Establishing Validity and Reliability for Locally Developed Instruments
The work of TEAMS is supported with funding provided by the National Science Foundation, Award Number DRL 1238120. Any opinions, suggestions, and conclusions or recommendations expressed in this presentation are those of the presenter and do not necessarily reflect the views of the National Science Foundation; NSF has not approved or endorsed its content.
Strengthening the quality of the MSP project evaluation and building the capacity of the evaluators by strengthening their skills related to evaluation design, methodology, analysis, and reporting.
Online Help-Desk for submitting requests

Website at: teams.mspnet.org

Webinar series targeted to specific evaluation topics

Tiered technical assistance for differentiated services

Instrument review and sharing
Establishing Validity and Reliability for Locally Developed Instruments

Webinar Presenter

Dr. Xin Wang
Objectives

• To review the definition of reliability and validity
• To review methods of determining the reliability and validity for locally developed instruments
• To discuss how to demonstrate that an instrument is clearly defined, has a direct interpretation, and measures the intended constructs
• To introduce a process for sharing reliability and validity information about locally developed instruments through TEAMS
Why Reliability and Validity Matter

Two of the primary criteria of evaluation in any measurement are:

• Whether we are measuring what we intend to measure—Validity

• Whether the same measurement process yields the same results each time—Reliability
Why Reliability and Validity Matter

• Reliability and validity in experimentation vs. in educational and psychological measurement
• The What Works Clearinghouse review standards:
  – Face validity
  – Reliability
  – Lack of over-alignment
Some Definitions

For the purpose of this webinar

• **Instrument**—a data collection tool that produces quantitative data
  – Tests and assessments
  – Surveys
  – Observation protocol (that yield quantitative data)

• Content presented in this webinar are applicable for data collection tools of these types
Reliability

Definition

The degree to which an instrument produces stable and consistent results
How to Measure Reliability

• Test-retest reliability
• Alternate-form reliability
• Internal consistency reliability
Test-Retest Reliability

• Use the instrument to collect data from the same respondents at two different points in time to see how stable the responses are

• Usually quantified with a correlation coefficient

• In general, $r$ values are considered good if $r \geq 0.70$
Test-Retest Reliability (cont’d)

• Test-retest specific items or the entire instrument

• Be careful about test-retest with items or scales that measure variables likely to change over a short period of time

• Test-retest over very short periods of time
Test-Retest Reliability (cont’d)

• Potential problem with test-retest is the practice effect

• Effects on your reliability estimates:
  – It inflates the reliability estimate
### Test-Retest Reliability (cont’d)

**Example:**

<table>
<thead>
<tr>
<th>ID#</th>
<th>Test</th>
<th>Retest</th>
</tr>
</thead>
<tbody>
<tr>
<td>A351</td>
<td>19</td>
<td>26</td>
</tr>
<tr>
<td>A352</td>
<td>20</td>
<td>23</td>
</tr>
<tr>
<td>A353</td>
<td>18</td>
<td>22</td>
</tr>
<tr>
<td>A354</td>
<td>25</td>
<td>26</td>
</tr>
<tr>
<td>A355</td>
<td>23</td>
<td>25</td>
</tr>
<tr>
<td>A356</td>
<td>18</td>
<td>23</td>
</tr>
<tr>
<td>A357</td>
<td>18</td>
<td>23</td>
</tr>
<tr>
<td>A358</td>
<td>20</td>
<td>17</td>
</tr>
<tr>
<td>A359</td>
<td>25</td>
<td>27</td>
</tr>
<tr>
<td>A3510</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>A3511</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>A3512</td>
<td>20</td>
<td>23</td>
</tr>
<tr>
<td>A3513</td>
<td>25</td>
<td>26</td>
</tr>
<tr>
<td>A3514</td>
<td>25</td>
<td>26</td>
</tr>
<tr>
<td>A3515</td>
<td>17</td>
<td>16</td>
</tr>
<tr>
<td>A3516</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>A3517</td>
<td>22</td>
<td>24</td>
</tr>
<tr>
<td>A3518</td>
<td>16</td>
<td>23</td>
</tr>
</tbody>
</table>
Test-Retest Reliability (cont’d)

To conduct a reliability analysis in SPSS

• Analyze – Correlate – Bivariate – click “Pearson” and “Flag significant correlations”

