A Review of STEM Research Instruments for Assessing Teacher Practices, Pedagogical Content Knowledge, and Content Knowledge

Drs. Daphne Minner and Alina Martinez

This work was conducted as part of the Community for Advancing Discovery Research in Education (CADRE). This material is based on work supported by the National Science Foundation under Grant No. DRL-0822241. Its contents are solely the responsibility of the authors and do not represent the official views of NSF.

Introduction

President Obama's administration has brought a renewed interest and focus on science, technology, engineering, and mathematics (STEM) education and related workforce issues. For example, the America COMPETES Reauthorization Act of 2010 (P.L. 111-358) calls for the National Science and Technology Council's (NSTC) Committee on STEM Education to create a 5-year Federal STEM education strategic plan. As an initial step in this strategic planning effort, the NSTC conducted a portfolio review of federal STEM education programs (NSTC, 2011). This report describes how 13 Federal agencies utilized $3.4 billion in fiscal year 2010 to support STEM education, out of the $1.1 trillion in annual U.S. spending on education. An independent audit conducted by the Government Accounting Office (GAO, 2012) found that across these 13 agencies, 209 STEM education programs were administered in fiscal year 2010. The Departments of Education (ED) and Health and Human Services (HHS) along with the National Science Foundation (NSF) had the largest fiscal investments, with NSF making the greatest investment (GAO, 2012). "Eighty percent of the funding supported STEM education investments were made by NSF, ED, and HHS" (NSTC, 2012, p.6). Across the NSF's six education research and development programs, Discovery Research K-12 (DR-K12) has the largest budget (NSTC, 2011).

The DR-K12 program seeks to significantly enhance the learning and teaching of STEM. The funded research projects focus on the "development, testing, deployment, effectiveness, and/or scale-up of innovative resources, models and tools" (NSF, 2011, p.2). As such, it is particularly important for the projects within this portfolio to use the soundest methods for testing the efficacy and ultimately effectiveness of the developed educational interventions. This paper presents findings from a review of the DR K-12 projects' proposed instruments across five cohorts of DR-K12 projects funded from 2008 to 2012. This collection of instruments represents commonly used tools for gathering information about educational innovations in the U.S. given that the DR-K12 portfolio is the nation's largest STEM educational intervention research and development fiscal investment.

There are two accompanying publications that were produced as a result of this review effort —*Compendium of research instruments for STEM education: Part 1 and Part 2*. Part 1 is on instruments to assess teacher practices, PCK, and content knowledge (Minner, D., Martinez, A., & Freeman, B., 2012) and Part 2 is on instrument to measure students' content knowledge, reasoning skills, and psychological attributes (Minner, D., Erickson, E., Wu, S., & Martinez, A., 2012). In each compendium we provide detailed information for each instrument on the constructs/variables that are measured, the target audience of the instrument, the subject domains assessed, information on obtaining the instrument, and references to related documents about reliability and validity evidence when it could be located. This paper includes data from Part 1, on teacher instruments and additional analysis of the constructs assessed related to teacher practices.

Methods

The driving research question for this instrument review was: *What are the instruments, constructs, and methods being used to study teacher outcomes within the DR-K12 portfolio*? For this review the research team decided to include only extant, named instruments as opposed to instruments being developed as part of a current grant proposal. This decision was made so that the information generated by this review would reflect assessment tools that are accessible to other researchers, and thus could contribute to knowledge building across studies of similar learning and teaching phenomena. Additionally, if an instrument is already in existence, it stands the most likelihood of having psychometric information generated across multiple settings, which is a fairer assessment of the technical quality of the tool. Three commonly assessed teacher outcomes were the target constructs for this review—teacher practices, pedagogical content knowledge (PCK), and content knowledge.

The review process involved three phases of work. The first phase included reviewing all the proposals that had been funded by the DR-K12 program since 2008 and for which materials were available, which netted 295 eligible projects[1]. Additional materials such as annual reports, publications, products, etc., where available, were reviewed as well, to extract the name of proposed teacher instruments and the constructs being measured. Once this initial dataset was constructed, a second phase of data collection was conducted for instrument-specific information about reliability and validity evidence, development and piloting, accessibility of the instrument, administration, and variables measured. This information was gathered through internet searches with the name of the instrument as the keyword. Information provided by the developer of an instrument was preferred over other sources and given preference if there was conflicting data. In some cases, the instrument was restricted use, so requests were made to the developer for this information. There were some instances of multiple versions of an instrument, in which case, the latest version was included in the dataset. All data was entered into an Excel spreadsheet then coded into descriptive categories so that frequency counts could be generated.

