Center for Education Policy Research

HARVARD UNIVERSITY

When Teachers Choose: Fairness and Authenticity in Teacher-Initiated Classroom Observations

American Educational Research Association, Annual Meeting 2015

This toolkit is a product of the Best Foot Forward project, a study of video technology in classroom observations. ©2015 President and Fellows of Harvard College.

Objections to Video

- Authenticity: teachers can pick their very best lesson, so administrators won't see what a teacher's typical lesson is actually like
- Fairness: scores will be unreasonably inflated, and it will become difficult to differentiate between teachers

Observation Setup

- Three observations across the year
  - Suggested, but not strict, deadlines
- Teachers could shoot as much as they wanted, but had to chose one video to submit to their administrator
- Gives us a pool of chosen and unchosen videos

Research Questions

- Did teachers score higher on videos that they chose to submit for observation than on videos they shot but chose not to submit?
- Are scores on chosen videos indicative of scores on unchosen videos?
- Do teachers think that their chosen lessons are of different quality than their typical ones?
- Do administrators believe that teachers are submitting different quality lessons?

Prior Research

Measures of Effective Teaching (MET) study in Hillsborough, Florida (Ho & Kane, 2013)

- Not for stakes; videos were shown to a teacher's own administrator, but were not used for official evaluation
- Find chosen/unchosen difference of 0.19 s.d. (0.072 on the 4 point scale used)
- Find disattenuated correlation of ~1 (taking into account that observations have measurement error)
- Find that chosen videos have both a higher reliability and wider distribution of teacher effects

Our Research Methods

Sent chosen and unchosen videos to 3rd party raters

- 197 videos, 900 ratings, 15 raters, and 60 teachers (30 elementary, 15 each MS math and ELA)
- Raters and teachers fully crossed within grade span
- Chosen and unchosen videos paired temporally

Our Research Methods

- All ratings were done using the CLASS rubric
- Evaluated on 4 domains (Emotional Support, Classroom Organization, Instructional Support, and Student Engagement)
- Raters certified prior to project and required to calibrate on 4 separate occasions during the project

Our Research Methods

Surveys:

– Given to both teachers and administrators in October and June

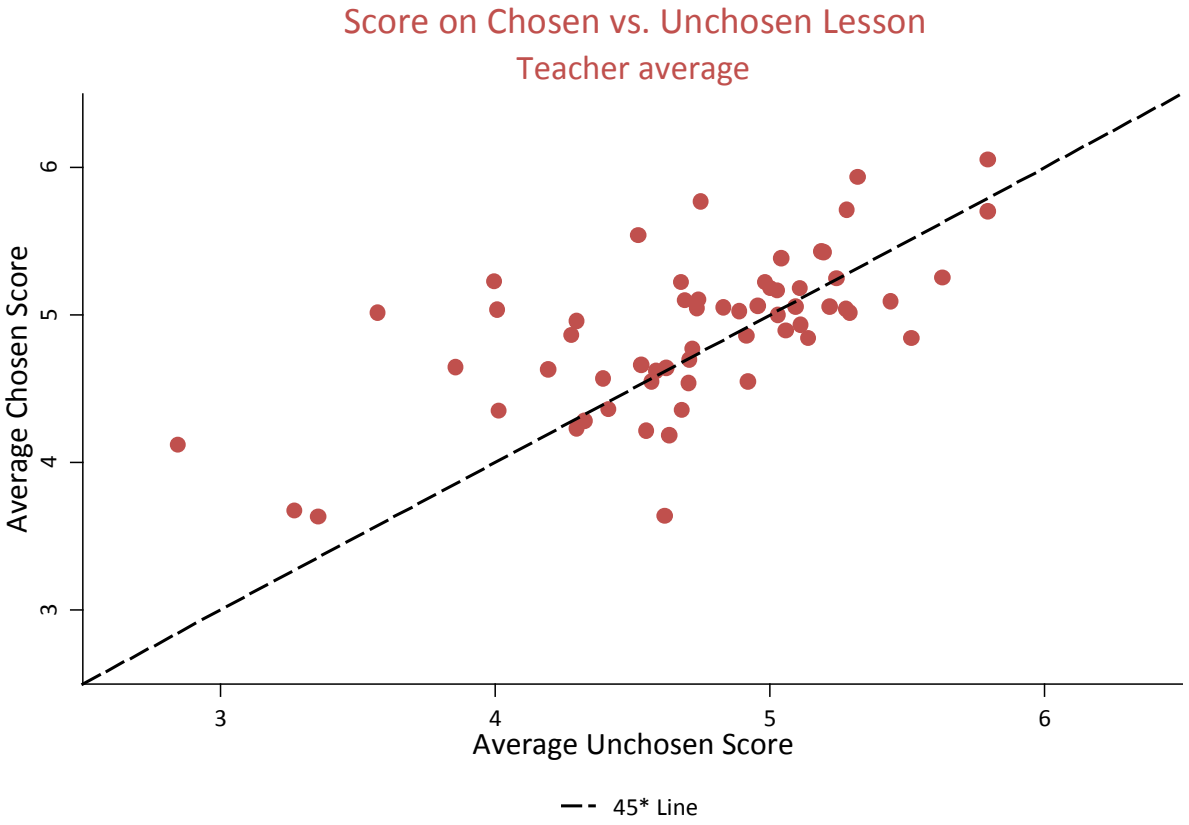
– Ask questions about authenticity and trust in the system

