TEAMS

Technical Evaluation Assistance in Mathematics & Science

Establishing Validity and Reliability for Locally Developed Instruments

TEAMS

Technical Evaluation Assistance in Mathematics & Science

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.

TEAMS

Technical Evaluation Assistance in Mathematics & Science

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

TEAMS

Technical Evaluation Assistance in Mathematics & Science

- 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:

ID#	Test	Retest
A351	19	26
A352	20	23
A353	18	22
A354	25	26
A355	23	25
A356	18	23
A357	18	23
A358	20	17
A359	25	27
A3510	17	19
A3511	18	20
A3512	20	23
A3513	25	26
A3514	25	26
A3515	17	16
A3516	27	27
A3517	22	24
A3518	16	23

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:

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

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)

RespondentID	q1	q2	q3	q4	q5	q6	q7	q8	q9
1904038991	7	7	7	7	7	7	7	7	7
1902279256	6	7	7	7	7	7	8	5	5
1901848884	7	7	6	7	7	7	7	7	6
1901114588	7	6	5	7	7	6	6	6	5
1900937813	8	7	7	7	6	6	7	6	8
1910386096	7	7	7	7	7	7	7	7	7
1901772970	8	8	6	8	8	8	8	7	6
1914444563	7	6	6	8	7	7	7	5	5
1904136696	8	7	7	8	7	7	7	7	7
1902218136	8	8	6	8	8	8	8	8	8
1901953875	7	6	5	7	7	7	6	6	7
1901663856	7	7	6	7	7	6	7	7	7
1901693338	5	6	2	6	5	5	6	5	5
1904459022	7	6	4	6	4	7	6	4	3
1901161460	6	6	6	6	6	5	6	5	5
1902192569	7	6	6	6	6	7	7	7	7
1902206958	7	6	3	6	6	5	6	6	6
1901095246	6	8	7	8	8	8	8	8	7
1901974062	7	7	7	7	7	7	7	7	7
1903858390	6	6	6	6	6	6	6	6	6
1902193864	8	8	7	8	7	7	8	7	7
1858968809	7	6	7	7	7	7	6	7	7
1858913494	6	7	5	6	6	5	6	6	6
1858910610	7	7	7	6	6	7	7	7	8
1858902557	6	6	6	6	6	6	6	6	6
1858897406	7	7	7	7	7	7	7	7	7
1858892268	6	6	6	6	6	5	5	6	5
1858888030	6	6	4	5	6	5	5	4	5
1858885363	8	7	5	6	8	6	7	6	6
1858885068	5	3	3	3	4	3	4	3	4
1858883294	8	6	6	6	6	6	6	6	7
1858881518	6	4	1	5	4	1	7	4	4

