

Inter-Rater Agreement Session

Janine Rudder & Valerie Randall

U.S. Department of Education

Tony Milanowski & Jackson Miller

Westat

Session Agenda

- What is inter-rater agreement, why is important and how can it be maximized? (Tony Milanowski & Jackson Miller)
- The TAP Approach (James Snyder)
- Mastery Charter (Tabenah Washington & Rebecca Schatzkin)
- Group Discussion: Identifying Inter-rater Agreement Challenges & Questions (Janine Rudder & Valerie Randall)

What is Inter-rater Agreement?

- The extent to which two or more raters agree on a performance rating, given that they had the opportunity to observe the same or highly similar performance and used the same rating tool.
- Is it the same as inter-rater reliability?
 - Colloquially, yes; technically, no
 - Inter-rater reliability concerns the extent to which the variation in ratings is due to ratee characteristics, while agreement involves the degree to which raters make identical ratings of ratees.

Why is it Important?

- Measurement accuracy
 - We want to measure practice, not any one evaluator's *perception* of practice
 - Many evaluators have trouble letting go of their pet theories of good practice when observing others
- Credibility to educators:
 - “My rating depends on who observes me” vs. “Evaluators apply the criteria evenly across educators”

How Do We Measure it?

- At what level should we assess it?
 - Overall rating
 - By dimension or standard
 - By sub-dimension of component of standard
 - What is the level to which the consequences are attached?
- What summary should we use?

Three Common Inter-rater Agreement Summary Measures

Measure	Concept	Advantages	High - Minimum	Comment
% Absolute Agreement	How often do raters agree on the exact rating?	Simple to understand; applies to almost any situation	90%-70%	75%, a rule of thumb; should have no ratings more than 1 level apart.
Cohen's kappa	How well do raters agree, corrected for chance agreement?	Corrects for chance agreement; ranges from 0 to 1	.75-.50	Some view .80 as high and .40 as borderline
Intra-Class Correlation	What proportion of rating variation is due to raters versus ratee behavior	Single index with percent of variance interpretation	.90-.67	Will be lower if there is low variation across grantees

Resources: <http://www.john-uebersax.com/stat/raw.htm>

Cohen (1960) A coefficient of agreement for nominal scales

Shrout & Fleiss (1979) Intra-class correlations: Uses in assessing rater reliability

How Can it be Maximized?

- Clarify the rubrics
- Define the task
 - Explain and give examples of how to interpret adjectives such as vague quantifiers (e.g., “frequent”, “extensive”)
 - Specify what evidence is to be collected and how collection should be done; define what evidence for differences in rubric levels would look like
- Train & retrain!
 - If you train to the point where agreement is high, raters can be regarded as equivalent

How can Inter-rater Agreement be Maximized?

- Monitor
 - Review a sample of rating documents; does evidence cited in the documents support the ratings given?
 - Have an expert outside observer accompany regular raters and independently rate a sample of observations. Compare the ratings.
 - Compare patterns of ratings across raters. Compare patterns to other outcome measures (e.g., value-added).
 - If problems are noted with raters, retrain or reassign.



The System for Teacher
and Student Advancement

A NEW DIRECTION FOR SUCCESS

Reliability in TAP Observations

Teacher Incentive Fund Grantee Meeting
August 23, 2011

Inter-rater Reliability in TAP

Consistency between the scores assigned by members of the leadership team at different times during the year resulting from the process of coming to consensus on collected evidence and assigned scores based on the TAP rubrics.

Initial Training and Ongoing Support of Inter-rater Reliability

Know it

The first step to creating inter-rater reliability is truly understanding the standard (rubric) being used to evaluate.

Assess it

In order to measure this understanding, you need to assess evaluators application of the rubric in a controlled environment.

Monitor/Address it

Once this baseline has been set, you need to provide ongoing support and training towards applying it successfully.

What are Effective Ways to Monitor and Address Inter-rater Reliability?

