Running Head: "UTEACH OBSERVATION PROTOCOL"

Development of the UTeach Observation Protocol: A Classroom Observation Instrument to Evaluate Mathematics and Science Teachers from the UTeach Preparation Program

Candace Walkington
University of Texas at Austin
Department of Curriculum and Instruction
1 University Station D5700
Austin, TX 78712, USA
c.walkington1@gmail.com

Prerna Arora
University of Texas at Austin
Department of Educational Psychology
1 University Station D5800
Austin, TX 78712, USA
arorap1@yahoo.com

Shasta Ihorn
University of Texas at Austin
Department of Educational Psychology
1 University Station D5800
Austin, TX 78712, USA
shasta.ihorn@gmail.com

Jessica Gordon
University of Texas at Austin
Department of Curriculum and Instruction
1 University Station D5700
Austin, TX 78712, USA
jdgordon@jsg.utexas.edu

Mary Walker
University of Texas at Austin
UTeach Natural Sciences
1 University Station G2550
Austin, TX 78712, USA
mwalker@austin.utexas.edu

Larry Abraham
University of Texas at Austin
Department of Curriculum and Instruction
1 University Station D5700
Austin, TX 78712, USA
l.abraham@mail.utexas.edu

Michael Marder
University of Texas at Austin
Department of Physics
1 University Station C1600
Austin, TX 78712, USA
marder@mail.utexas.edu

Abstract

The UTeach secondary mathematics and science teacher preparation program has recently undergone national replication based on its ability to increase recruitment and retention of mathematics and science teachers, of which there is a national shortage. The Noyce Scholarship program represents a primary effort of the National Science Foundation to increase the number of these teachers. However, despite the national scope of these efforts, there is no systematic investigation of the qualities of their graduates. We developed a classroom observation instrument, the UTeach Observation Protocol (UTOP), and conducted observations of novice secondary mathematics and science teachers. We discuss the development of this instrument and its structure and reliability based on these observations. We present preliminary findings comparing classroom observation ratings of graduates of UTeach and UTeach Noyce Scholars with graduates of other programs. We find that UTeach and non-UTeach novice teachers are rated comparably as first year teachers, but that the UTeach graduates show greater improvement over time. UTeach Noyce Scholars have significantly higher scores than other novice UTeach graduates and Non-UTeach graduates. Thus pre-service characteristics may identify quality teachers.

Development of the UTeach Observation Protocol: A Classroom Observation Instrument to Evaluate Mathematics and Science Teachers from the UTeach Preparation Program

Measurement of teacher quality lies at the heart of urgent debates about educational reform in the United States. Initiatives such as *Race to the Top* propose to measure teacher quality largely or completely through student standardized test score gains. Even the most vigorous advocates for the use of linear regressions on test scores to determine teacher quality acknowledge the need for multiple measures of teaching performance, particularly when decisions might lead to financial rewards or dismissal. But there is no agreement what the alternative measures might be.

The challenges have become particularly severe in the case of secondary mathematics and science. No content areas are viewed as more critical in the current debate on US competitiveness. Yet mathematics and science are not necessarily tested every year through standardized assessments, making it difficult to construct gain scores, particularly in science. Furthermore, the different sciences appear to draw on such different skills that in one study where Biology was used to calibrate skill in science, simply taking Physics emerged as a strong negative factor in student performance (Portals 2010). Committing technical errors in the identification of quality teachers would have severe consequences for a branch of the teaching profession already dominated by shortages of qualified professionals.

We believe that classroom observations provide the best complement to student standardized test scores in the attempt to measure teacher quality. However generic observation instruments aimed at all disciplines and employed by observers without disciplinary knowledge are not sufficient. Based on our review of existing classroom observation protocols, we believe there has not yet been an observational instrument with demonstrated validity and reliability that could plausibly provide useful measurements of mathematics and science teacher quality. Thus here we present the UTOP, an observational instrument building upon the Classroom Observation Protocol of Horizons Research (Horizons Research Inc., 1999). In this preliminary study we present evidence for the soundness of the instrument, as well as discussing difficulties yet to be overcome.

The UTeach secondary mathematics and science teacher preparation program was founded at the University of Texas at Austin in 1997 as a way to allow science and mathematics majors to obtain a degree in their discipline and teaching certification at the same time. Hallmarks of UTeach include a new set of education courses based upon research into how students learn mathematics and science, full-time employment of former secondary teachers to supervise and guide students, early and continuing field experience, compact degree plans, and permanent funding for essential program elements. Thus the UTeach program focuses on recruiting students with high content knowledge in the domain they will be teaching (math, science or computer science), and training them in strong research-based teaching strategies such as implementing formative assessment and questioning, promoting and modeling reflection and metacognition, facilitating classroom argumentation, applying the science of learning to instruction, methods for differentiation and diversity, and adopting advanced pedagogical techniques like project-based instruction. Approximately 90% of UTeach alumni enter teaching, and of those who enter 80% are still in schools after 5 years (Marder & Abraham, 2009); this is compared to a national norm of around 65% (SASS, 2009). UTeach has been the subject of both House and Senate testimony, and was recently identified as a model for teacher preparation by the President (Randall, 2010).

Our original motivation to begin this study came from the need to evaluate the performance of Noyce Scholars graduating from UTeach. Both Noyce and UTeach have national significance. Congress has regularly increased the funding of the Noyce Scholarship Program through the National Science Foundation since the Gathering Storm report (Augustine, 2006) inspired passage of the COMPETES Act. Yet the quality of teachers supported in this way has not been assessed. UTeach has received numerous grants that make it possible to replicate at universities across the US, most notably through a gift from ExxonMobil to the National Math and Science Initiative in 2007. Twenty-one universities currently implement the UTeach model, and over 7 more are likely to join during 2011. As with Noyce Scholars, the effectiveness of UTeach graduates has not been measured. That both of these programs have grown rapidly in the absence of such research testifies to the extreme shortage of science and mathematics teachers, but does not eliminate the need for more careful evaluation.

Theoretical Framework

Research on Teaching Practices

Teachers have a large impact on student achievement (Heck, 2008; Rivkin, Hanushek, & Kain, 2005; Rowan, Correnti, & Miller, 2002; Sanders & Rivers, 1996; Wright, Horn, & Sanders, 1997), but the effect has proved surprisingly difficult to measure. Sanders and Rivers (1996) report that in elementary mathematics, students taught by the least effective teachers over three years score on average 52-54 percentile points lower on standardized assessments than those taught by the most effective teachers over three years. Rowan et al. (2002) used the *Prospects* elementary data to estimate that teacher effects explain 8-18% of the variance in student mathematics achievement, and 52-72% of the variance in students' academic *growth* in mathematics. Rivkin et al. (2005) used a large Texas database of students in grades 5-7 to show that an increase of one standard deviation in teacher effectiveness corresponds to a 0.11 increase in the standard deviation of student mathematics achievement. Kane and Staiger (2008) studied differences in score gains of elementary teachers who received random classroom reassignments at the beginning of the school year, showing some correlation between student score gains teachers obtained during four preliminary years and gains measured during a final trial year. In all these cases, student standardized test score gains play an important role in the definition of effective teaching; thus finding that effective teaching increases test scores is partly a circular argument (Kupermintz 2002). The urgency of defining good teaching is further underscored by the growing popularity of Value-Added Measures in policy debates (Pianta & Hamre, 2009). While it is certainly true that students in some classrooms have higher scores and higher score gains than students in other classrooms, an important question is to what extent the differences can be attributed to teacher quality.

In 2007, Hiebert and Grouws reviewed the literature on how mathematics teaching practices impact achievement, writing that "documenting particular features of teaching that are consistently effective for students' learning has proven to be one of the greatest research challenges in education" (pg. 371). They found several teaching practices to be consistently related to achievement, identifying "opportunity to learn" as a key factor. Opportunity to learn is thought to be only partially under the

control of the teacher, but includes teacher considerations of prior knowledge, the nature and purpose of the tasks given, and the likelihood of student engagement in the material. Teaching practices targeted towards conceptual understanding where students and teacher discuss and make connections between concepts and students grapple with important mathematical ideas were also shown to be important to learning.

