Classroom Observation and Value-Added Models Give Complementary Information about Quality of Mathematics Teaching

Candace Walkington and M. Marder

October 9, 2013

Abstract

We developed the UTeach Observation Protocol (UTOP), which provides a systematic way to organize observations about teachers and students in a classroom, and provides numerical ratings of classroom quality in multiple dimensions. Through the Measures of Effective Teaching project we obtained UTOP ratings and comments on 982 videos of grades 4-8 mathematics classrooms. We also obtained results for each teacher in the videos from value-added models, which use changes in student test scores to evaluate teachers. We studied the connections between the UTOP ratings and the value-added model ratings. We were surprised by many findings. For example, the particular classroom attributes that lead to the largest student test score gains at sixth grade lead to the lowest test score gains at fifth grade. Our main conclusion is that classroom observation and value-added models supply complementary and separately valuable information on what happens in classrooms. Neither one nor the other can be used in isolation, nor does averaging the results together retain enough information. In the best classrooms, both observation results and student test-score gains are favorable.

Note: This material will appear as a chapter in a research volume on the Measures of Effective Teaching project. It is reproduced with permission of John Wiley & Sons Inc. We are grateful to the Bill and Melinda Gates Foundation for funding that enabled us to carry out this work. Partial funding for the development and pilot studies of the UTOP were provided by the National Science Foundation through a Noyce Scholarships grant, DUE-0630376. Inclusion of the UTOP in the MET project was organized by the National Math and Science Initiative in cooperation with Laying the Foundations, partially funded by the Carnegie Corporation. We particularly thank John Winn, Matthew Valerius, and Christy Hovanetz for their assistance. We also thank members of the original UTOP pilot team, including: Mary Walker, Larry Abraham, Prerna Arora, Shasta Ihorn, and Jessica Gordon. The opinions expressed in this chapter are the responsibility of the authors and are not necessarily shared by the National Science Foundation, Bill and Melinda Gates Foundation, National Math and Science Initiative, or Carnegie Corporation.

1 INTRODUCTION

Urgent debates about teaching effectiveness pervade media channels (Gladwell 2008; Meyers et al. 2006), policy documents (Gordon et al. 2006), and educational research (Pianta and Hamre 2009), as we move into an era of high-stakes testing with teacher, student, and school-level accountability measures. Teachers have a considerable impact on student achievement (Heck 2008; Rivkin et al. 2005; Rowan et al.

2002; Sanders and Rivers 1996; Wright et al. 1997); however the classroom behaviors that contribute to effective teaching have proven surprisingly difficult to measure. Initiatives such as *Race to the Top* propose to measure teacher quality making use of value-added models, where teachers are evaluated based on changes in their students' standardized test scores (Sanders and Rivers 1996; Rivkin et al. 2005). 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. But there is not yet agreement on what the alternative measures might be.

Many researchers have argued for the use of classroom observations: "Placing validated, standardized observational assessment of teachers' classroom instruction and interactions more squarely in the realm of large-scale education science and in protocols evaluating the impacts of teacher education could have tremendous downstream consequences in terms of traction on questions that vex the field" (Pianta and Hamre 2009, p. 109). Thus a significant effort is underway to determine the relationship between teacher value-added scores and classroom teaching behaviors measured by observation protocols (Gates Foundation 2012).

We have been associated with UTeach, a program to prepare secondary mathematics and science teachers at UT Austin. Working with other scientists, educational researchers, and master teachers, we developed a classroom observation protocol called the UTeach Observation Protocol (UTOP). Following a pilot study where the UTOP was used to evaluate UTeach graduates (reviewed briefly here), the National Math and Science Initiative enabled us to collaborate with the Measures of Effective Teaching (MET) project (Gates Foundation 2012) and rate a large set of video lessons from the MET database using the UTOP. This collaboration allowed UTOP ratings to be tied to teacher value-added scores on both standardized state assessments and on assessments designed to measure conceptual understanding. The goal of this chapter is to explore the connections we found between observation scores and value-added measures and explain the conclusions we reached.

1.1 Preconceptions about Teacher Quality

From our vantage point within UTeach, preparing mathematics and science majors to become teachers, we came to this study with some preconceptions:

- Teaching quality is not a single number. Excellent pre-service teachers possess many different qualities, some of which are best found from observation, some of which are best found from their writings, and some of which are best found from their performance on examinations.

- When we must make binary decisions (to certify or not certify), we do not use a cutoff on a single continuous metric. In UTeach there are numerous critical checkpoints where sufficiently poor performance bars recommendation for certification.

- The components of pre-service teaching quality can be measured separately and are separately actionable. A pre-service teacher who fails chemistry can retake chemistry. A pre-service teacher who focuses all of her attention during a teaching experience on two loud students at the front of the class can learn from the mistake and involve the whole class the next time.

- Changes in secondary student test scores are neither viable nor necessary for the evaluation of preservice teachers, since the contact of pre-service teachers with public school students is too infrequent for them plausibly to be held responsible for results on these exams, and the legal context holds the classroom teacher of record accountable.

In the school reform context, by contrast, the preconceptions are different. Teaching quality is defined in terms of changes in student test scores. From the start we perceived a difficulty. In the Measures of Effective Teaching project, “Measures” is plural but “Effective Teaching” is singular. The project name implies that there are many ways to measure, but they are aiming in the end at one thing. Yet if teacher quality is defined in terms of student achievement, and student achievement is defined as the score coming from value-added models, then by definition, measurements from value-added models are perfectly correlated with true teacher quality and student achievement. The other measures end up without a compelling technical role, except perhaps to provide diagnostic guidance on how to raise test scores.

We agree that ultimately schooling must be judged by its value to students. Even so, we do not grant exclusive status to annual rises on test scores. Some student learning objectives, such as the ability to work with a group, or give an oral presentation on a project, are better evaluated during observation than on paper tests. Furthermore, the value of schooling should be judged by the final results at the end of twelfth grade. If critics of “teaching to the test” are right, then there may be teaching practices at one grade that raise the scores that year, but put students at a disadvantage later down the educational line. In Section 4.3 we will discuss specific cases where this may be happening. For the moment, we simply present the possibility that separate measures have separate value, and explain why investigation cannot begin by uncritically privileging one over another.

2 THE DEVELOPMENT OF THE UTOP

We developed the UTOP starting in 2006 in response to a requirement to evaluate National Science Foundation-funded Noyce Scholars within UTeach. We began by first looking for classroom observation protocols that assessed teaching behaviors consistent with the goals and foci of UTeach; not finding any, we decided to modify the Classroom Observation Protocol (COP, Horizons Research 2000b,a). We added, subtracted, and modified indicators. The instrument has experienced several revisions, but we refer to them all as the UTeach Observation Protocol (UTOP).The original version of the UTOP contained 32 indicators, each rated on a 1-5 scale with “Don’t Know” and “Not Applicable” options. The indicators were organized into 4 sections: Classroom Environment, Lesson Structure, Implementation, and Mathematics/Science Content. Each of the 4 sections concluded with a 1-5 Synthesis Rating which was intended to capture the observers ’ overall rating of the teaching behaviors in that section, without necessarily being a numerical average. For example, if the teacher spent the class period communicating incorrect content, the synthesis rating could be rated to reflect this more strongly than a numerical average of the ratings in the Content section would permit. The protocol also included an extended post-observation teacher interview, which posed a variety of questions about the context and events of the lesson. For the current, full version of the UTOP , visit http://uteach.utexas.edu/UTOP.

We initially tested the UTOP in a pilot study where we conducted observations in the classrooms of UTeach alumni. In order to provide a comparison group for our graduates, efforts were made to find teachers from other preparation backgrounds working in the same schools, teaching similar classes in their content area. Thirty-six teachers were observed (21 UTeach alumni and 15 from other preparations backgrounds) over 5 semesters, with a total of 83 observations. Seven of the UTeach alumni had received Noyce Scholarships, and thus might be considered to represent the top tier of UTeach students in terms of academic qualifications and commitment to teach. All teachers had less than 5 years teaching experience, and most were in their first or second year of teaching. Two raters, one with a background in mathematics and one in science, were present at most observations.[1]

The results of this comparative study showed promise for the UTOP’s ability to differentiate between teachers from different preparation backgrounds. Figure 1 (A) shows the average scores in each of the 4

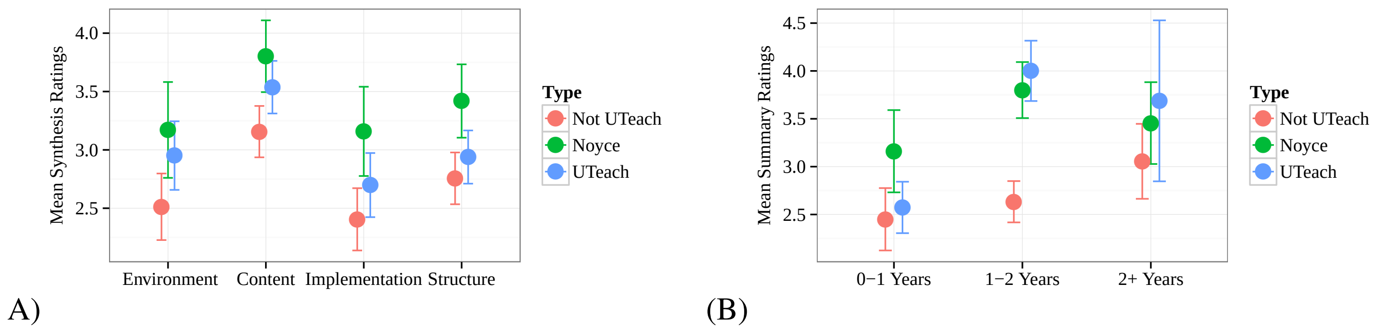

Figure 1: (A)Mean Synthesis Ratings (Table 2) for UTeach graduates who received Noyce Scholarships (N=7), UTeach graduates who did not receive Noyce Scholarships (N=14) and Non-UTeach graduates (N=15). Error bars are determined from the standard uncertainties produced by `lmer` function (Bates and Maechler 2010)in the R software package, specifying teacher identity as a random effect. (B) Mean Synthesis Ratings (Table 2) for Non-UTeach, UTeach (not Noyce), and UTeach Noyce Scholars as a function of number of years of classroom teaching experience. Standard errors are again produced by the `R` function `lmer` using teacher identity as a random effect. The small sample size of this pilot study means that no significant differences between teachers with more than two years of experience can be determined. Significant effects include the growth of non-Noyce UTeach graduates between their first and second years, and the difference between UTeach and non-UTeach graduates with 1-2 years of teaching experience.