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!*

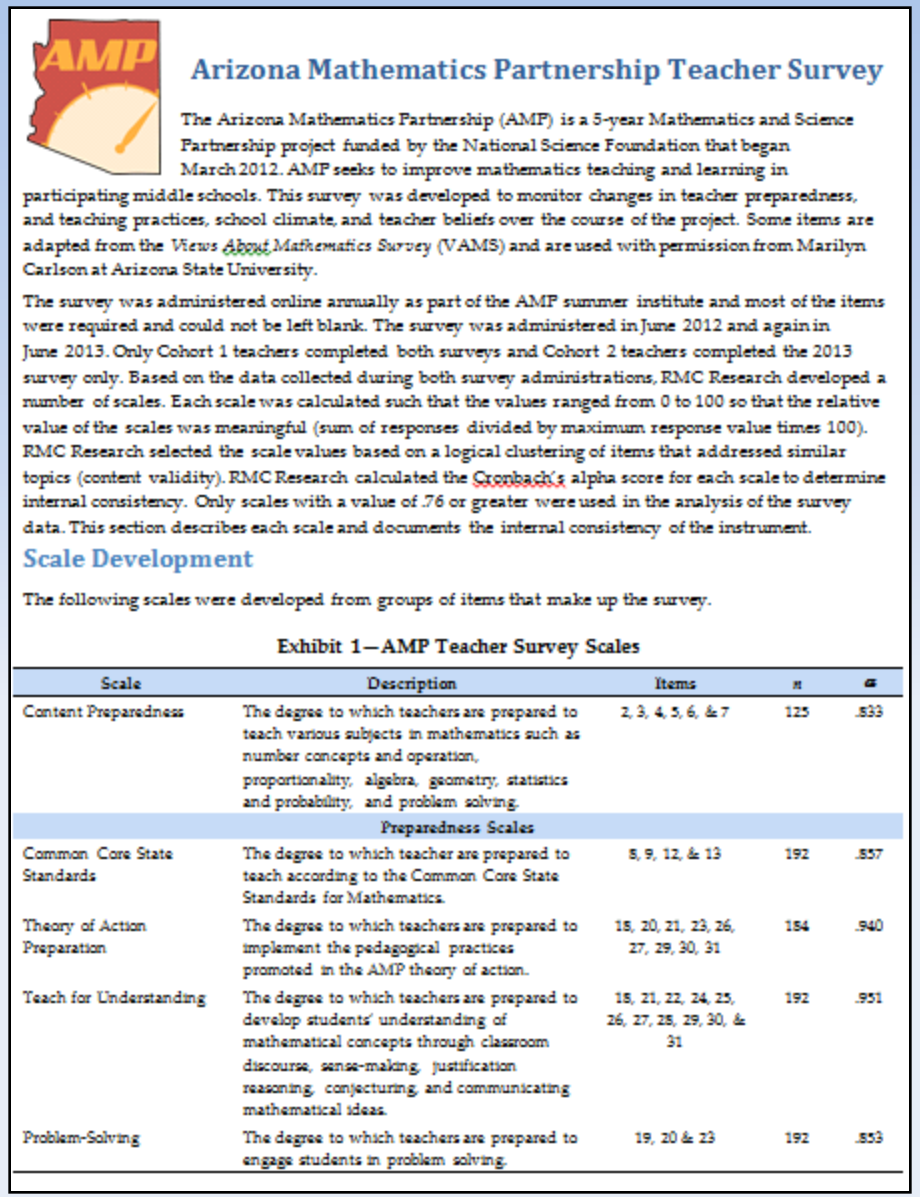
AMP Example

- 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/



Arizona Mathematics Partnership Teacher Survey

The Arizona Mathematics Partnership (AMP) is a 5-year Mathematics and Science Partnership project funded by the National Science Foundation that began March 2012. AMP seeks to improve mathematics teaching and learning in participating middle schools. This survey was developed to monitor changes in teacher preparedness, and teaching practices, school climate, and teacher beliefs over the course of the project. Some items are adapted from the *Views About Mathematics Survey* (VAMS) and are used with permission from Marilyn Carlson at Arizona State University.

The survey was administered online annually as part of the AMP summer institute and most of the items were required and could not be left blank. The survey was administered in June 2012 and again in June 2013. Only Cohort 1 teachers completed both surveys and Cohort 2 teachers completed the 2013 survey only. Based on the data collected during both survey administrations, RMC Research developed a number of scales. Each scale was calculated such that the values ranged from 0 to 100 so that the relative value of the scales was meaningful (sum of responses divided by maximum response value times 100). RMC Research selected the scale values based on a logical clustering of items that addressed similar topics (content validity). RMC Research calculated the Cronbach's alpha score for each scale to determine internal consistency. Only scales with a value of .76 or greater were used in the analysis of the survey data. This section describes each scale and documents the internal consistency of the instrument.

Scale Development

The following scales were developed from groups of items that make up the survey.