Information gathered during the second phase of data collection provided details that enabled a more fine-grained analysis of the substance of the tools and the psychometric evidence. The focus of this analysis was to assess the strengths and weaknesses in the measurement landscape for key educational constructs. The fifty-four PCK and Practice instruments were further differentiated into five categories of instruments: instructional practices, instruction plus one or two other constructs, instructional beliefs, system-wide reform focused, and discourse focused. The instruments in each of these categories are profiled in this paper. Most of the content knowledge instruments were developed for students and then adapted to assess teacher knowledge. As such, they are frequently developed by psychometricians for large-scale administration, and have undergone rigorous development and testing therefore they were not included in this more detailed analysis. The types of reliability indicators that we captured include: internal scale consistency alpha; interrater agreement as Kappa, percent agreement, or Spearman rank-order correlations. The list of content knowledge tests identified in this review is included in Appendix A and additional information is included in the *Compendium Part 1* (Minner, D., Martinez, A., & Freeman, B., 2012). Lastly, in the

[1] There were 36 projects where the project materials were not available for our review representing an 11% missing data rate. Since the research team only has access to the materials that Principal Investigators provided, the findings from this review should be considered only suggestive of trends within the portfolio of DR-K12 projects. Often PIs do not know exactly what they will end up using in the project, until the project is funded. Therefore, we use for convenience sake phrases like "projects used," but in fact we only know what they proposed to use or consider, not what they ended up using in their studies.

third phase, a content analysis of the indicators for instructional practices was conducted for a subset of the collected instruments.

Findings

Phase I: Investigator Identified Instruments

Seventy-five projects (25% of the DR-K12 portfolio) proposed to measure teacher practices, pedagogical content knowledge, or content knowledge as an outcome of the funded work. Seventy-one percent of these projects measured only one teacher outcome (n=32 projects measured practice; n=14 PCK; n=7 content); twenty-four percent measured two outcomes (n=4 PCK and content; n=8 PCK and practice; n=6 practice and content); and only five percent (n=4) measured all three types of outcomes. Across these 75 studies, eighty-two extant instruments were identified. The three most common instruments used for each outcome are listed in Table 1. For practices, a total of 42 instruments were identified, for PCK 24 instruments were identified, and for content knowledge it was 27 instruments. Some instruments were identified by the Principal Investigators as measuring more than one type of outcome.

Table 1: Number of studies that used the most frequently named instruments by construct

Instrument Name	Constructs		
	Practices	PCK	Content knowledge
Reformed Teaching Observation Protocol (modified)	15	2	1
Inside the Classroom Observation and Analytic Protocol	8		
Surveys of Enacted Curriculum (modified)	5	1	
Mathematical Knowledge for Teaching[2]	1	14	3
Knowledge of Algebra for Teaching (modified)		2	
Science Teaching Efficacy Belief Instrument		2	
Views of Nature of Science Form C			3
National Assessment of Educational Progress (modified)			3
Praxis content tests/ Earth & physical science (modified)			2
TIMSS content tests (modified)			2

Phase II: Instrument Details by Outcome Categories

The remaining findings are germane to the fifty-four PCK and Practice instruments that were further differentiated into the five categories noted above—instructional practices, instruction plus one or two other constructs, instructional beliefs, system-wide reform focused, and discourse focused. For each of these areas, we will describe the relative distribution of instruments by type of method employed, grade level, and conclude this section with reliability and validity information across the instruments.

Instructional Practices—There were eleven instruments that primarily assessed classroom instructional practices (see Table 2). These instruments (seven observation protocols, three scoring rubrics for educational products, one survey) were predominantly designed for pre-kindergarten through middle school teachers (n=6, 55%). There were slightly more focused on science (n=5, 45%) than mathematics (n=3, 27%) or technology (n=2, 18%). The science only observation

[2] Learning Mathematics for Teaching (LMT) is the name of the project, not an instrument. Content Knowledge for teaching Mathematics (CKT-M) and Mathematical knowledge for teaching (MKT) are the same and the current name is MKT. Therefore any study mentioning LMT, CKT-M, or MKT were counted here.