UTOP sections for UTeach graduates who were Noyce Scholars (green), UTeach graduates who were not Noyce Scholars (blue), and graduates of other teacher preparation pathways (red). Noyce Scholars scored highest on the UTOP, with UTeach Non-Noyce in the middle, and non-UTeach at the bottom. Figure 1 (B) shows how the scores of the three groups of teachers varied as a function of years of teaching experience. A conclusion suggested by the data is that the difference between Noyce Scholars and other new teachers is most evident in the first year, and that UTeach alumni overall stand out more because of their growth over the first three years of teaching than because of their teaching practices at the outset. We note that despite attempts to keep observers from knowing the background of the teachers they watched, attempts at blinding were not uniformly successful, and this may have biased the scores.

While our pilot study showed promise in establishing the reliability and validity of the UTOP, there were several shortcomings. First, the sample size was too small to establish many statistically significant results. Second, the observers in this study were from a small, university-based research team rather than from the general population of observers who might be interested in using the UTOP, so their reliability might have been higher than could realistically be obtained in other contexts. Third, the sample was composed of a relatively small collection of volunteers who offered informed consent, making it unclear how well the results could be generalized. Fourth, there was no information available on the value-added gains of the teachers involved in the study, making it impossible to draw any links with test-based measures of student achievement. We were fortunate to be able to overcome several of these limitations when the UTOP was adopted for use with the MET project.

We made several modifications to the UTOP when we partnered with MET after this pilot study was conducted. First, we reduced the number of indicators from 32 to 22, since there were indicators that could not be well-captured in a video context without access to a teacher interview. Second, we created detailed scoring rubrics that supplied guidelines for 1-5 ratings on each indicator, and removed many of the “Don’t Know” and “Not Applicable” options. Third, using the data from our pilot study we wrote sample vignettes of supporting evidence that could be used to justify 1-5 ratings on each indicator. A group of 99 teachers assembled by the National Math and Science Initiative and Laying the Foundation used this modified instrument to rate 982 video-taped mathematics lessons from the MET project database.

3 THE UTOP IN THE MET PROJECT

3.1 Research Questions

The Measures of Effective Teaching (MET) project is a large initiative to measure teaching effectiveness that involved the collection of video observation data and results of value-added model computations from 3,000 teacher volunteers from across the country. In 2009, with assistance from the National Math and Science Initiative, we were given the opportunity to rate a subset of the MET Project's classroom videos — 1001 videos of classrooms in grades 4-8 mathematics — using the UTOP. In this section, we briefly discuss some of the key findings relating to the UTOP that were found in MET's report on the classroom observation instruments (Gates Foundation 2012). After the MET study released reports on the relationship between observation and value-added scores for teachers from the study, we were provided access to the MET dataset and performed additional analyses of the same data.

We found ourselves with a number of questions.

1. Figure 3 of Gates Foundation (2012) showed that teachers with the best UTOP scores obtained higher value-added scores than teachers who scored highly on other observation instruments in the MET study. Thus the UTOP seemed particularly equipped, compared to the other instruments, to identify the strongest teachers. Furthermore, value-added scores plotted versus UTOP score were fairly flat (did not have a positive slope) for low to medium UTOP scores; the rise was at at the top end of UTOP scores. Put another way, high UTOP scores seemed to be associated with high value-added scores, but the UTOP did not discriminate well among teachers with low to medium value-added scores. Finally, the figure in the MET report provided no estimate of measurement uncertainty (error) of the value-added scores, making it difficult to determine whether apparent differences for different UTOP levels were statistically significant. We decided to investigate these points.

2. The UTOP measures 22 different aspects of classroom behavior, but in the MET report none of these individual behaviors were compared to teacher value-added gains. The report only compared a summary UTOP rating to teacher value-added scores. We wondered if any of these 22 behaviors stood out in connection with raising student test scores. If so, we might be able to use this information to help teachers improve.

3. Overall, the connections between value-added model scores and observation scores were fairly weak for the UTOP as well as the other observation instruments. There were many cases where value-added scores and observation scores were not in agreement. We sought a way to consider results from the UTOP and value-added models on an equal footing, and to examine carefully lessons where these two measures agreed or disagreed.

While our investigation proceeded essentially as planned, we modified our questions as analysis proceeded. Most importantly, we found that results from different grade levels were substantially different from one another and it often seemed best to disaggregate by grade before performing other analysis. We settled on the following sharpened research questions:

R1. What is the relationship between teacher value-added gains of teachers involved in the MET study and UTOP observation scores for these same teachers, for teachers with different levels of UTOP scores? In particular, how do the relationships between value-added models and UTOP scores change as one focuses on subgroups of teachers who have different levels of observation or value-added scores?

R2. What is the relationship between teacher value-added gains of teachers involved in the MET study and UTOP scores for these same teachers on the 22 individual UTOP indicators? In particular, how do the results depend upon grade level taught (4-8)?

R3. Using a subset of teaching behaviors from the UTOP that we present as a minimal or "consensus view" of effective teaching, what is the level of consistency between value-added measures of teaching effectiveness from teachers in the MET sample and UTOP classroom observation measures from these same teachers? What are the characteristics of classrooms where these measures strongly agree or strongly conflict?

3.2 Rating Method

Over the course of 7 weeks, 99 raters scored 1001 videos of grades 4-8 mathematics lessons from the MET video library using the UTOP. Of the 99 raters, 41 raters had backgrounds in science, and 58 were from mathematics backgrounds. These raters were highly qualified master teachers; the average rater had 19 years of teaching experience, and over half held Master's degrees or higher. For more information on rater background characteristics, see Gates Foundation (2012). All ratings were entered into an online version of the UTOP on SurveyMonkey.

The raters were trained through a process where they first watched and rated videos from the MET database in large and small groups, and discussed their ratings with other raters and UTOP developers. The raters were introduced to the scoring rubrics for each UTOP indicator, as well as the example vignettes of supporting evidence for each indicator. Trainees would later be given the standard "normed" UTOP ratings for each video, with supporting evidence cited for each rating. This training process lasted for a little under two days, at which point the raters were considered ready to rate on their own. However, as they rated videos, one third of the videos they rated would be double-scored, and they would have to discuss their ratings with the other rater who rated the same video after their original scores were entered. The two raters were asked to come to agreement on all indicator and synthesis ratings and record it. Of the 1001 videos, 331 were double-scored, and 10 were triple-scored. For the analyses presented here, if a lesson had two or three raters, we use the UTOP scores they agreed upon after discussion. If a lesson only had one rater, we use this rater's UTOP scores. There were 8 cases where two raters scored a lesson, but they did not subsequently discuss their ratings. In these cases, we use the average of the two rater's scores.

Video Version of UTOP The version of the UTOP used in the MET study contained 22 indicators, shown in Table 2 below. See Gates Foundation (2012) for descriptive statistics on how teachers in the sample scored across these 22 indicators. The short descriptions of the indicators are used in all figures in this chapter.

Indicator	***Short Description***	***Indicator Group***

Indicator	***Short Description***	***Indicator Group***
Section 1: Classroom Environment Synthesis	**Environment Synthesis**	**Synthesis**
1.1 The classroom environment encouraged students to generate ideas, questions, conjectures, and/or propositions that reflected engagement or exploration with important mathematics concepts.	Ideas	Innovative
1.2 Interactions reflected collegial working relationships among students.	Interactions	-
1.3 Based on conversations, interactions with the teacher, and/or work samples, students were intellectually engaged with important ideas relevant to the focus of the lesson.	Engagement	-
1.4 The majority of students were on task throughout the class.	On-Task	Consensus
1.5 The teacher's classroom management strategies enhanced the classroom environment.	Management	Consensus
1.7 The classroom environment established by the teacher reflected attention to issues of access, equity, and diversity for students (e.g. cooperative learning, language-appropriate strategies and materials).	Equity	Consensus
Section 2: Lesson Structure Synthesis	**Lesson Synthesis**	**Synthesis**
2.1 The lesson was well organized and structured.	Organized	Consensus
2.2 The lesson allowed students to engage with or explore important concepts in mathematics (instead of focusing on techniques that may only be useful in exams).	Important	Consensus
2.3 The lesson included an investigative or problem-based approach to important concepts in mathematics.	Inquiry	Innovative
2.4 The teacher obtained and employed resources appropriate for the lesson.	Resources	Consensus
2.5 The teacher was critical and reflective about his/her practice after the lesson, recognizing the strengths and weaknesses of their instruction.	Reflection	-
Section 3: Implementation Synthesis	**Implementation Synthesis**	**Synthesis**
3.1 The teacher's questioning strategies developed student conceptual understanding of important mathematics content.	Questioning	Innovative
3.2 The teacher used formative assessment effectively to be aware of the progress of all students.	Assessment	-
3.3 The teacher involved all students in the lesson.	Involvement	Consensus
3.4 An appropriate amount of time was devoted to each part of the lesson.	Timing	-
Section 4: Mathematics Content Synthesis	**Content Synthesis**	**Synthesis**
4.1 The mathematics content chosen was significant and worthwhile for this course.	Worthwhile	Consensus
4.2 During the observation, it was made explicit to students why the content is important to learn.	Explicit	Innovative
4.3 Content communicated through direct and non-direct instruction by the teacher is consistent with deep knowledge and fluency with the mathematics concepts of the lesson.	Fluent	Consensus
4.4 Teacher written content information was accurate.	Correct	-
4.5 Elements of mathematical abstraction were used appropriately in the lesson.	Abstraction	-

Indicator	***Short Description***	***Indicator Group***
4.6 Appropriate connections were made to other areas of mathematics or to other disciplines.	Connections	Innovative
4.7 During the lesson, there was discussion about the content topic's role in history or current events.	Society	Innovative
Summary: Mean of 4 Synthesis indicators	Summary	
Consensus: Mean of 9 Consensus indicators	Consensus	
Innovation: Mean of 6 Innovative indicators	Innovative	

Table 2: List of UTOP indicators used in MET video ratings. **Consensus** indicators are intended to be a subset that almost all reasonable observers would consider an essential component of effective teaching; 4.4 was omitted only because it was so frequently indicated as 'NA' when written materials were not visible. **Innovative** indicators are those that reflect qualities of classrooms valued within UTeach but not necessarily shared by all observers.