Classroom Observation Instruments

When the National Science Education Standards (NSES) for Teaching (NRC, 1996) and National Council of Teachers of Mathematics (NCTM) Professional Standards for Teaching (NCTM, 1991) were published, they called for new approaches to classroom pedagogy, content development, and lesson structure based on research on teaching and learning in mathematics and science. Based on these two sets of standards, the National Science Foundation (NSF) developed the Local System Change (LSC) Classroom Observation Protocol, or COP (Horizons Research Inc., 1999), in order to measure teaching quality as part of the evaluation of their Teacher Enhancement Initiative. We use a modified version of the COP, so now we will briefly discuss previous studies using this instrument.

The COP contains several sections where observers describe and classify the major activities, materials, and purposes of a math or science lesson, and then it provides four sections where observers rate various aspects of classroom instruction using a Likert (1-5) scale. Each of the four sections concludes with a synthesis rating, and at the end of the instrument the observer gives a capsule rating of the quality of the entire lesson on a 1-7 scale. In the "Looking Inside the Classroom" study, researchers utilized the COP to observe 364 mathematics and science teachers. Based on their 7-point capsule ratings of these lessons, they concluded that only 15% of K-12 math and science lessons in the U.S. are high quality, with 27% medium quality and 59% low quality. They found that some of the strengths of mathematics and science lessons in the U.S. include that they cover significant and worthwhile content, they have teachers who are confident in their ability to teach, and they have teachers that communicate accurate content information. However, they found that fewer than one in five lessons are intellectually rigorous, include teacher questioning targeted towards enhancing conceptual understanding, or provide

sense-making opportunities. They also concluded that some of the key characteristics that distinguished effective lessons from ineffective lessons include engaging students in the content, creating a respectful and rigorous classroom environment, ensuring access to learning for all students, using questioning strategies to promote understanding and engage in formative assessment, and helping students make sense of the content and make connections among ideas (Weiss, Pasley, Smith, Banilower, & Heck, 2003).

Although the *Looking Inside the Classroom* study did not use student learning measures when drawing conclusions about effective teaching practices, COP ratings in middle school science were later tied to student achievement and the lessening of achievement gaps (Johnson, Kahle, & Fargo, 2006). Lawrenz (2007) studied the effectiveness of teacher preparation programs funded through the NSF Collaboratives for Excellence in Teacher Preparation (CETP) program using a classroom observation protocol as part of the evaluation. This protocol was similar to the COP, but included only 12 indicators in addition to a 7-point overall capsule rating of the lesson. The observational data showed that CETP teachers were rated by external observers as more likely to use standards-based instructional practices in their classrooms, but that capsule ratings were not especially high for either CETP or non-CETP groups.

The COP used expert review of supporting evidence to establish its reliability, and their observers viewed and rated lessons individually after an initial training session. The training session for the COP involves participants giving a videotaped lesson an overall rating on a 7-point scale that ranges from "ineffective instruction" to "exemplary instruction." The COP's documentation states that 92% of training participants rate the lesson within one level (higher or lower) of the normed standard rating, and 57% are in exact agreement with the standard rating (Horizons Research Inc., 2000c). No data is given concerning indicator-level reliability, and observers using the COP observe individually after the training session. The COP team has also studied the internal consistency of the items on their protocol, finding Cronbach alpha levels ranging from 0.92 to 0.97 for each section (Horizons Research Inc., 2000c).

A recent research report compiled by Decision Information Resources Inc. (DIR) looked at the current classroom observation protocols in use, including the LSC Classroom Observation Protocol (Horizons Research Inc., 1999), the Expert Science Teaching Educational Evaluation Model (ESTEEM)

Classroom Observation Rubric (Burry-Stock & Oxford, 1994), the Reformed Teaching Observation Protocol (RTOP) (Piburn & Sawada, 2000), the *Inside the Classroom* Observation and Analytic Protocol (Horizons Research Inc., 2000a), the Diagnostic Classroom Observation instrument (Saginor, 2008), and the UTeach Observation Protocol (UTOP, 2009), and found that none of these instruments other than the RTOP had achieved appropriate levels of *either* validity or reliability (Nancy Dawson, personal communication, 2008). There are a number of general observational tools that have established reliability and validity, such as the CLASS protocol (Pianta & La Paro, 2004), which now has a secondary version, and Charlotte Danielson's framework for teaching (Danielson, 1996). However, these instruments do not take into account teaching behaviors specific to the disciplines of mathematics and science, such as placing content in the "big picture" of the domain, supporting sense-making about concepts through real world connections, and appropriately and powerfully making use of tools of abstraction. As a preparation program focused on recruiting teachers with s strong mathematics and science backgrounds, such content-specific indicators are essential to our purposes.

The fact that the RTOP is the only classroom observation instrument appropriate for secondary math and science classrooms that has obtained reliability is problematic, given that the RTOP is designed to measure use of inquiry approaches rather than quality of instruction. Accordingly, the RTOP places little emphasis on the accuracy and depth of the content being conveyed during a lesson. To illustrate this point, we used the RTOP to rate a videotape of a UTeach graduate teaching an engaging, inquiry-based lesson while communicating problematic content to the students. We concluded that the RTOP's indicators were insufficient to fully characterize a lesson with misleading content information and missed opportunities. Thus our goal was to continue the development of and attain sufficient reliability on a more balanced classroom observation instrument, the COP. Pianta and Hamre (2009) note that standardized observational instruments appropriate for use grades 9 through 12 are "the exception rather than the norm" (pg. 111).

Teacher Background and Student Achievement

In teacher education, much research has been conducted on how teacher background characteristics, such as certification status and attainment of a Master's degree, impact student achievement. The possibility of identifying promising teachers in advance has been called into question; in a recent Brookings Institution Report it is argued that paper qualifications such as teacher credentialing have "little predictive power in identifying effective teachers" (Gordon, Kane, & Staiger, 2006, p. 2). As a program dedicated to training and credentialing exemplary teachers, we are interested in investigating such claims.

Most of the evidence connecting teacher background to student achievement relies on techniques of linear modeling. Monk (1994), in a study of 10^{th} -12^{th} grade students and their teachers, found that teacher subject matter expertise in mathematics has a significant positive effect on student mathematics achievement, but that there are diminishing returns after the fifth undergraduate mathematics course taken by the teacher. Mathematics pedagogy courses taken by the teacher also have a significant positive impact on student achievement, but a major in mathematics has no effect and a Master's degree has a significantly negative impact on student achievement. In science, content course-taking and a science major have a significantly positive impact on student achievement, as do science pedagogy classes, but a Master's degree again has a significant negative impact.

Clotfelder, Ladd, and Vigdor (2010) looked at achievement on end-of-course exams in 5 subjects (including algebra, geometry, and biology) given in 9^{th}-10^{th} grade in North Carolina, finding that subject certification, traditional certification, and higher licensure exam scores all had a significant positive impact on student achievement, especially in mathematics. A one standard deviation difference in a teacher's math praxis exam score was associated with a 0.05 standard deviation increase in student achievement in algebra and geometry, while being certified in mathematics raised student achievement by 0.11 standard deviations. They concluded that overall, the students of a teacher with weak credentials would be expected to perform 0.23 standard deviations lower than the students of a teacher with strong credentials.

Here we examine the effect of teacher background by comparing teachers who graduated from the UTeach program with teachers who graduated from other programs. We also compare two sub-groups: UTeach graduates who received Noyce Scholarships and UTeach graduates who did not receive Noyce Scholarships. The observations we report in this paper followed from a commitment in our Noyce grant to study the teaching of our Noyce Scholars.

