Creating a System for Valid Classroom Observation

Catherine A. McClellan

R&D Director of Human CR Scoring

Some Contextual Words

- Tests/Assessments are not valid; neither are scores
- Reliability sets the maximum for validity
- Standardized assessment is not popular
- Certification and licensure are particularly contentious

THE TIMES
THE SUNDAY TIMES

From The Times
September 11, 2008

School test company ETS Europe 'lost £50m in fiasco'

Nicola Woolcock

The American company behind this summer's school tests fiasco blamed the Government yesterday and said it had been left £50 million out of pocket.

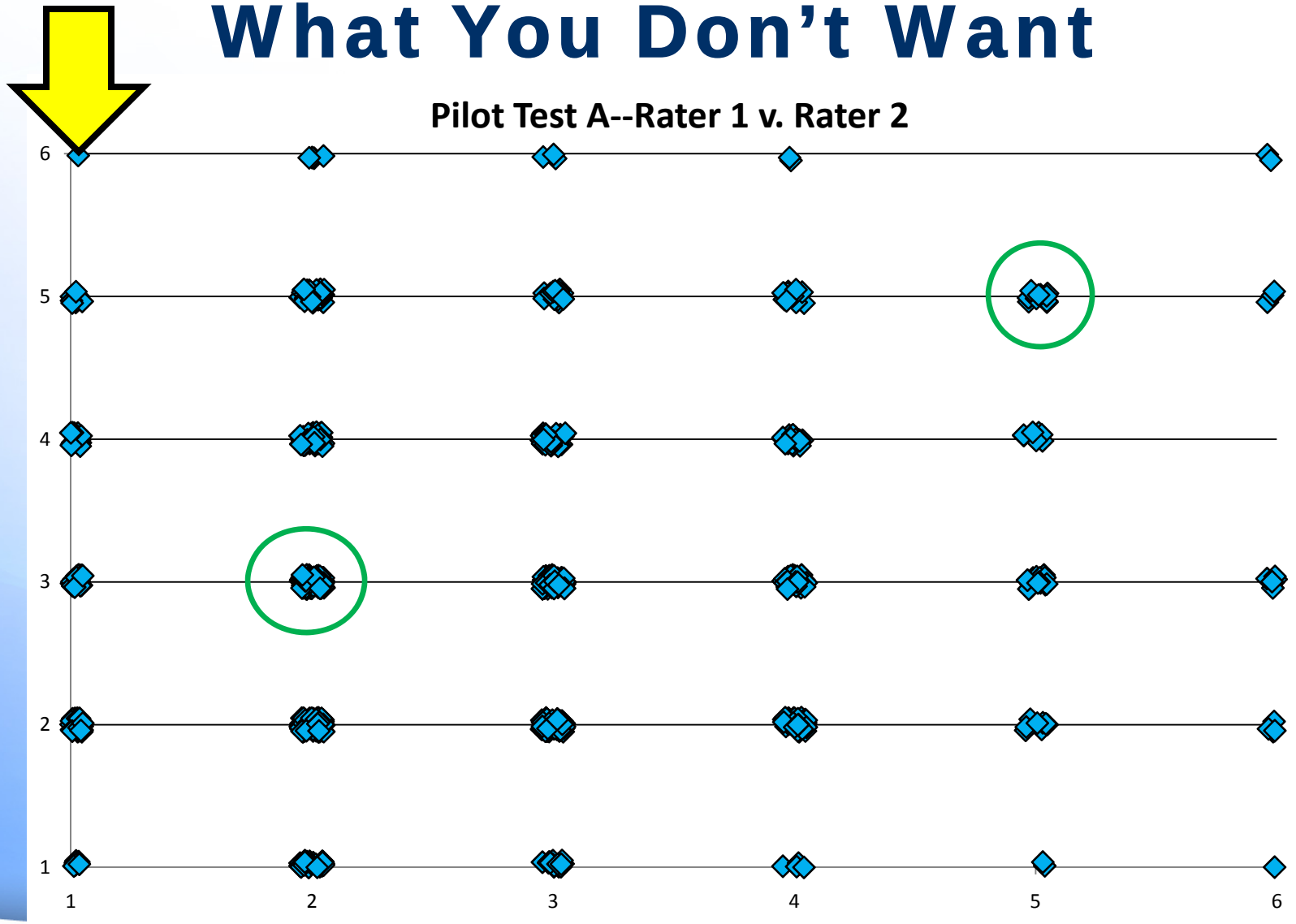
ETS Europe said that, rather than being sacked, it had instigated the termination of its £156 million five-year contract because it could not justify any more losses.