As can be seen from Table 2, for some of our analyses we found it useful to conceptually group different indicators. We selected a subset of 9 indicators from the UTOP that plausibly represent a consensus view of acceptable teaching (rightmost column in Table 2). That is, even if it is not taken as settled whether inquiry or direct instruction is to be preferred, it would be difficult to find anyone to defend poor performance on these particular indicators. For example, it would be hard to defend a lesson where much of the class is not paying attention, or where the teacher makes mathematical errors during an explanation. What we call the "Consensus score" for a lesson is the average of these 9 UTOP indicators. We also selected six indicators that represent UTeach faculty values but are not universally accepted as necessary components of effective teaching and labeled them "Innovative" – these are teaching behaviors that are commonly encouraged by mathematics educators, but they are not universally accepted.

Sample of Video Lessons The original sample was 1001 video lessons; however 6 were excluded because of severe audio or video problems, or because the video did not show a mathematics lesson. An additional 13 videos were omitted because they did not have corresponding value-added data, for a final video count of 982. The grade levels of the 982 video mathematics lessons are shown in Table 3. The 982 lessons were from the classrooms of 249 teachers. Two hundred and thirty-seven teachers had 4 videos of their classroom scored with the UTOP, while 10 teachers had 3 lessons and 2 teachers had 2 lessons. Video lessons typically lasted between 50 minutes and 1 hour, and all covered elementary or middle grades mathematics content. Teachers were volunteer participants from 6 school districts in 6 different states; for more demographic characteristics of the teachers see Gates Foundation (2012).

The 249 teachers in the sample taught a total of 409 class sections (i.e., consistent groups of students in a contained mathematics class); 89 teachers were filmed teaching only one class section, while 160 teachers were filmed teaching two class sections. For a given class section, teachers were filmed from 1 to 4 times: specifically 12 class sections were filmed once, 305 were filmed twice, 8 were filmed three times, and 84 were filmed four times. Each of these class sections corresponded to one set of value-added scores. We aggregated the scores from different UTOP observations of the same class section in three different ways, recording for each UTOP indicator the maximum, the minimum, and the average. In what follows, we have settled on use of the average, although it is not hard to construct arguments that for some indicators teachers should be accountable for the worst behavior that is seen, while for others

Grade Level	*Number of Videos*	*Number of Class Sections*	*Number of Teachers*
Grade 4	189	55	48
Grade 5	211	58	52
Grade 6	200	103	52
Grade 7	209	104	52
Grade 8	173	89	45

Table 3: Grade level of 982 video lessons scored on UTOP.

they should get credit for the best. We also aggregated all comments of all raters of all videos for each class section for use in our qualitative analysis.

3.3 Descriptive Statistics of UTOP and Value-Added Models

We computed averages (Table 4) and examined a variety of statistical models for UTOP indicators and value-added scores (Table 5). The value-added variables used in the analyses were either teacher value-added gains on State standardized tests in 2009 (the year before the video observations; State VAM 2009), in 2010 (the year of the video observations; State VAM 2010), or the teacher's value-added gains on the Balanced Assessment of Mathematics in 2010 (BAM 2010).

[2] The most important observations are:

1. The average UTOP scores in the MET sample are significantly lower than the scores observed in the UTOP pilot study (Figure 1). UTeach graduates with 1-2 years of experience in the pilot study had overall UTOP scores near 4, while typical UTOP scores in the MET sample are near 2.5, and drop going from elementary to middle school.
2. We found dependence of the UTOP scores on fraction of minority students and fraction of low-income students in the classroom; quality in the classroom went down as the fraction of minority students and low-income students went up. This is not true for any of the Value-Added Model scores, since the value-added models control for these demographic variables.

Both these results are disquieting. What sorts of lessons does one learn from the sample of MET teachers if on average they have so much room to improve? If observation scores systematically depend upon the concentration of student poverty, are observation scores fair? Should one systematically compensate?

Table 6 displays the distribution of observations organized by district and grade level. There are three districts for which several grade levels were not sampled at all. Upon seeing these results we were initially concerned, since as we will show in Section 4.2 there are strong dependencies upon grade level. The very inhomogeneous relation between district and grade level raised the possibility that the apparent dependence upon grade level was in fact a dependence upon district. To examine this scenario we repeated the entire analysis of Section 4.2 by restricting ourselves to Districts 4, 5, and 6, for which all available grades are sampled fairly evenly. The results were unchanged, except that the margins of error increased due to decreased sample size. A particular example is discussed in the caption to Figure 7. We do not otherwise mention this point further.

4 RESULTS

We now investigate each of the three research questions in turn. First, we examine the overall relationship between UTOP observation scores and teacher value-added gains for teachers in the MET sample,

Grade Level	*Avg. UTOP Summary (St. Dev)*	*Avg BAM 2010 (St. Dev)*	*Avg State VAM 2009 (St. Dev)*	*Avg State VAM 2010 (St. Dev)*
Grade 4	2.75 (0.671)	−0.036 (0.207)	0.029 (0.205)	0.002 (0.178)
Grade 5	2.76 (0.658)	0.065 (0.195)	0.009 (0.222)	0.023 (0.205)
Grade 6	2.51 (0.721)	−0.008 (0.274)	-0.008 (0.200)	−0.002 (0.253)
Grade 7	2.39 (0.649)	0.010 (0.292)	0.022 (0.113)	0.0003 (0.156)
Grade 8	2.33 (0.662)	−0.007 (0.231)	0.037 (0.165)	0.015 (0.170)

Table 4: UTOP Summary rating and average value-added gains for lessons in grades 4-8.

Group	Coefficient	Standard error
White	0.59	0.09
Gifted	0.41	0.15
Black	−0.29	0.09
Hispanic	−0.38	0.09
Low-Income	−0.43	0.10

Table 5: Simple regression models of UTOP Summary scores as a function of student demographic variables, all of which are fractions in the range [0,1]. The coefficients are slopes describing variation of the UTOP Summary across all grades as a function of the fraction of White, Gifted, Black, Hispanic, and Low-Income (eligible for free or reduced lunch) students in class.

examining the possibility of nonlinear relationships between the two measures. Second, we look at the relationship between UTOP scores on individual indicators and teacher value-added gains. And finally, we investigate both quantitatively and qualitatively the cases where observation scores and value-added models agree and disagree.

4.1 Overall Relationships between UTOP scores and Value-Added Method Scores

Our first research question addressed the relationship between value-added scores from teachers in the MET sample and UTOP scores from rating video lessons of these same teachers. We began this analysis by examining a relationship featured in the research paper from the second-year MET report, which displayed average value-added score as a function of teacher observation score ranked by percentile (the upper right panel of Figure 3 in Gates Foundation (2012). We were struck by the fact that for the best teachers, the UTOP was associated with higher gains on BAM 2010 than other observation instruments, a gain in valued-added score of around 0.11. Note that for all its value-added models, the MET project proposes that a gain of 0.25 corresponds approximately to a gain of 9 months of schooling.

To check this result, we first divided the 409 class sections up into 10 groups (deciles) based on the Summary rating for that class section, and plotted UTOP Summary decile versus BAM 2010. We also made similar plots for each of the 4 UTOP Synthesis ratings separately, and for the Consensus and Innovative indicator groups. We followed the plots with statistical analyses that examined whether different linear slopes and intercepts would be appropriate for different regions of these plots — if the relationship between UTOP and value-added would be significantly different depending on whether UTOP scores

District/Grade	4	5	6	7	8
1	0	0	26	32	30
2	10	10	8	0	0
3	8	12	0	0	0
4	17	16	32	28	10
5	10	10	17	20	19
6	10	10	20	24	30

Table 6: Numbers of classes observed by district and grade level. There is no grade for which all districts participated. There are three districts for which all grades were observed.

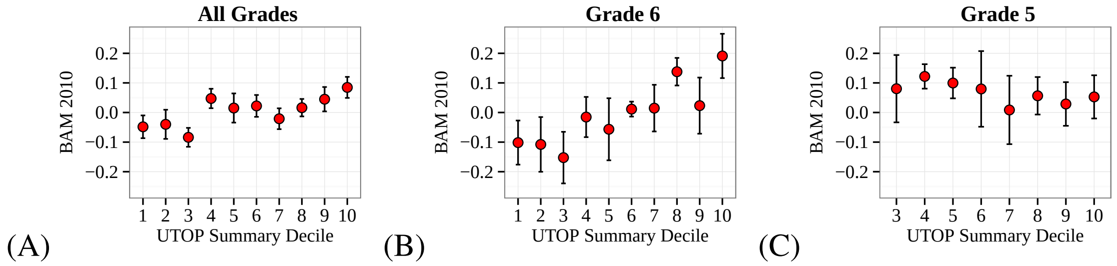

Figure 2: Average BAM value-added score versus decile ranking on UTOP Average Summary indicator (A) All grades. (B) Sixth grade only. (C) Fifth grade only.

were relatively high or relatively low. We tested for this difference by fitting linear models with grade level, UTOP score, UTOP scoring region, and the interaction of UTOP score and UTOP scoring region as predictors of value-added gains. UTOP scoring regions were obtained by dividing normalized UTOP scores into regions using a variety of intervals - including deciles, quantiles, thirds, and halves.

