MEASURES OF EFFECTIVE TEACHING: CAPTURING THE UTeach VISION IN CLASSROOM OBSERVATION

Candace Walkington
Wisconsin Center for Educational Research
University of Wisconsin - Madison
**MY BACKGROUND**

- Postdoctoral Fellow in Math Education at U-W Madison
- Ph.D. in Math Education UT-Austin
- B.S. and M.S. in Mathematics
- Taught courses for pre-service elementary and secondary teachers
- My research is on using learning sciences principles to design effective instructional interventions
- Researcher with UTeach/NMSI/MET examining teaching effectiveness
VIDEO #1

- 8th grade mathematics lesson on graphing linear equations
- Students working in groups
- Given 10 equations to graph, 5 with positive slope, 5 with negative slope
VIDEO #1: DISCUSSION POINTS

- What does this teacher do well?
- Are there any weaknesses or missed opportunities you see in his instruction?
- Based on the video data, is this an “effective” teacher? Why or why not?
VIDEO #2

- High school mathematics class learning about linear functions
- Students have been working on problem, now presenting to class

Describe the pattern. Assuming the sequence continues in the same way, how many dots are there at 100 minutes?

number of dots = 4m + 1
VIDEO #2: DISCUSSION POINTS

- What does the teacher do well?
- Do you see any weaknesses or missed opportunities in this lesson?
- How does this teacher’s style compare to the teacher in the other video?
  - Interactions with students
  - Framing/choice of task
- Based on the video data, is this an “effective” teacher?
UTOP PROJECT

Development and piloting of a classroom observation instrument (UTeach Observation Protocol) to measure characteristics of effective teaching in mathematics and science classrooms

Collaborators: Michael Marder, Mary Walker, Larry Abraham, Denise Ekberg, Gail Dickinson, & Kelli Allen, UTeach Natural Sciences, University of Texas at Austin

GRAs: Prerna Arora, Shasta Ihorn, & Jessica Gordon
MEASURING EFFECTIVE TEACHING

What does effective teaching look like when it happens?

“Documenting particular features of teaching that are consistently effective for students’ learning has proven to be one of the greatest research challenges in education” (Hiebert & Grouws, 2007)
MEASURING EFFECTIVE TEACHING

- What does effective teaching look like when it happens?
- Can the skills involved with being an effective teacher be successfully trained through a teacher preparation program?
- Can classroom observers be trained to make key distinctions in effective teaching practices?
CONTEXT OF WORK

- Measurement of teacher quality lies at the heart of current debates about educational reform
  - Challenges particularly severe in secondary mathematics and science – critical to debate on US competitiveness
  - Teacher preparation under scrutiny
**CONTEXT OF WORK**

- Initiatives like *Race to the Top* emphasize measuring teacher quality through student standardized test score gains
  - Unclear if tests measure all outcomes of education that we care about
  - Questions about reliability (Baker et al., 2010)
  - Not all subjects assessed
  - Value-added gains “black box” that does little to help us understand good teaching
“In teaching, the implications are even more profound. They suggest that we shouldn't be raising standards. We should be lowering them, because there is no point in raising standards if standards don't track with what we care about. Teaching should be open to anyone with a pulse and a college degree - and teachers should be judged after they have started their jobs, not before.”
CONTEXT OF WORK

The recent report “Measuring what Matters” urges for more accountability for teacher preparation programs:

- Student learning measures
- Persistence rates in teaching
- Licensure exam scores
- Feedback surveys
- **Classroom teaching performance**
  
  Implement high-quality observational assessments of classroom teaching by supporting efforts to link these assessments to student achievement and by developing rigorous training for classroom observers to ensure reliable findings.
THE UTeach Program

- Steady increase in number of students with strong STEM backgrounds going into teaching
- Replicated at 28 universities in 13 states
- 92% of graduates go into teaching, 82% remain 5 years later (compared to 65% nationally)

www.nationalmathandscience.org
BACKGROUND OF PROJECT

- Persistent requests to evaluate UTeach Graduates and Noyce Scholars
- UTeach boosts recruitment and retention, but are UTeach graduates effective teachers?
- Strongly felt need to go beyond value-added measures, and look into UTeach classrooms
- Insight into how they implement practices they learn in UTeach
- What are these practices?
Some Key Features of UTeach Philosophy