The Robert Noyce Teacher Scholarship Program was introduced by the NSF to recruit strong students into teaching and provide incentives for teaching in high-needs districts and schools. Including the population of Noyce Scholars in this study makes it possible to address whether high-quality teachers can be identified prior to teaching. Noyce Scholars have bearing upon this question because all of these students were selected before they began teaching according to criteria that correspond to widely held assumptions about what should make for strong teachers. All UTeach students who apply for Noyce Scholarships submit an application with three parts: a college transcript, two essays about entering the teaching profession, and letters of recommendation. Content strength in the discipline (GPA>3.0) is a mandatory consideration; the GPA of Noyce Scholars upon graduation varies from year to year between 3.4 and 3.6, compared with 3.25 of UTeach students overall. By comparing the teaching effectiveness of UTeach graduates who are Noyce Scholars to other UTeach graduates, we can examine whether characteristics such as undergraduate GPA are related to teacher quality.

Research Questions

A review of the literature on classroom observation and teacher effectiveness raised several important points. First, we note the lack of appropriate classroom observations instruments to evaluate secondary mathematics and science teachers while fully taking into account content knowledge and pedagogical content knowledge. Second, we observe that there are open questions in teacher education about how teacher background characteristics tie to classroom teaching behaviors and student achievement. As UTeach is a teacher preparation program with national significance, and the Noyce Scholarship program is also a teacher recruitment initiative with similar national importance, we felt we were in a unique position to develop an observation protocol and use classroom observations to explore

how preparation and background characteristics are tied to teaching practices. Thus we have three primary goals in this article:

1) We describe the UTOP, a classroom observation instrument we developed to observe mathematics and science teachers.
2) We present preliminary data concerning the performance of teachers from different preparation backgrounds, including UTeach graduates, as a function of time in the classroom and the economic status of the students they serve.
3) We examine the classroom observation outcomes of the specific subset of UTeach graduates who received Noyce Scholarships in order to examine how teacher background impacts teaching characteristics and to provide information on the value of providing financial support to strong students who commit to teach.

Method

Development of the UTOP

The LSC protocol (Horizons Research Inc., 1999), or COP, was the most balanced and research-based classroom observation instrument that we found appropriate for secondary math and science classrooms. However, we identified two problematic features of the COP that were also common in similar observation instruments based on the NSES and NCTM standards. The first was an undervaluing of the teacher's content knowledge and content knowledge for teaching, which are integrally important to classroom practice and student achievement (Shulman, 1986, 1987; Rowan et al., 1997; Hill, Rowan, & Ball, 2005). The second was that the COP was still somewhat slanted towards inquiry or investigative approaches, without fully acknowledging the possibility that students can actively construct knowledge while listening to a lecture, reading a passage, or participating in other forms of direct instruction (Schwartz & Bransford, 1998).

We modified or removed seven indicators from the COP to address these concerns, and also added an additional seven indicators, five relating to the content knowledge of the teacher in the

classroom. We also took the post-observation teacher interview questions from the *Inside the Classroom* Observation and Analytic Protocol (Horizons Research Inc., 2000b) and revised the questions to place a stronger emphasis on engaging the teacher in a discussion of the content of the lesson and their own content knowledge for teaching. A final change made to the COP was the removal of the overall 7-point rating of the lesson. Instead, the synthesis ratings from each of the four sections of the UTOP (Classroom Environment, Lesson Structure, Implementation, Math/Science Content) were intended to be used as summary scores for each observation.

The UTOP (see UTOP, 2009) includes 32 classroom observation indicators organized into four sections: Classroom Environment, Lesson Structure, Implementation, and Math/Science Content. The indicators are rated by observers on a 7-point scale: 1 to 5 Likert with Don't Know (DK) and Not Applicable (NA) options (for some items). Each of the four sections concludes with a *synthesis rating* on a 1 to 5 Likert scale, which is intended to be a summary or overall score based upon the observer's weighting of the relative importance of the indicator evidence. The UTOP also includes a post-observation teacher interview, which contains 12 open-ended questions to assist the observers in ascertaining the background and purposes of the lesson. Finally, the full UTOP includes both observer and teacher demographic forms, as well as spaces for the observer to fill in background information about the lesson.

Process for Conducting Observations

All observers were given the UTOP Training Manual, which was developed to describe each indicator, as well as how the rating scales apply to each item. Observers practiced using the UTOP on video math and science lessons, discussing their ratings and justifications in small groups. This was followed by the new observers visiting classroom sites and rating live lessons using the UTOP while paired with more experienced observers and then debriefing on indicator scores and supporting evidence. Even after observers were fully trained, they continued to conduct observations in pairs. Fully-trained observers continued to debrief with their partner observer over each lesson, discussing their scores and supporting evidence in order to come to a consensus on each indicator rating, whenever possible. At least

one observer present in each classroom was a general expert in the subject being taught (either math or science), and to avoid bias towards the UTeach program, in over 70% of observations at least one observer was "blind" to the educational background of the teacher.

To schedule observations, each pair of observers determined a date to visit a specific school site, and then contacted each participating teacher at that school via email. Observers tried to maintain an unobtrusive presence in the classroom during the scheduled time, taking detailed field notes on what was occurring during the lesson. After the lesson had concluded, the observers would make sure they collected any handouts or lesson plans used by the teacher and would conduct a post-observation teacher interview. After the observers had reviewed their field notes, the interview notes or audio file, and teacher-supplied handouts, they would each independently fill out a UTOP for the lesson. Each indicator rating on the UTOP is typically supported by several sentences to several paragraphs of evidence from the observation and interview.

Data Sources

Classroom observations using the UTOP were conducted continuously over 5 semesters, with a total of 83 observations. Each observation lasted either 50 or 90 minutes, and post-observation interviews occurred after 82 observations, with 66 taking place in person, 3 taking place via phone, and 13 taking place via email. Interviews lasted between 5 minutes and 1 hour, with the typical time span being 15 to 20 minutes. The majority of face-to-face and phone interviews (56 interviews) were audio recorded and transcribed.

The school district in which the majority (83%) of the observations were conducted is a large, urban district serving over 80,000 students, with 61% economically disadvantaged, 28% Limited English Proficient, 12% African American, 58% Hispanic, 3% Asian, and 26% White. The remaining observations were conducted in a nearby rural school district serving approximately 6,000 students, with 68% economically disadvantaged, 26% Limited English Proficient, 28% African American, 45% Hispanic, and 27% White. The observations took place at nine high schools, including one magnet school

and one application-based school operating on a lottery system. Our sample also includes observations from five middle schools, including one magnet school and one charter school.

By recruiting multiple novice teachers from each school environment teaching similar academic subjects, some of whom had graduated from the UTeach program and some of whom had not, we hoped to achieve some form of "matching" between our graduates and a comparison group. However, as observational research involves human subjects and must therefore include informed consent, obtaining a random sample was not possible. Our results must be interpreted with the caveat that our sample was composed of volunteers; both UTeach graduates and non-UTeach graduates opted out of participation, even though compensation was offered to the teachers. However, we did observe a wide range in teaching proficiency, and participants expressed widely varying degrees of self-confidence during their interviews. Therefore, it is not true that all under-confident and all unskilled teachers chose not to participate.

Of the 36 teachers who participated in the study, 17 were science teachers, 18 were math teachers, and 1 was a computer science teacher; were middle school teachers and 29 were high school teachers. Fourteen of the teachers taught in schools with 0-33% of students receiving free/reduced lunch, 16 taught in schools with 33-66% of students receiving free/reduced lunch, and 6 taught in schools with 66-100% of students receiving free/reduced lunch. Of the 36 participating teachers, 15 were from preparations backgrounds other than the UTeach program, while the remaining 21 were UTeach graduates. Of the 21 UTeach graduates, 7 were Noyce Scholarship recipients. Of the 15 teachers who had not graduated from the UTeach program, 7 were alternatively certified, 6 were certified through a 4-year university, and 2 have unknown certification backgrounds. All the teachers had less than 5 years of teaching experience; 27 of the observations were of first year teachers, 39 were of second year teachers, 14 were of third year teachers, and 3 were of fourth or fifth year teachers. Sixty of the 83 observations were of regular-level classes, while 30 of the 83 observations were of advanced-level classes. Advanced-level classes were classes designated by the school as "AP" or "Pre-AP," or classes that took place at a school that was a magnet school. Each teacher was observed between one and six times over the course of

their participation, with no more than two observations taking place in a semester. Table 1 gives a summary of the comparability of the observations of different groups examined in this study.