Exhibit 1—AMP Teacher Survey Scales

Scale	Description	Items	*n*	α
Content Preparedness	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.	2, 3, 4, 5, 6, & 7	125	.833
	Preparedness Scales			
Common Core State Standards	The degree to which teacher are prepared to teach according to the Common Core State Standards for Mathematics.	8, 9, 12, & 13	192	.857
Theory of Action Preparation	The degree to which teachers are prepared to implement the pedagogical practices promoted in the AMP theory of action.	18, 20, 21, 23, 26, 27, 29, 30, 31	184	.940
Teach for Understanding	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.	18, 21, 22, 24, 25, 26, 27, 28, 29, 30, & 31	192	.951
Problem-Solving	The degree to which teachers are prepared to engage students in problem solving.	19, 20 & 23	192	.853

Key Features

- Background about the development of the instrument
- Definitions of scales along with reliability information
- A print copy of the survey

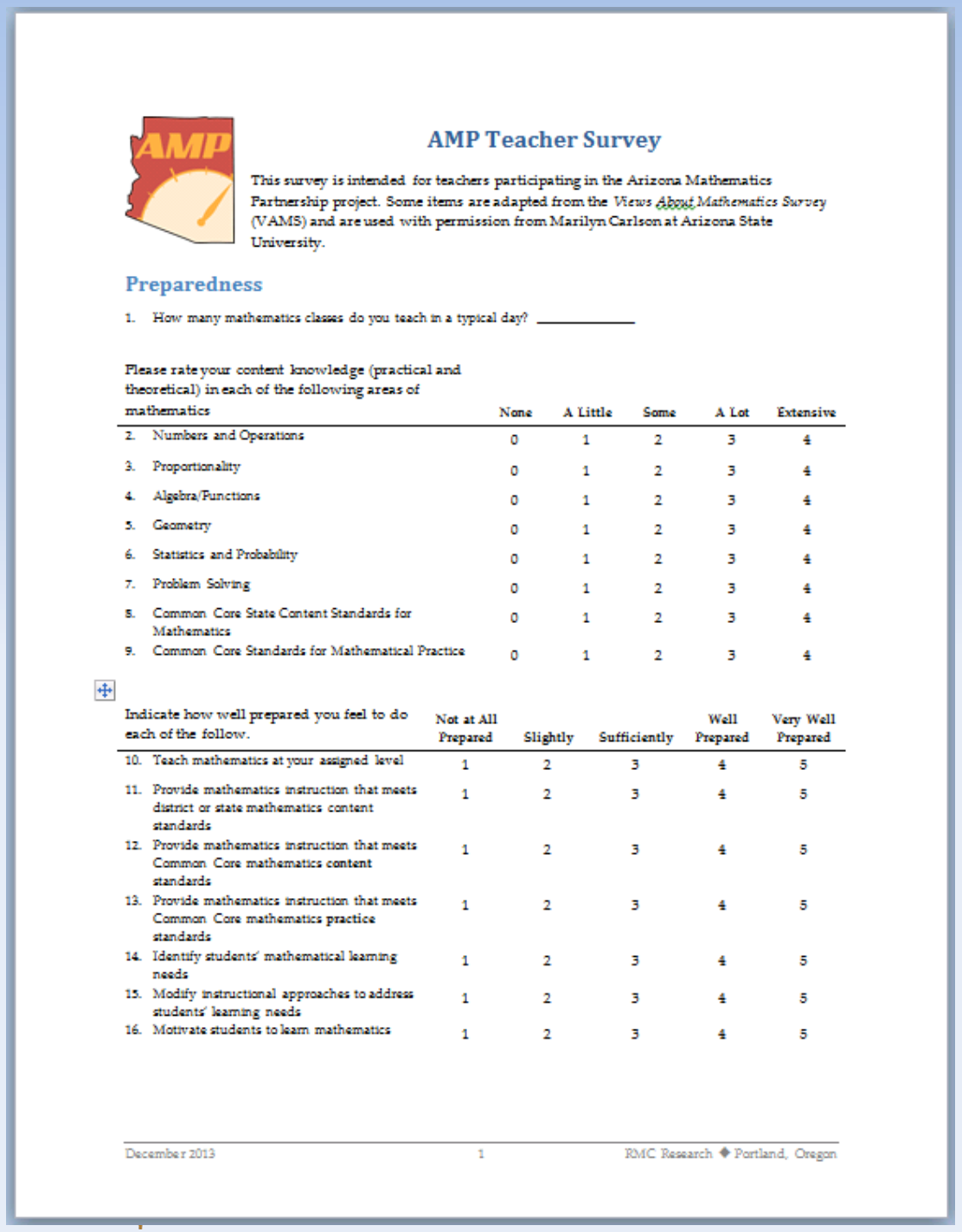
Scale Definition and Reliability

Exhibit 1—AMP Teacher Survey Scales

Scale	Description	Items	*n*	*α*
Content Preparedness	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.	2, 3, 4, 5, 6, & 7	125	.833
	Preparedness Scales			
Common Core State Standards	The degree to which teacher are prepared to teach according to the Common Core State Standards for Mathematics.	8, 9, 12, & 13	192	.857
Theory of Action Preparation	The degree to which teachers are prepared to implement the pedagogical practices promoted in the AMP theory of action.	18, 20, 21, 23, 26, 27, 29, 30, 31	184	.940
Teach for Understanding	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.	18, 21, 22, 24, 25, 26, 27, 28, 29, 30, & 31	192	.951
Problem-Solving	The degree to which teachers are prepared to engage students in problem solving.	19, 20 & 23	192	.853

The Survey

- The AMP Teacher Survey is administered online
- The published documentation includes a printed version



AMP Teacher Survey

