Measures of Effective Teaching: 5
- Observation Score Calibration and Validation, 2011

Bill and Melinda Gates Foundation

Scoring Design Memo
Terms of Use

The terms of use for this study can be found at:
http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/37090/terms

Information about Copyrighted Content

Some instruments administered as part of this study may contain in whole or substantially in part contents from copyrighted instruments. Reproductions of the instruments are provided as documentation for the analysis of the data associated with this collection. Restrictions on "fair use" apply to all copyrighted content. More information about the reproduction of copyrighted works by educators and librarians is available from the United States Copyright Office.

NOTICE
WARNING CONCERNING COPYRIGHT RESTRICTIONS

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use," that user may be liable for copyright infringement.
Memo

To: Thomas Kane; Steve Cantrell; Kerri Kerr
From: Catherine McClellan; Mark Atkinson; Thomas Van Essen
CC: Ivan Montoya; June Zumoff
Date: 8/3/2011
Re: Phase 1 & Phase 2 MET scoring design

This memo is provided to document joint proposals from Teachscape and ETS for a design for scoring all instruments in the main MET video scoring activity.

Phase 1 Scoring Designs

The following scoring designs are proposed for Phase 1. For each instrument listed below, the videos to be scored are the 2000 videos chosen as the “Plan B” set already scored on CLASS in an extra-contract activity. The videos are half math and half ELA. All have complete data for the study, comprising prior-year VAM from the district; current-year VAM from the study; student perception survey data, and all teachers in the data set have a complete set of 4 (single subject area) or 8 (self-contained) videos.

Note that QST has no videos in Phase 1, and the Phase 1 scoring for CLASS is complete (scored during “Plan B”). Both will have scoring designs in the next section for Phase 2.

Assumptions relevant to all instruments:

- Full online training and certification is assumed for all raters in the main MET scoring.
- Raters will complete a daily calibration assessment.
- Scoring begins at 00:00.
- Scoring for each segment is done immediately following viewing of that segment (rater presses pause, codes the segment, then watches and codes the next segment).
- Segments are coded as separate sets of scores (i.e. what occurs in one segment does not influence the codes of another segment).
- All segments of a lesson that are scored are done so by the same rater, except for CLASS (see Phase 2 CLASS design for details).
- Double-scoring will be implemented as scores being assigned to the same video on the same instrument and group of scales by different raters with no knowledge of each other’s score.
• Validity scoring will be implemented as seeding through live scoring at a fixed rate videos that have a “correct” score as determined by the AP and ETS content leads. These videos will be blind to the rater, in that they will have no features that distinguish them from the live scoring.

• If an instrument is scored in a Group of Scales, the set of scales that comprise the full instrument has been subset into smaller groups. These groupings were formed in collaboration with the AP and content leads, in an effort to limit the cognitive load that individual raters must manage as well as to separate scales that may be challenging to score simultaneously.

1. MQI
   o Score in 2 Groups of Scales: GoS1 consists of:
     • Errors & Imprecision,
     • Classroom Work Connected to Mathematics, and
     • Explicitness & Thoroughness. Note that of CWCM and E&T, only one can apply depending on the content and grade level of the class—the other will be N/A.
   o GoS2 consists of:
     • Student Participation in Meaning-Making and Reasoning,
     • Richness, and
     • Working with Students & Mathematics.
   o Score the 1000 Plan B math videos, plus 25% double-scoring.
   o Score the first four 7.5-minute segments of each video.
   o Record a holistic score on each scale after the four segment scores are recorded.
   o Record an Overall MQI score and a Lesson-Based Guess at MKT after all four segments’ scoring is complete.
   o Record audio and video quality scores after all other scores are complete.
   o This results in a rater recording 20 scores for a 30-minute portion of video viewing time on each of GoS1 and GoS2:
     • 3 scale scores for each of 4 segments,
     • 1 holistic score on each of 3 scales,
     • 1 Overall MQI score,
     • 1 Lesson-Based Guess at MKT, and
     • 3 audio/visual quality scores
2. PLATO

- Score in 2 Groups of Scales: GoS1 consists of:
  - Intellectual Challenge,
  - Classroom Discourse, and
  - Behavior Management; plus
  - Representations of Content and various Content Domain and Activity Structure indicators.