protocols (ISCOP, LFCPO, STIR) capture variables ranging from the lesson's temporal flow and percentage of time students spend in different types of groupings, to the extent of opportunity for students to engage in the various phases of the investigation cycle. The two scoring rubrics (TIDES, Scoop) are applied to lesson artifacts and instructional materials that the teacher provides students. They code for student grouping, structure of lessons, use of scientific resources, hands-on opportunities through investigation, cognitive depth of the materials, encouragement of the scientific discourse community, and opportunity for explanation/justification, and connections/applications to novel situations. The two mathematics only observation protocols (AFM, EMCO) are for PK and PK-6 classrooms and they assess general aspects instruction such as the type and depth of the mathematics in the instruction. The Quality of Instruction Measure describes the quality of instruction using proportion of lesson time spent on six dimensions: core mathematical ideas; representations matched to algorithms; conceptual and temporal links; elicitation of student thinking and teacher responsiveness; amount of student work contributed; and the kind of student work in the lesson. The two technology instruments (ETAP—survey and LoFTI—observation protocol) assess how technology is being used within the classroom context. The O-TOP is an observation protocol for post-secondary classrooms that captures the use of various instructional strategies to foster various high level thinking skills such as metacognition and divergent thinking.

Table 2. Details for instructional practice instruments

Acronym	Instrument Name	Instrument Details					
		Observation protocols	Scoring rubrics	Elem-middle level teachers	Science focus	Math focus	Technology focus
AFM	Assessment of the Facilitation of Mathematizing	✓		✓		✓	
EMCO	Early Mathematics Classroom Observation	✓		✓		✓	
ETAP	EdTech Assessment Profile			*			✓
ISCOP	Instructional Strategies Classroom Observation Protocol	✓		✓	✓		
LFCPO	Lesson Flow Classroom Observation Protocol	✓		✓	✓		
LoFTI	Looking for Technology Integration	✓		*			✓
O-TOP	OCEPT-Classroom Observation Protocol	✓			✓	✓	
STIR	Science Teacher Inquiry Rubric	✓		✓	✓		
TIDES	Transforming Instruction by Design in Earth Science--Teacher assignment quality rubrics		✓	*	✓		
Scoop	Scoop Notebook		✓	✓	✓		
	The Quality of Instruction Measure		✓	*		✓	

*unable to determine

Instructional Practices-Plus—There were eleven instruments that measured instructional practices in addition to one or two other constructs such as physical context, demographics, teacher content knowledge, or some aspect of classroom management (see Table 3). This more comprehensive nature is also reflected in the subject domain—most assess both mathematics and science; and the SIOP looks at more general teaching skills such as lesson planning and assessment. The ICOT is the only technology specific observation protocol that was identified in the proposals reviewed.

Table 3. Details for instructional practice-PLUS instruments

Acronym	Instrument Name	Instrument Details					
		Observation protocols	Interview protocol	survey	Science focus	Math focus	Technology focus
CETP-COP	The Collaboratives for Excellence in Teacher Preparation core evaluation classroom observation protocol	✓	✓		✓	✓	
EQUIP	Electronic Quality of Inquiry Protocol	✓			✓	✓	
ICOT	ISTE Classroom Observation Tool	✓					✓
KAT	Knowledge of Algebra for Teaching			✓		✓	
MQI	Mathematical Quality of Instruction	✓				✓	
PRAXIS	Praxis Teaching Foundations			✓	✓	✓	
PRISM	Preschool Rating Instrument for Science and Mathematics	✓			✓	✓	
SESAME	Self-Evaluation of Science and Math Education	*	*	*	✓	✓	
SIOP	Sheltered Instruction Observation Protocol	✓					
TIMSS	Third International Mathematics and Science Video Study (TIMSS)	✓		✓	✓	✓	
	Ohio Middle Level Mathematics and Science Education Bridging Study - Teacher Questionnaire			✓	✓	✓	

*unable to determine

Instructional Beliefs—Nine instruments were identified as measuring instructional beliefs (eight surveys and one interview) (see Table 4). Six (67%) of these were developed for science, with only IMBS and MTEBI for math. The TSES is a non-subject-specific instrument. The science beliefs measured centered around self-efficacy at teaching in general, at teaching science content and at teaching science investigation skills. The mathematics surveys similarly measured self-efficacy in various math domains and in teaching math.

