The Artifact Indicator Project:

Three Studies in the Use of STEM Classroom Artifacts

By

Jeanette Joyce

A dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Education

Written under the direction of

Drew H. Gitomer

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October 2017

ABSTRACT OF THE DISSERTATION

The Artifact Indicator Project:

Three Studies in the Use of STEM Classroom Artifacts

By

Jeanette Joyce

Dissertation Director:

Drew H. Gitomer

Calls for reform in STEM education have proliferated as nations strive to prepare students for the future global economy. The 21st century competencies described in the most recent reforms are represented in new standards (such as the Common Core State and Next Generation Science Standards). However, it is not enough to develop reforms through publication and legislation. What matters is how reform policies are interpreted by teachers and enacted in classrooms. Therefore, it becomes essential to have measures as indicators of how new reforms are reaching students and whether progress is being made toward reform goals. The following studies explore the possibility of using classroom artifacts in a complementary measure to classroom observations, achievement scores, and surveys. Classroom artifacts, which can include assigned tasks from teachers and the responding student work, are very useful in providing evidence about the instruction available to students. The proposed research extends the body of artifact work in several critical ways. First, study one provides a thematic synthesis of existing STEM artifact studies to develop a framework of design criteria. The second study will focus on

the design of a standards-based science classroom artifact indicator protocol, informed by findings from the first study. Study three explores the development of a standards-based math artifact indicator protocol, which differs in critical ways from the science domain. Findings will be helpful to the artifact research community as well as stakeholders in STEM education as we move toward reforms in classroom instruction that includes both content and practices.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. 1445632. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Table of Contents

List of tables

Article 1: Using Classroom Artifacts to Investigate STEM Instruction:

A Design Synthesis Study

Article 2: Using Classroom Artifacts to Track Enacted Science Reform:

The Artifact Indicator Protocol Study

Article 3: Using Classroom Artifacts to Assess Enacted Math Standards: The Artifact Indicator Protocol Study

List of figures

Article 1: Using Classroom Artifacts to Investigate STEM Instruction:

A Design Synthesis Study

Article 2: Using Classroom Artifacts to Track Enacted Science Reform:

The Artifact Indicator Protocol Study

Article 3: Using Classroom Artifacts to Assess Enacted Math Standards:

The Artifact Indicator Protocol Study

Introduction

Calls for reform in STEM education have proliferated as nations strive to prepare students for a future in the global economy. In a 2006 monograph, the RAND Corporation describes reform-oriented teaching as " …the development of complex cognitive skills and processes" (p. iii), and, over the past decade, reforms have emerged in both science and math that ask teachers to set instructional tasks for their students that embody this complexity. Consequently, stakeholders are seeking measures to answer the questions of to what extent teachers are able to implement new reforms and provide tasks that actually represent the intellectual depth asked for. Without such measures, stakeholders will be unable to gauge progress toward educational reform goals, or to give support to teachers and administrators in their quest to do so. This research investigates the potential of classroom artifacts in providing evidence that helps to answer these questions.

There have been calls for reform in STEM Education for nearly a century. Most recently, continued low US achievement on measures such as Trends in International Mathematics and Science Study (TIMMS) and National Assessment of Educational Progress (NAEP) has spurred further reforms (Nord et al, 2011). The Common Core State and Next Generation Science Standards were developed and are in part a response to the Heritage Foundation report (2009) stating the future economic growth in the U.S. was dependent on improvement in STEM education.

The 21st century competencies described in this most recent phase of reform are represented in new standards (such as the Common Core State and Next Generation Science Standards) for developing global citizens. According to Ananiadou & Claro

(2009) in their Organisation of Economic Co-operation and Development (OECD) report, "Developments in society and economy require that educational systems equip young people with new skills and competencies, which allow them to benefit from the emerging new forms of socialisation and to contribute actively to economic development under a system where the main asset is knowledge"(p. 5). One particular change in this set of reforms has been the extension of scrutiny to not just **content** but also **practices,** or processes, in which successful students must be proficient.

However, it is not enough to develop reforms through publication and legislation. What matters is how reform policies are interpreted by teachers and enacted in classrooms. This is in line with what Lipsky referred to as "street-level policy" (in Gilson, 2015), wherein ideas would be re-interpreted as they move from the halls of legislature to the halls of schools. Capps, Shemwell, and Young (2016) report that teachers can misunderstand new reforms and self-report that they are in compliance when tasks set for students are not truly aligned with standards. However, Allen and Penuel (2016) found that, with support, teachers were able to bring all new standards into classroom instruction. The challenge for stakeholders becomes how to elicit evidence of the extent to which standards represented in policies are actually being enacted in classrooms. Therefore, it becomes essential to have measures of how new reforms are reaching students and whether progress is being made toward reform goals. Such a set of measures would form an indicator system.

