Inter-Rater Reliability for Educator Effectiveness Observations

Stanley Rabinowitz, Ph.D.
CSAI/WestEd

August 29, 2013
Topeka, KS

THE CENTER ON STANDARDS & ASSESSMENT IMPLEMENTATION
WestEd CRESST

WestEd

Goals

1. Update on Kansas Educator Evaluation Protocol (KEEP) and other relevant accountability initiatives/issues

2. Understanding of key terms related to IRR

3. Factors affecting IRR related to instrumentation

4. Factors affecting IRR related to training

5. Determine follow-up needs and next steps

THE CENTER ON STANDARDS & ASSESSMENT IMPLEMENTATION
WestEd CRESST

WestEd

IRR Exercise

1. Think of a movie* you watched recently
2. Did you like the movie?
3. How would you rate the movie?
4. What characteristics did you use to rate the movie?
5. Did you give it an overall quality rating or use some quantitative scale?
6. Discuss different evaluative criteria that might make you change your rating
7. If you were to see it again do you expect you would give it a similar (identical) rating?
8. Was the movie "effective?"

THE CENTER ON STANDARDS & ASSESSMENT IMPLEMENTATION
WestEd CRESST

WestEd

Definition of Key Terms

- Reliability (Rxx) – Consistency of a rating/judgment/ decision

- Inter-rater Reliability (IRR) – Rate of agreement between/ among two or more raters or ratings (individuals x events)
  - measure of clarity among raters—are they all on the same page?
  - facilitates accurate data collection
  - ensures fairness across individuals and sites
  - legal defensibility
  - proper feedback to teachers
  - essential to support accountability
  - essential to support reform

Definition of Key Terms

- IRR not the same as Rxx (necessary but not sufficient)
- IRR not the same as Validity (necessary but not sufficient)
- IRR Benchmarks—what is sufficient?
  - Percent agreement
  - Research based (e.g., 80%)
  - Proportion reaching different ratings
  - Statistical measure (e.g., Cohen's kappa)

Factors Affecting IRR--Instrumentation

- Number of categories
- Scale used per category
  - Yes/No
  - Likert Numbers (1-5, 1-7, 1-100)—meaningful and realistic distinctions
  - Likert Terms (low medium high)
- Degree of innovation (familiarity)
- Clarity/consistency of terminology (e.g., demonstrates vs performs)
- Certain behaviors easier to get agreement
  - Likelihood of observed behavior
  - Subtlety of observed behavior– not seeing vs not present

Factors Affecting IRR—Instrumentation (cont.)

- Generic rubric plus category specific rubrics
- Support materials
  - Definitions
  - Examples
  - Exemplars
  - Videos
  - Artifacts
- Quality of Training

Factors Affecting IRR–Training

- Experience of Trainer
- Experience of Trainee (plus or minus)
- In-person vs web-based
- Use of exemplars and case studies (typical and unusual)
- Requirement for trainees to "qualify"
- One-time vs refresher (annual re-qualify)

Factors Affecting IRR–Training (cont.)

- Authenticity of training (context and setting)
- Training cycle:
  - Initial training
  - Calibration
  - Moderation
  - Adjustments
  - Implement
  - Adjustments
  - Retraining
- Quality of instrumentation and supports (instrumentation and training related)

Additional Factors Affecting IRR

- Stakes
- Publicity
- Timing—when and how often
- Moderation/External Audits
- Artifacts (videos)
- Who is doing the rating (principal vs external rater)
- Single vs. Multiple Observers

Next Steps

- Review of current instrumentation and training plans against success considerations
- Statewide rollout of observation protocol including training and IRR computation
- External validation and revision of rubrics and protocols—are the "correct" judgments being made?
- External validation and revision of entire system—incorporating lessons learned
- Follow-up TA from C3 and CSAI