I gratefully acknowledge the generous support from the Bill and Melinda Gates Foundation and that of Andy Porter and John Smithson in completing this work.
Motivation & Contribution

- State assessment results are increasingly being used for high-stakes decisions about teachers and schools.
- It is imperative to verify that improvements in assessment scores are within the power of teachers.
- Sensitivity is also relevant for the construction of weights in multiple-measures evaluation systems.
- This study is among the first
  - To compare sensitivity across subjects (math and ELA)
  - To include multiple widely-used measures of pedagogical quality in a sensitivity analysis
  - With such a large sample (~2000 teachers) and data of the type to be used in new state teacher evaluation systems
Research question

- To what extent are state assessments of student achievement sensitive to differences in observational and student survey ratings of instructional quality?
- How do optimal weights for multiple-measure composites differ across state tests with different levels of sensitivity?
Data

Bill and Melinda Gates Foundation’s Measures of Effective Teaching (MET) study

- Purpose: Design effective evaluation systems for teacher improvement
- N = 3,000+ teachers in grades 4-9 in six districts: Charlotte, Dallas, Denver, Hillsborough (FL), Memphis, New York
- Dallas and Denver excluded here due to smaller sample sizes
- Only grades 4-8 math and ELA used here

Measures

- Outcome variable: value-added measure on the state assessment based on averaging the residuals in a random-effects model
Measures, ctd.

Multiple measures of instruction:

- TRIPOD student survey (Ferguson)
- CLASS (Pianta)
- Framework for Teaching (FFT, Danielson)

Mathematics only

- Mathematical Quality of Instruction (MQI, Hill)

ELA only

- Protocol for Language Arts Teaching Observations (PLATO, Grossman)
Methods

- Simple techniques of the kind likely to be feasible in states and districts
  - Bivariate correlations of VAM scores with each of the instructional quality measures by state
  - Correlations removing outlier teachers on the DV
  - Searching for evidence of nonlinear relationships (quadratic, logarithmic transformations)
  - Subscale analyses
  - Grade level analyses
- Deriving optimal composites using regression
  - VAM
  - Equally-weighted average
## Results

Table 2
Correlations of State Test VAM Scores with Pedagogical Quality Measures by District and Subject

<table>
<thead>
<tr>
<th></th>
<th>ELA</th>
<th>Mathematics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>District 2</td>
<td>District 4</td>
</tr>
<tr>
<td><strong>TRIPOD</strong></td>
<td>0.281*</td>
<td>0.125*</td>
</tr>
<tr>
<td>265</td>
<td>265</td>
<td>256</td>
</tr>
<tr>
<td><strong>CLASS</strong></td>
<td>0.102</td>
<td>0.166*</td>
</tr>
<tr>
<td>191</td>
<td>190</td>
<td>177</td>
</tr>
<tr>
<td><strong>FFT</strong></td>
<td>0.023</td>
<td>0.123</td>
</tr>
<tr>
<td>191</td>
<td>190</td>
<td>177</td>
</tr>
<tr>
<td><strong>PLATO</strong></td>
<td>0.079</td>
<td>0.133</td>
</tr>
<tr>
<td>186</td>
<td>190</td>
<td>176</td>
</tr>
</tbody>
</table>

|                        | District 2 | District 4 | District 5 | District 6 | Overall |
| **TRIPOD**             | 0.206*     | 0.086      | 0.183*     | 0.192*     | 0.175*  |
| 237                    | 262        | 241        | 254        | 994        |
| **CLASS**              | 0.035      | 0.129      | 0.178*     | 0.212*     | 0.131*  |
| 174                    | 178        | 180        | 183        | 716        |
| **FFT**                | 0.032      | 0.102      | 0.24*      | 0.165*     | 0.136*  |
| 174                    | 180        | 180        | 183        | 717        |
| **MQI**                | 0.016      | 0.101      | 0.010      | 0.131      | 0.050   |
| 169                    | 180        | 179        | 183        | 711        |

*Note.* *p < .05. Values in each cell are pairwise Pearson correlations and sample sizes.
Results, nonlinear analyses

- Quadratic and logarithmic transformations were used for any state test that did not show sensitivity to this point.
  - No significant quadratic or logarithmic relationships were found. No evidence of nonlinear relationships.
## Results, subscale & grade analyses

<table>
<thead>
<tr>
<th>Table 5</th>
<th>Sensitivity of State Tests to Subscales of Instructional Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ELA</td>
</tr>
<tr>
<td></td>
<td>District 2</td>
</tr>
<tr>
<td>TRIPOD</td>
<td>0/12</td>
</tr>
<tr>
<td>CLASS</td>
<td>0/6</td>
</tr>
<tr>
<td>FFT</td>
<td>0/6</td>
</tr>
<tr>
<td>PLATO</td>
<td>0/6</td>
</tr>
<tr>
<td>MQI</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Fraction in each cell represents the proportion of subscales where correlation with VAM is significantly different from 0 at $p < .05$. Only tests that have shown no prior evidence of sensitivity are included.
How do differences in sensitivity relate to optimal weights in composite evaluation systems?

- If the goal is predicting VAM:
  - More sensitive assessments are given lower weight in the optimal composite

- If the goal is predicting an equally-weighted composite of VAM, TRIPOD, observational quality measure:
  - No apparent pattern in the relationship of sensitivity to the weights given.
  - VAM generally receives the smallest weight

- Weights on VAM generally higher in mathematics than ELA
Discussion

- One interpretation: most of the state assessments show modest sensitivity to one or more pedagogical quality measures.
  - Exception: District 4 ELA which shows sensitivity to only one subscale of one instrument and at one grade level.
- A second interpretation: each of the assessments is insensitive to at least one, and often multiple measures of pedagogical quality
  - Exception: District 6 mathematics
  - Generally looks like more sensitivity in mathematics than in ELA, though subject differences are not statistically significant.
Where do we go from here?

- These tests not constructed to be specific to these particular quality measures.
- Policymakers need to decide what “good” instruction looks like.
  - Foreground sensitivity in the assessment and accountability argument.
  - If tests are constructed to be sensitive to good instruction but are not, that’s a problem.
  - Of course, there’s also content …
- Researchers need to work to determine what makes some tests more sensitive to instructional content and/or quality than others