Video Scoring

• Catherine McClellan
Measures of Effective Teaching

- Bill & Melinda Gates Foundation
- 2-year, $50 million project; ~3000 teachers in 7 school districts
  - 1. Student achievement gains on state standardized assessments and supplemental assessments
  - 2. Classroom observations and teacher reflections
  - 3. Teachers’ pedagogical content knowledge
  - 4. Student perceptions of the classroom instructional environment
  - 5. Teachers’ perceptions of working conditions and instructional support at their schools
ETS Role with Observations

• Subcontractor to Teachscape to score the classroom observation 360° videos
• 2 non-content specific: CLASS and Framework for Teaching
• 3 content-based: MQI; PLATO, and QST
• Challenges:
  – Unusual video format
  – Large-scale, distributed online scoring
  – Non-expert raters
  – “Assessment” use
  – Cost and time
Project Constraints

- Score approximately 23,000 videos
- Complete approximately 64,000 scorings on 5 observation protocols
- Employing more than 500 raters
- Complete interim data deliverables
- Meet budget and schedule
- Deliver high-quality data
Staffing Scoring

• Find lots of raters
  – Job description
  – Specialized skills needed

• Train the raters
  – Build training

• Verify scoring skills
  – Build assessments
Staffing Scoring

• Find and train the scoring leadership
  – No experienced raters
• Schedule the work
  – For hundreds or thousands of raters
  – Manually
  – Across systems
• Replace losses due to attrition
Consider the Instruments

- Number of scales and domains
- Number of score levels
- Distribution of scores
- Cognitive load for a non-expert
- Relevance and utility given video capture
- Purpose of MET vs. typical use
Design and Build Scoring Software

• Parallel design of scoring and design of software
• Need for very early unplanned viewer
• Changes in specifications in flow
• Interaction with partner web sites
• Expertise bottleneck (me!)
Create Scoring Support

• Rater skills assessments and quality monitoring
  – Certification
  – Calibration
  – Validity

• Training examples
  – Benchmarks
  – Rangefinders

• Practice videos

• Rationales
The Source

• Master Coding
  – 50 videos were selected to be representative of the full sample for each of the 5 instruments
  – These videos were master-coded by each AP and supplemented by the ETS content leads, providing
    • correct scores
    • time-stamped indications of evidence occurrence in each segment
    • comments and evidence
Very Frequent Evidence, Uniform Occurrence

Frequency of Comments vs Minute Number Within Video
Frequent Evidence, Clustered Occurrence

Frequency of Comments

Minute Number Within Video
Moderately Frequent Evidence, Very Clustered Occurrence

Frequency of Comments vs. Minute Number Within Video

-2 0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 82 84 86 88
Infrequent Evidence, Cyclical Occurrence

Minute Number Within Video

Frequency of Comments

-2 0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78

-2 0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78

Confidential and Proprietary. Copyright © 2011 Educational Testing Service. All rights reserved.
Infrequent Evidence, Little Structure

Frequency of Comments

Minute Number Within Video

- Infrequent Evidence, Little Structure
Special Challenges of Video

• Format of captures
• Rater bias
• Audio and video quality
• User errors
  – Technical
  – Other
• “Performing”
Scoring Teaching Observations

• Teaching is complex
• Different instruments focus on different aspects
• Instruments are complex
• Rater training is time-consuming
• Inter-rater agreement standards may prove challenging
THANK YOU!

Catherine McClellan
R&D Director of Human CR Scoring
CMcclellan@ets.org