- Organized, well-managed, on-task classroom
- Attention to issues of diversity and access
- Incorporating inquiry/investigative learning
- Using technology for teaching and learning
- Fluid and accurate communication of content
- Fostering student-student collaboration
- Formative assessment of student progress
- Applications and inter-disciplinary connections
- Critical practices of self-reflection
- Facilitating classroom discussion and “student talk”
Background of Project

Few classroom observation instruments with established reliability appropriate for UTeach vision and evaluation goals:

- Content-specific to math and science teaching
- Appropriate for wide range of grade levels (K-college)
- Flexible use of teaching styles, including inquiry/investigation and direct instruction
- Advanced pedagogical strategies (questioning techniques, formative assessment, etc.)
- Strong focus on how content expertise contributes to effective teaching
CLASSROOM OBSERVATION PROTOCOLS

Charlotte Danielson’s Framework

Shulman (1986): What about content knowledge, and PCK?

RTOP


CLASS

COP/LSC Protocol

A Study of K–12 Mathematics and Science Education in the United States
DESCRIPTION OF UTOP

- Modified Horizon Research Inc.’s COP (*Inside the Classroom* Study)
  - Based on reform standards (NSES, NRC, NCTM)
  - Tailored to UTeach vision
  - No published indicator or synthesis-level reliability
  - No scoring rubrics
- Modified indicators mapped well to:
  - UTeach Holistic Framework
  - UTeach Portfolio Expectations
  - UTeach Apprentice Teaching Observation Instruments
DESCRIPTION OF UTOP

- Original version had 32 indicators (teaching behaviors) in 4 sections
  - Classroom Environment
  - Lesson Structure
  - Implementation
  - Mathematics/Science Content
- 1-5 scale, DK/NA options
- Section Synthesis Ratings
- Teacher interview
<table>
<thead>
<tr>
<th>Rating</th>
<th>Indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The classroom environment encouraged students to generate ideas, questions, conjectures, and/or propositions that reflected engagement or exploration with important mathematics and science concepts.</td>
</tr>
<tr>
<td>Evidence:</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>Rubric</td>
</tr>
<tr>
<td>1.2</td>
<td>Interactions reflected collegial working relationships among students. (e.g. students worked together productively and talked with each other about the lesson). <em>It's possible that this indicator was not applicable to the observed lesson. You may rate NA in this case.</em></td>
</tr>
<tr>
<td>Evidence:</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>Rubric</td>
</tr>
<tr>
<td>1.3</td>
<td>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.</td>
</tr>
<tr>
<td>Evidence:</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>Rubric</td>
</tr>
<tr>
<td>1.4</td>
<td>The majority of students were on task throughout the class.</td>
</tr>
<tr>
<td>Evidence:</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>Rubric</td>
</tr>
<tr>
<td>1.5</td>
<td>The teacher’s classroom management strategies enhanced the classroom environment.</td>
</tr>
<tr>
<td>Evidence:</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>Rubric</td>
</tr>
<tr>
<td>1.6</td>
<td>The classroom is organized appropriately such that students can work in groups easily, get to lab materials as needed, teacher can move to each student of student group, etc.</td>
</tr>
<tr>
<td>Evidence:</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>Rubric</td>
</tr>
<tr>
<td>1.7</td>
<td>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, attentiveness to student needs).</td>
</tr>
<tr>
<td>Evidence:</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>Rubric</td>
</tr>
<tr>
<td>Rating</td>
<td>Indicator</td>
</tr>
<tr>
<td>--------</td>
<td>-----------</td>
</tr>
<tr>
<td>1.1</td>
<td>The classroom environment encouraged students to generate ideas, questions, conjectures, and/or propositions that reflected engagement or exploration with important mathematics and science concepts.</td>
</tr>
</tbody>
</table>

**Evidence:**

1.2 Interactions reflected collegial working relationships among students. (e.g. students worked together productively and talked with each other about the lesson).  
*It's possible that this indicator was not applicable to the observed lesson. You may rate NA in this case.*

**Description**

**Rubric**

**Specific Rating Examples**

---

This indicator assesses the degree to which students have learned to be collegial, respectful, cooperative, and interactive when working in groups. Evidence of collegial working relationships among students includes collaborative discussions about topics relevant to the lesson and successful distributing of roles and responsibilities within each group…

---

This indicator should be rated a 1 if there is group work during the lesson, but the group work is highly unproductive. This could include behavior where the majority of the groups are socializing, off-task, arguing, or ignoring each other, as well as regular instances of students copying and/or certain group members doing all of the work.