To Monitor Inter-rater Reliability

To Address Inter-rater Reliability

System: C.O.D.E.
Comprehensive Online Data Entry

Observations Reports Administration Forums Calculate!

Reports

- [School Meeting Schedule by District \(Table\)](#)
At-a-glance of clusters /leadership team meetings schedule by school/district. This report is very effective for leaders of multiple locations.
- [School Meeting Schedule \(Table\)](#)
At-a-glance of clusters /leadership team meetings schedule by school/district. This report is very effective for leaders of multiple locations.
- [Observer Averages by Rubric Domain \(Table\)](#)
Provides Observer's average score of total number of Observations by rubric indicator/domain to target inter-rater reliability. To maximize the effectiveness of this report, use it in conjunction with the "Observer Schedule of Observations by Teacher" report.
- [Teacher Averages by Rubric Domain \(Table\)](#)
Total Observations completed or observed by teacher by rubric indicator/domain to target inter-rater reliability. This information can be used for improvement areas. This information should be used in conjunction with the "Observer Schedule of Observations by Teacher" report.
- [Teacher Levels Averages by Rubric Domain \(Table\)](#)
Total number of Observations by rubric indicator/domain by teacher level. This information can be used for improvement areas. This information should be used in conjunction with the "Observer Schedule of Observations by Teacher" report.
- [Grade Level Averages by Rubric Domain \(Table\)](#)
Total number of Observations by rubric indicator/domain by grade level. This report is used for improvement areas. This information should be used in conjunction with the "Observer Schedule of Observations by Teacher" report.

Test School C1 - Overall Averages by Rubric Indicator (2011-2012)

Rubric Indicator	Average Score
Instructional Plans	0.33
Student Work	0.45
Assessment	0.28
Exemplars	0.54
Managing Student Behavior	0.35
Environment	0.53
Assessing Culture	0.4
Standards and Objectives	0.47
Measuring Student Learning	0.28
Establishing Instructional Context	0.24
Lesson Structure and Pacing	0.36
Activities and Materials	0.31
Questions	0.29
Assessment Feedback	0.24
Grouping Students	0.33
Teacher Content Knowledge	0.39
Teacher Knowledge of Systems	0.35
Modeling	0.48
Student Learning	0.28