Table 1

Characteristics of Teacher Observations in Sample, based on Preparation Backgrounds

	UTeach Noyce Graduates (N=7 teachers, 21 observations)	UTeach Non-Noyce Graduates (N=14 teachers, 31 observations)	Non-UTeach Graduates (N=15 teachers, 31 observations)
Subjects Taught	2 middle school science teachers 1 physics teacher 3 algebra I/II/geometry teachers 1 computer science teacher	3 biology teachers 4 physics/chemistry teachers 2 middle school math teachers 5 high school math teachers (algebra I/II/geometry/pre-cal)	3 middle school science teachers 3 biology teachers 1 biology and chemistry teacher 1 middle school math teacher 7 high school math teachers (algebra I/II/geometry/calculus)
School Type	9 observations at 3 public high schools 4 observations at 2 public middle schools 3 observations at 1 magnet high school 5 observations at 1 magnet middle school	16 observations at 3 public high schools 2 observations at 1 public middle school 11 observations at 1 application-based high school 2 observations at 1 charter middle school	17 observations at 4 public high schools 4 observations at 2 public middle schools 5 observations at 1 application-based high school 2 observations at 1 magnet high school 3 observations at 1 magnet middle school
School Poverty Level	8 observations at low-poverty schools[1] 6 observations at medium-poverty schools[2] 7 observations at high-poverty schools[3]	10 observations at low-poverty schools 19 observations at medium-poverty schools 2 observations at high-poverty schools	18 observations at low-poverty schools 9 observations at medium-poverty schools 4 observations at high-poverty schools
Level of Classes	11 observations of regular-level classes 10 observations of advanced-level classes	23 observations of regular-level classes 8 observations of advanced-level classes	19 observations of regular-level classes 12 observations of advanced-level classes
Teaching Experience	5 observations of 1st year teachers 9 observations of 2nd year teachers 7 observations of 3rd year teachers	15 observations of 1st year teachers 14 observations of 2nd year teachers 2 observations of 3rd year teachers	7 observations of 1st year teachers 16 observations of 2nd year teachers 5 observations of 3rd year teachers 3 observations of 4th and 5th year teachers

UTOP Reliability & Internal Consistency

[1] Schools with 0-33% free/reduced lunch
[2] Schools with 33-66% free/reduced lunch
[3] Schools with 66-100% free/reduced lunch

The UTOP was used by pairs of expert observers observing math and science lessons together and then rating them independently. The observers obtained an average weighted kappa of 0.41 on the 32 individual indicators rated on the 7-point (1-5 with DK/NA options) scale, which meets the criteria for moderate agreement (Landis & Koch, 1977). On the synthesis or summary ratings for each of the four sections our observers obtained an overall weighted kappa of 0.63, which is considered substantial agreement. Given that neither the COP or the RTOP had established any sort of item-level reliability, we find these results to be a promising beginning to further development of the UTOP.

The UTOP shows similar levels of internal consistency to the COP. Cronbach alpha was calculated for each section of the UTOP using the 1-5 Likert rating scale data, and omitting data where DK or NA had been selected. Indicators 3.10 (The teacher's instructional strategies included safe, environmentally appropriate, and ethical implementation of lab procedures*)* and 4.8 (Mathematics and science were portrayed as a body of knowledge influenced by human decisions and influencing human society*)* were omitted from this analysis, due to a high number of missing data points and many instances of observers selecting DK and NA for these items, as well as the results of the factorial analysis detailed in the next section which strongly suggested the deletion of these items. The Cronbach alpha values for each of the four sections ranged from 0.905 to 0.962.

We were also interested in examining the factorial structure of the indicators on the UTOP. Preliminary results showed the Content section of the UTOP forming a widely-spread cluster, with the other indicators being divided into two additional clusters depending on whether they were primarily teacher-centered or student-centered. However, the data reported here is based on 83 observations, which is not a sufficient sample size to draw conclusions from such a structural analysis. Current work with the UTOP is using larger sample sizes to focus on this issue.

Results

An analysis of the synthesis ratings in each of the four sections of the UTOP (Classroom Environment, Lesson Structure, Implementation, and Math/Science Content) forms the core of the

preliminary data we will present on teaching quality. The average synthesis rating for each of the four UTOP sections is given in Table 2. We first look at comparative graphs of UTOP scores for different subgroups, and then turn to a discussion of the statistical significance of the differences that are found.

Table 2

Average UTOP Synthesis Ratings (N=83 observations) by math/science subgroups

Section of UTOP	Science Teacher Observations (N= 39)		Math & Computer Science Teacher Observations (N= 44)	
	Avg Synthesis Rating	Standard Deviation	Avg Synthesis Rating	Standard Deviation
1: Classroom Environment	2.84	1.20	2.64	1.22
2: Lesson Structure	3.00	1.10	2.92	0.98
3: Implementation	2.64	1.30	2.62	1.14
4: Math/Science Content	3.41	1.17	3.29	0.96

Comparing UTeach, Non-UTeach, and Noyce Scholar UTOP Ratings Graphically

The unit of analysis we use in the graphs and t-tests that we present here is the teacher (N=36), rather than observation (N=83). Each teacher's four synthesis ratings on the four UTOP sections were averaged across all observations of that teacher, and this set of four averages per teacher was used for statistical analysis. Further, the teacher's scores across these four UTOP sections were also averaged to provide an overall average synthesis rating or "Composite UTOP Rating for each teacher. Although using the observation as the unit of analysis gave more statistical power and yielded more significant results, we could not justify treating multiple observations of the same teacher as statistically independent of one another. Thus we took a conservative approach by averaging the scores for each teacher, rather than treating them as independent. The one exception to this general rule is an examination of change in teaching characteristics over time, which we will discuss shortly.

Students who graduated from the UTeach program were rated higher on the UTOP by observers than the comparison group. Figure 1 shows the Composite UTOP ratings of Non-UTeach teachers (i.e. teachers who did not go through the UTeach program; N=15), UTeach Noyce Scholars (i.e. UTeach graduates who received Noyce Scholarships; N=7), and UTeach Non-Noyce teachers (i.e. UTeach

graduates who did not receive Noyce Scholarships; N=14). Error bars in Figure 1 and the remainder of the figures in this paper display standard error of the mean. The Noyce Scholars score highest in each section, with the UTeach Non-Noyce group in the middle and the non-UTeach group at the bottom.

Since the UTeach program places a strong emphasis on preparing math and science teachers to be successful when working with diverse populations, we were interested in examining lesson ratings in schools of different economic levels. Figure 2 shows that UTeach teachers appear to be consistently scored higher in terms of composite UTOP rating when compared to observations of Non-UTeach graduates, across school economic levels.

[Insert Figures 1 and 2 around here]

We also examined growth in synthesis ratings as a function of teacher experience, since one of our initial beliefs was that UTeach graduates might display different patterns of growth from the comparison group as they develop as teachers. Averaging observation scores for each teacher would not make sense for growth over time data; we had teachers that were observed repeatedly during their first, second, and third years of teaching. Instead, we averaged observations of teachers that occurred in the same semester for the purposes of the graphs presented here, which resulted in a sample size of 71 observations rather than 83 observations. For the purposes of the t-tests presented later, we averaged observations of the same teacher conducted in the same year, to take a more conservative approach, which gave a sample size of 50 observations.

Figure 3 shows the Composite UTOP rating of UTeach observations and Non-UTeach observations for teachers having between 0 and 2.5 years of teaching experience (left). In the graph on the right side of Figure 3, UTeach Noyce Scholars are shown separately from the other two groups. As Figure 3 shows, while UTeach and Non-UTeach graduates begin in their first year of teaching with similar ratings, UTeach graduates seem to show more pronounced patterns of growth over their first 3 years in the classroom. Figure 4 shows the average synthesis ratings for each of the four UTOP sections for Non-UTeach observations (left) and UTeach observations (right) as a function of years of teaching experience. These graphs suggest differing patterns of change in all four sections of the UTOP as the teachers develop

through their novice years in the profession. We found it particularly interesting that the Math/Science Content ratings seem to show a pattern of growth for UTeach observations, while they show a pattern of decline for Non-UTeach observations.