This indicator should be rated a 2 if …

---

Rating of 3 Example: The students were put into debate groups for this class period - one group would debate another group, while the rest of the student groups were in the audience. The groups worked together smoothly - the students were able to pick who was doing what part of the debate, coordinate their arguments, and split the time slots when necessary. The audience also would occasionally compare their notes during breaks…

---

1.6 The classroom is organized appropriately such that students can work in groups easily, get to lab materials as needed, teacher can move to each student of student group, etc.

**Evidence:**

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, attentiveness to student needs).
**UT Pilot Study**

- Over 5 semesters, conducted 83 observation of 3 groups of teachers:
  - UTeach Noyce Scholar Graduates (N=7)
  - UTeach Non-Noyce Graduates (N=14)
  - Non-UTeach Graduates (N=15)
- Novice teachers (most 0-3 years exp)
- Math, science, and computer science classes
- 9 high schools, 5 middle schools, 2 districts
- 50-90 minute observation, 1-2 times per semester
- 2 observers present
COMPARATIVE ANALYSIS

- **UTeach Noyce**
- **UTeach Non-Noyce**
- **Non-UTeach**

### Horizontal Axis: Section of UTOP
- Classroom Environment
- Lesson Structure
- Implementation
- Math/Science Content

### Vertical Axis: Average Synthesis Rating
- Scale from 1 to 5

The graph illustrates the comparative analysis of teaching environments across different sections of UTOP for UTeach Noyce, UTeach Non-Noyce, and Non-UTeach. The ratings are visualized with distinct markers for each group, showing variations in classroom environment, lesson structure, implementation, and math/science content.
PILOT RESULTS – PREPARATION BACKGROUND

- **Noyce Scholars** rated significantly higher on UTOP than other groups, \( p < .01 \) when in regular-level classes (advanced class scores near ceiling)
- **Alternative Certified** teachers rated significantly lower on UTOP than other teachers \( p < .05 \)
- Small sample sizes, non-random sampling
COMPARATIVE ANALYSIS

![Graph showing comparative analysis of Composite UTOP Rating across different years of teaching experience and UTeach categories.]

- UTeach Noyce
- UTeach Non-Noyce
- Non-UTeach

Years of Teaching Experience:
- 0-0.5
- 1-1.5
- 2-2.5

Composite UTOP Rating:
- 0
- 1
- 2
- 3
- 4
- 5
PILOT RESULTS – TEACHING EXPERIENCE

- Teaching experience NS predictor for Non-UTeach ($p = 0.869$) and Noyce ($p = 0.533$)
- Teaching experience significant predictor for UTeach Non-Noyce ($p < .05$)
- UTeachers grow more on UTOP scores over time, after starting out at similar level to Non-UTeach.
THEMES FROM QUALITATIVE DATA

- UTeach graduates struggle initially
  - Lack familiarity with curriculum
  - Issues with classroom management
  - Difficulty implementing the vision of UTeach
- As time goes on, things get better – implement some UTeach practices, and see successes
- School environment important for opportunities
  - Explained large portion of variance
  - Influenced scores +0.5 vs. -0.5
Teacher 27 is a UTeach graduate teaching Integrated Physics and Chemistry at a public Austin high school.

- The students talked over the teacher and interrupted her in other ways while she was speaking to the class.
- The teacher described how she had tried 4 different methods to try to get students to write science reports, and that this was the latest in her string of failures.
Case Study of One UTech Teacher’s Growth over Time

Teacher 27 is a UTech graduate teaching Integrated Physics and Chemistry at a public Austin high school.

- The students did not get off-task – the lesson was structured into small chunks that kept students engaged.
- The unit had lots of applications to everyday life.
- The teacher did not use formative assessment - she elicited few student responses.
**Case Study of One UTeacher’s Growth over Time**

Teacher 27 is a UTeach graduate teaching Integrated Physics and Chemistry at a public Austin high school.