Table 4. Details for instructional belief instruments

Acronym	Instrument Name	Instrument Details					
		Observation protocols	Interview protocol	survey	Science focus	Math focus	Technology focus
IMBS	Indiana Mathematics Beliefs Scale			✓		✓	
MTEBI	Mathematics Teaching Efficacy Belief Instrument			✓		✓	
PSI-T	Principles of Scientific Inquiry-Teacher			✓	✓		
SETAKIST	Self-Efficacy Teaching and Knowledge Instrument for Science Teachers			✓	✓		
STEBI	Science Teaching Efficacy Belief Instrument			✓	✓		
TBI	Teacher Belief Interview		✓		✓		
TSES	Teachers' Sense of Efficacy Scale			✓			
TSI	Teaching Science as Inquiry			✓	✓		
VNOS-C	Views of Nature of Science Form C			✓	✓		

*unable to determine

System-Wide Reform Focused—This set of ten instruments capture instructional practices, instructional beliefs, the administrative/policy context influencing instruction, student and teacher demographics, and the content being taught (see Table 5). These instruments have been used to investigate the effect of education system reform efforts. Half of the instruments capture mathematics and science instruction, four are either mathematics or science-specific, and the FFT is non-domain-specific. In this set of instruments is where we first saw the emergence of English Language Arts-specific items included on two of the instruments (SEC, SII). This set of instruments predominately employs survey administration, with 60 percent of the instruments using this approach. There are two observation protocols (LSC, COP), one interview (CIP), and one rubric (FFT).

Table 5. Instruments assessing multiple dimensions related to system-wide reform efforts.

Acronym	Instrument Name	Constructs											
		Practices	Content	Beliefs	Management	Assessment	Social	Physical	Admin context	Demographic	planning	science	math
CIP	Inside the Classroom Teacher Interview Protocol			✓			✓	✓	✓			✓	✓
COP	Inside the Classroom Observation and Analytic Protocol	✓	✓		✓		✓	✓	✓			✓	✓
CTRI	Coaching /Teacher Reflection Impact Surveys	✓		✓	✓					✓			✓
FFT	Danielson's Framework for Teaching Domains	✓	✓		✓	✓	✓	✓			✓		
--	Inside the Classroom Teacher Questionnaire: Math or Science version	✓	✓	✓	✓					✓		✓	✓
LSC	LSC Core Evaluation Classroom Observation Protocol	✓	✓				✓				✓	✓	✓
SEC	Surveys of Enacted Curriculum	✓		✓		✓			✓	✓		✓	✓
SII	Study of Instructional Improvement	✓	✓	✓					✓	✓			✓
TIMSS-R	TIMSS-R Science Teacher Questionnaire			✓					✓	✓	✓	✓	
TIMSS-R	TIMSS-R Mathematics Teacher Questionnaire			✓					✓	✓	✓		✓

Discourse Focused—There were thirteen instruments in this category that looks at instructional practices as well as the social aspects of the classroom community, including classroom management (see Table 6). All of these instruments are observation protocols and three employ additional methods (ISIOP—survey, IQA—scoring rubric, ELLCO—interview). Fourth-six percent of the instruments (n=6) are non-domain-specific and instead focus of the assessment of the teacher-student interaction in terms of the verbal discourse and the emotional support exhibited in the classroom. There are four instruments in this set that have English Language Arts-specific items (DAISI, IQA, ELLCO, Mathematics Classroom Observation Protocol). Three instruments measure mathematics-specific discourse (Mathematics Classroom Observation Protocol, COEMET, IQA) three measure science-specific discourse (ISIOP, DAISI, Science Classroom Observation Guide), and one measures both (RTOP).

Table 6. Instruments assessing multiple dimensions related to classroom discourse environment.