• Move “Test” and “Retest” to “Variables” then click “OK”

• Pearson Correlation Result: $r = 0.728$
Alternate-Form Reliability

• Items are reworded or their order is changed to produce two instruments that are similar but not identical

• Determines the similarity of two different versions of the same instrument
Alternate-Form Reliability (cont’d)

• You can measure alternate-form reliability at the same time point or separate time points

• If you have a large enough sample, you can randomly assign participants to two groups and administer a different version of the instrument to each group
### Alternate-Form Reliability (cont’d)

- **Example:**

<table>
<thead>
<tr>
<th>Bonnie’s Argument (Form A)</th>
<th>Bonnie’s Argument (Form B)</th>
<th>Write your answer below each question</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Claim:</strong> When you add any two even numbers, your answer is always even.</td>
<td><strong>Claim:</strong> When you add any two odd numbers, your answer is always even.</td>
<td>1. Is Bonnie’s argument viable? Explain.</td>
</tr>
<tr>
<td><strong>Bonnie’s argument:</strong></td>
<td><strong>Bonnie’s argument:</strong></td>
<td>1. Does Bonnie’s argument work for all even numbers or just for some even numbers? Explain.</td>
</tr>
<tr>
<td>2 + 2 = 4 4 + 2 = 6</td>
<td>1 + 3 = 4 3 + 1 = 4</td>
<td>1. What are the strengths of Bonnie’s argument?</td>
</tr>
<tr>
<td>2 + 4 = 6 4 + 4 = 8</td>
<td>3 + 5 = 8 5 + 3 = 8</td>
<td>1. How could Bonnie’s argument be improved?</td>
</tr>
<tr>
<td>2 + 6 = 8 4 + 6 = 10</td>
<td>5 + 7 = 12 7 + 5 = 12</td>
<td></td>
</tr>
<tr>
<td>So the sum is always even.</td>
<td>So the sum is always even.</td>
<td></td>
</tr>
</tbody>
</table>

5/21/2014
Alternate-Form Reliability (cont’d)

To calculate

• Administer the two versions of the instrument to the same participants within a short period of time
• Correlate the results of the two versions using Pearson’s Correlation
Internal Consistency Reliability

• Internal consistency reliability is a measure of how inter-correlated the items or group of items of an instrument are:
  – Cronbach’s coefficient alpha
  – Kuder-Richardson formula 20 (KR-20)
  – Split-half reliability (Spearman-Brown coefficient)
Internal Consistency Reliability (cont’d)

• Example: Measures of teacher preparedness to teach mathematics
• Questions: How would you rate your level of preparedness on a scale from 1 (unsatisfactory) to 8 (exceptional) related to each of the following statements?
  1. Provide mathematics instruction that meets appropriate standards
  2. Teach problem solving strategies
  3. Teach mathematics with the use of manipulative materials, such as algebra tiles, geometric shapes, and so on
  4. Sequencing mathematics instruction to meet instructional goals
  5. Select and/or adapt instructional materials to implement your written curriculum
  6. Make connections within mathematics and between mathematics and other subject areas
  7. Providing a challenging curriculum for all students you teach
  8. Using a variety of assessment strategies
  9. Using results from student assessment to inform practice
## Internal Consistency Reliability (cont’d)

<table>
<thead>
<tr>
<th>Respondent ID</th>
<th>q1</th>
<th>q2</th>
<th>q3</th>
<th>q4</th>
<th>q5</th>
<th>q6</th>
<th>q7</th>
<th>q8</th>
<th>q9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1904038991</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>1902279255</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>1901849884</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>190111468</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>1900937813</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>1910388096</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>1901772970</td>
<td>8</td>
<td>8</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>1914444563</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>1904136696</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>1902218135</td>
<td>8</td>
<td>8</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>1901953875</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>1901663866</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>1901693338</td>
<td>5</td>
<td>6</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>1904459022</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>7</td>
<td>6</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>1901161460</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>1902192659</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>1902209568</td>
<td>7</td>
<td>6</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>1901095246</td>
<td>6</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>1901974062</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>1903658990</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>1902193864</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>1958668809</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>1958913494</td>
<td>6</td>
<td>7</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>1959910610</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>1959902567</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>1958897406</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>1958892646</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>1958889030</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>1958885363</td>
<td>8</td>
<td>7</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>1958885068</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>1958883294</td>
<td>8</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>1958881518</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>7</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>
Internal Consistency Reliability (cont’d)