- GoS2 consists of:
  - Modeling,
  - Strategy Use & Instruction, and
  - Time Management; plus
  - Representations of Content and various EL Instructional Technique indicators.

- Score the 1000 Plan B ELA videos, plus 25% double-scoring.
- Score the first two 15-minute segments of each video.
- Record audio and video quality scores after all other scores are complete.

- This results in a rater recording 37 scores on GoS 1 for a 30-minute portion of video viewing time:
  - 4 scale scores on each of 2 segments
  - Representation of Content on each of 2 segments
  - 7 Content Domain scores on each of 2 segments
  - 5 Activity Structure scores on each of 2 segments
  - 3 audio/visual quality scores

- This results in a rater recording 19 scores on GoS 2 for a 30-minute portion of video viewing time:
  - 4 scale scores on each of 2 segments
  - Representation of Content on each of 2 segments
  - 3 EL Instructional Technique scores on each of 2 segments
  - 3 audio/visual quality scores
3. Framework for Teaching

- Score in 3 Groups of Scales. GoS1 consists of:
  - Creating an Environment of Respect & Rapport and
  - Using Questioning & Discussion Techniques.

- GoS2 consists of:
  - Establishing a Culture for Learning,
  - Managing Classroom Procedures, and
  - Communicating with Students.

- GoS3 consists of:
  - Managing Student Behavior,
  - Engaging Students in Learning, and
  - Using Assessment in Instruction.

- Score the 2000 Plan B videos, plus 12.5% double-scoring
  - Score GoS1 in minutes 0-12 and 25-35. Score GoS2 in minutes 0-15 and 30-35.
    Score GoS3 in minutes 5-15 and 25-35.
  - Record audio and video quality scores after all other scores are complete.
  - This results in a rater recording 5 scores on GoS 1 for a 22-minute portion of video viewing time:
    - 1 score on each of 2 scales
    - 3 audio/visual quality scores
  - This results in a rater recording 6 scores on GoS2 for a 20-minute portion of video viewing time:
    - 1 score on each of 3 scales
    - 3 audio/visual quality scores
  - This results in a rater recording 6 scores on GoS3 for a 20-minute portion of video viewing time:
    - 1 score on each of 3 scales
    - 3 audio/visual quality scores

These are the designs agreed upon in the design conversations with the AP and the Gates Foundation for each instrument.
“The Pause”

The possibility has been discussed of a period between Phases 1 & 2, during which the primary analysts—Tom Kane, Doug Staiger, and Dan McCaffrey—would complete some preliminary analyses of the data. The period of analysis is generally referred to as “The Pause” and will be so called herein. The intent was that the results of these analyses would be considered before scoring designs for Phase 2 were finalized. However, in order to meet the data delivery deadlines specified by the Gates Foundation for Year 1 and Year 2 data set, it is not possible to stop work during The Pause. Scoring must continue straight through if the Year 1 deadlines, especially, are to be met. This set of constraints, along with the realities of the budgets available for the work, have led to the necessity of selecting a video sample and preliminary scoring designs prior to The Pause. These are the designs described in the next section.

Phase 2 scoring designs

The following scoring designs are proposed for Phase 2. For each instrument listed below, the videos to be scored are the “randomization” videos. These videos are the ones selected for random assignment for students and teachers in Year 2 of the MET study. Effectively, they are the “intent to treat” group. The Venn diagram in Figure 1 shows the various categories of these videos, in terms of membership in Year 1, Year 2, the randomization sample, and the Plan B sample. Boxes in blue are totals for the sample noted, and boxes in white a sub-samples of one of the overall samples.
Assumptions relevant to all instruments:

- Full online training and certification is assumed for all raters in the main MET scoring.
- Raters will complete a daily calibration assessment.
- Scoring begins at 00:00.
- Scoring for each segment is done immediately following viewing of that segment (rater presses pause, codes the segment, then watches and codes the next segment).
- Segments are coded as separate sets of scores (i.e. what occurs in one segment does not influence the codes of another segment).
- All segments of a lesson that are scored are done so by the same rater.
The sample sizes for each instrument are presented in the table below.