Acronym	Instrument Name	Constructs			
		Practices	Management	Social	Physical
CLASS	The Classroom Assessment Scoring System	✓	✓	✓	
--	Classroom Snapshot	✓	✓	✓	✓
CLO	Classroom Lesson Observation Instrument	✓	✓	✓	
COEMET	Classroom Observation of Early Mathematics Environment and Teaching	✓	✓	✓	
DAISI	The Dialogic Activity in Science Instruction	✓		✓	
EAS	The Emergent Academic Snapshot	✓	✓	✓	
ELLCO	The Early Language and Literacy Classroom Observation	✓	✓		✓
IQA	Instructional Quality Assessment	✓		✓	
ISIOP	Inquiring into Science Instruction Observation Protocol	✓	✓	✓	✓
--	Mathematics Classroom Observation Protocol	✓		✓	
RTOP	Reformed Teaching Observation Protocol	✓		✓	
--	Science Classroom Observation Guide (NCOSP)	✓		✓	
SPC	Standards performance continuum	✓		✓	

Reliability and Validity—In assessing this collection of 54 instruments for reliability and validity evidence overall, we found a rather alarming level of missing information. For reliability evidence 38 percent (19/50) of the eligible instruments (4 were n/a) have missing information (see Tables 7 & 8); for validity evidence 51 percent (27/53) have missing information (1 was n/a). In Table 7 we see that by method type, the low frequency of interview and rubric instruments dramatically influences the percentage of missing information. Therefore, for users of these types of protocols, it is particularly important to obtain instrument-specific information prior to deployment and to pilot them with your own study participants. In comparing observation protocols to surveys, there was proportionally more missing information for the observation protocols.

Table 7. Number and percentage of instruments measuring PCK and practice constructs for each method type by reliability and validity indicators

Method type*	**Reliability Evidence Level (%)**					**Validity Evidence Type****						
	Miss.	Low	Accept	Good	N/A	Miss	Cont	Const	Pred	Concur	Discr	N/A
Observation (n=29)	9 (31)	1 (3)	8 (30)	11 (38)	--	15 (52)	8 (28)	4 (14)	6 (21)	4 (14)	2 (7)	--
Interview (n=2)	1 (50)	--	--	--	1 (50)	2 (100)	--	--	--	--	--	--
Rubric (n=4)	3 (75)	1 (25)	--	--	--	3 (75)	--	--	--	1 (25)	--	--
Survey (n=18)	5 (28)	--	5 (28)	5 (28)	3 (17)	6 (33)	5 (28)	6 (33)	3 (17)	--	--	1 (6)

*one instrument unobtainable and unable to determine method type from existing descriptions.
**Cont=content; Const=construct; Pred=predictive; Concur=concurrent; Discr=discriminant

The reliability and validity information that was available for this sample of instruments, indicates that there is a greater proportion of observation protocols with higher reliability levels (in the good range being 0.80 or higher) than surveys (38% vs. 28%). In looking at the balance of evidence by instrument foci, stronger evidence of multiple types of validity using a single instrument exists for protocols assessing discourse variables and instructional beliefs than the other three categories of foci. Overall, thirteen of the 50 eligible instruments (26%) had evidence of acceptable levels of reliability (range of 0.60-0.79) and sixteen (32%) had good levels. In terms of validity evidence across the fifty-three eligible instruments, thirteen (23%) had addressed content validity, ten (19%) construct validity, nine (17%) predictive validity, five (9%) concurrent validity, and two (4%) discriminant validity. See *Compendium Part 1,* for detailed reliability and validity information by instrument (Minner, D., Martinez, A., & Freeman, B., 2012).

For the eleven instructional practice instruments, one had low reliability evidence, and four (36%) had acceptable or good evidence. For only two instruments was the team able to find validity evidence (see Table 8). For Instructional practices plus, five (45%) instruments had evidence of acceptable or good reliability and four provided validity evidence (36%). For beliefs sixty-seven percent of these instruments had evidence of either acceptable or good reliability, and 78 percent had demonstrated evidence of validity. System-wide reform sixty-six percent of the instruments had acceptable or good reliability and 56 percent had validity evidence. Of the discourse-focused instruments, sixty-two percent of these instruments (n=8) have acceptable or good reliability and evidence of validity. Fifty-four percent of these instruments had demonstrated evidence of more than one type of validity, more than any other category of instruments.