Figure 2(A) shows the results of dividing teachers into deciles according to the Summary UTOP score and plotting the average BAM 2010 Value-Added score for each decile. We find the same difference shown in the MET Report between top and bottom teachers, and a relatively clear positive relationship between BAM value-added score and UTOP Summary score for teachers in the 7th to 10th decile of UTOP scores. However, the relationship between UTOP Summary score and BAM Value-Added score is not particularly smooth, and standard errors on the order of 0.05 limit the number of cases in which difference between teachers in different deciles could be called significant. Figure 2(B) shows the same relationship, but restricted to sixth grade where we find that the correlation between UTOP and BAM value-added is the strongest. Indeed now the difference between lowest- and highest-ranked teachers is 0.3, or “11 months of schooling”. For fifth grade, as shown in Figure 2(C), value-added scores if anything decrease as UTOP scores increase, but none of the differences is significant. We tested whether we could use statistical models to fit different slopes to different regions of the graph in Figure 2A — perhaps allowing for a flatter slope for the first two-thirds of the UTOP deciles, and a steeper slope for the final one-third. However, the uncertainties were too high to permit us to detect any significant differences of this kind.

We next proceeded to examine the various UTOP composite and Synthesis ratings in the same manner. Figure 3 shows BAM 2010 value-added scores versus the Consensus, Summary, and Innovation groupings of UTOP indicators. The strongest correlation with BAM value-added scores arises for the Consensus indicator that focuses on competence in routine classroom practices, and the weakest correlation arises for the indicators we associated with innovative teaching practices. One reason that the

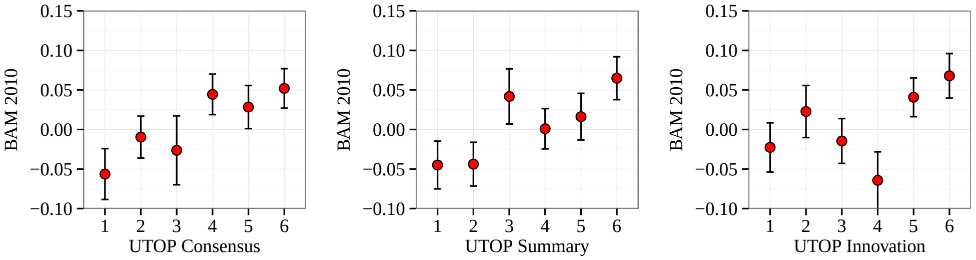

Figure 3: Average BAM 2010 value-added score versus quantile ranking on UTOP Consensus, Summary, and Innovation indicators

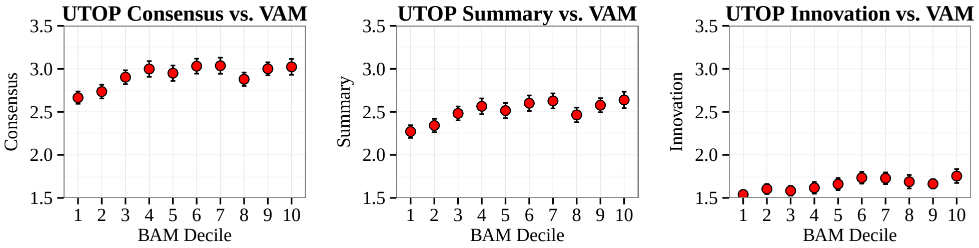

Figure 4: Average UTOP Consensus, Summary, and Innovation scores versus BAM value-added score decile rankings.

Innovation indicator leads to such inconclusive results becomes more apparent from Figure 4, which inverts the axes and plots mean UTOP composite scores versus BAM decile. What now is obvious is that the overall scores on the Innovation indicator are extremely low. One cannot expect to learn how these supposedly innovative practices affect value-added scores unless one has a reasonable sample of teachers performing them well.

Figure 5 shows BAM 2010 value-added as a function of UTOP scoring quantile on the 4 Synthesis ratings. The Classroom Environment and Lesson Structure Synthesis ratings show a somewhat steady growth of value-added score with UTOP score. Attempts using statistical models to fit different slopes to different regions of the graphs for Synthesis, Consensus, and Innovative ratings were not successful.

We conclude that the teachers with the highest UTOP scores do indeed have higher value-added scores than those with the lowest. However, the measurement uncertainties are large, and attempts to identify a pattern in how value-added scores depend upon UTOP scores were not successful.

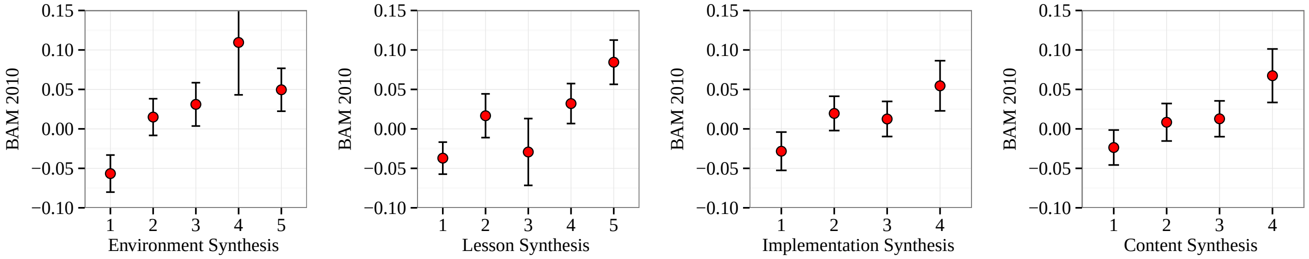

Figure 5: BAM value-added score versus ranking on UTOP Synthesis ratings. The use of four or five groups results from attempting to group into quintiles with the R function `cut2`.

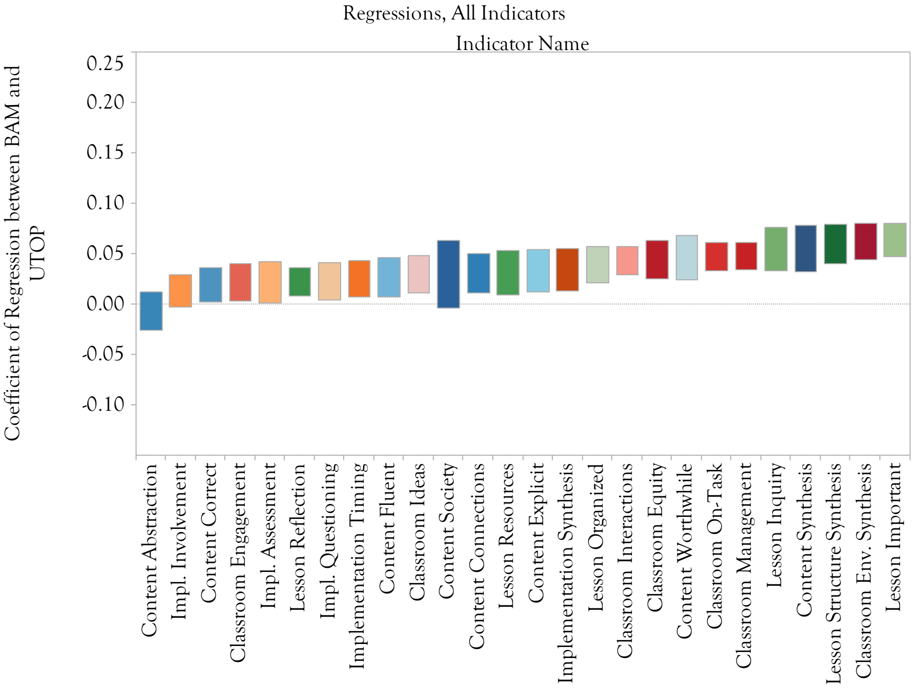

Figure 6: Regression slopes for all UTOP indicators showing their relationship to value-added on BAM 2010. Each bar has a height of two standard uncertainties.

4.2 Correlation between UTOP indicators and Value-Added Scores

Our second research question concerned the relationship between value-added gains and scores on different individual UTOP indicators. We wanted to know whether particular teaching behaviors or classroom characteristics were associated with improved student test scores[3]. Table 7 displays coefficients showing all the cases in which improvement of a particular UTOP indicator was associated with a statistically significant increase in value-added model score. Figure 6(A) shows the same data for BAM value-added only, in graphical form[4]

Table 7 highlights in red those indicators that achieved the highest level of statistical significance for all three value-added measures. Three of these indicators (On-Task, Management, Classroom Environment Synthesis) describe orderly, supportive, well-managed classrooms. The fourth (Important) is the indicator that specifically examines whether a classroom session focuses upon test preparation, or on important mathematical ideas. Thus, if we have to say in general, independent of grade level, what features of classroom performance to focus on in the hopes of raising test scores, these are the ones. .

Most of the UTOP indicators have a significant and positive relationship with value-added scores on the 2009 and 2010 State assessment. The value-added scores derived from state tests seem to emphasize procedural skill somewhat more than the scores from BAM. For example, value-added scores have a significant relationship with **Inquiry** in the BAM results, but not for the state tests, while **Timing** shows significant correlation for the state tests and not for BAM. In the following analysis we focus on BAM, partly because we value its emphasis on conceptual understanding, and partly because it was a single exam administered across the whole study, rather than a collection of disparate examinations from different states.