<table>
<thead>
<tr>
<th></th>
<th>CLASS</th>
<th>MQI</th>
<th>PLATO</th>
<th>FFT</th>
<th>QST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summer pilot</td>
<td>750</td>
<td></td>
<td>Phase 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP double-score</td>
<td>250</td>
<td>Plan B</td>
<td>1000</td>
<td>Plan B</td>
<td>2000</td>
</tr>
<tr>
<td>Plan B</td>
<td>2000</td>
<td>PB double-score</td>
<td>250</td>
<td>PB double-score</td>
<td>0</td>
</tr>
<tr>
<td>PB double-score</td>
<td>500</td>
<td>Phase 1 grand total</td>
<td>1250</td>
<td>Phase 1 grand total</td>
<td>2250</td>
</tr>
<tr>
<td>Phase 2</td>
<td></td>
<td>Phase 2</td>
<td></td>
<td>Phase 2</td>
<td></td>
</tr>
<tr>
<td>Y1 rand non-PB</td>
<td>5852</td>
<td>Y1 rand non-PB</td>
<td>2679</td>
<td>Y1 rand non-PB</td>
<td>3020</td>
</tr>
<tr>
<td>Y1 rand PB</td>
<td></td>
<td>Y1 rand PB</td>
<td>672</td>
<td>Y1 rand PB</td>
<td>873</td>
</tr>
<tr>
<td>Y1 non-rand PB</td>
<td></td>
<td>Y1 non-rand PB</td>
<td>1340</td>
<td>Y1 non-rand PB</td>
<td>1304</td>
</tr>
<tr>
<td>Y1 subtotal</td>
<td>5852</td>
<td>Y1 subtotal</td>
<td>3351</td>
<td>Y1 subtotal</td>
<td>3652</td>
</tr>
<tr>
<td>Y1 d-s &amp; val</td>
<td>585</td>
<td>Y1 d-s &amp; val</td>
<td>335</td>
<td>Y1 d-s &amp; val</td>
<td>365</td>
</tr>
<tr>
<td>Y1 grand total</td>
<td>6437</td>
<td>Y1 grand total</td>
<td>3686</td>
<td>Y1 grand total</td>
<td>4017</td>
</tr>
<tr>
<td>Y2 rand</td>
<td>6257</td>
<td>Y2 rand</td>
<td>2991</td>
<td>Y2 rand</td>
<td>3209</td>
</tr>
<tr>
<td>Y2 d-s &amp; val</td>
<td>626</td>
<td>Y2 d-s &amp; val</td>
<td>299</td>
<td>Y2 d-s &amp; val</td>
<td>321</td>
</tr>
<tr>
<td>Y2 total</td>
<td>6883</td>
<td>Y2 total</td>
<td>3290</td>
<td>Y2 total</td>
<td>3530</td>
</tr>
<tr>
<td>All Phase 2</td>
<td>13320</td>
<td>All Phase 2</td>
<td>6976</td>
<td>All Phase 2</td>
<td>14754</td>
</tr>
</tbody>
</table>