[Insert Figures 3 and 4 around here]

Statistical Significance of Comparison Results

Welch's *t*-tests were used to investigate the significance of the trends found in the figures presented in the previous section. *T*-tests were performed on each teacher's average ratings. We decided to focus on *t*-tests rather than a more advanced statistical method for two reasons. First, our data is not crossed; the number of times each teacher was observed varied between one and six visits. Second, we believe the assumption of linearity is used too freely in statistical analysis in education, where many relationships are inherently non-linear. Since we present a number of *t*-tests, we present our results with the caveat that by using a 5% level of significance, we expect roughly 1 in 20 of our tests to be false positives. The *t* tests were in two sets. A first set compared teacher characteristics without regard to years of teaching. There were 15 separate tests in this set, which yielded three comparisons with $p<.01$ and three more with $p<.05$. The odds of these all being false positives is less than 10^{-8}. A second set of tests was performed disaggregating by numbers of years of teaching. For these comparisons only composite UTOP indicators were used, and we performed 15 *t*-tests, of which 4 showed significance with $p<.05$.

A summary of the significant results appears in Table 3. In many cases absence of significance is due to small subgroup sample size. The main results are:

- Noyce Scholars are rated higher than Non-UTeach teachers, ($p<0.05$ and $p<.01$) in all four sections of the UTOP and Composite UTOP rating ($p<.01$).
- Noyce Scholars are rated significantly higher ($p<0.05$) than UTeach Non-Noyce teachers in the Content section of the UTOP. Noyce Scholar observations are rated significantly higher ($p<0.05$) than UTeach Non-Noyce observations for teachers in their first year of teaching, in Composite UTOP rating and Lesson Structure.

- UTeach observations are rated significantly higher than Non-UTeach observations Composite UTOP ratings for teachers in their second year of teaching. There are no significant differences between groups in the third year of teaching, possibly due to smaller subgroup sample sizes.
- UTeach observations of third year teachers are significantly higher ($p<.05$) than UTeach observations of first year teachers, showing growth over time. This is not the case for Non-UTeach observations.
- No differences by economic subgroup of students at the observed teacher's school reached significance.

Most importantly, no significant differences according to any subgroup division were found that favored Non-UTeach over UTeach (all), Non-UTeach over Noyce, or Non-UTeach over UTeach Non-Noyce. This means that according to our *t*-test analysis, UTeach graduates were always rated higher or not significantly different than the comparison group. Noyce Scholars were always rated either higher or not significantly different from the comparison groups for all subgroup divisions.

Table 3

Summary of significant t-test results

Measure	Condition	Alternative Hypothesis	Significance
Composite UTOP Rating & Synthesis Ratings for Content and Lesson Structure	All	Noyce Scholars>Non-UTeach	*p<.01*
Synthesis Ratings for Class Environment and Implementation	All	Noyce Scholars>Non-UTeach	*p<.05*
Synthesis Rating for Content	All	Noyce Scholars>UTeach Non-Noyce	*p<.05*
Composite UTOP Rating	Second year of teaching	UTeach (all)>Non-UTeach	*p<.05*
Composite UTOP Rating	First and third year of teaching, UTeach	UTeach (first year)<UTeach (third year)	*p<.05*
Composite UTOP Rating	Second year of teaching	Noyce Scholars>Non-UTeach	*p<.05*
Composite UTOP Rating	First year of teaching	Noyce Scholars>UTeach Non-Noyce	*p<.05*

Regression Analysis

We also used hierarchical linear modeling to fit a 3-level random intercept model to the data (Snijders & Bosker, 1999). The model used for the analysis was constructed such that observations were nested within teachers, and teachers were nested within schools. Random intercept terms included teacher, school, and course (Algebra I, Calculus, Biology etc.). Fixed effects included the teacher's preparation background (UTeach Noyce, UTeach Non-Noyce, or Non-UTeach), years of teaching experience, whether the class was regular level or advanced level, and whether the school was a middle school or a high school. Measures relating to poverty level of students in the schools did not reach significance in the model. Dummy coding was used for factor variables. The dependent variable was average UTOP synthesis rating. The general form of a 3-level random intercept model is:

$$Y_{ijk} = \gamma_{000} + \gamma_{100}x_{ijk} + V_{00k} + U_{0jk} + R_{ijk}$$

In this model, *i* is the level one variable (e.g. the observation), *j* is the level two variable (e.g. the teacher), and *k* is the level three variable (e.g. the school). The γ_{000} term is the overall intercept, and the *x* value is a predictor with estimated coefficient γ_{100}. The V_{00k} term is the level 3 random intercept, which is independent and identically distributed with mean 0 and variance ϑ^2. The U_{0jk} term is the level 2 random intercept, which is also independent and identically distributed with mean 0 and variance τ^2. Finally, the R_{ijk} term is the error term, which is independent and identically distributed with mean 0 and variance σ^2.

Results are shown in Table 4. Significance levels (*p*-values) are estimated using the normal distribution. The reference categories are a Non-UTeach teacher with no teaching experience, teaching a regular-level high school class. As can be seen from the intercept value, such a teacher is predicted to have an average synthesis rating of 2.068.

The results in Table 4 complement the *t*-test results. Noyce Scholars have significantly higher UTOP scores than Non-UTeach teachers (score difference of 1.37, $p<.01$) and UTeach Non-Noyce teachers (score difference of 1.370+.292 = 1.662, $p<.01$), but only if they are teaching regular-level classes. When Noyce Scholars teach advanced classes, all significant differences disappear between

Noyce and Non-UTeach (score difference of 1.370-1.470 = -0.10, p=0.838), and Noyce and UTeach Non-Noyce (score difference of 1.662-1.470 -.137= 0.055, p=0.930).

Table 4

Summary of regression results

Random Effects:			
	Type	Variance	St Dev
Teacher ID	Intercept	0.0202	0.1420
Course	Intercept	0.1247	0.3532
School	Intercept	0.3881	0.6230
Fixed Effects:			
	Coefficient	Error	*t*-value
(Intercept)	2.068	0.3866	5.35***
Non-UTeach	Ref.		
Noyce Scholar	1.370	0.4971	2.757**
UTeach Non-Noyce	-0.292	0.4546	-0.642
Regular-Level Class	Ref.		
Advanced Class	0.699	0.3423	2.041*
Years of Experience	0.023	0.1408	0.165
High School Class	Ref.		
Middle School Class	0.946	0.3568	2.65*
Noyce Scholar*Advanced Class	-1.470	0.4639	-3.169**
UTeach Non-Noyce*Advanced Class	0.137	0.4797	0.285
Noyce Scholar*Years of Experience	0.123	0.2777	0.445
UTeach Non-Noyce*Years of Experience	0.857	0.2821	3.039**

* p<.05, ** p<.01, *** p<.001

Years of teaching experience has no significant impact on UTOP scores for either Non-UTeach teachers (p=0.869) or Noyce Scholars (p=0.533), but gaining more experience seems to significantly improve the UTOP scores of UTeach Non-Noyce teachers (p<.05). UTeach Non-Noyce teachers gain an estimated 0.88 (0.023+0.857) points on their average UTOP rating, per year of experience. This finding is based on novice teachers, in their first 3 years of teaching.

The coefficient describing the interaction of Noyce Scholars and Advanced Classes is negative, and significantly different from zero. Thus it may appear that Noyce Scholars are worse at teaching advanced classes than the other teachers in our sample. However, this observation results from a misinterpretation of the coefficients in the linear model. As shown in Figure 5, the correct interpretation is that only Noyce Scholars in our sample performed equally well with both regular and advanced classes. The negative coefficient in the hierarchical linear model arises because on the whole composite scores are higher for advanced classes, and to account for the fact that scores do not rise for the Noyce Scholars the model gives them a negative coefficient.