The table of regression coefficients is overall not very illuminating because patterns do not always leap out and because the different value-added models do not always agree. In Figure 7 we focus on

UTOP Indicator	*BAM 2010*	*State VAM 2010*	*State VAM 2009*
1.1 Ideas	NS	.0467 (.0138) ***	.0432 (.0126) *
1.2 Interactions	.0427 (.0140) **	.0361 (.0113)**	.0410 (.0100) ***
1.3 Engagement	NS	.0398 (.0138) **	.0335 (.0126) **
1.4 On-Task	.0473 (.0141) ***	.0563 (.0106)***	.0477 (.0096) ***
1.5 Management	.0476 (.0136) ***	.0562 (.0103)***	.0468 (.0093) ***
1.7 Equity	.0442 (.0189) *	.0414 (.0144)**	.0527 (.0129) ***
2.1 Organized	.0393 (.0183) *	.0492 (.0139)***	.0427 (.0126) ***
2.2 Important	.0638 (.0166) ***	.0438 (.0128)***	.0452 (.0115) ***
2.3 Inquiry	.0545 (.0218) *	.0334 (.0168) *	.0398 (.0151) **
2.4 Resources	NS	NS	.00588 (.00150) ***
2.5 Reflection	NS	NS	NS
3.1 Questioning	NS	.0403 (.0141) **	NS
3.2 Assessment	NS	.0352 (.0157) *	.0365 (.0142) *
3.3 Involvement	NS	.0402 (.0123) **	.0412 (.0113) ***
3.4 Timing	NS	.0509 (.0138) ***	.0368 (.0126) **
4.1 Worthwhile	.0464 (.0219) *	.0434 (.0168) *	.0440 (.0153) **
4.2 Explicit	NS	.0437 (.0161)**	.0431 (.0145) **
4.3 Fluent	NS	.0439 (.0149) **	.0505 (.0134) ***
4.4 Correct	NS	NS	NS
4.5 Abstraction	NS	NS	.0316 (.0123) *
4.6 Connections	.0468 (.0196) *	.0554 (.0149) ***	.0440 (.0136) *
4.7 Society	NS	NS	.0663 (.0232) **
Section 1: Classroom Environment Synthesis	.0622 (.0183) ***	.0618 (.0139) ***	.0521 (.0127) ***
Section 2: Lesson Structure Synthesis	.0595 (.0198) **	.0612 (.0151) ***	.0559 (.0137) ***
Section 3: Implementation Synthesis	NS	.0671 (.0156) ***	.0543 (.0142) ***
Section 4: Mathematics Content Synthesis	.0552 (.0232) *	.0565 (.0177) **	NS

Table 7: Tabulation of all significant correlations between individual UTOP indicators/Synthesis ratings and value-added scores from the Balanced Assessment of Mathematics (BAM) and 2010 and 2009 State mathematics exams. Each column shows the significant regression slope coefficients, while parentheses show the uncertainty in the slope estimate. The asterisks indicate the level of significance (* = $p < .05$, ** = $p < .01$, *** = $p < .001$). NS means "Not Significant." Entries in red showed the highest level of statistical significance for all three value-added models.

a single indicator, 2.3, which concerns investigative or problem-based approaches to mathematics and examine it in more detail. Here are some lessons from the figure:

1. The graphs associated with the three different exams are strikingly similar to one another. The three exams seem fundamentally to be measuring the same thing.

2. Test scores vary quite differently with inquiry practices in different grade levels: negatively at fifth grade, positively in sixth grade. This is one reason why when all grades are lumped together this indicator does not emerge as strongly associated with value-added scores.

3. There is not an extremely obvious visual pattern related to the color of the dots. Therefore, the relationship between poverty concentration and classroom performance, although it exists, is weak.

4. The individual class sections form a cloud surrounding the regression line, rather than clustering tightly upon it. This is the meaning of a correlation coefficient $R \approx 0.3$

Having found that the relationship between UTOP scores and value-added scores depends upon grade level, we turn to a more systematic investigation of this point. We focus on fifth and sixth grades, which are more different from each other than any two other grade pairs. Figure 8, displays regression slopes for each UTOP indicator. The main messages from this figure are

1. By and large, the better fifth-grade teachers look according to the UTOP, the lower their student test score gains[5].

2. By and large, the better sixth-grade teachers look according to the UTOP, the higher their student test score gains.

3. In many cases, the teaching practices associated with the best score gains at sixth grade (for example Inquiry) are associated with the lowest score gains at fifth grade, and vice versa.

A similar pattern persists for many other indicators and for the three separate value-added models; correlations with UTOP are most positive at sixth grade, have an intermediate value in fourth and seventh grade, are weaker in eighth grade, and weakest of all in fifth grade[6].

What are some possible explanations? In our sample, all the fifth grade scores are from self-contained elementary schools, while all sixth grade scores are from middle schools. In middle school, 6th grade students are making a difficult transition, leaving the support of self-contained classrooms. The teaching behaviors valued by the UTOP may be especially important during this transition. The nature of the mathematical content itself also varies according to grade level — 5th grade may be a year where students are still focusing on concrete, more calculational aspects of mathematics. Sixth grade, by contrast, can potentially mark the beginning of a transition to higher-level, more abstract mathematical content. This may be a key point in students' mathematical development where the behaviors on the UTOP are most important. While UTOP scores had the strongest relationships to value-added in 6th grade, they seemed to have especially weak or even negative relationships in 5th and 8th grade. We note that in some states 5th and 8th grade state tests have particularly high stakes in that they constitute key exit exams that the students cannot progress to the next grade level without passing. Regardless of the official state testing procedures, in the 5th and 8th grade years at all schools in the sample students are being evaluated on whether they are ready to exit one school setting (elementary or middle school), and enter another (middle or high school). We investigate the 5th versus 6th grade difference qualitatively at the end of the next section.

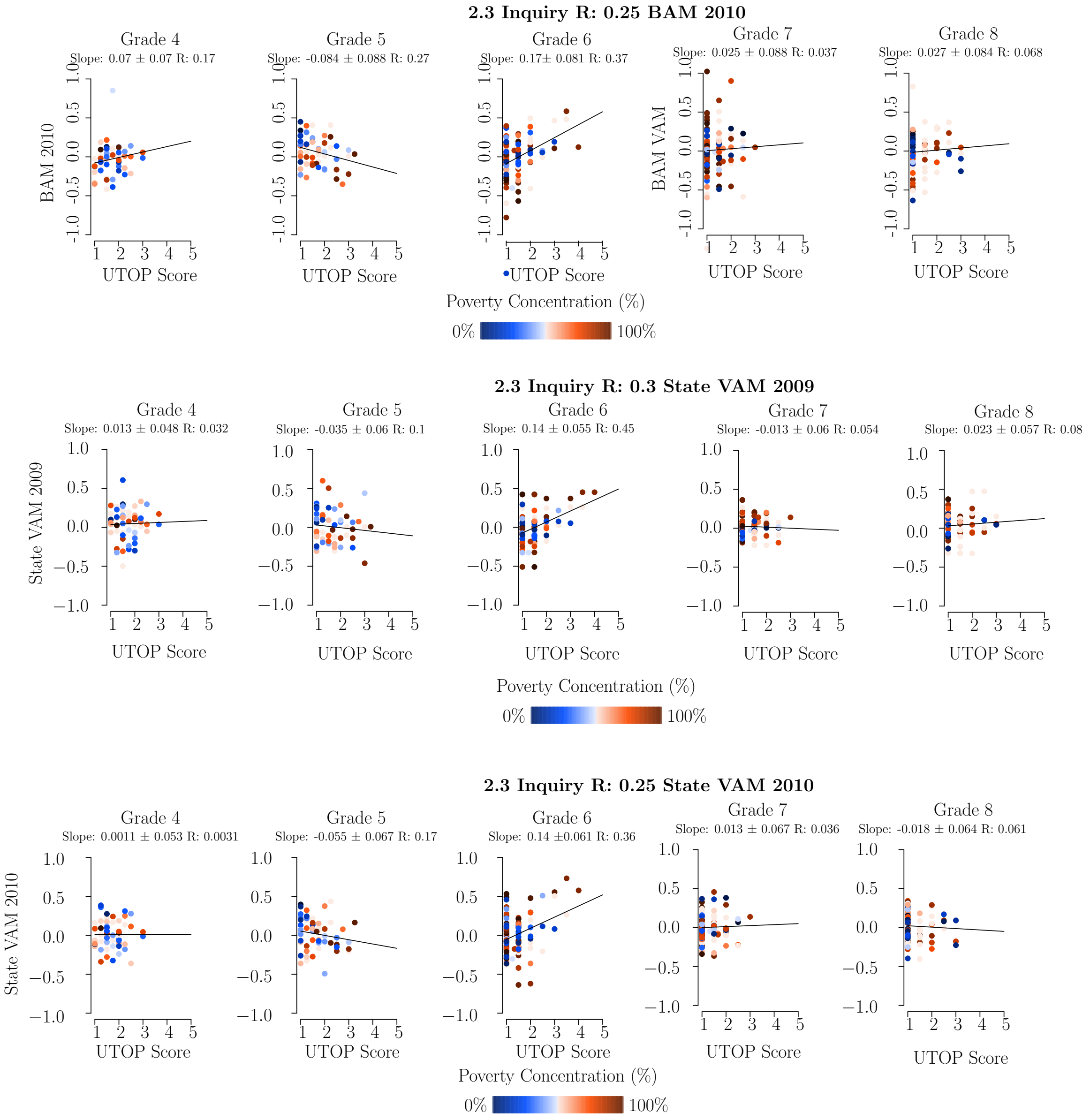

Figure 7: Scatter plots showing the relationship between the UTOP **Lesson Inquiry** Indicator *("The lesson included an investigative or problem-based approach to important concepts in mathematics")* and BAM 2010, State 2010, and State 2009 scores. Regression lines are shown, together with estimates for uncertainty of the slope. Separate teacher classrooms are treated completely independently; including teacher identity as a random effect makes no appreciable difference. When restricted only to districts 4, 5, and 6 (see Table 6 and accompanying discussion) the regression slopes change only in the second decimal place. For example, the slope for sixth grade becomes 0.16±0.098 with $R = 0.4$, the slope at grade 7 becomes 0.06±0.1 with $R = 0.095$. This and subsequent plots color-code classes according to student poverty concentration. The overall lesson of the color coding is that in all cases the relationship between UTOP score and poverty is weak, as patterns do not stand out.

