The New Teacher Project

Transforming Teacher Evaluations

NCTE Webinar – October 2011

Nearly everyone believes that that teacher evaluations are broken. The crucial question now is: How can we fix them?

Problems With Current Evaluation Systems

Infrequent: Teachers can go years between evaluations.

Unfocused: Student academic progress is rarely a factor.

Undifferentiated: Nearly all teachers are rated good or great.

Unhelpful: Teachers say evaluations don't give them useful feedback.

Inconsequential: Ratings rarely factor into employment decisions.

The result: We treat teachers like interchangeable parts.

"Everyone agrees that **teacher evaluation is broken.** Ninety-nine percent of teachers are rated satisfactory and most evaluations ignore the most important measure of a teacher's success - which is how much their students have learned."

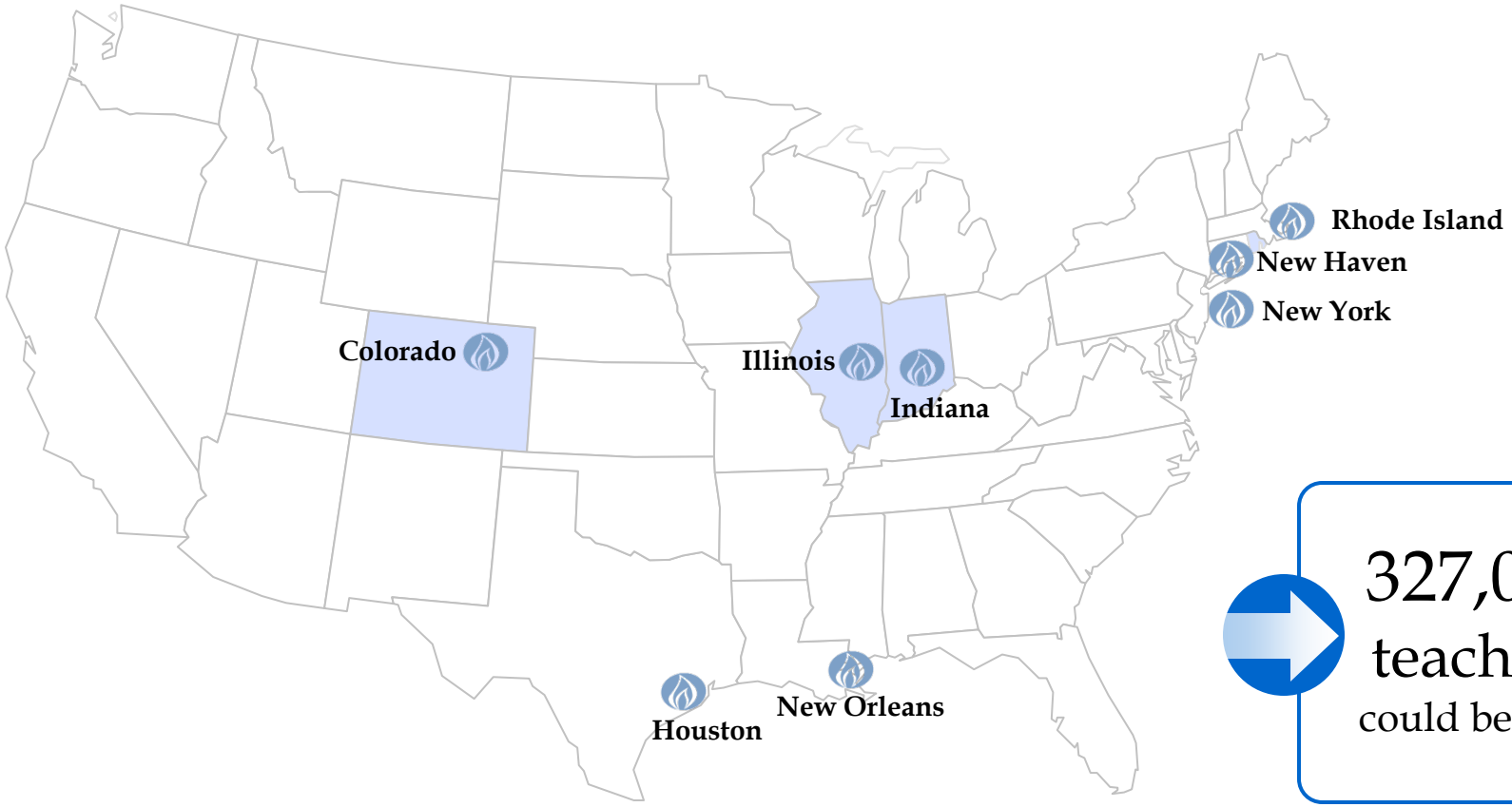
U.S. Education Secretary Arne Duncan
Remarks at the Natl. Press Club, July 2010

"**Our system of evaluating teachers has never been adequate**… [It] has failed to achieve what must be our goal: continuously improving and informing teaching."

