	38	63	75	25	50	63	50	63	**49**
#19	19	53	26	63	0	0	5	0	0	58	68	47	53	0	21	11	26	16	21	11	0	11	0	0	**21**
#20	29	3	52	17	3	10	31	67	76	48	24	31	10	32	73	66	31	34	32	28	25	28	10	7	**32**
#21	26	15	85	48	4	46	85	31	69	96	85	81	88	88	58	88	80	79	96	88	15	54	15		**63**
#22	22	5	50	23	18	55	19	14	45	73	50	57	43	0	0	50	29	45	48	0	5	52	0	9	**30**
#23	20	5	0	60	0	15	58	20	20	33	18	5	15	0	0	45	0	70	75	85	10	45	0	65	**28**
#24	17	18	47	0	47	65	59	47	47	76	69	76	82	65	73	25	88	53	31	93	6	18	41	12	**49**
#25	16	6	0	6	44	0		6	63	93	56	0	0	44	6	50	0	81	38	19	6	0	0	0	**24**
#26	9	0	11	13	22	22	11	0	0	78	67	0	0	0	50	56	67	100	75	100	0	0	0	100	**34**
#27	11	0	27	27	0	60	73	0	82	45	27	36	45	18	27	0	36	9	45	0	9	55	0	55	**29**
#28	22	9	5	59	23	82	90	18	23	77	86	50	91	50	0	45	18	77	73	0	29	50	29	0	**43**
#29	12	50	75	25	42	42	50	25	64	92	67	75	50	50	8	92	33	67	100	58	27	50	25	58	**53**
#30	18	6	0	29	28	17	6	6	0	18	6	6	28	56	28	56	56	56	56	44	50	61	22	12	**28**
#31	11	11	40	0	0	50	0	20	0	44	0	80	100	30	80	10	50	70	22	70	30	89	50	0	**37**
#32	25	4	4	0	12	28	0	4	36	20	21	32	0	58	0	36	64	44	36	44	0	13	4	8	**20**
#33	17	18	50	65	12	0	0	6	0	65	94	0	76	0	0	47	18	41	76	59	12	47	24	0	**31**
#34	10	10	60	50	50	60	90	90	60	10	20	20	70	30	0	40	50	33	60	60	20	40	10	0	**41**
#35	7	0	0	14	29	0	33	43	14	43	29	14	0	14	33	86	100	100	86	100	43	43	29	43	**39**
#36	9	0	22	0	11	22	33	11	67	56	78	11	22	67	33	67	78	89	56	44	0	33	0	38	**36**
#37	28	7	36	36	4	89	71	14	29	86	46	0	0	93	7	36	54	36	36	43	85	79	4	75	**42**
#38	31	29	65	32	3	45	42	53	87	74	73	16	52	90	0	39	10	42	61	68	61	48	6	65	**46**
#39	16	0	0	0	20	6	86	0	60	81	88	67	94	56	0	63	14	80	25	38	0	6	0	40	**36**
645	**All**	**11**	**29**	**29**	**13**	**33**	**39**	**21**	**39**	**55**	**45**	**34**	**39**	**44**	**25**	**43**	**36**	**47**	**46**	**39**	**19**	**34**	**12**	**28**	**33**

Note. ID is the Class ID #; N is the class size; column headings (#1–#23) correspond to QAS items (see Table 4.1); items grouped according to confirmatory factor analysis of QAS student items in Appendix Table A.6; Avg. is the average class proportion across items; numbers highlighted in **Red** indicate greater proportion of students; boxed percentages referred to in text.

Appendix Table A.8: *Proportion of Students Within a Classroom Who Rated QAS Item as Occurring as Frequently as Their Teacher's Post-Survey Rating (%)*