Findings

- Mean scores of chosen videos are higher by about a quarter of a standard deviation (0.17 points on the 7 point CLASS scale)

- But scores on chosen and unchosen videos are highly correlated (~0.75) after accounting for measurement error

Findings


Score on Chosen vs. Unchosen Lesson
Teacher average
Average Chosen Score
3
4
5
6
Average Unchosen Score
3
4
5
6
45* Line

Findings

- Unlike MET, we find that unchosen videos had higher reliability, at 49% compared to 41% for chosen videos
- Unchosen videos also have higher teacher-level variance, by 39%.

Findings

Large differences by grade span

- Much higher reliability in elementary
  - Generally MS has a much higher portion of lesson variance, but lower overall teacher + lesson variance
- Elementary has much stronger chosen-unchosen correlation (elementary at ~1, MS at ~0.3)

Teacher Survey Findings

Question	Control Mean	Treatment Difference
How confident are you that your classroom observation rating this year will be an accurate assessment of your teaching?	0.537	0.058 (0.060)
Thinking about the lessons that were used during your classroom observations this year, how much better or worse was the quality of your instruction when compared with a typical day?	0.093	0.049 (0.035)
Overall, how fair was the classroom observation process this year?	0.583	0.121*** (0.046)

Note: Standard errors are reported in parenthesis, and allow for clustering within school. *, **, *** indicate significance at the 90%, 95%, and 99% level, respectively. Results are reported in terms of the fraction of teachers selecting the top two responses on the Likert scales used (e.g. “Moderately fair” or “Extremely fair” for the last question)

Administrator Survey Findings

Question	Control Mean	Treatment Difference
Thinking about the teachers who were part of the study this year, how confident are you that your classroom observation provided an accurate rating of their teaching?	0.761	-0.053 (0.089)
Thinking about the teachers who were part of the study this year, how much better or worse was their teaching during the lessons used for their classroom observations than when they were not being observed?	0.304	-0.073 (0.093)

Note: Standard errors are reported in parenthesis, and allow for clustering within school. *, **, *** indicate significance at the 90%, 95%, and 99% level, respectively. Results are reported in terms of the fraction of admins selecting the top two responses on the Likert scales used (e.g. “Moderately better than usual” or “Much better than usual” for the last question).

Synthesis

- Fairness: Teachers perceive the process as more fair, and ranking of teachers is largely preserved
- Authenticity: teachers and administrators do not report a difference in authenticity, but they do receive higher scores

Future Directions

- Evaluating all lessons, not just the subset the teacher recorded
  - Even shot-but-not-submitted videos may not be indicative of typical lessons
- Extending study to novice and junior teachers
- Identifying what, if any, teacher characteristics predict quality differential of chosen videos

Appendix –
Disattenuated Correlation Formula

$$\rho = Corr(Score\downarrow chosen,i,r\ ,\ Score\downarrow unchosen,i,r'\ )/\sqrt{rel\downarrow chosen * rel\downarrow unchosen}$$

- $Score_{chosen,i,r}$ is the score of a chosen video from teacher *i* by rater *r*
- $Score_{unchosen,i,r'}$ is the score of an unchosen video from teacher *i* by a different rater *r'*
- rel_{chosen} and $rel_{unchosen}$ are the reliability of chosen and unchosen videos

Appendix – Variance Components

Study	Percent of Variance					
	Teacher	Section	Lesson	Rater	Rater by Teacher	Residual
Kane & Staiger (2012)	37	4	10	6	43	
Ho & Kane (2013)	39		7	13	17	23
Administrators	45		5	10	15	24
Peers	27		11	17	21	22
Best Foot Forward	37		21	10	7	25
Elementary	49		15	8	6	12
Middle School	19		30	14	8	30