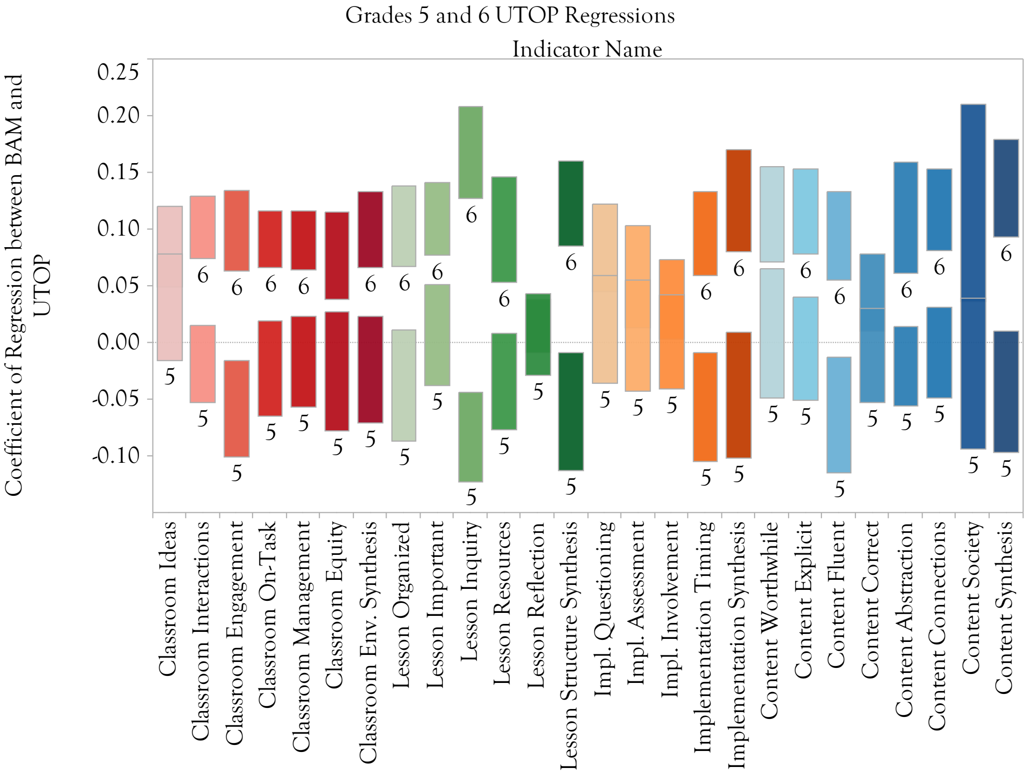

Figure 8: Regression slopes for all UTOP indicators with respect to BAM value-added scores in 5th and 6th grade. The height of each bar is two standard uncertainties.

4.3 Qualitative Analysis of UTOP and VAM Scores Treated on Equal Footing

Our final research question led us to grapple directly with the relatively low level of consistency between observation scores and value-added scores. One possible conclusion would be that observation scores are weak and unreliable measures of student achievement, and in view of their expense should be discarded. This conclusion is partly created by a rhetorical device, the use of the words *student achievement* for scores from value-added models. Some additional terminology can help emphasize that computer-scored tests are not the only valuable student outcome. Our terminology was suggested by the observation that sometimes teachers can raise student test scores by teaching them content errors that are helpful on exams[7], but *this is not acceptable*.

In Table 2, we defined UTOP indicators that represent a "consensus" view of effective teaching. Figure 9 plots the BAM 2010 Value-Added scores versus the Consensus indicator. Patterns similar to those seen for other indicators, such as in Figure 7 are evident here too. Because the Consensus indicator was chosen to represent classroom attributes that almost any reasonable person would view as essential, we used it to define Acceptable and Unacceptable teaching.

Specifically, we define Acceptable Teaching to correspond to an average score of 3 or more on the Consensus indicators and Unacceptable Teaching to be a score less than 3. By contrast, we define Effective Teaching to correspond to a positive score from Value-Added Models on BAM 2010 and Ineffective Teaching to correspond to negative scores. Thus we end up with four possibilities for each class section rated with the UTOP and BAM value-added scores: Unacceptable Ineffective Teaching, Unacceptable

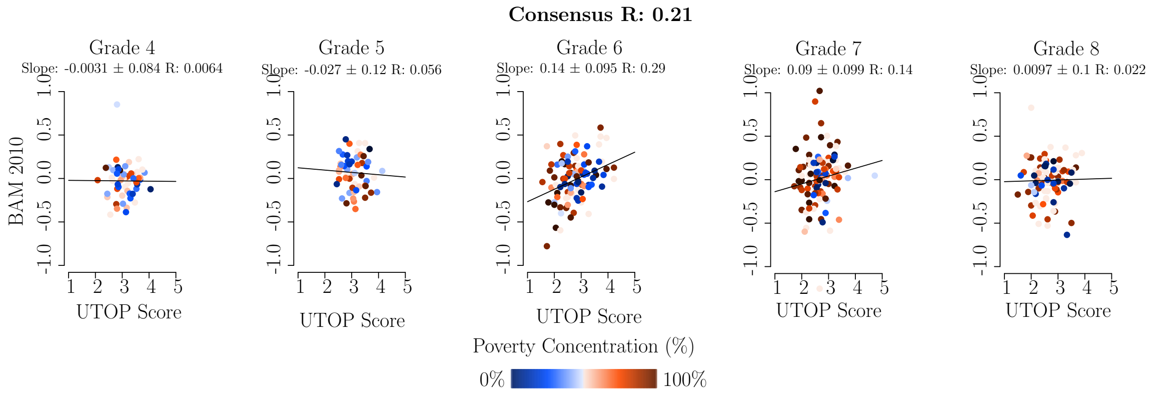

Figure 9: Scatter plots showing the relationship between the UTOP **Consensus** Indicator for Acceptable teaching and BAM 2010 value-added. Regression lines are shown, together with estimates for uncertainty of the slope. Separate teacher classrooms are treated completely independently; including teacher identity as a random effect makes no appreciable difference.

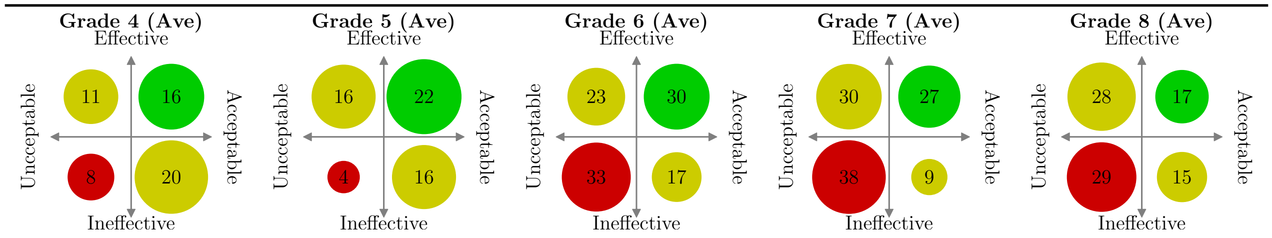

Figure 10: Quadrant plots showing distributions of class sections according to the four possible combinations of Acceptable/Unacceptable and Effective/Ineffective teaching. The area of each circle is proportional to the percentage of class sections at that grade level in each quadrant, while the number of class sections is indicated inside.

Effective Teaching, Acceptable Ineffective Teaching, and Acceptable Effective Teaching.

The distribution of teachers according to these quadrants is shown by grade level in Figure 10. Many things stand out in this figure:

1. Acceptable but Ineffective teachers are most likely in 4th and 5th grade, compared to 6th, 7th, and 8th grade.

2. Effective but Unacceptable teachers are most likely in 5th, 7th and 8th grade.

3. Unacceptable Ineffective teachers are most likely in 6th, 7th, and 8th grade, while Acceptable Effective teachers are most likely in 4th and 5th grade. Grade school looks much better than middle school, viewed this way.

Having divided all class sections into these four quadrants, we proceeded to carry out a qualitative investigation, making use of the sometimes extensive comments provided by UTOP raters for each indicator. For example, for the UTOP indicator about classroom management, raters typically wrote supporting evidence that cited specific instances of misbehavior by students, and particular techniques the teacher used in order to control behavior. To select class sections for the analysis, we calculated the average BAM 2010 and average UTOP Consensus rating for each class section and normalized both measures.

We defined Deviance as the difference between normalized BAM 2010 and normalized UTOP Consensus score. We selected the 6 class sections that had the highest and lowest Deviance – those with high normalized BAM and low normalized UTOP (Unacceptable Effective) and those with a low normalized BAM and high normalized UTOP (Acceptable Ineffective). For comparison purposes, we also looked at class sections with especially high BAM and UTOP (Acceptable Effective), and especially low BAM and UTOP (Unacceptable Ineffective)[8].

We now provide a narrative that describes the characteristics of the classrooms with the highest and lowest Deviance between observation scores and value-added scores in the four quadrants[9].

Quadrant 1: High BAM, Low UTOP (Unacceptable Effective Teachers) These classrooms were sometimes characterized by an explicit focus on standardized test preparation during the lesson. In some cases test preparation was isolated to the warm-up, in other cases it was the focus of the lesson. One observer writes "[The] teacher moved from warm up to independent practice. Students understood lesson was to practice concepts that could appear on tests." Some observers noted that these lessons focused on training students to perform procedures without understanding, rather than conceptual development of significant mathematical ideas. There was also little attempt at formative assessment or higher-level questioning in these classrooms. "There was no exploration in this lesson. The student[s] went through the procedure of solving Pythagorean theorem problems. The teacher asked fill in the blank questions."

The teacher made major content-related errors in one or more observations of half of these class sections — for example, one observer wrote that "The teacher [made] many errors as she communicated the content of the course. She emphasized that the base is the straight line (not line segment) that is on the bottom of the triangle and that height of the triangle had to be measured from the top of the triangle to the bottom. She also indicated that it is a straight line rather than a slanted line." The other observer from the same lesson noted that the teacher told students the height was the tallest part of the triangle, resulting in much student confusion. Teachers in other class sections simply seemed to be unable to deal with student confusion surrounding the mathematics content, and as a result students seemed to leave the lesson without a firm grasp on the concepts. One observer writes that "Teacher didn't deal with all the students who were having problems. He didn't address or try to correct their understanding. He just showed more problems to work."