				Individual Work			Interactive Work			Discussion		Reports & Projects				Experimental Work					Assessments				
ID	N	#1	#14	#2	#6	#7	#3	#8	#9	#4	#5	#10	#11	#12	#13	#15	#16	#17	#18	#19	#20	#21	#22	#23	Avg.
#1	23	52	45	0	0	22	65	52	41	41	36	22	17	14	95	26	22	30	35	35	26	26	4	9	**31**
#2	11	36	36	27	36	0	27	18	9	0	60	9	45	9	55	27	36	36	18	27	0	27	36	36	**27**
#3	14	0	7	69	93	57	43	36	64	64	50	0	71	36	43	14	57	71	64	57	36	64	7	71	**47**
#4	28	11	36	43	21	11	36	15	50	4	50	61	7	29	21	25	50	39	32	14	30	32	36	29	**30**
#5	7	0	0	17	29	14	57	14	43	14	0	14	0	0	33	86	43	29	43	14	14	57	14	14	**24**
#6	18	17	25	33	33	18	56	22	44	28	28	31	35	19	17	28	33	28	59	39	33	6	33	28	**30**
#7	11	9	27	9	9	18	18	45	36	36	36	27	45	36	55	64	36	9	18	27	36	18	55	0	**29**
#8	18	11	39	6	44	24	44	28	22	28	17	39	17	12	17	50	28	53	33	22	0	6	6	17	**24**
#9	16	7	31	6	25	44	88	19	21	13	19	53	13	50	69	31	25	33	19	31	31	31	44	63	**33**
#10	8	0	50	25	25	0	43	13	38	29	38	38	0	38	0	25	0	13	25	38	0	0	38	50	**23**
#11	21	24	14	43	24	33	15	42	19	33	19	29	5	19	19	10	33	30	33	38	24	24	33	33	**26**
#12	13	8	17	38	54	23	23	8	38	31	23	8	23	54	46	25	8	31	31	8	8	23	38	15	**25**
#13	10	20	0	50	20	0	0	30	40	40	30	20	10	22	20	0	11	10	10	30	50	30	30	30	**22**
#14	9	22	0	33	44	56	33	11	11	0	22	33	56	33	44	22	33	56	33	25	11	11	11	44	**28**
#15	26	31	44	38	50	23	27	69	12	52	12	27	58	12	58	62	31	50	23	19	12	28	44	12	**34**
#16	17	12	19	35	24	24	47	12	7	41	29	24	29	12	38	41	24	29	12	6	0	24	12	12	**22**
#17	12	67	50	42	50	58	82	75	17	92	58	58	67	42	8	50	67	50	58	83	25	50	17	42	**52**
#18	8	0	25	38	50	38	25	25	25	25	63	25	25	25	13	38	0	13	25	0	50	25	13	25	**26**
#19	19	47	16	32	16	5	68	5	0	37	26	53	16	47	53	47	32	42	21	21	21	21	0	21	**28**
#20	29	38	7	45	55	34	31	22	21	38	31	34	21	21	8	17	10	28	21	24	29	21	34	3	**26**
#21	26	27	8	28	23	38	8	50	12	4	15	12	8	4	27	12	20	21	0	8	4	17	8		**16**
#22	22	23	36	9	45	36	38	18	18	18	32	29	29	23	64	41	29	36	33	14	38	24	23	27	**30**
#23	20	5	5	25	35	10	37	60	15	28	0	45	45	50	74	40	25	25	20	15	10	10	40	15	**28**
#24	17	35	29	100	29	35	29	41	24	24	25	12	18	24	13	50	13	40	56	7	29	41	35	53	**33**
#25	16	44	38	25	38	0		0	25	7	13	19	19	44	38	38	38	19	19	56	25	38	56	69	**30**
#26	9	0	22	13	0	22	44	11	22	22	33	33	0	0	25	11	11	0	25	0	0	13	0	0	**13**
#27	11	0	9	27	27	10	18	9	9	27	9	27	36	18	27	73	9	91	27	64	9	18	18	18	**25**
#28	22	36	19	27	36	9	5	45	32	23	14	9	5	18	59	32	36	14	18	27	24	18	29	5	**23**
#29	12	0	8	50	42	17	25	33	36	8	17	17	50	25	8	8	50	33	0	33	45	25	33	33	**26**
#30	18	22	28	29	28	39	17	11	17	18	18	39	22	39	28	22	28	28	28	39	33	22	22	41	**27**
#31	11	33	10	44	30	20	70	50	50	22	11	20	0	50	10	80	30	20	11	30	20	0	50	0	**29**
#32	25	36	52	44	60	40	78	50	28	4	33	32	40	25	96	36	16	36	32	36	13	35	16	60	**39**
#33	17	18	38	29	41	0	63	24	19	35	6	29	24	35	18	41	12	47	24	29	24	41	35	29	**29**
#34	10	30	10	50	40	10	0	0	20	60	10	70	30	30	70	30	40	33	10	20	30	20	20	20	**28**
#35	7	14	29	14	14	0	0	43	29	43	29	14	14	29	17	14	0	0	14	0	14	14	29	43	**18**
#36	9	11	11	22	56	11	22	0	22	44	22	22	11	22	11	22	11	11	44	44	11	56	11	13	**22**
#37	28	21	25	46	11	11	21	54	36	7	36	25	25	4	75	43	18	54	54	43	15	7	71	7	**31**
#38	31	32	23	42	57	13	29	27	6	16	17	42	42	10	58	42	13	45	35	26	13	10	45	16	**29**
#39	16	25	0	56	20	19	7	20	33	19	13	27	6	25	50	31	29	20	38	25	20	13	6	40	**24**
645	**All**	**24**	**24**	**35**	**35**	**22**	**36**	**31**	**25**	**26**	**25**	**30**	**26**	**24**	**42**	**34**	**26**	**34**	**29**	**28**	**21**	**24**	**29**	**26**	**28**

