UTeach, UTeach Replication and the UTOP

Strategies to increase the quantity, long-term retention and diversity of science and mathematics teachers and tools to measure the effectiveness of high quality science and mathematics teaching

Mary Ann Rankin, Candace Walkington, & Mary H. Walker

2011 SMTI Conference, Portland, OR
The UTOP Project

• Background of Project & Goals
• Description of UTOP
• Overview of Pilot Work at UT
• Overview of NMSI/MET Studies
• Lessons Learned
  • UTeach Practices & Content Knowledge Issues
  • Instrument Structure, Consistency, & Reliability
  • Rater Training & Background
• Future Directions
Persistent requests to evaluate UTeach graduates beyond descriptive statistics.
  - Including Noyce Scholars

Value-added assessments are insufficient as sole measure of teaching effectiveness
  - Can inaccurately classify teachers (Kane & Staiger, 2008)
  - Questions about what tests really measure
  - Not all subjects tested
Measures of Teaching Effectiveness

Conceptualize multi-directional relationship between value-added, observation, and other evaluations - these methods should validate & inform each other.
Key Questions

- What classroom behaviors are associated with effective teaching?
- How can classroom observers be trained to reliably differentiate strong teachers from weak teachers?
Background of Project

- Few instruments with established reliability/validity appropriate for evaluating UTeach goals:
  - Flexible use of teaching styles, including, but not limited to, inquiry/investigation
  - Advanced pedagogical strategies (questioning techniques, problem-based learning, etc.)
  - Content-specific to math & science teaching
  - Strong focus on content knowledge, and how content expertise contributes to effective teaching
  - Appropriate for wide range of grade levels (K-college)
Description of UTOP

- Modified Horizon Research Inc.’s COP (*Inside the Classroom Study*) to fit these goals.
  - Based on reform standards (NSES, NRC, NCTM)
  - 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 in 4 sections:
  - Classroom Environment
  - Lesson Structure
  - Implementation
  - Math/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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evidence:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2 Interactions reflected collegial working relationships among students. (e.g. students worked together productively and talked with each other about the lesson).</td>
</tr>
</tbody>
</table>

*It's possible that this indicator was not applicable to the observed lesson. You may rate NA in this case.*

<table>
<thead>
<tr>
<th>Evidence:</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evidence:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.4 The majority of students were on task throughout the class.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evidence:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5 The teacher’s classroom management strategies enhanced the classroom environment.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evidence:</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evidence:</th>
</tr>
</thead>
<tbody>
<tr>
<td>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).</td>
</tr>
</tbody>
</table>
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
UT Pilot Study

- 2 observers present at each observation (semi-blind), debrief and come to consensus
- Weighted kappa - agreement statistic that takes into account:
  - Sometimes raters agree by chance alone
  - If one rater scores 4, and the other scores 5, this isn’t as bad as scoring 1 and 5
- Pre-consensus synthesis rating kappa = 0.63 (substantial)
- Pre-consensus item kappa = 0.51 (moderate)
UT Pilot Study

- UTOP Ratings analyzed using:
  - Comparative graphs with error bars (SEM)
  - T-tests
  - Hierarchical Linear Model
- Focus on Synthesis Ratings & Average Synthesis Rating
- Background characteristics of teacher and school
Comparative Analysis

![Graph showing the average synthesis rating for different sections of UTOP (Classroom Environment, Lesson Structure, Implementation, Math/Science Content) for UTeach Noyce, UTeach Non-Noyce, and Non-UTeach.]
HLM Regression Model

- **3 Level Model**: Observations nested within teachers nested within schools
- **DV**: Average Synthesis Rating
- **Random Effects**: Teacher, School, Course
- **Predictors**: Preparation, Experience, Regular/Advanced Class, Middle/High School
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.
  - 0.88 per year for novice years
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 for these groups
High school mathematics class learning about linear functions