[Insert Figure 5 around here]

Thus the results of the regression analysis confirm and extend certain findings from the t-tests. We found that Noyce Scholars have significantly higher UTOP scores than the comparison groups, but only when teaching regular-level classes. This may be due to ceiling effects for advanced classes. We also found the UTeach Non-Noyce teachers have significant growth over time, as they progress through their novice years of teaching. This is compared to a non-significant relationship between teaching experience and UTOP score for Noyce Scholars (likely due to ceiling effects) and Non-UTeach graduates.

An additional predictor was added into the model to describe whether the teacher being observed was alternatively certified, or certified through a four-year university. We did not have certification information for two of the teachers who had been observed a total of 4 times, so for this analysis the number of observations was reduced to 79. Results showed that being alternatively certified significantly decreased UTOP average synthesis rating (t=-2.363, p<.05) by 0.787 points (SE= 0.333). The coefficients and significance levels for other predictors remained similar to the model in Table 4. Thus the model suggests that alternatively certified teachers have lower scores on the teaching behaviors assessed on the UTOP.

Observer Bias

A potential weakness of this study is that it was organized by UTeach program officials with an obvious interest in the success of the program, and observations were conducted by graduate students paid

through program funds. Some of these came from the Noyce Scholarship program, which itself has an interest in establishing the effectiveness of its graduates to obtain continued federal support, and some came from UTeach discretionary monies. These sources of potential bias were evident to the study designers from the outset, and therefore we took specific steps to reduce the potential for bias even if it could not be removed completely.

1) We attempted to employ observers who had no prior knowledge of the UTeach program or Noyce Scholarship program. Two of the primary observers have never had any contact with UTeach apart from their observational role and were never exposed to faculty opinions of the UTeach program philosophy. The other two primary observers did at some point serve as teaching assistants for UTeach classes and thus were familiar with program goals.
2) We tried to ensure that observers were unaware of which teachers were graduates of the UTeach program or supported by the Noyce. Prior to the observation, teachers were sent an email instructing them not to reveal their preparation background to the observers, directly or indirectly.

Looking at the observation data for each of the four primary observers, we saw that all the observers on average rated UTeach graduates at the same level (3 on a scale of 1-5) despite their differing levels of involvement with the UTeach program. However, average ratings of teachers not from the UTeach program were more variable, ranging between 2.25 and 2.7, depending on the observer. This raised the possibility that although observers may not show bias *towards* UTeach graduates, they may be showing bias *against* non-UTeach graduates. We determined based on this data that it would be warranted to go back and specify whether each observer was blind while conducting an observation, and compare the scores of blinded observers against non-blinded observers for UTeach observations versus non-UTeach observations.

We found that our data may display an observer bias against Non-UTeach observations. When observers were blind to the fact that they were observing a Non-UTeach graduate, the average rating was 2.8; when observers knew they were observing a Non-UTeach graduate, the average rating was 2.5. This difference does not reach the level of statistical significance, and may have been partially a result of the

fact that 46% of the blind observations were of first year teachers, while only 16% of the not blind observations were of first year teachers. We would expect the differences between UTeach and non-UTeach to be less pronounced during the first year in the classroom, which may have contributed to the smaller difference seen in the blind observations. Also, 34% of the not blind observations were of Noyce Scholars, while only 17% of the blind observations were of Noyce Scholars. If our hypothesis about Noyce Scholars being superior to other UTeach graduates is correct, we would also except to see more pronounced differences in the not blind observations between UTeach and non-UTeach, as Noyce Scholars fell into the UTeach group.

Significance

We have investigated the teaching practices of UTeach graduates and UTeach Noyce Scholars using classroom observation measures. After reviewing current observation instruments for math and science classes, we determined that an instrument needed to be developed that was appropriate for use in a wide range of settings from kindergarten to college, that recognized a range of practices as "good" teaching, limiting the bias towards specific pedagogical approaches, and that fully took into account how math and science content knowledge interact with teacher quality. We developed the UTeach Observation Protocol (UTOP) by modifying Horizon's Classroom Observation Protocol according to the needs we identified from our review of other instruments. Preliminary findings show that the UTOP has promising levels of inter-rater reliability and internal consistency. Although the development of the UTOP will be ongoing based on these results, we presented some preliminary findings on the teaching practices and observation ratings of UTeach graduates, UTeach graduates who received Noyce Scholarships, and graduates from preparation programs other than UTeach, including alternative certification programs.

We found that Noyce Scholars are overall rated significantly higher on the UTOP than the other groups, suggesting that the background characteristics that led to these teachers being awarded Noyce Scholarships may be indicators of teacher effectiveness. However, it is important to note that not all qualified students in the UTeach program apply for Noyce Scholarships. In addition to having a high

GPA, Noyce recipients had to go through the process of applying for the scholarship, which encompassed making a commitment to teach, soliciting letters of recommendation, and writing a successful essay. We therefore know that all Noyce Scholars were strong students, while UTeach Non-Noyce graduates consisted in a mix of stronger and weaker students. Thus it is particularly striking that significant differences could be detected between the two groups. We are presuming (and so far as we know it is true) that similar considerations have gone into choosing Noyce Scholarship recipients across the United States. Thus our results should have bearing on Congress' preferred vehicle for encouraging mathematics and science majors to enter teaching.

We also found that growth on UTOP ratings during a teacher's first 3 years in the classroom appeared to be different for UTeach graduates and the comparison group, with UTeach graduates showing a stronger pattern of upward growth despite starting at a comparable rating level during their first year. Although the differing patterns of growth are apparent in the ratings of all four UTOP sections, they seem to be most pronounced in the Math/Science Content section. Finally, we found that when teacher UTOP ratings were considered with respect to the demographics of the student populations they serve, that UTeach graduates had a higher average synthesis rating than the comparison group across demographic categories, although these differences do not reach significance.

There are quite a few caveats to the results outlined above, as we have learned many lessons about conducting classroom observations over the past three years. As observational research involves human subjects, we had to obtain informed consent from all participants; thus our sample may not be representative of the population we wish to measure. Further, although we tried to "match" teachers in similar classroom settings teaching similar courses with similar years of experience, this level of matching proved difficult to maintain, especially in smaller math and science departments where there were fewer research volunteers. We also had planned for all of our observations to be conducted by observers who were "blind" to the educational background of the teachers being observed, but we did not succeed on all occasions (e.g. some teachers disclosed training without prompting), and this may have had

an impact on our results. Finally, we have engaged in a continued effort to obtain inter-rater reliability on the UTOP, which is a high-inference instrument.

We plan to continue to revise the UTOP using the reliability and consistency results reported in this paper. We also plan to continue to conduct observations at the middle school, high school, and college level in order to refine the instrument and continue to gather data on the teaching practices of UTeach graduates, as well of graduates of other teacher preparation programs. Observers using the UTOP typically cite large amounts of supporting evidence for each indicator, which in combination with the teacher interview transcriptions has provided us with the possibility of engaging in a rich, qualitative analysis to complement our quantitative analysis on teaching practices. This will allow us to better understand the struggles of our graduates as they progress through their novice years of teaching, and to design our teacher preparation program to fit their needs and take into account their strengths and weaknesses as identified by observers. We also plan to obtain the standardized test scores of our graduate's students to conduct comparative analysis, and use in conjunction with classroom observation data.

This line of research is intended to contribute to the field's understanding of what constitutes effective teaching, in a way that moves beyond the "black box" of using value-added assessments of standardized test scores and allows excellence in teaching to be connected to behaviors and interactions within the classroom. Given that teacher credentials, particularly relating to content background, have become a national focus, developing classroom observation instruments that can assess how content knowledge arises in classroom contexts is an important focus for educational research. Further, as both the UTeach program and the Noyce Scholarship program continue to expand and gain national recognition, it is critical that a research agenda be developed that assesses the quality of students coming from these programs. Finally, there are many unanswered questions circulating in educational policy and teacher education arenas about how teacher background characteristics, including preparation background, contribute to teaching behaviors and ultimately to student achievement. The study reported here contributes to the discussion of all of these important and timely issues in education.

Acknowledgements

The support of the National Science Foundation through the Noyce Scholarship Program, EHR 0630376, is gratefully acknowledged. The opinions expressed in this paper are those of the authors and are not endorsed by the National Science Foundation.