To calculate Cronbach’s Alpha (for Likert-scale items) or KR-20 (for dichotomous items) in SPSS

• Analyse – Scale – Reliability Analysis – Move the variables into the “Items” box
• Select “Alpha” in “Model” box
• Click “Statistics”-under the “Descriptive for”-click “Scale” & “Scale if item deleted”; and then ”-click “Correlations” under the “Inter-Item
• Click “Continue” and then “OK”

Report:
• Cronbach’s alpha = 0.929
Internal consistency reliability (cont’d)

To calculate Spearman-Brown coefficient

• Analyse – Scale – Reliability Analysis – Move the variables into the “Items” box
• Select “Split-half” in “Model” box
• Click “Statistics”-under the “Descriptive for”-click “Scale” & “Scale if item deleted”; and then click “Correlations” under the “Inter-Item”
• Click “Continue” and then “OK”

Report:

• Spearman-Brown coefficient (unequal length) = 0.939
Questions!

Please use your chat box to submit questions for Xin regarding Reliability.
Validity

Definition

• The degree to which an instrument measures what it sets out to measure

• The degree to which evidence and theory support the interpretation of [results] entailed by proposed uses

(Standards for Educational and Psychological Testing, 1999)
How to Measure Validity

• Face validity
• Content validity
• Criterion validity
• Construct validity
Face Validity

• Refers to the degree to which an instrument appears to measure what it purports to measure
  – To assess: Ask instrument users and intended audience to evaluate whether the instrument appears to measure the construct of interest
  – Many do not consider this as a measure of validity
Content Validity

• Sampling the entire domain of the construct it was designed to measure

• Subjective measure of how appropriate the items seem to content experts
  – Usually consists of an organized review of the instrument’s contents
  – Still very qualitative
Content Validity (cont’d)

• To assess:
  – Gather a panel of judges
  – Give the judges a table of specifications of the amount of content covered in the domain
  – Give the judges the instrument
  – Judges draw a conclusion as to whether the proportion of content covered matches the proportion of content in the domain
Criterion Validity

• Correlation between the instrument and a criterion.
• Measure of how well one instrument stacks up against another instrument or predictor
  – **Criterion:** Other accepted measures of the construct or measures of other constructs similar in nature.
• A criterion can consist of any standard with which your instrument should be related
Criterion Validity (cont’d)

• Three types:
  – **Convergent Validity**: High correlations with instruments that measure similar constructs taken at the same time
  – **Divergent Validity**: Low correlations with instruments that measure different constructs taken at the same time
  – **Predictive Validity**: High correlation with a criterion in the future
  – Assess with correlation coefficient
Construct Validity

• Most valuable and most difficult measure of validity

• Appropriateness of inferences drawn from results regarding an individual’s status of the psychological construct of interest

• Two considerations:
  – Construct underrepresentation: An instrument does not measure all of the important aspects of the construct
  – Construct irrelevant variance: Results are affected by other unrelated processes
Construct Validity (cont’d)

• Homogeneity: The instrument measures a single construct
  – Evidence: High internal consistency - calculated by split-half reliability

• Convergence: Instrument is related to other measures of the same construct and related constructs
  – Evidence: High correlations with other measures
Questions!

Please use your chat box to submit questions for Xin regarding Validity.
Leveraging Our Efforts Through TEAMS

Webinar Presenter: Dave Weaver
The Evaluator’s Dilemma

Use existing instruments or develop you own?
Using Existing Instruments

• Information about instrument reliability is not often available

• If it is:
  – It is generally for the entire instrument
    • The instrument may address constructs not relevant to your project
      – Not valid for your project
      – Increased burden

• Adapting the instrument negates reliability rating
Developing Your Own Instrument

- Project specific instruments are more likely to have better content validity
  - Better aligned with the project
- Evaluators must start from scratch to develop scales and test reliability
  - Generally requires instrument modification during early stages
  - Increased burden
Tired of Recreating the Wheel

What is the Solution?
TEAMS Suggests

Instruments developed with public funds would be more useful to other evaluators if the developers of the instruments would publish reliability and validity information at the construct level!
AMP Example

