An Introduction to NCTQ's Project Management System: RevStat

A system designed to support the reliability of ratings in NCTQ's *Teacher Prep Review*

Ensuring reliability in NCTQ's earlier, smaller-scale teacher prep studies:

1. In-house, supervised evaluation.
2. Allowing IHE's to preview their findings and/or ratings in time to make corrections before publication.

Ensuring reliability in NCTQ's *Teacher Prep Review*:

1. Supervision of off-site evaluation through in-house Team Leaders using the **RevStat** management system.
2. A limited "**Due Diligence**" process for the purpose of identifying systematic flaws in ratings methodologies.
3. A post-publication, public on-line "**Forum**" for resolution of rating issues.

Why RevStat?

- Given the scale of the *Teacher Prep Review* and the need to monitor as many as 80 off-site analysts, RevStat is essential to maintaining accuracy of data collection, processing, and analysis.

- RevStat disciplines all rating processes. NCTQ's custom-designed central database generates regular data on ratings by standard. Team Leaders are trained to answer key questions about the data in regular RevStat reports and initiate further research and action as necessary.

What is the “-Stat” process?

A “-Stat” process is used by NCTQ and other organizations for purposes of:

- Performance management and quality control.
- Creating a framework for structured team learning.
- Ensuring that all process data is short-cycle and allows for instant adaptation as necessary.

When was the “-Stat” process developed?

- In the early 1980s, Bill Bratton, Chief of the Boston Transit Police, started mapping crime on maps.
- He held bi-weekly meetings with captains to monitor trends and allocate resources.

New York City adopted the process and named it “CompStat”:

- In 1994 Jack Maple, Bill Bratton, and Rudy Giuliani began to use crime statistics to target resources and hold front line officers accountable for results.
- They were awarded the Harvard Innovation in Government Award.

The Baltimore City Schools has developed SchoolStat:

- Baltimore City School System began implementing School-Stat over central office operations in September 2001.
- This was the first application of the Stat process to school systems.

Why is RevStat a Good Match for this Project?

- Much of the work involved in the review is done by “virtual teams”—teams of off-site analysts managed by in-house staff members.
- There is a large workflow:
  - Each rating is the end point of a complex and inter-woven sequence of processing and analysis.
  - RevStat’s database records data at each work step, allowing managers to identify hard to see trouble spots with complex origins.
- We aim for complete reliability:
  - We need to have rigorous quality control; RevStat reports help ensure reliability.

RevStat reminds NCTQ staff to ask and answer four key questions on an ongoing basis:

- Is the *Teacher Prep Review* on track for completion?
- What are the magnitude and sources of disagreements between analysts in ratings and what can they tell us about ratings processes?
- How faithfully is each analyst working according to established protocols? Are rating distributions by any individual, or the team as a whole, drifting?
- Are we periodically reconnecting analysts to the purpose of our work and reinforcing our mission to improve teacher prep?

Structure and Schedule of RevStat

On a daily basis, NCTQ staff provide non-stop support and oversight:

- Five Team Leaders field questions related to analysis and resolve analytical issues with senior staff as necessary.

On a regular basis (usually every other week), Team Leaders:

- Monitor "gross" agreement by examining variances “flagged” by database (as defined by differences in indicator or final ratings on standards).
- Monitor individual analyst and team rating distributions over any given time period.
- Identify rating "drift" over time by individual analyst and/or by team.
- Facilitate weekly or alternate week team conference calls/webinars about ratings issues based on RevStat data.

Triggers for in-depth examination of variances:

- An agreement rate of less than 90% for the majority of standards for which there are two independent analyst.
- Analyst rating distributions that differ significantly at any given score level when comparing analysts to team averages.
- Individual and/or team ratings distributions that differ substantially from a baseline period.
- Four standards with the potential for identical ratings to mask different evaluations receives special attention from Team Leaders: Common Core Elementary Content, Lesson Planning, Student Teaching, and Outcomes.

"Disagreement" rates currently reflect...