Table 8. Number and percentage of instruments by focus, reliability, and validity indicators

Instrument Focus (n=54)	**Reliability Evidence Level (%)**					**Validity Evidence Type***					
	Miss	Low	Accept	Good	N/A	Miss	Content	Const	Pred	Concur	Discr
Instructional Practices (11)	6 (55)	1 (9)	3 (27)	1 (9)	--	9 (82)	1 (9)	--	--	1 (9)	--
Instruction plus… (11)	6 (55)	--	2 (18)	3 (27)	--	7 (64)	2 (18)	2 (18)	--	1 (9)	--
Instructional Beliefs (9)	--	--	3 (33)	3 (33)	3 (33)	2 (22)	4 (44)	5 (56)	--	--	--
System-wide reform (10)	3 (33)	--	3 (33)	3 (33)	1 (11)	4 (44) 1 n/a	2 (22)	1 (11)	3 (33)	--	--
Discourse (13)	4 (31)	1 (8)	2 (15)	6 (46)	--	5 (38)	4 (31)	2 (15)	6 (46)	3 (23)	2 (15)

*Const=construct; Pred=predictive; Concur=concurrent; Discr=discriminant. An instrument may have generated evidence of more than one type of validity.

Phase III: Conceptualizing Instructional Practice

In this final analysis, the instruments that met the following inclusion criteria were identified: (1) acceptable to good reliability evidence was available; (2) some form of validity evidence was available; and (3) the instrument items were available for independent review, free of charge. The items in the fifteen instruments that met these criteria were reviewed more closely to determine the various aspects of practice that are measured and a content analysis was performed producing the matrix in Appendix B. The practice indicators were subsequently grouped into four superordinate categories to describe a possible latent construct underpinning the individual items: (1) classroom climate (the social/emotional aspects of the classroom), (2) content-related factors, (3) sense-making responsibility, and (4) pedagogical content knowledge (see Table 9). The items that show the most consistency across instruments (i.e. 50% or more of the instruments included the item) fall within the constructs ***of content-related factors:***

- content storyline--connecting the concepts coherently
- adept content understanding (error free) used to facilitate students developing big ideas
- elicitations of prior knowledge, misconceptions
- connections to real life, contextualizing content;

opportunities for student sense-making:

- level 1: teaching by telling; reinforcement of knowledge; demonstrations to show students the correct information
- level 2: more facilitation of student sense-making; strategies not used consistently or skillfully
- level 3: facilitation of student knowledge construction; questioning is responsive to student ideas to encourage reasoning; and

pedagogical content knowledge:

- students engage directly with phenomenon, hands-on, collecting data, doing math
- students formulate explanations
- students evaluate, justify, clarify, represent thinking to peers
- habits of mind for problem solving, scientific thinking, multiple approaches considered
- metacognition encouraged.

Table 9. Practice indicator items assessed in the 15 instruments reviewed

Practice Indicator Items	Superordinate categories	# instruments that contain this item*	% instruments that contain this item*
teacher actively and positively engaged with students/ encouraged participation	climate	5	33
teacher acknowledge and reinforced student efforts/ provide praise for persistence	climate	5	33
student collaboration encouraged	climate	6	40
student discourse shapes lesson significantly	climate	5	33
teacher show curiosity and enthusiasm for content	climate	2	13
classroom management and social climate factors	climate	3	20
level of student engagement	climate	6	40
lesson purpose provided to students	content	4	27
teacher knowledgeable and confident about subject	content	4	27
content storyline--connecting the concepts coherently	content	7	47
adept content understanding (error free) used to facilitate students developing big ideas	content	8	53
content checklist	content	3	20
resource richness for lesson	content	3	20
multiple representations of concepts used	content	4	27
elicits prior knowledge, misconceptions	content	10	67
connects to real life, contextualizes content	content	10	67
interdisciplinary connections made	content	4	27
type of cognitive activity--low (recall) to high (construct) checklist	sense-making	2	13
teacher directed vs. student directed facilitation	sense-making	6	40
level 1: teaching by telling; reinforcement of knowledge; demonstrations to show students the correct information	sense-making	7	47
level 2: more facilitation of student sense-making; strategies not used consistently or skillfully	sense-making	7	47
level 3: facilitation of student knowledge construction; questioning is responsive to student ideas to encourage reasoning	sense-making	9	60
students engage with scientifically oriented questions	PCK-Science	3	20/ (38)*
students plan investigations to gather evidence	PCK-Science	5	33/ (63)*
students encouraged to explore concepts prior to explanation offered by teacher	PCK-Science	3	20/ (38)*
students engage directly with phenomenon, hands-on, collecting data, doing math	PCK	7	47
students formulate explanations	PCK	12	80
students evaluate, justify, clarify, represent thinking to peers	PCK	12	80
students encouraged to use science /math terms	PCK	5	33
students use representations/ abstractions	PCK	5	33
habits of mind for problem solving, scientific thinking, multiple approaches considered	PCK	10	67
metacognition encouraged	PCK	8	53
assessment practices used	PCK	5	33
type of instructional modes/ activities noted	PCK	4	27