References

Burry-Stock, J. & Oxford, R. (1994). Expert science teaching educational evaluation model (ESTEEM): Measuring excellence in science teaching for professional development. *Journal of Personnel Evaluation in Education*, **8**(3), 267-297.

Clotfelter, C., Ladd H., & Vigdor, J. (2010). Teacher credentials and student achievement in high school: A cross subject analysis with student fixed effects. *The Journal of Human Resources*, **45**(3), 655-681.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: ASCD.

Gordon, R., Kane, T., & Staiger, D. (2006). Identifying effective teachers using performance on the job. Discussion Paper 2006-01. The Hamilton Project.

Heck, R. H. (2008). Teacher effectiveness and student achievement: Investigating a multilevel cross-classified model. *Journal of Educational Administration,* **47**, 227-249.

Hiebert, J. & Grouws, D. (2007). The effects of classroom mathematics teaching on students' learning. In D. Grouws (Ed.), *Handbook of Research on Mathematics: Teaching and Learning* (pp. 371-400). New York: MacMillian.

Hill, H., Rowan, B. & Ball, D. (2005): Effects of teachers' mathematics knowledge for teaching on student achievement. *American Educational Research Journal,* **42**(2), 371-406.

Horizons Research Inc. (1999). Local Systemic Change through Teacher Enhancement Classroom Observation Protocol. Retrieved June 2009 from http://www.horizon-research.com/instruments/lsc/cop.php.

Horizons Research Inc. (2000a). *Inside the Classroom* Observation and Analytic Protocol. Retrieved June 2009 from http://www.horizon-research.com/instruments/clas/cop.php.

Horizons Research Inc. (2000b). *Inside the Classroom* Interview Protocol. Retrieved June 2009 from http://www.horizon-research.com/instruments/clas/interview.php.

Horizons Research Inc. (2000c). Validity and reliability information for the LSC Classroom Observation Protocol. Retrieved June 2009 from http://www.horizon-research.com/reports/.

Johnson, C. & Kahle, J. (2006). Effective teaching results in increased science achievement for all students. *Wiley InterScience* (www.interscience.wiley.com).

Kane, T. & Staiger, D. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. NBER Working Paper No. 14607.

Kupermintz, H. (2002). Teacher effects as a measure of teacher effectiveness: Construct validity considerations in TVAAS (Tennessee Value-Added Assessment System). *CSE Technical Report 563*. University of California, Los Angeles.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159-174.

Lawrenz, F., Huffman, D., & Gravely, A. (2007). Impact of the Collaboratives for Excellence in Teacher Preparation Program. *Journal of Research in Science Teaching*, **44**(9), 1348-1369.

Marder, M. & Abraham, L. (2009). Student data, Spring 2009. Retrieved June 2009 from https://uteach.utexas.edu/go/uteachweb/Data-and-Documents/UTeach-Student-and-Graduate-Data.

National Council for Teachers of Mathematics (1991). *Professional standards for teaching mathematics*. Reston, VA.

National Research Council, National Academy of Sciences. (1996). *National science education standards*. Washington, DC.

Pianta, R. C., & La Paro, K. M. (2004). Classroom Assessment Scoring System (CLASS). Unpublished measure, University of Virginia.

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

Piburn, M. & Sawada, D. (2000). Reformed Teaching Observation Protocol (RTOP): Reference manual. ACEPT**:** 1-41. Retrieved June 2009 from http://cresmet.asu.edu/prods/rtop_files/RTOP_Reference_Manual.pdf.

Portals (2010). *Portal Report: Teacher Preparation and Student Test Scores in North Carolina*. Retrieved November 2010 from http://publicpolicy.unc.edu/?q=node/266

Randall, K. (January 7, 2010). *UTeach expansion recognized as President Obama spotlights importance of teachers in improving U.S. innovation*. Retrieved February 2011 from http://www.edb.utexas.edu/education/news/2010/uteachobama

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

Rowan, B., Chiang, F., & Miller, R. (1997). Using research on employees' performance to study the effects of teachers on students' achievement. American Sociological Association. **70:** 256-284.

Rowan, B., Correnti, R., & Miller, R. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the *Prospects* study of elementary schools. *Teachers College Record*, **104**(8), 1525-1567.

Saginor, N. (2008). *Diagnostic Classroom Observation: Moving Beyond Best Practice*. Thousand Oaks, CA, Corwin Press.

Sanders, W. & Rivers, J. (1996). Cumulative and residual effects of teachers on future student academic achievement. Knoxville: University of Tennessee Value-Added Research and Assessment Center.

SASS (2009). From SASS data table "Percentage distribution of public school teachers by stayer, mover, and leaver status." Retrieved August 2010 from http://nces.ed.gov/surveys/sass/tables/tfs_2005_02.asp

Schwartz, D. & Bransford, J. (1998). A time for telling. *Cognition and Instruction*, **16**(4), 475-522.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *American Educational Research Association Journal*, **15**, 4-14.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, **57**(1), 1-27.

Snijders, T. & Bosker, R. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage Publications.

UTOP (2009). http://uteach.utexas.edu/ResearchMethods/UTOP_2009.doc

Weiss, I., Pasley, J., Smith, P., Banilower, E., & Heck, D. (2003). Looking inside the classroom: A study of k-12 mathematics and science education in the United States. Chapel Hill, NC: Horizon Research, Inc. Retrieved June 2009 from http://www.horizon-research.com/insidetheclassroom/reports/looking/