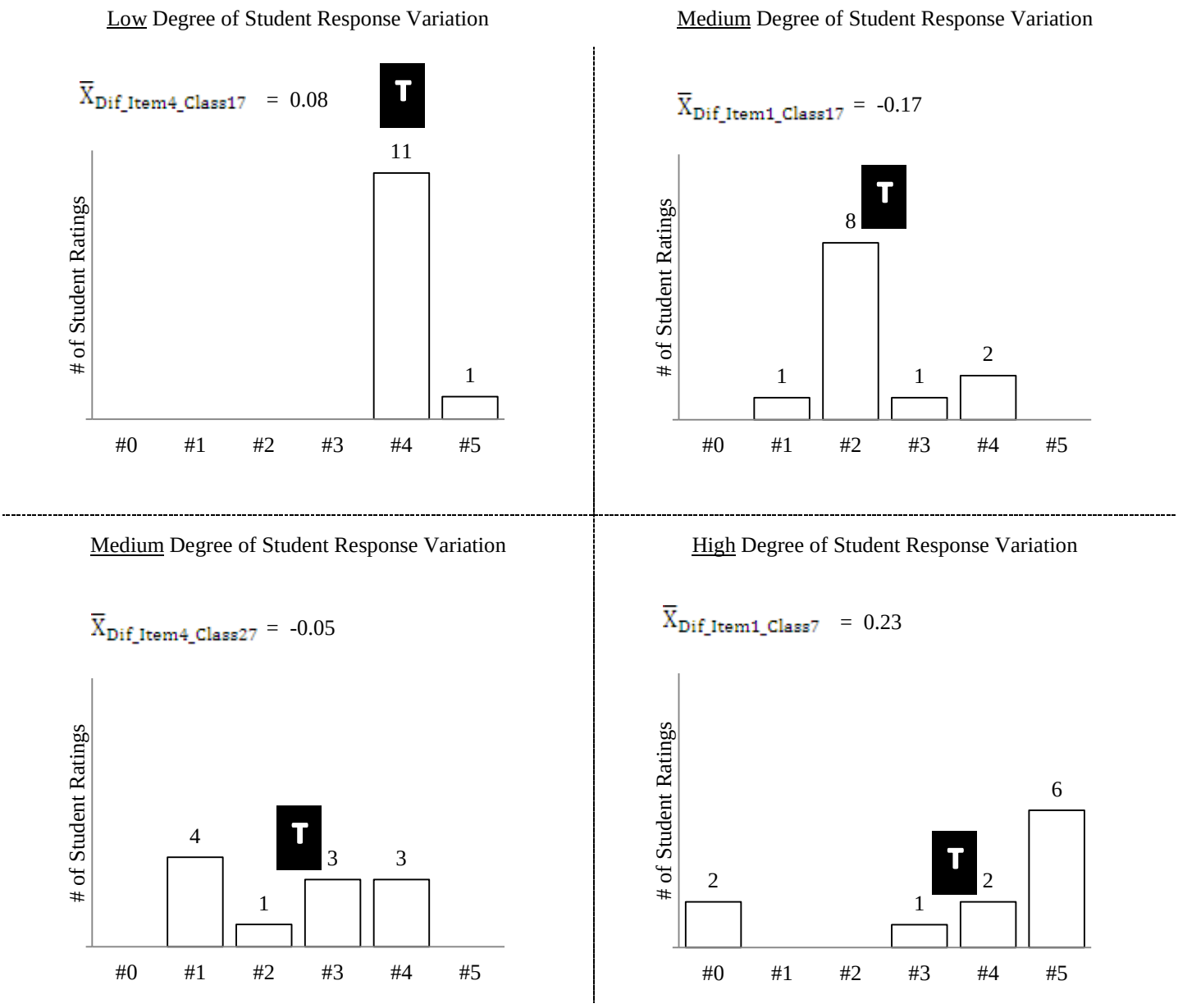
Note. ID is the Class ID #; N is the class size; column headings (#1–#23) correspond to QAS items (see Table 4.1); items grouped according to confirmatory factor analysis of QAS student items in Appendix Table A.6; Avg. is the average class teacher–student agreement across items; numbers highlighted in **Red** indicate smaller proportion of students; numbers highlighted in **Green** indicate greater proportion; boxed percentages referred to in text.

Appendix Table A.9: *Proportion of Students Within a Classroom Who Rated QAS Item as Occurring More Frequently than Their Teacher's Post-Survey Rating (%)*

				Individual Work			Interactive Work			Discussion		Reports & Projects				Experimental Work					Assessments				
ID	N	#1	#14	#2	#6	#7	#3	#8	#9	#4	#5	#10	#11	#12	#13	#15	#16	#17	#18	#19	#20	#21	#22	#23	Avg
#1	23	43	36	100	100	17	35	35	59	36	41	30	35	19	5	13	22	22	22	17	70	52	96	4	**40**
#2	11	64	27	0	55	82	36	73	18	100	20	82	18	64	45	9	0	9	27	27	100	55	55	45	**44**
#3	14	100	93	31	0	7	50	0	0	7	43	0	7	64	21	86	21	29	7	29	64	29	93	29	**35**
#4	28	89	61	54	79	89	7	85	36	0	0	29	54	0	43	0	32	21	21	57	0	4	64	50	**38**
#5	7	100	100	0	71	86	0	86	43	43	57	71	86	17	17	0	57	57	57	71	86	43	86	86	**57**
#6	18	67	6	50	61	82	44	61	44	22	0	31	47	56	11	28	22	11	12	28	67	78	61	39	**40**
#7	11	73	64	91	91	64	45	45	45	36	45	9	9	45	36	9	9	36	55	55	55	73	45	100	**49**
#8	18	78	44	89	28	76	33	56	72	6	22	50	39	88	67	28	50	18	28	67	100	94	88	83	**57**
#9	16	93	25	6	75	19	0	81	14	0	0	13	0	50	31	0	13	0	25	44	63	44	50	38	**30**
#10	8	75	38	50	63	50	57	50	25	0	0	50	88	38	63	63	88	25	0	63	100	86	63	50	**51**
#11	21	71	52	38	71	19	40	42	43	52	71	52	95	19	14	29	29	5	48	19	71	29	19	0	**40**
#12	13	92	83	46	23	46	38	92	15	46	46	0	0	8	46	75	92	69	38	75	75	0	0	0	**44**
#13	10	80	100	20	80	80	10	40	10	10	10	20	30	44	50	100	89	90	90	70	40	60	60	70	**54**
#14	9	0	0	0	0	22	11	89	89	0	0	22	11	67	44	0	22	22	56	75	89	78	89	33	**36**
#15	26	65	56	35	23	27	19	19	27	40	85	54	27	16	8	35	62	27	46	54	88	40	56	84	**43**
#16	17	76	56	24	71	41	24	53	60	12	47	65	47	18	6	41	47	24	59	24	94	76	88	71	**49**
#17	12	25	50	8	50	0	9	0	8	8	42	25	8	25	92	50	8	50	8	8	25	42	8	17	**25**
#18	8	100	25	0	13	13	0	0	50	0	0	13	50	38	38	25	63	25	0	75	0	13	38	13	**26**
#19	19	0	58	5	84	95	26	95	100	5	5	0	32	53	26	42	42	42	58	68	79	68	100	79	**51**
#20	29	59	41	38	41	55	38	11	3	14	45	34	69	46	19	17	59	38	46	48	46	52	55	90	**42**
#21	26	58	8	24	73	15	8	19	19	0	0	8	4	8	15	0	0	0	4	4	81	29	77		**21**
#22	22	73	14	68	36	9	43	68	36	9	18	14	29	77	36	9	43	18	19	86	57	24	77	64	**40**
#23	20	90	95	15	65	75	5	20	65	39	82	50	40	50	26	15	75	5	5	0	80	45	60	20	**44**
#24	17	47	24	0	24	0	12	12	29	0	6	12	0	12	13	25	0	7	13	0	65	41	24	35	**17**
#25	16	50	63	69	19	100		94	13	0	31	81	81	13	56	13	63	0	44	25	69	63	44	31	**46**
#26	9	100	67	75	78	56	44	89	78	0	0	67	100	100	25	33	22	0	0	0	100	88	100	0	**53**
#27	11	100	64	45	73	30	9	91	9	27	64	36	18	64	45	27	55	0	27	36	82	27	82	27	**45**
#28	22	55	76	14	41	9	5	36	45	0	0	41	5	32	41	23	45	9	9	73	48	32	43	95	**34**
#29	12	50	17	25	17	42	25	42	0	0	17	8	0	25	83	0	17	0	0	8	27	25	42	8	**21**
#30	18	72	72	41	44	44	78	83	83	65	76	56	50	6	44	22	17	17	17	17	17	17	56	47	**45**
#31	11	56	50	56	70	30	30	30	50	33	89	0	0	20	10	10	20	10	67	0	50	11	0	100	**34**
#32	25	60	43	56	28	32	22	46	36	76	46	36	60	17	4	28	20	20	32	20	88	52	80	32	**41**
#33	17	65	13	6	47	100	38	71	81	0	0	71	0	65	82	12	71	12	0	12	65	12	41	71	**40**
#34	10	60	30	0	10	30	10	10	20	30	70	10	0	40	30	30	10	33	30	20	50	40	70	80	**31**
#35	7	86	71	71	57	100	67	14	57	14	43	71	86	57	50	0	0	0	0	0	43	43	43	14	**43**
#36	9	89	67	78	33	67	44	89	11	0	0	67	67	11	56	11	11	0	0	11	89	11	89	50	**41**
#37	28	71	39	18	86	0	7	32	36	7	18	75	75	4	18	21	29	11	11	14	0	14	25	18	**27**
#38	31	39	13	26	40	42	29	20	6	10	10	42	6	0	42	19	77	13	3	6	26	42	48	19	**25**
#39	16	75	100	44	60	75	7	80	7	0	0	7	0	19	50	6	57	0	38	38	80	81	94	20	**41**
645	**All**	**65**	**47**	**37**	**52**	**44**	**25**	**48**	**36**	**19**	**29**	**36**	**35**	**32**	**33**	**23**	**38**	**19**	**25**	**34**	**60**	**42**	**59**	**46**	**38**