*percentage in parentheses is of science-focused or science *and* math-focused protocols (n=8 total)—see Appendix B

Conclusion

This review of the state of measurement tools for STEM educational interventions indicates that as a community of scholars, more effort needs to be made to provide relevant psychometric information on the tools we develop and use. Without the basic information about what is needed to achieve an acceptable level of interrater reliability, users of these observation protocols, interview protocols, and scoring rubrics do not have the necessary information to make informed choices about the implementation of these tools in their own work. Information about survey scale coherence, as well as content and construct validity is essential to move the field forward in reaching a community consensus on operational definitions of key outcome variables. Predictive, concurrent and discriminant validity evidence is what policy makers expect our tools provide so that informed decisions about the efficacy and effectiveness of interventions can be made soundly. With the level of missing information, just over half of the instruments have evidence of acceptable or good levels of reliable implementation and scale consistency, and less than a third having associated validity evidence, there is a good deal of work yet to accomplish.

In terms of how teaching practices are measured, this review found some consensus regarding key elements of instruction related to content, how to teach specific content with targeted techniques (PCK), and defining different levels of student sense-making opportunities provided by the teacher. However, how social climate is captured in instruments still varies quite a bit with some instruments placing more emphasis on the discourse between students, others preferring to focus on student–to–teacher interactions and the emotional tone set by the teacher.

References

GAO, *Science, Technology, Engineering, and Mathematics Education: Strategic planning needed to better manage overlapping programs across multiple agencies*, GAO-12-108 (Washington, D.C.: Jan. 20, 2012).

Minner, D., Martinez, A., & Freeman, B. (2012). *Compendium of research instruments for STEM education. Part 1: Teacher practices, PCK, and content knowledge.* Cambridge, MA: Abt Associates. Retrieved from Education Development Center, CADRE website: http://cadrek12.org/sites/default/files/Compendium%20of%20STEM%20instruments%20Part%201.pdf

Minner, D., Erickson, E., Wu, S., & Martinez, A. (2012). *Compendium of research instruments for STEM education. Part 2: Measuring students' content knowledge, reasoning skills, and psychological attributes.* Cambridge, MA: Abt Associates. Retrieved from Education Development Center, CADRE website: http://www.cadrek12.org/sites/default/files/Compendium%20of%20STEM%20instruments%20Part%202_11-20-12.pdf

NSF (2011). Discovery Research K-12 (DR-K12): Program Solicitation 11-588.

NSTC (2011, December). *The Federal Science, Technology, Engineering, and Mathematics (STEM) Education Portfolio.* http://www.whitehouse.gov/sites/default/files/microsites/ostp/costem__federal_stem_education_portfolio_report.pdf

NSTC (2012, February). *Coordinating Federal Science, Technology, Engineering, and Mathematics (STEM) Education Investments: Progress Report.* http://www.whitehouse.gov/sites/default/files/microsites/ostp/nstc_federal_stem_education_coordination_report.pdf

Appendix A: Content Knowledge Instruments Identified to Assess Teachers

Acronym	Instrument Name
ACT	American College Testing
AP	Advanced Placement
CST	Content Specialty Test Earth Science for New York Teacher Certification
DTAMS	Diagnostic Teacher Assessments in Mathematics and Science
DTAMS	Diagnostic Mathematics Assessments for Elementary Teachers--algebra
FACETS	Diagnoser Tools
FCI	Force Concept Inventory
GRE	Graduate Record Exam
ITBS	Iowa Test of Basic Skills
MAP	Missouri Assessment Program
MKT	Mathematical Knowledge for Teaching
M-SCAN	The Mathematics Scan
MOSART	Misconceptions-Oriented Standards-Based Assessment Resources for Teachers
NAEP	National Assessment of Educational Progress
PISA	Program for International Student Assessment
PRAXIS	content tests/ Earth & physical science
Regents	New York State Regents Exam
SAT	Stanford Achievement Test
TAGLIT	Taking a Good Look at Instructional Technology
TIMSS	content tests
West-E	Washington Educator Skills Test-Endorsements
WESTEST	Science WESTEST
	American Chemical Society Division of Chemical Education Examinations Institute
	IL Certification Testing System Study Guide-Science: Biology
	Classroom Test of Scientific Reasoning (Lawson)