1. CLASS Phase 2 design will be the same as the Phase 1/Plan B design.
   a. Score the full randomization sample, minus the videos in the randomization sample that are also in Plan B. Score with 5% double-scoring and 5% validity scoring. No Plan B/Phase 1 videos will be re-scored on CLASS in Phase 2.
   b. CLASS-Upper Elementary (CLASS-UE) consists of videos from grades 4-6 and CLASS-Secondary (CLASS-SEC) consists of videos from grades 7-9.
   c. Score in 1 GoS with all 12 CLASS scales.
   d. Score the first two 15-minute segments of each video
   e. Two different raters will score the first and second segments of a single CLASS video.
   f. Record audio and video quality scores after all other scores are complete.
   g. This results in each rater recording 15 scores for a 15-minute portion of video viewing time; for each video with two 15-minute segments, 30 total scores will be recorded by two unique raters.
2. MQI Phase 2 design
   
a. Score in 1 Groups of Scales consisting of all MQI scales.

b. Score the full randomization sample of math videos and re-score the portion of the Plan B/Phase 1 videos that are not already in the Phase 2 set. Score with 5% double-scoring and 5% validity scoring.

c. Score the first four 7.5-minute segments of each video.

d. Record a holistic score on each scale after the segment scores are recorded.

e. Record an Overall MQI score and a Lesson-Based Guess at MKT after all segments’ scoring is complete.

f. Record audio and video quality scores after all other scores are complete.

g. This results in a rater recording 33 scores for a 30-minute portion of video viewing time.

3. PLATO
   
a. Score in 1 Group of Scales consisting of: Intellectual Challenge, Classroom Discourse, Behavior Management, Modeling, Strategy Use & Instruction, Time Management and Representations of Content. No scores will be captured for Content Domain, EL Instructional Technique, or Activity Structure indicators. [NOTE: if it is decided that some or all of the PLATO ALD scales will be used in Phase 2, the estimates for cost and schedule will require revision.]

b. Score the full randomization sample of ELA videos and re-score the portion of the Plan B/Phase 1 videos that are not already in the Phase 2 set. Score with 5% double-scoring and 5% validity scoring.

c. Score the first two 15-minute segments of each video, recording scores after each 15-minute segment.

d. Record audio and video quality scores after all other scores are complete.

e. This results in a rater recording 17 scores for a 30-minute portion of video viewing time.

4. Framework for Teaching
   
a. Score in 1 Group of Scales consisting of all FfT scales.

b. Score the full randomization sample, and re-score the portion of the Plan B/Phase 1 videos that are not already in the Phase 2 set. Score with 5% double-scoring and 5% validity scoring.

c. Score in minutes 0-15 and 25-35; this is the union of the segments from Phase 1.

d. Record audio and video quality scores after all other scores are complete.
e. This results in a rater recording 11 scores for a 25-minute portion of video viewing time.

5. QST

a. Score in 3 Groups of Scales. GoS1 consists of:
   
   i. Sets the Context and Focuses Learning on Key Science Concepts,
   
   ii. Uses Representations,
   
   iii. Demonstrates Content Knowledge, and
   
   iv. Provides Feedback for Learning.

b. GoS2 consists of:
   
   i. Promotes Students’ Interest and Motivation to Learn Science,
   
   ii. Assigns Tasks to Promote Learning and Addresses the Task Demands,
   
   iii. Uses Modes of Teaching Science Concepts, and
   
   iv. Elicits Evidence of Students’ Knowledge and Conceptual Understanding.

c. GoS3 consists of:
   
   i. Initiates the Investigation,
   
   ii. Provides Guidelines for Conducting the Investigation and Gathering Data,
   
   iii. Guides Analysis and Interpretation of Data, and
   
   iv. Elicits Evidence of Students’ Knowledge and Conceptual Understanding.

d. Score all biology videos. Score with 10% double-scoring and 10% validity scoring.

e. Score GoS1 and GoS2 on the first two 15-minute segments of each video, recording scores after each 15-minute segment. Score GoS3 on the entire laboratory videos, recording scores after each 15-minute segment.

f. Record audio and video quality scores after all other scores are complete.

g. This results in the GoS1 and GoS2 raters recording 11 scores for a 30-minute portion of video viewing time. For the GoS3 raters, the labs will be viewed in their entirety, and may vary somewhat in length. Assuming a 1-hour lab, GoS3 raters will record 19 scores for a 60-minute portion of video viewing time.