Two of its vice-presidents faced a cross-party committee of MPs that is trying to discover who is accountable for delays and mistakes in marking this year's Key Stage tests, taken by 11 and 14-year-olds in England.

"Standardized"?

- *It should not matter to the examinee which form of the test he/she takes*
  - Good test development
  - Equating
  - Scale scores
  - Stability and comparability over time
- *It must not matter to the examinee which rater scores his/her test responses*
  - Good item and rubric development
  - Equating (or not)
  - Scale scores (or not)
  - Stability and comparability over time (or not)

What You Don't Want



ETS' Typical Scoring Model

- Raters work in teams of 6-12 under the direction of a Scoring Leader
- Raters work fixed shifts; they submit their availability and are scheduled to work
- Distributed and online scoring is our main model
  - We do some paper & pencil, and some on site, but much less
- We score about 35-40 million pieces every year with ~50,000 current raters
- Scoring goes on 18 hours per day, 7 days per week, 364 days per year

What I Worry About

- Rater bias
- Rater influence
- Rater selection
- Rater training
- Rater skills assessment
  - Initial
  - Ongoing
- Monitoring scoring
  - Agreement within session
  - Agreement over time
  - Accuracy

Rater Bias

- No one believes that s/he is biased—and we all are
- You *cannot* remove all biases from scoring
- Make raters conscious of their preferences so that control can be exercised during scoring
- ETS has the luxury of disqualifying raters who know the candidate from scoring that response—you probably don't

Rater Influence

- This concern comes into play—or doesn't—depending on the inferences to be made

- We control this in three dimensions:
  - "Touches" per candidate
  - Proportion of pool scored by a single rater
  - Over-pairing of specific raters in double-scoring

Rater Selection

- What makes a good rater?
  - Meticulous and detailed
  - Obedient
- How do we choose them? It depends
  - Clients may set requirements
  - Generally some level of content expertise is required
  - Proxy measures like employment, education, and licensure status
- Good SLs know the rubric and have counseling and interpersonal skills

Rater Training

- With near unanimity, everyone prefers face-to-face training
- We have found that online training works just as well or better
- The more complex the scoring task, the more challenging getting training right becomes
- If you get training wrong, everything else will fall apart in ways that cannot be fixed

Rater Skills Assessment

- Initial assessment is often conducted at the end of training (“certification”), using a standard of agreement with “correct” scores provided by experts
- Most ETS programs require raters to pass a short scoring accuracy test (“calibration”) before every shift
- SLs review rater scoring (“backscore”) throughout every shift; SLs speak with every rater on the team every shift

Monitoring Scoring

- Consistency within session: double-scoring
  - Interrater agreement is popular, but a weak measure
  - Raters can agree and both be wrong
  - If raters disagree, there is no guarantee that either is correct and no simple way to determine which one is
  - Agreement measures corrected for chance offer a better, but still imperfect, picture
- Consistency over time: trend scoring
  - Similar statistics to double-scoring, and with the same problems
  - Structure of the trend set has strong influence on power to detect problems

Monitoring Scoring (2)

- Accuracy: certification & calibration
  - Raters who are aware they are being tested perform quite differently than “normal” scoring
- Accuracy: validity scoring
  - Pre-scored responses are seeded into live scoring
  - Invisible to the raters; visible to the SLs and staff
- Accuracy: backscoring
  - Dependent on the expertise of the SL
  - Best if SL skills are also subject to regular evaluation

Scoring Design

- ***Please*** read Heather's paper on G-studies
- If you have limitations on time, money, expertise, or resources in conducting studies yourself, think long and hard about how your evaluations are designed:
  - How often?
  - By whom?
  - What level of reliability?
  - Measured how?
- Scrutiny will be *intense* and may be by a court

Some Cautions

- Take care in asking a single instrument to perform too many functions—you often end up with something that does everything badly and nothing well
- The ideal design or instrument to support analysis may not be the same as one to support skills diagnosis and professional development
- When adding tasks to staff with a full load, something else has to come out—the job of leadership is to decide what

Some Cautions (2)

- Consider NCLB and how student evaluation has changed
  - Most states hired companies to create “custom” tests for their students and curricula
  - 10 years later, we have come to realize that we want a single core of student assessments that are comparable across all states
  - We spent a lot of time and money before reaching that conclusion
- The content of student assessments is, by and large, well-defined and agreed on
- Teacher observation decidedly is *not*