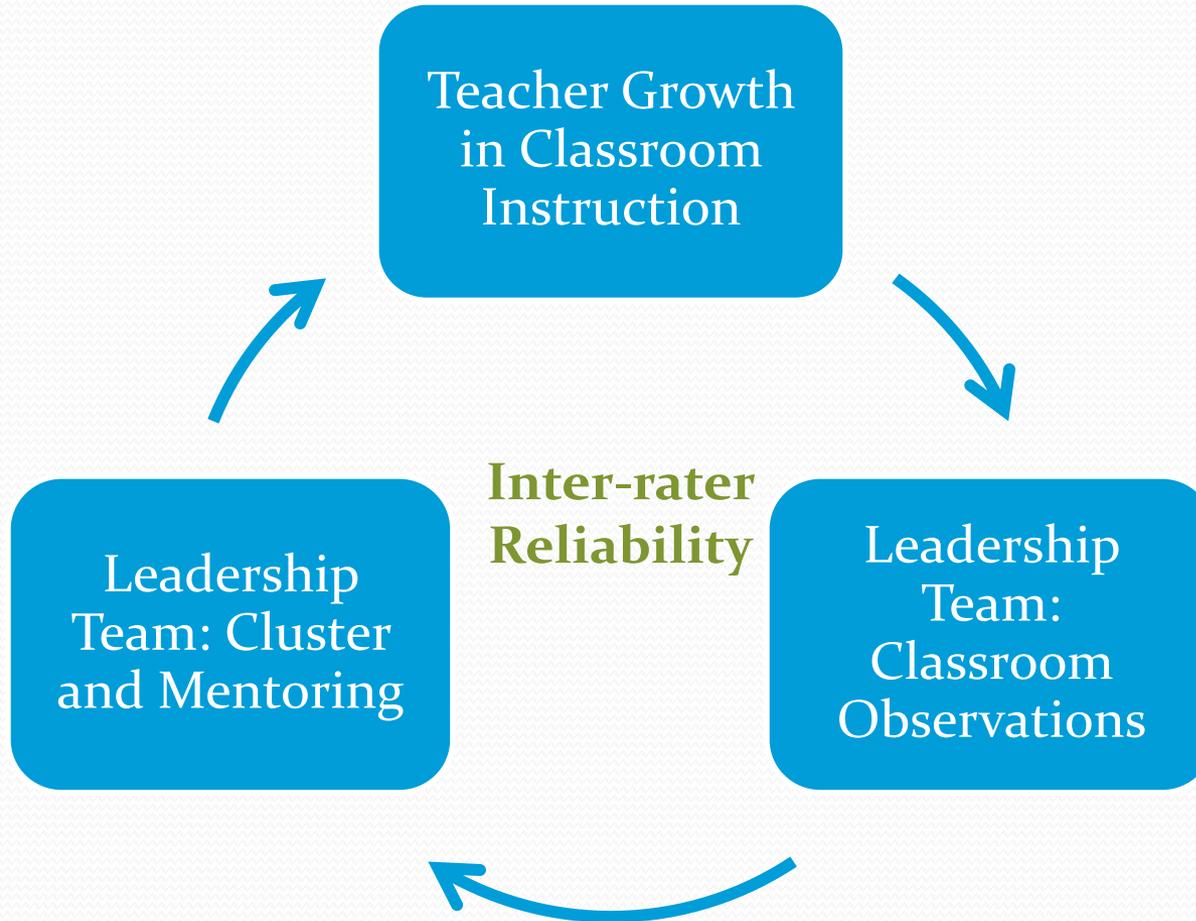
Language Arts - 1.3a. Fourth Grade (Identify Adjectives and Write Descriptive Paragraph) Lesson

Video.mpg

Language Arts: Grade 4

01:00:03:25

TAP Inter-rater Reliability in Practice: A Process Based upon Continuous Improvement