Appendix B: Item Matrix by Instrument for Instructional Practices

original construct category from Phase II analysis	content domains	acronym	teacher actively and positively engaged with children/ encouraged participation	teacher acknowledge and reinforced student efforts/ provide praise for persistence	student collaboration encouraged	student discourse shapes lesson significantly	teacher show curiosity and enthusiasm for content	classroom management and social climate factors	level of student engagement	lesson purpose provided to students	teacher knowledgeable and confident about subject	content storyline, connecting concepts coherently	adept content understanding (error free) used to facilitate student developing big ideas	content checklist	resource richness for lesson	multiple representations of concepts used	elicits prior knowledge, misconceptions	connects to real life, contextualizes content	interdisciplinary connections made
practices	Math	AFM											X					X	
practices	Science	ISCOP								X		X					X	X	
practices	Math & science	O-TOP	X	X	X						X		X			X	X	X	X
practices	Science	STIR																	
practices	Science	Scoop			X	X				X		X	X		X		X	X	
practice plus	Math & science	CETP-COP			X				X		X		X				X	X	X
practice plus	Math & science	EQUIP			X	X			X			X	X				X		
practice plus	Technology	ICOT												X					
practice plus	Math	MQI										X	X			X	X		
discourse	General	CLASS	X	X				X	X	X		X			X	X		X	
discourse	Math	COEMET	X	X	X		X	X	X		X		X	X	X			X	
discourse	Math & ELA	IQA				X			X								X		
discourse	Science	ISIOP	X	X	X		X	X	X	X		X	X	X			X	X	X
discourse	Math & science	RTOP	X	X		X					X	X				X	X	X	X
discourse	General	SPC				X											X	X	
	# instruments		5	5	6	5	2	3	6	4	4	7	8	3	3	4	10	10	4
		%	0.33	0.33	0.4	0.33	0.13	0.2	0.4	0.27	0.27	0.47	0.53	0.2	0.2	0.27	0.67	0.67	0.27

original construct category from Phase II analysis	content domains	acronym	type of cognitive activity--low (recall) to high (construct)	teacher directed vs. student directed facilitation	level 1: teaching by telling; reinforcement of knowledge; demonstrations to show	level 2: more facilitation of student sense-making; strategies not used consistently or skillfully	level 3: facilitates student knowledge construction; questioning responds to student ideas, encourages reasoning	students engage with scientifically oriented questions	students plan investigations to gather evidence	students encouraged to explore concepts prior to explanation offered by teacher	students engage directly with phenomenon, hands-on, collecting data, doing math	students formulate explanations	students evaluate, justify, clarify, represent thinking to peers	students encouraged to use science /math terms	students use representations/ abstractions	habits of mind for problem solving, scientific thinking, multiple approaches considered	metacognition encouraged	assessment practices used	type of instructional modes/ activities noted
practices	Math	AFM		x	X	X	X												
practices	Science	ISCOP									x	x	x	x	x	x			
practices	Math & science	O-TOP					x					x	x			x	x		
practices	Science	STIR		x				x	x		x	x	x						
practices	Science	Scoop		x	X	X	X	x	x		x	x	x	x	x	x	x	x	
practice plus	Math & science	CETP-COP	x		x	x	x					x	x		x	x	x	x	x
practice plus	Math & science	EQUIP	x	x	X	X	X		x	x	x	x	x			x		x	
practice plus	Technology	ICOT																	x
practice plus	Math	MQI			X	X	X				x	x	x	x	x	x	x	x	x
discourse	General	CLASS										x	x			x	x		
discourse	Math	COEMET									x	x	x	x		x	x		
discourse	Math & ELA	IQA										x	x						
discourse	Science	ISIOP		x	X	X	X	x	x	x	x	x	x	x		x	x	x	x
discourse	Math & science	RTOP					x		x	x		x	x		x	x	x		
discourse	General	SPC		x	x	x	x												
	# instruments			2	6	7	7	9	5	3	7	12	12	5	5	10	8	5	4
		%	0.13	0.4	0.5	0.5	0.6	0.2	0.33	0.2	0.47	0.80	0.80	0.33	0.33	0.67	0.53	0.33	0.27

X=protocols that provide specific teacher moves that fit into each of these levels of teaching skill