Note. ID is the Class ID #; N is the class size; column headings (#1–#23) correspond to QAS items (see Table 4.1); items grouped according to confirmatory factor analysis of QAS student items in Appendix Table A.6; Avg. is the average class proportion across items; numbers highlighted in **Red** indicate greater proportion of students; boxed percentages referred to in text.

Appendix Figure A.9: *Classes with Similar Unstandardized Differences in Means on Items #1 and #4 yet Different Underlying Patterns in Student Response Variation*



Note. **T** represents teacher rating.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95–135.

Achenbach, T. M. (2006). As others see us: clinical and research implications of cross-informant correlations for psychopathology. *Current Directions in Psychological Science*, *15*(2), 94–98.

Ahearne, M., Haumann, T., Kraus, F., & Wieseke, J. (2013). It's a matter of congruence: How interpersonal identification between sales managers and salespersons shapes sales success. *Journal of the Academy of Marketing Science*, *41*(6), 625–648.

Aleamoni, L.M. (1981). Student ratings of instruction. In J. Millman (Ed.), *Handbook of teacher evaluation* (p. 110–145). Beverly Hills, CA: Sage.

Aleamoni, L.M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, *13*, 153–166.

Assouline, M., & Meir, E. I. (1987). Meta-analysis of the relationship between congruence and well-being measures. *Journal of Vocational Behavior*, *31*(3), 319–332.

Baker, E. L., Barton, P. B., Darling-Hammond, L. Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., Shepard, L. A. (2010). Problems with the use of student test scores to evaluate teachers. *Economic Policy Institute Briefing Paper*, #278. Washington, DC: Economic Policy Institute. Retrieved August 9, 2014.

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

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

Bao, Y., Dolan, S. L., & Tzafrir, S. S. (2012). Value congruence in organizations: literature review, theoretical perspectives, and future directions. *ESADE Business School Research Paper*, *239*.

Beck, K. H., Hartos, J. L., & Simons-Morton, B. G. (2006). Relation of parent–teen agreement on restrictions to teen risky driving over 9 months. *American Journal of Health Behavior*, *30*, 533–543.

Benton, S. L., & Cashin, W. E. (2012). Student ratings of teaching: A summary of research and literature (IDEA Paper no. 50). Manhattan, KS: The IDEA Center.

Bingham, R. D., Haubrich, P. A., & White, S. B. (1993). Explaining teacher/principal differences in evaluating schools. *Journal of Educational Administration*, *31*(1).

Blazar, D., & Kraft, M. A. (2015). Teacher and teaching effects on students' academic behaviors and mindsets (Working Paper 41). Cambridge, MA: Mathematica Policy Research.

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

Boston, C. (2002). The concept of formative assessment. *Practical Assessment, Research and Evaluation*, *8*, 9. Retrieved from http://PAREonline.net/getvn.asp?v=8&n=9.

Box, G. E., & Draper, N. R. (1987). *Empirical model-building and response surfaces* (Vol. 424). New York: Wiley.