• Arizona Mathematics Partnership
  – NSF Targeted MSP Project
  – Maricopa Community Colleges
    • Lead: Scottsdale Community College
  – 7 School Districts
  – Middle School Mathematics
  – Beginning Year 3 of 5
AMP Teacher Survey
Available in the Resource section of the TEAMS Website:

http://teams.mspnet.org/
Key Features

• Background about the development of the instrument
• Definitions of scales along with reliability information
• A print copy of the survey
## Scale Definition and Reliability

### Exhibit 1—AMP Teacher Survey Scales

<table>
<thead>
<tr>
<th>Scale</th>
<th>Description</th>
<th>Items</th>
<th>$n$</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content Preparedness</td>
<td>The degree to which teachers are prepared to teach various subjects in mathematics such as number concepts and operation, proportionality, algebra, geometry, statistics and probability, and problem solving.</td>
<td>2, 3, 4, 5, 6, &amp; 7</td>
<td>125</td>
<td>.833</td>
</tr>
<tr>
<td>Common Core State Standards</td>
<td>The degree to which teacher are prepared to teach according to the Common Core State Standards for Mathematics.</td>
<td>8, 9, 12, &amp; 13</td>
<td>192</td>
<td>.857</td>
</tr>
<tr>
<td>Theory of Action Preparation</td>
<td>The degree to which teachers are prepared to implement the pedagogical practices promoted in the AMP theory of action.</td>
<td>18, 20, 21, 23, 26, 27, 29, 30, 31</td>
<td>184</td>
<td>.940</td>
</tr>
<tr>
<td>Teach for Understanding</td>
<td>The degree to which teachers are prepared to develop students’ understanding of mathematical concepts through classroom discourse, sense-making, justification reasoning, conjecturing, and communicating mathematical ideas.</td>
<td>18, 21, 22, 24, 25, 26, 27, 28, 29, 30, &amp; 31</td>
<td>192</td>
<td>.951</td>
</tr>
<tr>
<td>Problem-Solving</td>
<td>The degree to which teachers are prepared to engage students in problem solving.</td>
<td>19, 20 &amp; 23</td>
<td>192</td>
<td>.853</td>
</tr>
</tbody>
</table>
The Survey

• The AMP Teacher Survey is administered online

• The published documentation includes a printed version
How Is This Helpful

If published information about publicly available evaluation instruments was available at the construct level, then...

• Evaluators could develop reliable instruments by choosing groups of items that are proven reliable measures for constructs relevant to their projects

• The instruments would be more valid measures of the project

• Instruments would be less burdensome for constituents

• More likely to detect impact
How TEAMS Can Help!

If you have instruments developed with public funds that should be shared, and you would like to make them more useful to other evaluators using this approach, you could:

• Put the information in a format similar to the AMP example and send it to TEAMS for website posting

  OR

• Request assistance from TEAMS
TEAMS Services

• Assistance with scale identification and definition
• Internal consistency analysis
• Content validity verification
• Assistance formatting results
• Dissemination of instruments with reliability and validity data at the construct level
Questions!

Please use your chat box to submit questions for Dave
<table>
<thead>
<tr>
<th>John T. Sutton, PI</th>
<th>Dave Weaver, Co-PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xin Wang, Research Associate</td>
<td>RMC Research Corporation</td>
</tr>
<tr>
<td>RMC Research Corporation</td>
<td>111 SW Columbia Street</td>
</tr>
<tr>
<td>633 17th Street</td>
<td>Suite 1030</td>
</tr>
<tr>
<td>Suite 2100</td>
<td>Portland, OR 97201-5883</td>
</tr>
<tr>
<td>Denver, CO 80202-1620</td>
<td></td>
</tr>
<tr>
<td>Phone: 303-825-3636</td>
<td>Phone: 503-223-8248</td>
</tr>
<tr>
<td>Toll Free: 800-922-3636</td>
<td>Toll Free: 800-788-1887</td>
</tr>
<tr>
<td>Fax: 303-825-1626</td>
<td>Fax: 503-223-8399</td>
</tr>
<tr>
<td>Email: <a href="mailto:sutton@rmcrcres.com">sutton@rmcrcres.com</a></td>
<td>Email: <a href="mailto:dweaver@rmccorp.com">dweaver@rmccorp.com</a></td>
</tr>
<tr>
<td><a href="mailto:wang@rmcrcres.com">wang@rmcrcres.com</a></td>
<td></td>
</tr>
</tbody>
</table>