Students have been working on problem, now presenting to class

Clip: 1:35-7:17

Linear function: \( y = 4x + 1 \)
Video #1

Discussion Points:

◦ What is the nature of the learning task that the teacher has chosen?
◦ What teaching behaviors is this teacher using that seem effective?
◦ Do you see any weaknesses or missed opportunities in this lesson?
NMSI/MET Studies

- Two additional UTOP studies conducted in partnership with the Gates Foundation’s *Measures of Effective Teaching* project, and NMSI
- Connect teaching behaviors to teacher value-added
- No UTeachers in this study – purpose was to validate and refine the UTOP
- No value-added results have been released
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
- 6 school districts in 6 states
- Collected via Teachscape video
- Many videos double-scored
NMSI/MET Studies

- Developed rubrics (1-5) for each indicator and examples of supporting evidence for each rating level ([online manual](#))
- Removed indicators that would be difficult to assess in video medium, without a teacher interview
- Training consisted of raters watching and rating 4-5 videos (one in pre-training webinar), group discussions, reviewing normed ratings
Lessons Learned: Rater Training

- **Pilot**: 4 GRAs trained through live pair observations over weeks/months
- **Study 1**: 17 raters trained on videos over 4 days, used only half of the UTOP.
- **Study 2**: 99 raters (divided into 2 groups) trained over 1.5 days, used entire UTOP.
Lessons Learned: UTeach Practices

- Most of the 4-8 math video lessons from this national sample did not score highly on the UTOP.
- Few/no examples of what the UTeach program would consider “exemplary” teaching.
- No quality teaching examples to use for training.
- Study 1 raters still giving some videos high scores.
- UTOP being used as “norm-referenced” instead of “criterion-referenced.”
Lessons Learned: UTeach Practices

- 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
- UTOP designed to deeply assess content-specific issues
## Problematic Content Examples

| Inaccurate | “A marker has no volume because you can’t fit anything inside of it.”
|            | “You can use the commutative property of division.” |
| Grade Level Issues | $\sqrt{x^2} = x$
|            | $x(x-1)/x = (x-1)$ |
| Formulaic/ Not Generalizable | “Remember when solving proportion word problems, that ‘of’ means divide.”
|            | “For fractions in this class, remember the denominator is always the bigger number.” |
| Problematic | - Using the same variable name in single problem to represent different quantities
|            | - Misuse of equals sign: $7 + 3 = 10 + 2 = 12$
|            | $Dan = 4$ |

Examples are either fictionalized, or occurred enough times that teacher is not identifiable.
Video #2

- 8th grade mathematics lesson on graphing linear equations
- Given 10 equations to graph, 5 with positive slope, 5 with negative slope
- Working in groups
  - Clip 1: 1:46-2:26
  - Clip 2: 6:39-10:16
  - Clip 3: 26:20-27:20
Video #2

- Discussion Points:
  - What does this teacher do well?
  - Are there any weaknesses or missed opportunities you see in his instruction?
  - How does this teacher’s style compare to the teacher in the other video?
    - Interactions with students
    - Framing/choice of task
Lessons Learned: UTeach Practices

- Little emphasis on conceptual understanding
- Mostly direct instruction mixed with (ineffective) group work
- Many classrooms tightly managed, little “student talk”
Lessons Learned: UTeach Practices

- Surface-level engagement often seen, but deep conceptual thinking about significant mathematics ideas rare
- Instrument (and observers) were able to capture this distinction ("hands on" vs. "minds on") – accentuated in manual & in training
Factor Analysis of UTOP

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

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

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

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

Cluster 4: Making Content Connections
- To real world (authentic)
- To other disciplines
  - To “big picture”
- To history/current events
## Dimensions of Teaching

<table>
<thead>
<tr>
<th>Content-General</th>
<th>Standard Teaching Behaviors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cluster 1: Fostering Surface Engagement</td>
</tr>
</tbody>
</table>
## Dimensions of Teaching