AFT President Randi Weingarten
Remarks at the Natl Press Club, January 2010

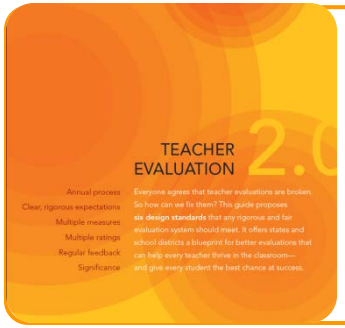
TNTP is collaborating with a number of districts and states on teacher evaluation design and implementation.

TNTP Evaluation Design and Implementation



327,000 teachers could benefit.

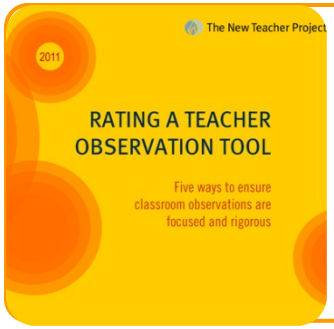
Since October 2010, we have published three publicly available guides to improving teacher evaluation systems.



Teacher Evaluation 2.0

October 2010

Six design standards that any rigorous and fair teacher evaluation system should meet.



Rating a Teacher Observation Tool

February 2011

Five simple questions to help policymakers pick a strong observation framework.

Smart Spending for Better Teacher Evaluation Systems

June 2011

Investments states should make to implement a new teacher evaluation system successfully.

TNTP has played a number of roles in past and current partnerships.

Services	
Stakeholder facilitation	At both district and state level, incorporation of input into design models. This is time consuming and expensive, but often necessary.
Rubric design and modification	We have created rubrics for classroom observation from scratch and updated existing tools.
Appraiser training	Initial system rollout, ongoing co-observations, re-training. This work has been mostly district-level.
Communications	Creation of custom messages and delivery strategies to ensure that users understand the new system well.
Data monitoring	Collection of real time data on implementation that can be used to correct course.
SEA and LEA capacity	Re-organizing roles and responsibilities for agency staff to support implementation of new evaluations long term

Early conclusions

1. **Improved evaluations are likely to show that mediocre instruction is much more common than has been publicly acknowledged. Basic instructional skills are not in evidence in a large number of classrooms.**

2. **Implementation depends on working closely with administrators of all levels to hold a high standard for classroom instruction. The tendency to inflate is extremely strong, based on a combination of past culture and administrator skill.**

3. **Incorporating evidence of student learning is hard work but can be done – if we acknowledge that professional judgment is an important part of measuring student learning. There is confusion among practitioners about whether new evaluations are meant to improve or obviate professional judgment. Only systems that do the former will succeed.**

How we think about: Training Appraisers

- Training is not a one shot deal. Takes significant investment in up front training and then continuous follow-up, evaluator development, and coaching. We are focusing more on the concept of evaluator development. Norming on focused video segments is easier but doesn't translate smoothly to real schools. If given the choice between investing more resources in up-front training to achieve the highest level of statewide consistency vs investing fewer resources up front and saving more for ongoing re-training, choose the latter.

How we think about: Inter-rater Reliability

- It is possible to make too much of this. While inaccurate evaluations will quickly (and rightly) expose new systems to criticism and challenge, a narrow focus on the degree to which all administrators rate the same will take us down the wrong path. First, the most likely scenario in which there is high IRR is one in which everyone is inflating. Second, most differences in evaluator ratings are inconsequential, meaning they do not affect summative decisions or actions. Third, even when we have invested very large resources in policing IRR, there are some schools that tend to be outliers, for local reasons that are understandable and extremely difficult to overcome.

- A better way to think about this might be a strike zone. Baseball umpires do not all have the same strike zone, and players know this very well. However, players respect an umpire who establishes the parameters of the zone and enforces it consistently. Now, there are limits to the amount of idiosyncrasy that can be tolerated, but within those limits, perfect consistency is not necessary. Help raters to develop and implement a strike zone.

- Expect outliers and be ready to bring them back to the strike zone.

How we think about: System Capacity

- The most important leverage point in delivering better evaluations is the group of people who supervise the principals – whether they are superintendents, assistant superintendents, etc. Their job is to cultivate strong professional judgment. They can do this by ensuring that principals have every reason to deliver good, accurate evaluations, and by stepping in to correct course when it is not happening. Pressure or accountability that does not come from the principal's direct manager will have substantially less power.

- Therefore, building internal capacity at the district level is essential. Outside vendors can play a useful role in generating and disseminating data so it can be acted on quickly, and by generating supports. But without a focus on principals making smart judgments, as opposed to a fool-proof system, we'll end up in about the same place we started.