For More Information about TAP™ or
the Best Practices Center

NIET

NATIONAL INSTITUTE FOR
EXCELLENCE IN TEACHING

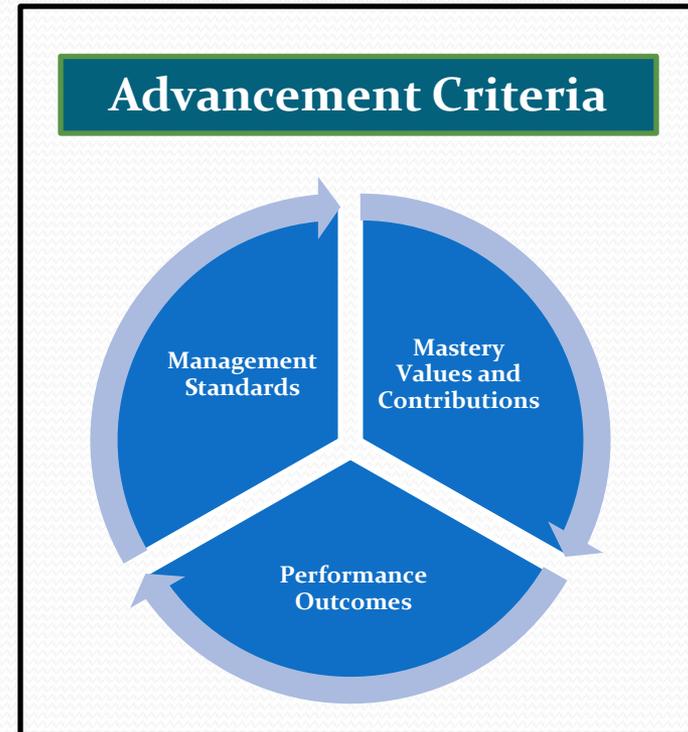
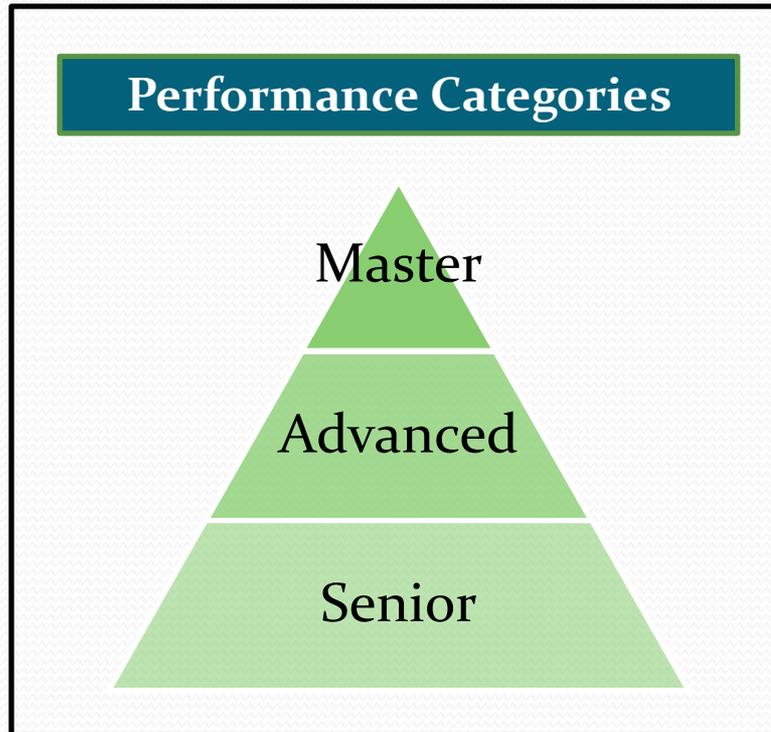
www.niet.org

Mastery Charter Principal Evaluation

- M₃ (Mastery Management Model) System for Evaluation
- Aligned System
 - Training and PD
 - Cycle of Feedback
 - Peer Leadership Review
 - Consistent Regional Support
 - Show me the Data
 - Formal Evaluations (MY/EOY)

Performance Based

Management Management Model



**Annual
Performance
Reviews**





Mastery Charter Schools
Excellence. No Excuses.

Mastery Charter Principal Evaluation

STEP 1: Training and PD

- Apprentice School Leaders or APs prior to appointment
- 3 day management orientation/training (July)
- 2.5 hour principal training every 3 weeks (Sept-June)

STEP 2: Cycle of Consistent Feedback

- Expectations set in writing prior to academic year
- Peer Leadership Review (at least 2x/year)
 - Academic Focus
 - Culture Team Focus
- Regional Support
 - Weekly visits by RD (no more than 1:6 ratio)

Mastery Charter Principal Evaluation

STEP 3: Show Me the Data

- Regional report with key indicators (every 3 weeks)
- Comparison progress with other principals based on Mission Metrics targets
- Teacher/Staff Surveys (every six weeks online)

STEP 4: Formal Evaluation

- Mid-Year Feedback and Review
- End of Year Feedback and Review
- Appeal process



Mastery Charter Principal Evaluation Questions?

Tabenah Washington, Grants Manager

Tabenah.washington@masterycharter.org

Rebecca Schatzkin, Director of Human Resources

Rebecca.schatzkin@masterycharter.org

IRA Challenges and Solutions

Discussion

- Introductions and brief description of your knowledge of inter-rater agreement pertaining to teacher and principal evaluations.
- In table groups, identify 2-3 issues or questions you have on achieving agreement or measuring it.
- Collaboratively devise ways to address these issues.
- Share and discuss challenges and solutions as a whole group.