Braun, H., Chudowsky, N., & Koenig, J. A. (2010). Getting value out of value-added: Report of a workshop. Washington, DC: National Academies Press. Retrieved March 3, 2015, from http://www.nap.edu/openbook.php?record_id=12820&page=1.

Brekelmans, M., & Wubbels, T. (1991). Student and teacher perceptions of interpersonal teacher behavior: A Dutch perspective. *The study of learning environments*, *5*(1).

Brookman-Frazee L., Haine, R. A., Gabayan, E. N., & Garland, A. F. (2008). Predicting frequency of treatment visits in community-based youth psychotherapy. *Psychological Services*, *5*, 126–138.

Burstein, L., McDonnell, L. M., Van Winkle, J., Ormseth, T., Mirocha, J., & Guitón, G. (1995). *Validating national curriculum indicators*. Santa Monica, CA: Rand Corporation.

Cable, D. M., & Judge, T. A. (1996). Person–organization fit, job choice decisions, and organizational entry. *Organizational behavior and human decision processes*, *67*(3), 294–311.

Campbell, J., Kyriakides, L., Muijs, D., & Robinson, W. (2004). *Assessing teacher effectiveness: developing a differentiated model*. Psychology Press.

Cantrell, S., & Kane, T. J. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study*. Measures of Effective Teaching (MET) Project. Seattle, WA: Bill & Melinda Gates Foundation.

Chambers, J., De Los Reyes, I. B. & O'Neil, C. (2013) How much are districts spending to implement teacher evaluation systems? Case studies of Hillsborough County Public Schools, Memphis City Schools, and Pittsburgh Public Schools. RAND Education and American Institutes for Research. Working Paper: WR-989-BMGF.

Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). *Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh Public Schools*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. NBER Working Paper No. 17699. *National Bureau of Economic Research*.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). How and why do teacher credentials matter for student achievement? Washington, DC: Urban Institute, National Center for Analysis of Longitudinal Data in Education Research. (ERIC Document Reproduction Service No. ED509655).

Cochran-Smith, M. (2010). Forward. In M. Kennedy (Ed.), *Teacher assessment and the quest for teacher quality*: *A handbook*. San Francisco, CA: Jossey-Bass.

Correnti, R., & Martínez, J. F. (2012). Conceptual, methodological, and policy issues in the study of teaching: Implications for improving instructional practice at scale. *Educational Assessment*, *17*(2–3), 51–61.

Costin, F., Greenough, W. T., & Menges, R. J. (1971). Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research*, *41*(5), 511–535.

Cronbach, L. J. (1954). Report on a psychometric mission to Clinicia. *Psychometrika*, *19*(4), 263–270.

Cronbach, L. J. (1958). Proposals leading to analytic treatment of social perception scores. *Person perception and interpersonal behavior*, *353*, 379.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological bulletin*, *74*(1), 68.

Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological bulletin*, *50*(6), 456.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

De Los Reyes, A. (2011). More than measurement error: Discovering meaning behind informant discrepancies in clinical assessments of children and adolescents. *Journal of Clinical Child & Adolescent Psychology*, *40*(1), 1–9.

De Los Reyes, A., & Kazdin, A. E. (2004). Measuring informant discrepancies in clinical child research. *Psychological assessment*, *16*(3), 330.

De Los Reyes, A., & Kazdin, A. E. (2006). Informant discrepancies in assessing child dysfunction relate to dysfunction within mother–child interactions. *Journal of Child and Family Studies*, 15, 643–661.

De Los Reyes, A., Goodman, K. L., Kliewer, W., & Reid-Quiñones, K. R. (2010). The longitudinal consistency of mother–child reporting discrepancies of parental monitoring and their ability to predict child delinquent behaviors two years later. *Journal of Youth and Adolescence*, *39*, 1417–1430.

Decker, G. (2012, November). Student surveys seen as unlikely evaluations element, for now. *Chalkbeat*. Retrieved from http://ny.chalkbeat.org/2012/11/28/student-surveys-seen-as-unlikely-addition-to-evaluations-for-now/#.Vso3LnQrJhE.

den Brok, P., Bergen, T., & Brekelmans, M. (2006). Convergence and divergence between students' and teachers' perceptions of instructional behaviour in Dutch secondary education. In D. L. Fisher & M. S. Khine (Eds.), *Contemporary approaches to research on learning environments*: *World views* (pp. 125–160). Singapore: World Scientific.

DePascale, C. A. (2012). Managing multiple measures. *Principal*, *91*(5), 6–10.

Derlin, R., & Schneider, G. T. (1994). Understanding job satisfaction principals and teachers, urban and suburban. *Urban Education*, *29*(1), 63–88.

Desimone, L. M. (2006). Consider the source response differences among teachers, principals, and districts on survey questions about their education policy environment. *Educational Policy*, *20*(4), 640–676.

Desimone, L. M., Smith, T. M., & Frisvold, D. E. (2010). Survey measures of classroom instruction comparing student and teacher reports. *Educational Policy*, *24*(2), 267–329.

Doran, H. C., & Lockwood, J. R. (2006). Fitting value-added models in R. *Journal of Educational and Behavioral Statistics*, *31*(2), 205–230.

Edwards, J. R. (1991). *Person-job fit: A conceptual integration, literature review, and methodological critique*. John Wiley & Sons.

Edwards, J. R. (1994). The study of congruence in organizational behavior research: critique and a proposed alternative. *Organizational behavior and human decision processes*, *58*(1), 51–100.

Edwards, J. R. (2001). Ten difference score myths. *Organizational Research Methods*, *4*(3), 265–287.

Edwards, J. R. (2002). Alternatives to difference scores: Polynomial regression and response surface methodology. *Advances in measurement and data analysis*, 350–400.