<table>
<thead>
<tr>
<th>Content-General</th>
<th>Standard Teaching Behaviors</th>
<th>Advanced Teaching Behaviors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>Cluster 1: Fostering Surface Engagement</em></td>
<td><em>Cluster 2: Fostering Deep Conceptual Understanding</em></td>
</tr>
</tbody>
</table>
## Dimensions of Teaching

<table>
<thead>
<tr>
<th></th>
<th>Standard Teaching Behaviors</th>
<th>Advanced Teaching Behaviors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content-General</strong></td>
<td>Cluster 1: Fostering Surface Engagement</td>
<td>Cluster 2: Fostering Deep Conceptual Understanding</td>
</tr>
<tr>
<td><strong>Content-Specific</strong></td>
<td>Cluster 3: Content Accuracy &amp; Fluidity</td>
<td></td>
</tr>
</tbody>
</table>
## Dimensions of Teaching

<table>
<thead>
<tr>
<th></th>
<th>Standard Teaching Behaviors</th>
<th>Advanced Teaching Behaviors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content-General</strong></td>
<td><em>Cluster 1: Fostering Surface Engagement</em></td>
<td><em>Cluster 2: Fostering Deep Conceptual Understanding</em></td>
</tr>
<tr>
<td><strong>Content-Specific</strong></td>
<td><em>Cluster 3: Content Accuracy &amp; Fluidity</em></td>
<td><em>Cluster 4: Making Content Connections</em></td>
</tr>
</tbody>
</table>
Dimensions of Teaching Assessed by UTOP

Bars show standard error of the mean.
Factor Analysis of UTOP

- UTeachers scored higher in all 4 clusters
- Most pronounced differences in:
  - Fostering deep conceptual understanding
  - Content-focused indicators
- These behaviors seem to be somewhat rare in general teacher pop, but were more often successfully employed by UTeachers
- UTOP well-designed to assess these types of teaching behaviors
Lessons Learned: Rater Training

- Raters did not always agree on classifications of teaching behaviors measured by UTOP
- Identified and reviewed lessons with “severe disagreement”:
  - Extremely charismatic teacher using low-level teaching strategies (most common)
  - Weak teacher using some elements of reform approaches (or maybe just discourse of reform)
  - Investigative lesson where mathematical content is left implicit/localized
- Re-norming webinars and interventions for raters with consistent issues differentiating rating levels
Lessons Learned: Rater Background

- When we conducted the pilot work at UT, a key question we confronted was, **who is qualified to rate lessons with the UTOP?**
  - UTOP focuses on Content Knowledge
- Conducted reliability analysis of double-scored videos for 99 raters from Study 2
  - Different raters had vastly different reliability (kappas range from 0.122 to 0.631)
Lessons Learned: Rater Background

- Raters from science background have lower reliability (weighted kappa) than raters from math background ($p < .05$)
  - Higher standards for engagement
  - No overall content section differences

- Raters with education/science education/other degree have lower reliability than raters with math/science/math ed degree ($p < .01$)
  - Higher standards for engagement
  - Less likely to catch content issues
UTOP Overview

- UTOP distinguishes between surface-level and deep-level engagement, and raters must be able to make this distinction.
- UTOP assesses content-specific aspects of pedagogy, and should be used by content experts.
- UTeachers excel most at facilitating deep-level engagement, and content-specific behaviors.
- UTeachers improve on UTOP behaviors with more experience.
- UTOP scores depend on context of lesson and characteristics of observer – multiple observations, 2 observers present.
Lessons Learned: Rater Training

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

- Video library of teaching behaviors assessed on UTOP
- Use of UTOP to examine teaching practices of UTeach graduates
- Use of UTOP as part of teacher professional development
- Use of UTOP for evaluation of university teaching.
- Use of UTOP to develop lessons (PhET)
Questions, Comments, Suggestions?