This survey is intended for teachers participating in the Arizona Mathematics Partnership project. Some items are adapted from the *Views About Mathematics Survey* (VAMS) and are used with permission from Marilyn Carlson at Arizona State University.

Preparedness

1. How many mathematics classes do you teach in a typical day? _____

Please rate your content knowledge (practical and theoretical) in each of the following areas of mathematics	None	A Little	Some	A Lot	Extensive
2. Numbers and Operations	0	1	2	3	4
3. Proportionality	0	1	2	3	4
4. Algebra/Functions	0	1	2	3	4
5. Geometry	0	1	2	3	4
6. Statistics and Probability	0	1	2	3	4
7. Problem Solving	0	1	2	3	4
8. Common Core State Content Standards for Mathematics	0	1	2	3	4
9. Common Core Standards for Mathematical Practice	0	1	2	3	4

Indicate how well prepared you feel to do each of the follow.	Not at All Prepared	Slightly	Sufficiently	Well Prepared	Very Well Prepared
10. Teach mathematics at your assigned level	1	2	3	4	5
11. Provide mathematics instruction that meets district or state mathematics content standards	1	2	3	4	5
12. Provide mathematics instruction that meets Common Core mathematics content standards	1	2	3	4	5
13. Provide mathematics instruction that meets Common Core mathematics practice standards	1	2	3	4	5
14. Identify students' mathematical learning needs	1	2	3	4	5
15. Modify instructional approaches to address students' learning needs	1	2	3	4	5
16. Motivate students to learn mathematics	1	2	3	4	5

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

TEAMS

Technical Evaluation Assistance in Mathematics & Science

John T. Sutton, PI Xin Wang, Research Associate	Dave Weaver, Co-PI
RMC Research Corporation 633 17th Street Suite 2100 Denver, CO 80202-1620	RMC Research Corporation 111 SW Columbia Street Suite 1030 Portland, OR 97201-5883
Phone: 303-825-3636 Toll Free: 800-922-3636 Fax: 303-825-1626 Email: sutton@rmcres.com wang@rmcres.com	Phone: 503-223-8248 Toll Free: 800-788-1887 Fax: 503-223-8399 Email: dweaver@rmccorp.com