Ferdinand, R. F., van der Ende, J., & Verhulst, F. C. (2006). Prognostic value of parent–adolescent disagreement in a referred sample. *European child & adolescent psychiatry*, *15*(3), 156–162.

Ferguson, R. (2010). *Student perceptions of teaching effectiveness*. Discussion brief from the National Center for Teacher Effectiveness and the Achievement Gap Initiative, Harvard University, Cambridge, MA.

Follman, J. (1992). Secondary school students' ratings of teacher effectiveness. *The high school journal*, *75*(3), 168–178.

Follman, J. (1995). Elementary public school pupil rating of teacher effectiveness. *Child Study Journal*, *25*(1), 57–78.

Gil D. H. (1987). Instructional evaluation as a feedback process. *Techniques for Evaluating and Improving Instruction*, *1987*(31), 57–64.

Glueck, C. L. (2013). *Measuring parent–teacher expectation congruence and examining student outcomes* (doctoral dissertation). Retrieved from https://getd.libs.uga.edu. University of Georgia.

Glueck, C. L., & Reschly, A. L. (2014). Examining congruence within school–family partnerships: Definition, importance, and current measurement approaches. *Psychology in the Schools*, *51*(3), 296–315.

Goe, L., Bell, C., & Little, O. (2008). Approaches to evaluating teacher effectiveness: a research synthesis. Washington, DC: National Comprehensive Center for Teacher Quality.

Guion, K., Mrug, S., & Windle, M. (2009). Predictive value of informant discrepancies in reports of parenting: relations to early adolescents' adjustment. *Journal of Abnormal Child Psychology, 37*(1), 17–30.

Haertel, E. H. (2013). *Reliability and validity of inferences about teachers based on student test scores.* The 14th William H. Angoff Memorial Lecture. Princeton, NJ: Educational Testing Service.

Hanover Education. (2013, February. Student perception surveys and teacher assessments. Retrieved from https://dese.mo.gov/sites/default/files/Hanover-Research-Student-Surveys.pdf.

Hanushek, E. A., & Rivkin, S. G. (2006). Teacher quality. *Handbook of the Economics of Education, 2,* 1051–1078.

Hanushek, E.A. (2011). Valuing teachers: how much is a good teacher worth. *Education Next*. *11*(3), 41–45.

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

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. *Second handbook of research on mathematics teaching and learning*, *1*, 371–404.

Houseman, G. M. (2007). *Explaining the discrepancy between principals' and teachers' perceptions of the principal leader behavior* (doctoral dissertation). Retrieved from Digital Library and Archives. Virginia Polytechnic Institute and State University.

iKnow My Class Survey (2016). Thousand Oaks, CA: Quaglia School Voice. Retrieved from http://svsurveys.corwin.com/myClass.jsp?loc=US.

Intuneness. (2017). In Merriam-Webster.com. Retrieved from https://www.merriam-webster.com/dictionary/in%20tune.

Israel, P., Thomsen, P. H., Langeveld, J. H., & Stormark, K. M. (2007). Parent–youth discrepancy in the assessment and treatment of youth in usual clinical care setting: consequences to parent involvement. *European Child & Adolescent Psychiatry*, *16*(2), 138–148.

Jakobsson, U., & Westergren, A. (2005). Statistical methods for assessing agreement for ordinal data. *Scandinavian Journal of Caring Sciences*, *19*(4), 427–431.

Jansen, K. J., & Kristof-Brown, A. L. (2005). Marching to the beat of a different drummer: Examining the impact of pacing congruence. *Organizational Behavior and Human Decision Processes*, 97(2), 93–105.

Johns, G. (1981). Difference score measures of organizational behavior variables: A critique. *Organizational Behavior and Human Performance, 27*(3), 443–463.

Kamo, Y. (2000). "He said, she said": Assessing discrepancies in husbands' and wives' reports on the division of household labor. *Social Science Research, 29*(4), 459–476.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Measures of Effective Teaching (MET) Project. Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review, 27*(6), 615–631.

Kane, T., Kerr, K., & Pianta, R. (2014). *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*. New York, NY: John Wiley & Sons.

Kane, T.J. & Cantrell, S. (2010). *Learning about teaching: Initial findings from the measures of effective teaching project*. Measures of Effective Teaching (MET) Project. Seattle, WA: Bill & Melinda Gates Foundation.

Kennedy, M. M. (1999). Approximations to indicators of student outcomes. *Educational Evaluation and Policy Analysis, 21*(4), 345–363.

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9*(3), 231–251.

Kyriakides, L. (2001). Measurement of Teaching in Cyprus: Limitations of current practice. *Proceedings of the 4th Annual Conference of the Cyprus Educational Association.* Nicosia, Cyprus.

Kyriakides, L. (2005). Drawing from teacher effectiveness research and research into teacher interpersonal behaviour to establish a teacher evaluation system: A study on the use of student ratings to evaluate teacher behaviour. *The Journal of Classroom Interaction, 40*(2), 44–66.

Laird, R. D., & Weems, C. F. (2011). The equivalence of regression models using difference scores and models using separate scores for each informant: implications for the study of informant discrepancies. *Psychological Assessment, 23*, 388–397.

Larson, M. D., Norman, S. M., Hughes, L. W., & Avey, J. B. (2013). Psychological capital: a new lens for understanding employee fit and attitudes. *International Journal of Leadership Studies*, *8*(1), 28–43.

Lauver, K. J., & Kristof-Brown, A. (2001). Distinguishing between employees' perceptions of person–job and person–organization fit. *Journal of Vocational Behavior*, *59*(3), 454–470.

Lavigne, A. L., & Good, T. L. (2013). *Teacher and student evaluation: Moving beyond the failure of school reform*. New York, NY: Routledge.