In 4 of these 6 class sections, at least one observer noted the teacher treating their students disrespectfully. One observer writes "There were several occasions where students were confused and asked for help and the teacher either dismissed the student or became confrontational with the student" while another describes a different classroom where "[The teacher's] demeanor toward some students was adversarial, demeaning, and inappropriate. 'You're just making a mess. You're like a big disaster. What are you doing!'" Overall, these lessons often had a major weakness – namely disrespectful behavior, content mistakes, or very low student engagement with the concepts – that compromised their UTOP scores.

Quadrant 2: Low BAM, High UTOP (Acceptable Ineffective Teachers) These classrooms were also sometimes characterized by an explicit focus on standardized test preparation. "This lesson was review for an EOG. The students were not uncovering any new concepts. The students were just performing calculations given the formulas." In nearly all of the class sections (5 out of 6), observers commented that instruction was procedural, and focused on direct instruction and guided practice. One observer writes "The quality of the interactions and engagements was consistent, but the intellectual level stayed at a low level for most of the students. While students stayed active most of the time, neither students nor the teacher stretched or challenged the concepts in this lesson."

The content being covered in these classrooms was somewhat dull and delivered in a traditional format. One observer noted that "The lesson structure was very monotonous." In half of these class sections, instruction was largely symbolic and mechanical, with few connections to the contexts in which mathematics arises. One observer wrote that "Her engagements with the students are not mathematical in nature, they are designed to keep the classroom under control and on-task. She calls on students to volunteer answers, but it is as if they are filling in the blanks... There is no spark of interest, no freedom to explore and ask questions, etc."

These lessons were usually orderly with few management problems. One observer writes "This was a direct instruction lesson, but one that was run to a point that the students were on task. This teacher keeps the class with him and has management techniques that maintain student involvement." Most of these teachers were fluid and accurate when communicating mathematical content. One observer writes "The examples, discussions, explanations and verbal content were excellent and showed the deep knowledge and fluency of the lesson which was solving equations and inequalities. She was able to explain, answer questions without any hesitation or incorrect statements," while another describes, "The teacher moved through the calculations smoothly and without mistake...the lecture was very concise and correct." Overall, these lessons could be described as "orderly but unambitious," or "coherent but unimaginative."

Quadrant 3: Low BAM, Low UTOP (Unacceptable Ineffective Teachers) These classrooms were often focused on standardized test preparation. One observer writes "There was no opportunity to engage or explore any math concepts.... this was entirely test prep review... the teacher talked a great deal about learning the material so that they could pass the state test..." Nearly all of these lessons consisted of tasks where students were given simple, procedural, drill problems. "There was no way for students to engage or explore the content. The problems were very basic with only one or two steps." No significant connections were made to the real world, history, or to other disciplines, and the significance of the content was not made explicit to students. "There was no engagement. The problems were written on a transparency, they were simply numbers, not related to anything else. Students were asked to read an explanation from the board and then work the problem."

Nearly all of these classrooms had major issues with student behavior – students were yelling, fighting, and being disruptive, or were simply bored and refused to engage in the lesson. One observer describes how "This room was a zoo.... yelling answers at the top of their lungs... does not make intellectually engaged." Another observer writes that "The teacher attempts to manage the students and has some strategies but fails to follow through on any of them... the teacher continues lecturing while the students are talking." These lessons had significant amounts of wasted time, with one observer describing how "This lesson was way too time consuming, he did not efficiently move from one problem to the next."

Students' roles were often passive and involved copying the teacher's procedures and answering low-level questions. One observer writes "This was a review lesson in preparation for a state test and students followed along answering the teacher's recall type questions" while another notes that "The students seem to know that the teacher is expecting choral responses... students answered in unison." There were few issues with teacher content knowledge — this may be due to the fact that most of these classrooms were so poorly managed, or so procedural and without student input, that there was little opportunity for the teacher to demonstrate content knowledge. One observer noted that "The teacher never worked a problem or explained to a student how to solve a problem. She only focused on discipline and trying to get kids to do their corrections without helping them know how to do the corrections." By nearly any measure or definition, these classrooms displayed very ineffective teaching.

Quadrant 4: High BAM, High UTOP (Acceptable Effective Teachers) These classrooms had little focus on standardized exam preparation — standardized testing was rarely mentioned. These teachers all had orderly, well-managed classrooms where students were always on-task and engaged. One observer describes how "The teacher had a good rapport with her class and almost no visible instances of students being off task or needing correction of misbehavior. Most all the students were involved in the lesson and interested in participating in discussion and working the problems to get the correct answer." Students in all of these classrooms were often encouraged to generate contributions and solutions: "The students in this classroom were constantly generating ideas, conjectures, and propositions....the teacher asked for multiple strategies and she explored each one before she went on... the environment was friendly and she encouraged all answers to be looked at and explored."

The majority of these class sections was engaged in one or more lessons that involved investigation or problem-based learning: "The structure of this lesson was excellent... it led the students through an exploration where they had to try and solve problems on their own with different strategies before being taught about [t]he rules and the procedures." The majority of these class sections also included lessons that focused on making authentic, real world connections. One observer describes how the teacher "...guided them through real-world examples from their own lives to make the concepts of area and perimeter to come alive for the student[s]" while another wrote that "These concepts went farther than focusing on simply getting an answer correct and focused on life skills."

At least half of these class sections displayed other important characteristics. Observers noted instances of students being engaged in mathematical justification, explanation, or proof. One observer wrote that: "Students would cite concepts they had learned using evidence to support their position, and they would respond to the other students' points," while another described how "Students are asked to explain their answers." These teachers also made specific moves to make the importance of the content explicit to students. "The teacher led them through a series of questions to show them why Pythagorean Theorem is so important. She did not start with showing them the formula first but instead showed them why it would be so difficult to solve certain problems without it." Finally, observers noted that these teachers engaged in high quality formative assessment of student progress: "Teacher frequently stops video for formative assessment, asking questions that review the material instead of just asking if they understand." These classrooms were characterized by students engaged in learning and grappling with important mathematics concepts.

Summary of Quadrants Figure 11 summarizes some key characteristics of the six outlying class sections in each of the four quadrants:

High Value-Added in Fifth Versus Sixth Grade During the course of our prior analyses, we had found large differences in the relationship between scores on individual UTOP indicators and teacher value-added measures by grade level. One particularly striking comparison was 5th versus 6th grade. We conclude these analyses with a qualitative analysis of 5th and 6th grade classrooms where the teacher has especially high BAM 2010 value-added scores. These classrooms were selected by looking for the 6 class sections with with the highest BAM 2010 value-added in 5th grade, and the top 6 class sections in 6th grade.

In the top 6 BAM value-added class sections in fifth grade, instruction was weak to mediocre, as measured by the scores and supporting evidence on the UTOP. In 4 of the 6 class sections, the students were described as generally being on-task. This is not surprising given the descriptions of Acceptable Effective and Unacceptable Effective above as orderly classes. We did not find an explicit focus on standardized testing in these classrooms, although the type of instruction described is quite consistent

	Unacceptable	Acceptable
Effective	• **Focus**: Test prep & procedures • **Content**: Teacher errors • **Management**: Orderly, but sometimes hostile	• **Focus**: Conceptual understanding; investigation • **Content**: Connected to "real world" and important • **Management**: Orderly, & on-task
Ineffective	• **Focus**: Test prep & procedures • **Content**: Little covered, few/no connections • **Management**: "This classroom was a zoo," students passive	• **Focus**: Test prep & procedures; lessons are dull • **Content**: Fluidly communicated • **Management**: Orderly, & on-task

Figure 11: Key characteristics of the six outlying class sections in each of the four quadrants of Effective and Acceptable teaching.

with traditional testing drill strategies. The observers noted little intellectual engagement (3 of 6 class sections), few or no real-world connections (5 of 6 class sections) or mentions of the importance of the content (3 of 6 class sections). Observers found incorrect or poorly communicated mathematical content (3 of 6 class sections), and significant amounts of wasted time (3 of 6 class sections). Some of these classrooms were quite poor, some were mediocre, and none exhibited teaching practices accepted in the mathematics education community or recommended by educational research.

The story with the top 6 BAM value-added class sections in sixth grade was somewhat different. Three of the 6 top class sections in sixth grade had actually fallen into the outlier analysis of our "Acceptable Effective" category, and had been analyzed in the previous section. These were very strong classrooms in terms of both BAM value-added and UTOP scores, where students were engaged with important mathematical ideas and participating in authentic problem-solving. They were consistent with current views of research on effective instruction in elementary mathematics. However, the other 3 class sections were the opposite — these classes had very poor instruction, and unacceptable teachers, as measured by the UTOP. Students in these classrooms were off-task or bored (3 of 3 class sections), and the teacher sometimes focused on test preparation and used rote and keyword type approaches to mathematics learning (2 of 3 class sections). One observer writes "There was a lot of body language and lack of participation that indicated either boredom or lack of understanding/care about the content... I noticed at least two on each side of the camera who were slumped in their desks with heads in hands." In these class sections there were also poor questioning techniques (3 of 3) and poor formative assessment (2 of 3) - "A few students asked questions, but they were procedural, no higher order questions were asked or proposed. Students were engaged, but only rotely solving by a process without understanding what or why the process resulted in a solution." There was little student generation of ideas or questions (3 of 3), and few real world connections (2 of 3). An observer wrote that ,"The examples were non existent in a real life setting and was a rote exercise in the mechanical simplifying of a simple linear equation."

These observations at fifth and sixth grade demonstrate the danger of using value-added assessments as the "gold-standard" of teaching effectiveness. Although in the sixth grade data high BAM we see some classrooms where productive learning of mathematics seems to be occurring, the supposedly exemplary instruction singled out by valued-added scores on their own is highly variable, and in some cases indefensible.

5 DISCUSSION AND CONCLUSION

We conclude by repeating some of the points with the greatest potential implications for policy and practice.

1. The typical classroom practices observed in the MET study were rather weak, as judged by the UTOP, and therefore one should be hesitant to draw strong conclusions about what exemplary schools look like from this project. Practices that many mathematics educators might label Innovative were almost never observed (Figure 4).