- The teacher gave clear instructions for the project the students were engaged in. The classroom environment was warm, positive, and calm.
- The teacher checked in with every single student multiple times during the lesson, providing scaffolding as needed, and eliciting their thinking with “why” questions.
SUMMARY: PILOT STUDY

- Developed classroom observation protocol based on vision of UTeach, as well as current research and standards in science and math education.
- UTeach graduates start out relatively weak, but gain momentum over time as they implement UTeach practices.
- **Key Question:** Is the UTOP a valid and reliable instrument that measures important components of effective teaching?
NMSI/MET Studies

- Two additional UTOP studies conducted in partnership with the Gates Foundation’s *Measures of Effective Teaching* project, and the National Math and Science Initiative
- No UTeachers in these studies – validate and refine UTOP
- Examine reliability, consistency, factor structure – make any necessary modifications
- Connect teaching behaviors on UTOP to teacher value-added gains
THE MET PROJECT

- 3000 teachers from 7 school districts, 7 states
- Various subjects (mathematics, English, science) and grade levels
- Multiple measures of effectiveness (observations, value-added, student surveys, teacher exams)
- Multiple video lessons of each teacher (~4)
- Multiple classroom observation instruments
  - Charlotte Danielson’s FFT
  - CLASS protocol
  - MQI Rubric
  - UTOP
NMSI/MET Studies

- **Study 1**: 17 raters (AP Math teachers) scored 235 video lessons of 119 teachers
- **Study 2**: 99 raters (math and science master teachers with LTF), scored 994 video lessons of 250 teachers
- All lessons grades 4-8 mathematics
- Many videos double-scored
NMSI/MET Studies

- Training consisted of watching and rating 4-5 videos, group discussions, reviewing normed ratings

(Image of Teachscape viewer removed)
RESULTS: STUDY 1

- Most of the 4-8 math video lessons from this national sample did not score highly on the UTOP
- Few examples of what the UTeach program would consider “exemplary” teaching
- Revisions to manual & training
- Introduced outside videos of exemplary teaching
Many middle school math teachers teaching inaccurate content, using formulaic/key word type approaches.

- **5/5** training videos we (semi-randomly) selected contained at least 1 instance of the teacher communicating incorrect content.

- Raters identified problematic content issues in around one half of all lessons.
RESULTS: STUDY 2

- Little emphasis on conceptual understanding
- Mostly direct instruction mixed with (ineffective) group work
- Many classrooms tightly managed, little “student talk”
Surface-level engagement often seen, but deep conceptual thinking about significant mathematical ideas rare

Instrument (and observers) were able to capture this distinction ("hands on" vs. "minds on") – accentuated in manual & in training

**RESULTS: STUDY 2**

- Resource Use
- Classroom Management
- On Task
- Intellectual Engagement
- Higher-order Questions
- Content Connections
- Investigative Learning
PSYCHOMETRIC ANALYSIS OF UTOP

- **Construct Validity**: Accurately measures theoretical, non-observable construct or trait
  - Factor Analysis

- **Reliability**: Provides repeatable, consistent results when administered
  - HLM models looking at impact of rater and lesson

- **Criterion-Related Validity**: Useful for predicting other measures
  - Connection to teacher value-added
FACTOR ANALYSIS OF UTOP

- Uses correlation/covariation between different items to determine how related
- Reveal a smaller set of latent, unobserved variables or “factors” that underlie or explain the larger set of variables
- What macro-constructs relating to effective teaching behaviors are being measured by the indicators on the UTOP?
FACTOR 1: FOSTERING SURFACE LEVEL ENGAGEMENT

- Classroom management
- Majority “on task”
- Group-work dynamic
- Time management
- Lesson Organization
- Appropriate Resources
- Issues of equity & access
- Teacher critical of lesson
Factor 2: Fostering Deep, Conceptual Understanding

- Students generate ideas/conjectures
- Students intellectually engaged
- Students explore content
- Use of higher-order questions
- Use of inquiry/investigation
FACTOR 3: CONTENT ACCURACY & FLUIDITY

- Accurate written content information
- Accurate & fluid verbal communication of content
- Appropriate use of abstraction
FACTOR 4: CONTENT CONNECTIONS

- Connect content to “real world” and other disciplines
- Connect content to history & current events
- Connect content to the “big picture” of the discipline
**Factor Analysis of UTOP**

**Cluster 1: Fostering Surface Engagement**
- On task & involved
- Class management
  - Group work
- Lesson organization

**Cluster 2: Fostering Deep Conceptual Understanding**
- Inquiry/investigation
- Higher-order questioning
- Intellectual engagement

**Cluster 3: Content Accuracy and Fluidity**
- Verbal & written accuracy/fluidity
- Effective use of abstraction

**Cluster 4: Making Content Connections**
- To real world (authentic)
  - To “big picture”
- To history/current events
FACTOR ANALYSIS OF UTOP

Cluster 1: Fostering Surface Engagement
- On task & involved
- Class management
- Group work
- Lesson organization