Laurie, J. W. (1966). Convergent job expectations and ratings of industrial foreman. *Journal of Applied Psychology*, 50, 97–101.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766.

Lin, T. J., Lee, M. H., & Tsai, C. C. (2014). The commonalities and dissonances between high-school students' and their science teachers' conceptions of science learning and conceptions of science assessment: a Taiwanese sample study. *International Journal of Science Education*, *36*(3), 382–405.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology*, *34*(2), 120–131.

Markow, D. & Scheer, M. (2003). *The MetLife survey of the American teacher: An examination of school leadership*. New York: Harris Interactive Inc. Retrieved from http://files.eric.ed.gov/fulltext/ED505002.pdf.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, *76*(5), 707.

Marsh, H. W., & Hocevar, D. (1991). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and Teacher Education*, *7*(4), 303–314.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, *47*(2), 106–124.

Martinez, J. F., Borko, H., & Stecher, B. M. (2012). Measuring instructional practice in science using classroom artifacts: Lessons learned from two validation studies. *Journal of Research in Science Teaching*, *49*(1), 38–67.

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis*, *21*(1), 29–45.

McCaffrey, D. F., & Lockwood, J. R. (2011). Missing data in value-added modeling of teacher effects. *Annals of Applied Statistics*, *5*, 773–797.

McCaffrey, J. R., Lockwood, D. F., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value added models for teacher accountability* [Monograph]. Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org/pubs/monographs/2004/RAND_MG158.pdf.

McKeachie W. J. (1987). Can evaluating instruction improve teaching? *Techniques for Evaluating and Improving Instruction, 31*, 3–8.

Measures of Effective Teaching (MET) Project (2012). *Asking students about teaching: student perception surveys and their implementation*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://k12education.gatesfoundation.org/resource/asking-students-about-teaching-student-perception-surveys-and-their-implementation.

Measures of Effective Teaching (MET) Project Q & A (2012). *"They are the experts": a national teacher of the year talks about student surveys*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/Student_Survey_Teacher_QandA.pdf.

Mehrens, W. A. (1990). Combining evaluation data from multiple sources. In J. Millman and L. Darling-Hammond, (Eds.), *Teacher Evaluation* (pp. 322–344). Newbury Park, CA: Sage.

Michalos, A. C. (1986). Job satisfaction, marital satisfaction, and the quality of life: A review and a preview. In F.M. Andrews (Ed.), *Research on the Quality of Life* (pp. 57–83). Ann Arbor: University of Michigan, Survey Research Center.

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A composite estimator of effective teaching*. Measures of Effective Teaching (MET) Project. Seattle, WA: Bill & Melinda Gates Foundation.

Miller, S. A., & Davis, T. L. (1992). Beliefs about children: A comparative study of mothers, teachers, peers, and self. *Child Development*, *63*(5), 1251–1265.

Montgomery, J. L., & Baker, W. (2007). Teacher-written feedback: Student perceptions, teacher self-assessment, and actual teacher performance. *Journal of Second Language Writing*, *16*(2), 82–99.

Moorman, R. H., & Podsakoff, P. M. (1992). A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behaviour research. *Journal of Occupational and Organizational Psychology*, *65*(2), 131–149.

Mount, M. K., & Muchinsky, P. M. (1978). Person–environment congruence and employee job satisfaction: A test of Holland's theory. *Journal of Vocational Behavior*, *13*(1), 84–100.

Murray, H. G. (1997). Does evaluation of teaching lead to improvement of teaching? *The International Journal for Academic Development*, *2*(1), 8–23.

Muthén, L. K., & Muthén, B. O. (1998–2010). Mplus Version 6 [Computer software]. *Los Angeles, CA: Muthén & Muthén*.

My Student Survey (2016). Nashville, Tennessee: STeP survey. Retrieved from http://mystudentsurvey.com/our-surveys/step-survey/#.

Ostroff, C., Shin, Y., & Kinicki, A. J. (2005). Multiple perspectives of congruence: Relationships between value congruence and employee attitudes. *Journal of Organizational Behavior*, *26*(6), 591–623.

Partee, G. L. (2012). *Using multiple evaluation measures to improve teacher effectiveness: State strategies from round 2 of No Child Left Behind Act waivers*. Washington, DC: Center for American Progress. Retrieved from https://www.americanprogress.org/issues/education/reports/2012/12/18/48368/using-multiple-evaluation-measures-to-improve-teacher-effectiveness.

Patel, N., & Stevens, S. (2010). Parent–teacher–student discrepancies in academic ability beliefs: influences on parent involvement. *School Community Journal*, *20*(2), 115–136.

Peterson, K. D. (1987). Teacher evaluation with multiple and variable lines of evidence. *American Educational Research Journal*, *24*(2), 311–317.

Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices*. Thousand Oaks, CA: Corwin Press.

Peterson, K. D., Wahlquist, C., & Bone, K. (2000). Student surveys for school teacher evaluation. *Journal of Personnel Evaluation in Education*, *14*(2), 135–153.

Pham-Gia, T., & Hung, T. L. (2001). The mean and median absolute deviations. *Mathematical and Computer Modeling*, *34*(7), 921–936.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109–119.

Popham, W. J. (2013). *Evaluating America's teachers: Mission possible*? Thousand Oaks, CA: Corwin.

Race to the Top Program: Guidance and Frequently Asked Questions (2010). Washington, D.C.: U.S. Department of Education. Retrieved from https://www2.ed.gov/programs/racetothetop/faq.pdf.