2. The classroom attributes most reliably and consistently associated with increased value-added scores were supportive, orderly, well-managed classrooms that focus on important content rather than explicit preparation for tests (Table 7).

3. The classroom practices favored by the UTOP were more often seen in elementary school than in middle school. The drop in quality in middle school is striking (Figure 10). The professional development needs of teachers in various grade levels may be substantially different.

4. There is some potentially encouraging news about middle school. The highest value-added scores in sixth grade correlate well with high scores on many specific UTOP indicators (Figure 8). This does not prove that that helping teachers improve their practice in accord with the UTOP will raise scores, but it certainly leaves open the possibility. At 6th grade, the UTOP indicates that raising student scores requires both overall excellence across all domains, and a willingness to use innovative methods to engage and motivate students.

5. Teachers from the UTeach program, whose study included coursework specifically aimed at instilling practices measured by the UTOP, score considerably higher on the UTOP than other teachers we examined. (Compare Figure 1(B) with the Summary scores in Figure 4). Therefore it is plausible that professional development which uses the UTOP to identify areas with potential for teacher growth, and then provides relevant coursework similar to that used in the UTeach program, could raise UTOP scores [10].

6. An evaluation system based on averages of value-added and observation scores can gloss over undesirable outcomes. Teachers making content errors or creating a hostile environment can obtain good value-added scores. Teachers subjecting students to correct but boring lessons can obtain acceptable UTOP scores. Good teaching needs to be both acceptable and effective. The best teachers focus on conceptual understanding rather than explicit preparation for tests, while maintaining orderly engaged classrooms (Table 7 and Figure 11).

A common criticism of teacher evaluation processes is that they rate almost all teachers as excellent and provide little useful feedback (New Teacher Project 2009). Yet UTOP observations in this study were more demanding than the value-added measures. In grades 5-8, more teachers were rated Unacceptable than were rated Ineffective (Figure 10). Furthermore, observation is in a better position to raise certain sorts of uncomfortable questions than value-added models. Our UTOP measurements found that teaching became systematically worse in moving from elementary to middle school. Value-added models compare teachers in similar environments to each other and may not raise such possibilities. Similarly, if worse teachers are systematically placed in classrooms of students from low-income families, value-added models can react by controlling for environmental effects so that the effect disappears. Observation scores typically do not (Table 5).

Classroom observation has the potential to be more ambitious than value-added modeling. Let us take the example of the **Society** indicator, which asks for a discussion of a topic's role in history or society. This was the lowest-ranked indicator for the MET sample, with an average score of 1.1. Our own view is that it is valuable for students to understand something of the social and historical significance of mathematics, whether or not this understanding is rewarded in today's mathematics examinations. We do not want students to become good citizens because it will raise their mathematics scores; we want students strong in mathematics because it will help make them good citizens. UTeach graduates in our pilot study had an average on this indicator of 3.5, perhaps because UTeach has a semester-long class focused on this topic. A similar course should help other teachers improve on this indicator. We offer this example to illustrate the way that observation scores and professional development could interact, and the way that an observation instrument can more flexibly broaden the goals of school than exams.

But we do not conclude that analysis of student test scores is superfluous. Neither measure, UTOP scores nor BAM value-added, appears on its own adequate to capture good teaching. Analyses of classrooms where the two measures differed revealed that acceptable UTOP scores can be assigned to teachers with strong content knowledge and good management skills, but who lack the pedagogical skills to effectively engage students in mathematics learning. High BAM value-added scores can be obtained by teachers with weak content knowledge, who treat their students disrespectfully, or who also lack the pedagogical skills to engage students in mathematics learning. When both of these measures were in accord, the results were much more compelling. Few could argue that the classrooms with both low BAM and low UTOP scores belonged to teachers not in need of substantial assistance. The classrooms with high BAM and high UTOP scores displayed many of the characteristics favored by reform movements in education, based on research on how children learn. Classroom observation and value-added scores should be considered together, on equal footing, to help teachers develop as they progress in their careers, as measures of effective teaching.

References

Bates, D. and Maechler, M. (2010), lme4: Linear mixed-effects models using s4 classes. R package version 0.999375-35, URL `http://CRAN.R-project.org/package=lme4`

Braun, V. and Clarke, V. (2006), Using thematic analysis in psychology, *Qualitative Research in Psychology*, **3**(2), 77–101

Gates Foundation (2012), Gathering feedback for teaching: Combining high quality observations with student surveys and achievement gains. Research paper., URL `http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf`. Retrieved 9 December 2012

Gladwell, M. (2008), Most likely to succeed: How can we hire teachers when we can't tell who's right for the job?, *New Yorker*, **Dec. 15**, URL `http://www.newyorker.com/reporting/2008/12/15/081215fa_fact_gladwell`. Retrieved October 2010

Gordon, R., Kane, T. J., and Staiger, D. O. (2006), Identifying effective teachers using performance on the job, URL `www.brookings.edu/views/papers/200604hamilton_1.pdf`. Retrieved November, 2010

Granger, E. M., Bevis, T. H., Saka, Y., Southerland, S. A., Sampson, V., and Tate, R. L. (2012), The efficacy of student-centered instruction in supporting science learning, *Science*, **338**, 105–108

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

Horizons Research (2000a), Inside the classroom interview protocol, URL `http://www.horizon-research.com/instruments/clas/interview.php`. Retrieved June 2009

Horizons Research (2000b), Validity and reliability information for the lsc classroom observation protocol, URL `http://www.horizon-research.com/instruments/clas/cop.php`. Retrieved June 2009

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

Meyers, L., Gamst, G., and Guarino, A. (2006), *Applied Multivariate Research*, Thousand Oaks, CA: Sage Publications

New Teacher Project (2009), The Widget Effect, URL `http://widgeteffect.org.RetrievedJune2013`

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

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

Rowan, B., Correnti, R., , and 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

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

Silverstein, S. C., Dubner, J., Miller, J., Glied, S., and Loike, J. D. (2009), Teachers' participation in research programs improves their students' achievement in science, *Science*, **326**, 440–442

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

Notes

[1] Raters obtained a weighted kappa agreement of 0.41 on individual UTOP indicators, which is moderate agreement, and 0.63 on the synthesis ratings, which is substantial agreement (Landis and Koch 1977)

[2] In some models, we used UTOP scores to predict BAM value-added scores or value-added scores from the State tests. Grade level was also often a predictor in the models, and the data set we used to construct the models was always aggregated to the class section level. Note that BAM was administered only in 2010, so the pre-score for the BAM value-added models was the State Test in 2009. Descriptive information about UTOP ratings and value-added gains as they vary by grade level are shown in Table 4.

[3] We fit linear regression models with grade level and UTOP indicator or Synthesis ratings as predictors. We also included a term for the interaction of the UTOP score with grade level (4-8). This made for a base set of 26 different models, as there are 22 indicators on the UTOP and 4 Synthesis ratings. We fit this set of models for three different dependent variables: State VAM 2009, State VAM 2010, and BAM 2010. We tested for the significance of the interaction term, display the results visually by using the regression coefficients to show patterns that emerged, and use the standard error of the regression coefficients to show the uncertainty.

[4] The BAM value-added of the class sections had mean of .005 and standard deviation of .255; values ranged from -1.254 to 1.021

[5] Few of the regression coefficients are significantly different from one another, and few of them are significantly different from zero. Pooling indicators together to reduce the level of uncertainty one can conclude that the relation between improved performance on the UTOP and increased BAM value-added scores is negative. Overall, the distribution of value-added scores was higher in fifth grade than in any other grade. t tests give significant differences (p<.05) for the averages of BAM scores between 4th and 5th, 5th and 6th, and 5th and 8th grades.

[6] We examined many interaction models and grouping of grades. The grouping that seemed best to fit the data was to assemble the 4th, 5th, 7th, and 8th grade lessons into one group, and the 6th grade lessons into another group. We looked at models that had indicator score, grade level group, and the interaction of these two terms as predictors. We found that for 16 of the 22 UTOP indicators, the interaction term was significant ($p < .05$; Indicators 1.1, 1.2, 1.3, 1.4, 1.5, 2.1, 2.3, 3.1, 3.4, 4.2, 4.3, 4.5, 4.6) or marginally significant ($p = .0740$, .0644, .0533; Indicators 2.2, 2.4, and 4.1, respectively). The interaction term was also significant for three of the Synthesis Ratings - Lesson Structure, Implementation, and Mathematics Content (p = .0302, .0147, and .0150, respectively). In all cases, the trend was the same — the indicator or Synthesis rating had a stronger, positive relationship with value-added for 6th grade class sections, compared to other grade levels. We ran these some models with 2009 and 2010 State VAM as the dependent measure, and the results were similar.

[7] For example, “In a fraction, the denominator is the largest integer.”

[8] We selected these sections by summing the normalized UTOP Consensus rating and the normalized BAM 2010, and choosing the highest and lowest sums. Note that results were similar if we instead used the normalized UTOP Summary and the normalized BAM 2010, rather than UTOP Consensus rating. A problem we encountered is that a number of the Unacceptable Effective and Acceptable Ineffective class section outliers actually had the greatest sum as well as the greatest difference. This is because one measure (BAM or UTOP) was anomalously high or low. We omitted these class sections when choosing Acceptable Effective and Unacceptable Ineffective class sections for extra examination.

[9] Once 6 extreme class sections were selected for each of the 4 quadrants, we compiled the UTOP supporting evidence for all lessons from that class section. There were most commonly 2 separate lessons for each class section (the total ranged from 1 to 4), and UTOP observations were typically carried out by 2 different observers (although the number ranged from 1 to 7). We used thematic analysis techniques (Braun and Clarke 2006). Themes were given more prominence if they were observed in multiple lessons and cited by multiple raters; however for class sections that had only one lesson and one rater, this was not possible. The major themes coded for each class section were then compared within each quadrant to determine which themes consistently arose across all or many class sections.

[10] Some studies provide direct evidence that innovative practices favored by the UTOP can raise value-added scores as well (Silverstein et al. 2009; Granger et al. 2012)