Cluster 2: Fostering Deep Conceptual Understanding
- Inquiry/investigation
- Higher-order questioning
- Intellectual engagement

Cluster 3: Content Accuracy and Fluidity
- Verbal & written accuracy/fluidity
- Effective use of abstraction

Cluster 4: Making Content Connections
- To real world (authentic)
- To “big picture”
- To history/current events

General Pedagogy
- Group work
- Lesson organization

Pedagogical Content Knowledge
DIMENSIONS OF TEACHING ASSESSED BY UTOP

Bars show standard error of the mean.
DIMENSIONS OF TEACHING ASSESSED BY UTOP

Bars show standard error of the mean
INSTRUMENT RELIABILITY

- Can we **consistently** measure teaching **effectiveness** on the UTOP, beyond the biases of individual raters, or the characteristics of particular lessons?
- HLM regression models (g-study)
- Goal: **60-80%** of the variance in teacher scores on the instrument attributable to the **stable** characteristics of the **individual teacher**
  - Classroom observation is **necessarily** subjective, and varied based on lesson content
INSTRUMENT RELIABILITY

UTOP Overall (Average Synthesis)

<table>
<thead>
<tr>
<th></th>
<th>% total variance explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesson</td>
<td>13.3%</td>
</tr>
<tr>
<td>Teacher</td>
<td>32.77%</td>
</tr>
<tr>
<td>Rater</td>
<td>0%</td>
</tr>
<tr>
<td>Residual (Rater x Lesson)</td>
<td>53.9%</td>
</tr>
</tbody>
</table>

One observer comes in once per year and evaluates the teacher.
INSTRUMENT RELIABILITY

○ **Decision-study:**
  
  - Determine how many times per year a teacher needs to be observed, by how many different raters, to get a stable estimate

○ **Implementation Results:**
  
  - 2 observations per year, with 2 raters present at each observation (UT Pilot study)
  - 4 observations per year, with 1 rater present at each observation (MET Study)
## Instrument Reliability

**Schedule:** 4 observations/year, 4 different raters

<table>
<thead>
<tr>
<th>Category</th>
<th>Reliability Coeff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classroom Environment</td>
<td>67%</td>
</tr>
<tr>
<td>Lesson Structure</td>
<td>62%</td>
</tr>
<tr>
<td>Implementation</td>
<td>64%</td>
</tr>
<tr>
<td>Mathematics Content</td>
<td>40%</td>
</tr>
<tr>
<td>Overall (Avg Syn)</td>
<td>66%</td>
</tr>
</tbody>
</table>
RATER RELIABILITY

- Identified and reviewed lessons with “severe disagreement”
  - Charismatic teacher using low-level teaching strategies
  - Weak teacher using some elements of reform approaches
  - Using “discourse of reform”
- Re-norming webinars and interventions
- Issues with content expertise – reliability analysis of 99 raters
  - Degrees in fields other than math or science related to lower content reliability ($p < .01$)
RESULTS SUMMARY

- UTeachers & Noyce Scholars show excellence in classroom, especially over time
- UTeach vision of effective teaching is multi-dimensional – 4 critical factors
- School site important for opportunities to implement UTeach practices
- Need multiple observations, multiple raters
- Appropriate rater content expertise, and focused training to making critical differentiations (i.e., surface level vs. deep level engagement)
**Implications**

- UTTeachers are effective teachers based on UTOP measures, but a larger study is needed to verify these findings.
- The UTOP measures important characteristics of effective teaching, and can be implemented successfully by raters with modest amount of training – ready for dissemination.
I want to tell you that I feel that my teaching has greatly improved as a result of being trained in UTOP. I have been teaching for over 25 years, 18 at the college level. When I started teaching high school seven years ago, I was certified, so I didn't have to go through any type of mentoring program. Everyone assumed because of my age and experience that I knew what I was doing. I have muddled through, learning as I went, but in the past few weeks I have seen a big change, mostly in the way I question my students. I don't know if this was ever seen as a "side effect" of being a UTOP rater, but it certainly has been wonderful for me.
FUTURE DIRECTIONS

- Use of UTOP to compare classrooms at project-based school (with UTeach graduates) to those at traditional school, same low income school district
- Use of UTOP to evaluate teaching with physics-based simulations (PhET)
- Use of UTOP for evaluation of university teaching
- Value-added correlations for UTOP will be publicly released this year by MET
QUESTIONS, COMMENTS, SUGGESTIONS?