State and Local Efforts to Investigate the Validity and Reliability of Scores from Teacher Evaluation Systems

Corinne Herlihy
High-Stakes Conference
University of Arizona
October 12, 2012
Overview

• Introduction and purpose
• Sample and Methods
• What do we mean by valid and reliable?
• Findings
  ▫ Reliable and Valid
  ▫ Evaluators, Instruments and Sampling Lessons
  ▫ Decisions, PD for Teachers, and Changes in School Culture
Purpose

• We examine states’ efforts to investigate and ensure the reliability and validity of scores emerging from new teacher evaluation systems
• We focus on areas known to be of concern in the production of high-quality observational scores
  ▫ The choice of the observational instrument
  ▫ Rater training and certification
  ▫ The number of lessons collected per teacher per year
Sample of 17 States

• Each state received a RTTT grant or NCLB waiver before July 1, 2012;
• Conducted a pilot-test of its new teacher evaluation system during or before the 2012-13 school year;
• Had statutory language describing a teacher evaluation system which satisfied the requirements of RTTT or NCLB waiver;
• And did not have any pending legislation, as of 7/1/12, which would substantially change the system
Methods

• Document collection and analyses to answer basic questions about state-level guidance
  ▫ Legislation and government guidelines,
  ▫ RTTT applications, and
  ▫ NCLB flexibility requests, etc.

• Led to creation of a 25-question interview protocol

• Interviews with 13 state DOE staff from 12 states, typically a director, coordinator or executive officer of the state’s efforts to implement new teacher evaluation system
Reliable and Valid

What do we mean by reliable and valid?
What do states mean for scores to be reliable and valid?
What efforts are underway to ensure the reliability and validity of scores?
What would you like to see in order to assure reliable and valid scores?
**Making the Grade in New York City**

Do the progress reports really help develop good schools? Or do they simply confuse parents and irritate teachers and principals?

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>SCORE</th>
<th>GRADE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student Progress</td>
<td>18.1</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td>out of 60</td>
<td></td>
</tr>
<tr>
<td>Student Performance</td>
<td>8.5</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td>out of 25</td>
<td></td>
</tr>
<tr>
<td>School Environment</td>
<td>7.8</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>out of 15</td>
<td></td>
</tr>
<tr>
<td>Closing the Achievement Gap</td>
<td>2.5</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>(16 max)</td>
<td></td>
</tr>
<tr>
<td>Overall Score</td>
<td>36.9</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>out of 100</td>
<td></td>
</tr>
</tbody>
</table>
What do we mean by reliable and valid?

• Score validity
  ▫ Construct validity
    • Face validity
    • Factor analyses or other construct identification procedures
    • Criterion-related validity
  ▫ Consequential validity
    • Both for the individuals in the system
    • And for the system itself – are there positive or negative unintended consequences?
What do states mean for scores to be reliable and valid?

- 9 states responded
- The most common answer (3 states) was to see congruence between observation and student-assessment-based metrics
- Two answers focused on elements of the system that would improve validity and reliability
  - a distribution of scores that better reflects reality
  - raters who are more grounded in benchmarks and evidence
- Some answers were very vague, only loosely related to standards for validity and reliability
What efforts are underway to ensure the validity and reliability of the scores?

- 12 states responded
- Responses varied widely across states
- Taken together states are attending to many of the issues involved in establishing reliability and validity of assessment systems
- Rare to see a state attending to more than a handful of issues at a time, and some states are attending to only one
What efforts are underway to ensure the validity and reliability of the scores?

- Enhanced training of evaluators to meet standards
- Empirical investigations (finished or planned)
  - Correlate VAM and observation scores
  - Investigate inter-rater reliability
  - Factor analyses of data
  - Studies of validity and reliability studies
- Auditing/monitoring techniques planned
- Selection of instruments with evidence of reliability and validity
What would you like to see in order to assure reliable and valid scores?