Raudenbush, S. W. (2013). What do we know about using value-added to compare teachers who work in different schools? *Carnegie Foundation for the Advancement of Teaching: Carnegie Knowledge Network Knowledge Brief 10.*

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Raudenbush, S. W., & Jean, M. (2012). How should educators interpret value-added scores? *Carnegie Foundation for the Advancement of Teaching: Carnegie Knowledge Network Knowledge Brief 1*.

Raudenbush, S.W., Bryk, A.S, & Congdon, R. (2004). HLM 6 for Windows [Computer software]. *Lincolnwood, IL: Scientific Software International*.

Renk, K. (2005). Cross-informant ratings of the behavior of children and adolescents: The "gold standard". *Journal of Child and Family Studies*, *14*(4), 457–468.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, *94*(2), 247–252.

Rogers, C. R. (1987). Rogers, Kohut, and Erickson: A personal perspective on some similarities and differences. *The evolution of psychotherapy*, 179–187.

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability the difference score in the measurement of change. *Journal of Educational Measurement*, *20*(4), 335–343.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *92*(3), 726.

Rothstein, J. (2010). Teacher quality in educational production: tracking, decay, and student achievement. *The Quarterly Journal of Economics. 125*(1), 175–214.

Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: Lessons from the study of instructional improvement. *Educational Researcher*, *38*(2), 120–131.

Sanders, W. L., Wright, S. P., & Horn, S. P. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, *11*(1), 57–67.

Sax, G. (1997). *Principles of educational and psychological measurement and evaluation.* Belmont, CA: Wadsworth Publishing Company.

Schramm, W. (1956). Educators and communication research. *Educational Leadership*, *13*(8), 503–509.

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

Schweig, J. (2014b). *Multilevel factor analysis and student ratings of instructional practice* (doctoral dissertation). Retrieved from Escholarship.org. University of California, Los Angeles.

Seidel, T. (2006). The role of student characteristics in studying micro teaching–learning environments. *Learning Environments Research*, *9*(3), 253–271.

Shanock, L. R., Baran, B. E., Gentry, W. A., Pattison, S. C., & Heggestad, E. D. (2010). Polynomial regression with response surface analysis: A powerful approach for examining moderation and overcoming limitations of difference scores. *Journal of Business and Psychology*, *25*(4), 543–554.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard educational review*, *57*(1), 1–23.

Silva, M. C. (2012). The scholarship of teaching as science and as art. *Journal of Nursing Education*, *51*(11), 599–601.

Sirotnik, K. A. (1980). Psychometric implications of the unit-of-analysis problem (with examples from the measurement of organizational climate). *Journal of Educational Measurement, 17*(4), 245–282.

Staiger, D. O., & Kane, T. J. (2014). Making decisions with imprecise performance measures: The relationship between annual student achievement gains and a teacher's career value-added. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 144–169). San Francisco, CA: Jossey-Bass.

Stevens, J. J. (1987). Using student ratings to improve instruction. *Techniques for Evaluating and Improving Instruction, 1987*(31), 33–38.

Stiggins, R. J., & Duke, D. L. (1988). *The case for commitment to teacher growth: Research on teacher evaluation*. Albany: State University of New York Press.

Stone, C. A. (1997). Correspondences among parent, teacher, and student perceptions of adolescents' learning disabilities. *Journal of Learning Disabilities, 30*(6), 660–669.

Stronge, J. H., & Ostrander, L. P. (1997). Client surveys in teacher evaluation. *Evaluating teaching: A guide to current thinking and best practice*, 129–161.

Taut, S., & Sun, Y. (2014). The development and implementation of a national, standards-based, multi-method teacher performance assessment system in Chile. *Education Policy Analysis Archives, 22*(71), 1–30.

Teacher Evaluation 2.0 (2010) (Technical Report). New York, NY: The New Teacher Project. Retrieved from https://tntp.org/assets/documents/Teacher-Evaluation-Oct10F.pdf.

Thorndike, E.L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25–29.

Toch, T., & Rothman, R. (2008). *Rush to judgment: Teacher evaluation in public education*. Washington, DC: Education Sector.

Tripod Student Surveys [tripoded.com], personal email communication, September 4, 2015.

Turban, D. B., & Jones, A. P. (1988). Supervisor–subordinate similarity: types, effects, and mechanisms. *Journal of Applied Psychology, 73*(2), 228.

Virtanen, V., & Lindblom-Ylanne, S. (2010). University students' and teachers' conceptions of teaching and learning in the biosciences. *Instructional Science, 38*(4), 355–370.

Walkington, C., & Marder, M. (2014). Classroom observation and value-added models give complementary information about quality of mathematics teaching. In T. J. Kane, K. A. Kerr, & R.C. Pianta (Eds.) *Designing Teacher Evaluation Systems: New guidance from the Measures of Effective Teaching Project.* (pp. 234–277). San Francisco: Josey-Bass.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. New York, NY: The New Teacher Project.

Wilkerson, D. J., Manatt, R. P., Rogers, M. A., & Maughan, R. (2000). Validation of student, principal, and self-ratings in 360 feedback for teacher evaluation. *Journal of Personnel Evaluation in Education*, *14*(2), 179–192.

Worrell, F. C., & Kuterbach, L. D. (2001). The use of student ratings of teacher behaviors with academically talented high school students. *Journal of Secondary Gifted Education*, *14*(4), 236–247.

YouthTruth Student Survey: Design & Methodology (2016). San Francisco, CA: YouthTruth Student Survey. Retrieved from http://www.youthtruthsurvey.org/wp-content/uploads/2016/01/YouthTruth-Design-and-Methodology-Report-2016.pdf.

Zyphur, M. J., Zammuto, R. F., & Zhang, Z. (2016). Multilevel latent polynomial regression for modeling (in) congruence across organizational groups: The case of organizational culture research. *Organizational Research Methods*, *19*(1), 53–79.