Final ratings that differ by <u>any</u> amount (because ratings use 2 or 3 -part scales or because database has specially set "flag"):	Final ratings that differ by <u>two or more</u> rating levels (because ratings use 5-part scale):
Selection Criteria	Early Reading*
English Language Learners*	Common Core Elementary Math*
Struggling Readers*	Common Core Elementary Content
Common Core Middle School Content	Assessment and Data
Common Core High School Content	Instructional Design for Special Ed.**
Classroom Management	Common Core Content for Special Ed.
Secondary Methods	
Outcomes	
Lesson Planning	
Student Teaching	

*Only one analyst with 10% oversample to assess reliability; evaluated by subject specialists.

**Each program evaluated by two subject specialists.

Approximately once a month, NCTQ's entire *Teacher Prep Review* team meets to discuss RevStat data:

- Each team leader compiles RevStat reports for standards for which analysis is in progress:
  - Reliability
  - Rating Distribution
  - Progress to Completion
- Team discusses and establishes steps necessary to address any issues.

From raw data to analysis: Example "Reliability" RevStat report

Date: August 20, 2012

Reliability Report: Selection Criteria

NOTE: IF ANY RESPONSES ARE NEGATIVE, A STATEMENT REGARDING THE POSSIBLE CAUSE(S) AND PLAN FOR REMEDIATION SHOULD BE PROVIDED.

Is the aggregate proportion of agreement among analysts greater than or equal to 90% over the past one or two weeks? Yes, the overall agreement rate is 92%.

Is the proportion of agreement for each analyst on the team greater than or equal to 90% over the past one or two weeks? Yes. Our overall agreement rate is 94.75% and our overall accuracy rate is 97%. Please see below for details.

Analyst 1: Over the last two weeks, Analyst 1 completed 212 programs with Disagreement/Exceeds Variance (EV) ratings for 15 programs. One of these EV ratings was with Analyst 2, one was with Analyst 3, and 13 were with Analyst 4. This is a 93% overall agreement rating without determining which analyst was responsible for the EV rating. When those determinations are made, Analyst 1's accuracy rating increases to 96%.

Are the pairings of analysts for ratings distributed proportionally? Yes.

From raw data to analysis: Example "Analyst Rating Distribution" RevStat report

Date: September 14, 2012

Analyst Rating Distribution Report: Selection Criteria

NOTE: IF ANY RESPONSES ARE NEGATIVE, A STATEMENT REGARDING THE POSSIBLE CAUSE(S) AND PLAN FOR REMEDIATION SHOULD BE PROVIDED.

For each level of rating and for the last two weeks, does the proportion of each level of rating by each analyst appear to differ significantly at any level of rating from the team as a whole?

The proportion of "4s" of each of the four analysts is within 5 percentage points of the team average; the proportions of "0s" and "2s" differ more, but in no case do they deviate more than 9 percent and that in only one case. These deviations appear to be entirely explained by rating variances already examined and the fact that the analysts differ in rating activity in states in which state context is relevant to and impacts the selection criteria.

What is the baseline period? July 6- August 7

Does the distribution of ratings of each analyst and the team as a whole seem to be approximately those of the baseline period? While the proportion of "4" ratings is fairly constant, there appears to have been a more significant decrease in "0s" and increase in"2s."

PLAN FOR REMEDIATION: Examination of how evaluation of state's programs may explain this change in distribution.

The RevStat process also involves two other types of reports:

- The number of programs eligible for evaluation for “strong design” designation.
- The status of analyst training and/or development of data features relevant to the rating process.

Beyond the usual periodic review of reliability reports, trouble-shooting discussions will involve...

	Teacher Prep Review Leadership Team	Teacher Prep Review Director	Teacher Prep Review Team Leaders	Teacher Prep Review Rating Teams
NCTQ Leadership Team	✓			
Teacher Prep Review Leadership Team		✓	✓	
Teacher Prep Review Director			✓	
Teacher Prep Review Team Leaders				✓

Teacher Prep Review 2.0 and beyond

With the guidance of the Audit Panel advising us on rating processes, we plan to continue to improve the RevStat process for each successive edition of the *Teacher Prep Review*.