- 8 states responded
- Responses widely varied across states, and, where we interviewed more than one official per state, variable within states as well.
- Two states were concerned with technical aspects of the production of VAM scores
- Two states expressed concern about the reliability and validity of assessments in non-tested subjects
- Two states suggested a process for improving reliability and validity
- Other state concerns involved widening the pool of evaluators
Evaluators, Instruments and Sampling

Who is required to conduct teacher observations?
How did states go about choosing the instrument?
How did states arrive at the number of required observations?
Are there plans to report reliability and validity of scores?
Questions about the evaluators

- Have you thought about how to account for possible subjectivity from raters? For example, principals/ coworkers assigning overly harsh or lenient scores?
- Is your state/office involved in the training of raters of classroom instruction? How?
- Is your state/office involved in the certification of raters of instruction?
- Is your state/office involved in the monitoring of raters once they pass certification?
- Do raters periodically have to recertify?
What we learned about evaluators

• 17 states responded
• Principals are primary and lone evaluators
• All evaluators will participate in some type of training
• Training usually focuses on familiarizing raters with the evaluation process, rubrics and types of evidence, although some training also includes how to provide effective feedback
Evaluators

Who is required to conduct teacher observations?
Principal/Administrator: (DE, FL, LA, MD, TN, NY, IL)
Principal/Administrator OR instructional leader, other designee: (CO, IN, MA, OH, RI)
Principal/Administrator and peer (NC)
No specification (AZ, CT, GA, OK)

Do states require multiple raters?
Yes (NC)
Recommend (IN)
Conditional, only for an evaluation of ineffectual (MD)
No (AZ, CO, CT, DE, FL, GA, IL, LA, MA, NY, OH, OK, RI, TN)
What we learned about evaluators

• Rater training is both centralized and un-centralized, and in-house and by contract firms
• Although some states talk about certification, very few describe an object evaluation process for raters
• Many states will likely award certification for attending training sessions
• Very few states specify a timeline for requiring rater recertification
• Many states plan to provide ongoing professional development and support to raters but this is rarely an objective monitoring process
Evaluators

Who is responsible for training raters?
State training (DE, GA, IL, MD, MA, OH, TN)
District training (CO, CT, FL, IN, LA, NY)
No specification or Other (AZ, NC, OK, RI)

Do states require raters to meet objective certification criteria (other than simply attending a training workshop)?
Yes (IL, LA, OH, TN, NC, OK)
No (AZ, CO, DE, GA, CT, FL, IN, MA, MD, NY, RI)
Questions about observational instruments

• How did you go about choosing the instrument?
• What did you know about the tool before you adopted it?
  ▫ Did you have any information about reliability of validity?
• What do you know about the way it has been adapted by districts?
What we learned about instruments

• 11 states responded to question about how the instrument(s) was chosen
  ▫ Most instruments were chosen by a committee or task force to align with state standards
  ▫ Five states created their own tool; others uses an existing/modified tool(s)

• 9 states responded to questions about how the instrument has been adapted by districts
  ▫ Unless the state model was mandated, few states knew of district implementation plans
Type of Evaluation System

- Entirely state-based system (DE, GA)
- State model(s) with some district choice (IL, MD, MA, NC, OH, OK, RI, TN)
- Mainly district developed (AZ, CO, CT, FL, IN, LA, NY)
What we learned about instruments

• 10 states responded
• Most states reviewed some information about the validity and reliability of the instruments chosen
  ▫ Two states did their own study
  ▫ One state plans to do its own study
  ▫ Three states looked at data provided by outside sources
  ▫ Four states did not know about the reliability or validity of the instrument
Questions about sampling of lessons

• We know that statutes require X observations per teacher per year. Can you tell me a little about how you arrived at that number?
• Was any kind of study conducted using your instrument in order to determine the optimal / most cost efficient number of observations? If so, when? If not, why not?
• Is there any flexibility around this number? Off-the-books guidance to districts around the number of observations?
• Have you determined if any of these observations will be unannounced? Why or why not?
What we learned about sampling of lessons

- 10 states provided information about how the number of observations was determined
- None indicated that there was a formal study to determine the optimal number of observations
- Four states mentioned using research from other studies (MET was mentioned three times) to determine the number
- Five states indicated that the number was arrived at through a negotiation process with teachers' unions
- Five of the states indicated that the number was chosen, at least in part for practical and logistical reasons
- Two states said that raters are encouraged to conduct more observations if needed
States by Number of Observations Required

- 0 (OK)
- 1 (FL, DE)
- 2 (AZ, CO, GA, IL, IN, LA, MD, NY, OH)
- 3 (CT, MA, NC, RI)
- 4 (TN)
Are there plans to report the reliability and validity of scores?