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

Figures

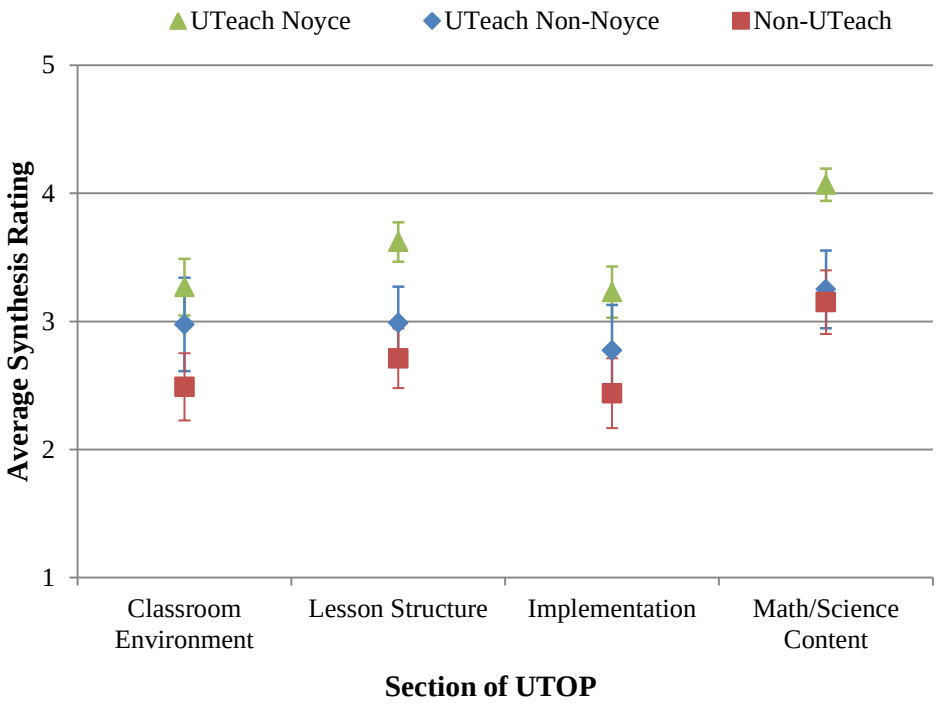

Figure 1

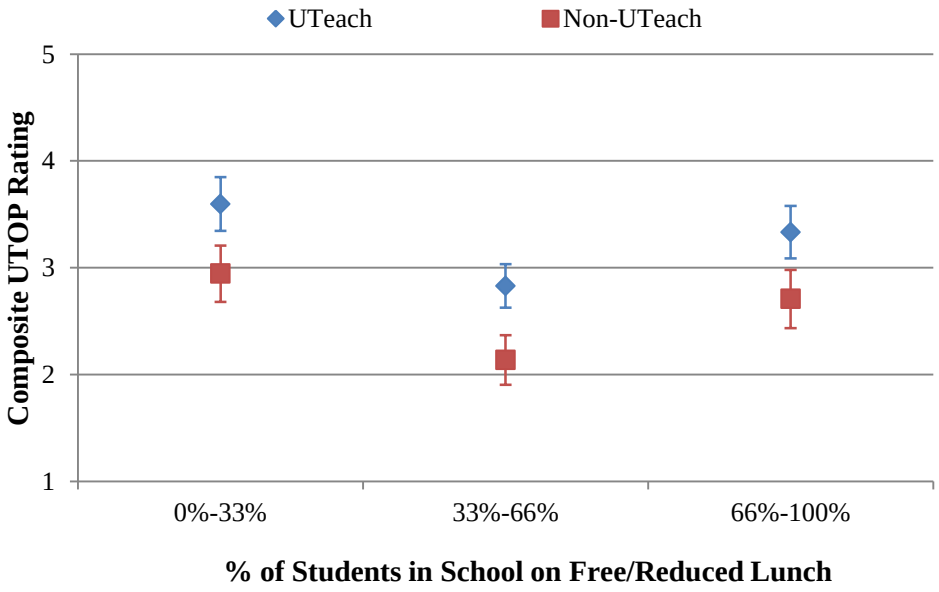

Figure 2

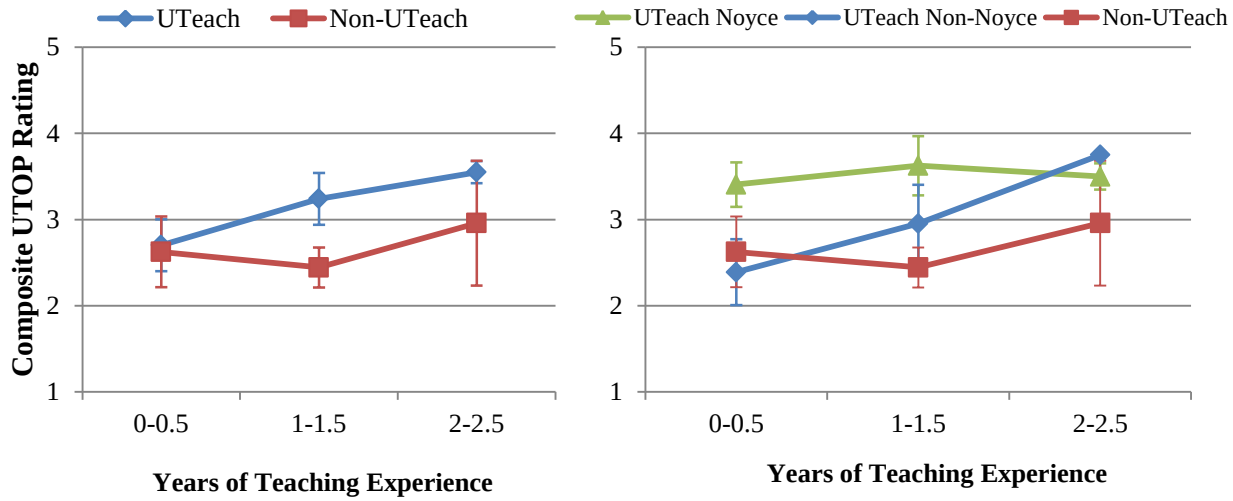

Figure 3

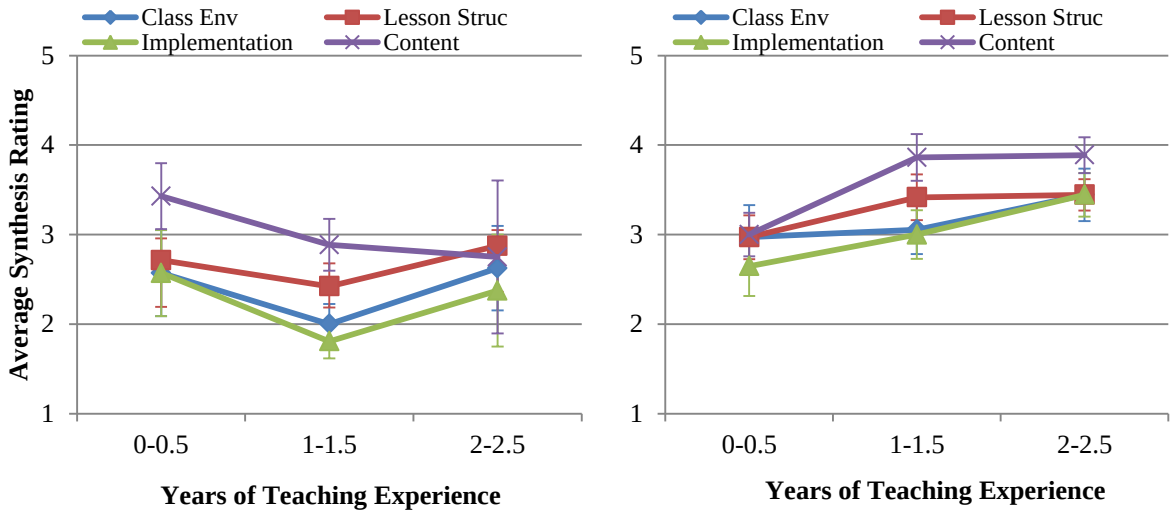

Figure 4

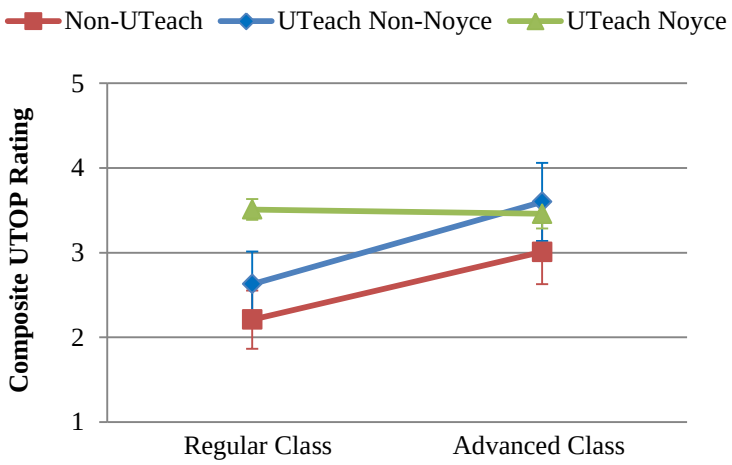
Non-UTeach
UTeach Non-Noyce
UTeach Noyce
Composite UTOP Rating
5
4
3
2
1
Regular Class
Advanced Class

Figure 5

Figure Captions

Figure 1. Average Sythesis Ratings in four UTOP Sections, by Non-UTeach, UTeach Noyce, and UTeach Non-Noyce subgroups. Multiple observations of particular teachers are averaged together, and different teachers are treated as independent.

Figure 2. Composite UTOP Rating of Non-UTeach teachers and UTeach teachers as a function of school economic level. Multiple observations of particular teachers are averaged together, and different teachers are treated as independent.

Figure 3. (Left) Composite UTOP Ratings of UTeach observations and non-UTeach observations as a function of years of teaching experience. (Right) Composite UTOP Ratings of UTeach Noyce observations, UTeach Non-Noyce observations, and non-UTeach observations, as a function of years of teaching experience. Observations of a particular teacher are averaged together if they occurred in the same semester; otherwise they are treated as independent.

Figure 4. (Left) Average Synthesis Ratings of Non-UTeach graduate observations in the 4 UTOP sections as a function of years of teaching experience (Right) Average Synthesis Ratings of UTeach graduate observations in the 4 UTOP sections as a function of years of teaching experience. Observations of a particular teacher are averaged together if they occurred in the same semester; otherwise they are treated as independent.

Figure 5. Composite UTOP Ratings of UTeach Noyce, UTeach Non-Noyce, and Non-UTeach graduates teaching regular and advanced classes