- 11 states responded
- Few states have plans to publish reliabilities, though many states plan to conduct research on the new systems
- Only 3 states (of 17 in full sample) have legislation that requires a report on aspects of reliability and validity
State specifies inquiry into overall scores

CT, IL, NC, OK, TN

CO, GA, MA, NY, TN

LA, OH

Inquiry into value added scores

Inquiry into observation system
Decisions, PD for teachers, change in school culture

What purposes will the overall teacher evaluation scores be put?
Are there efforts to ensure those decisions are correctly made?
What kinds of PD are teachers receiving about the new systems?
Do you think the evaluation system may change the culture in schools?
What purposes will the overall teacher evaluation scores be put?

- 17 states responded
- All but one state (AZ) specify consequences of poor performance on a teacher evaluation, ranging from not receiving raises, participating in assistance programs and remediation plans to dismissal and termination.
- All but one state (CT) specify some consequence for excellent performance, ranging from bonuses, salary increases, non-probationary status, tenure, and fewer observations in the future.
Are there efforts to ensure those decisions are correctly made?

- Only 3 states responded
  - Few states have specified ways to ensure correct decision making.
- Most states say they are waiting to see how the new system works and may make adjustments
- “Yes, again, that’s work over time. That’s part of—back to those validity arguments that we’re going to be making. So right now our big check is are we correctly identifying effective teachers. And has someone been misidentified? And so there are procedural safeguards in place for that. And we’ll be following the data over time. But we haven’t—last year was a gradual implementation of the model. We’re about to launch the first full year of evaluations, so we’ll be doing that work.”
What kinds of PD are teachers receiving about the new systems?

- 12 states responded
- 10 indicated that professional development and information for teachers was largely left to districts
- One state had been directly involved in informing teachers by delivering presentations and talks to different schools
- One state that had recently passed legislation charging the state to provide training on the evaluation system to all teachers
What kinds of PD are teachers receiving about the new systems?

- There were four states that were using variations of "train the trainer" models.
- Two states described providing building level principals "meetings in a box" to disseminate information about the evaluation systems.
- Four states mentioned that they were providing information about the evaluation and training opportunities through online platforms. Webinars were mentioned by officials from two states.
Do you think the evaluation system may change the culture in schools?

- 7 states responded
- All respondents acknowledged to some degree that the new evaluation systems would identify fewer teachers in the top tier than the previous system did
- Four of the states specifically mention that the transition will be difficult
- Four states talk about the difficulties in redefining the ratings of "satisfactory" and "proficient" so that the expectation isn't that all teachers should pile up at "exemplary“ category
Do you think the evaluation system may change the culture in schools?

• In one state, there are plans for principals themselves to be evaluated on their ability to be good evaluators, which may, the official said, help to temper any perverse incentives principals could have to assign inflated scores.

• Finally, one state framed the shift to identifying fewer exemplary teachers as quite positive, saying,
  ▫ "It’s going to come down to embracing the principles behind the model and being able to communicate and message that throughout schools and districts while really fostering a sense of trust, that this is not about - it’s not a gotcha, but this is about identifying, recognizing truly excellent practice, ...seeing this as a collective effort to get better... knowledge is power."
Conclusions

• An important goal for research and policy-makers is to investigate whether the right decisions are made based on evaluation scores
• Few seem to be considering the negative unintended consequences of a system that may generally be perceived of as being arbitrary
• There has been huge change in a very short time, and there may be good reasons why states are falling short in the effort to balance better systems and available resources—but, states are not necessarily making informed trade-offs
• While most states seem to understand that implementation is key, they may not understand its implications for validity and reliability
If you have questions or would like more information, please contact:

Corinne Herlihy
Project Director
National Center for Teacher Effectiveness
Harvard Graduate School of Education
50 Church Street, 4th Floor
Cambridge, MA 02138
617-496-9817
corinne_herlihy@gse.harvard.edu