In 2013, a National Research Council report, *Monitoring Progress Toward Successful K-12 Education: A Nation Advancing?,* called for a national indicator system that could be used to improve STEM education. The report described 14 indicators that

were needed to guide improvement. Congress then directed the National Science Foundation to begin implementing a progress monitoring system for the indicators. In response to this directive, there has been a call for development of new instruments to be used in an indicator system. "A monitoring and reporting system designed around these indicators would be unique in its focus on key aspects of teaching and learning and could enable education leaders, researchers, and policy makers to better understand and improve national, state, and local STEM education for all students" (National Research Council, 2013, p. 3). The call is for an indicator system to describe the implementation of new college and career readiness standards into daily classroom tasks (Committee on the Evaluation Framework for Successful K-12 STEM Education; National Research Council, 2013; Means, Mislevy, Smith, Peters, & Gerard, 2016). One of these indicators that was identified as a priority was Indicator #5: *Classroom coverage of content and practices in CCSS and NGSS.*

Such an indicator would examine teaching in a different way than previous research in teaching quality. There currently exists a body of work describing various types of evaluation of instruction, including large-scale indicators like NAEP. These evaluations often make use of student achievement measures, observational measures, and survey measures. Each of these can make a useful contribution to understanding what is happening in classrooms and the extent to which reforms are implemented in different ways. Achievement measures can provide information on what students have mastered, although, such measures can lag behind reform initiatives, particularly in historically "untested" subjects like science, (Buckendahl, Plake, Impara, & Irwin, 2000; Martone & Sireci, 2009). Observations can capture student-teacher interactions, even

when captured by videotape (Casabianca, McCaffrey, Gitomer, Bell, Hamre, & Pianta, 2013). When a teacher interacts with students, there is much unseen history that has passed between them before, and is informing this particular moment. For example, a long pause after a question could indicate that the teacher hasn't adequately prepared the class to answer, or that the class has internalized the expectation that they must give thoughtful responses, and will be given the space to adequately formulate these. An observer who is only seeing this moment may lack the context to properly interpret the exchange. Another often-used technique, self-report through survey, can be cost-effective, and can shed light onto teachers' perception of their practice. All of these methods have their affordances. The following studies explore the possibility of using classroom artifacts as a complementary measure.

Classroom artifacts, which can include both assigned tasks from teachers and the responding student work, are very useful in providing evidence about the nature of instruction available to students. Beginning with early portfolio assessments (Campbell, Kapinus, & Beatty, 1995; Koretz, Stecher, Klein, & McCaffrey, 1994; LeMahieu, Gitomer, & Eresh, 1995) to more recent attempts to develop protocols for assessment of teachers and learners (Borko, Stecher, & Kuffner, 2007; Matsumura & Pascal, 2003), these tangible traces of classroom interaction have been studied as useful measures of teaching quality and student work. Research has found that the intellectual demand of classroom artifacts can be measured reliably (Borko et al, 2005; Clare & Aschbacher 2001; Matsumura et al, 2008) and that demand is connected to student outcomes (Matsumura and Pascal, 2003; Mitchell et al, 2005; Newmann et al, 2001). In work looking at middle school mathematics and science classes, Borko, Stecher, and

colleagues found that a reliable view into classroom practices could be gained from an examination of teacher assignments and student work (Borko, Stecher, & Kuffner, 2007; Borko, Stecher, Alonzo, Moncure, & McClam, 2005).

The proposed research extends the body of work using artifacts in several critical ways. First, it provides a thematic synthesis of existing STEM artifact studies and how these have been designed. This will allow for the expansion of artifact research, so that future researchers can gain insight into lessons learn and consider how previous work has influenced findings and inferences made. This review will form the foundation for the second and third studies.

Because " (C)onstruction of knowledge, disciplinary inquiry and the audiences and purpose have meaning specific to respective disciplines,"(Alexander & Judy, 1988, p. 354) it is important to develop and test discipline specific protocols. Specifically, I will investigate the potential of extending the to-date uses of artifact study to a new use, as a component of an indicator system that has the capacity to describe the alignment of classroom work and assessments with the current college and career readiness standards. The second study explores the feasibility of developing a protocol in the domain of science for this purpose, which differs substantially from the previous work done with classroom artifacts that has focused on assigning scores to individuals and not on describing a status of a system. The third study is a related but distinct study of developing such a protocol for math.

The first study of the dissertation will involve a comprehensive review of existing frameworks for use with artifacts in order to develop key design characteristics in the development of domain-specific artifact protocols. Particular attention will be given to

sampling and rating designs. This synthesis will provide a framework for the design of the indicator protocol system, which will both build on previous work and extend artifact use in a novel direction. The synthesis will also provide a critical resource for future artifact research.

The second study will focus on design of a standards-based classroom artifact protocol in science, informed by findings from the first study. According to a recent RAND/Asia Society report, "public school systems are [increasingly] expected to promote a wide variety of skills and accomplishments in their students, including both academic achievement and the development of broader competencies…" which "… are seen as critical components of college and career readiness" (Soland, Hamilton, & Stecher, 2013, p. 1). These 21st century competencies are represented in new standards for science, such as the Next Generation Science Standards, and incorporate not only content, but also the development of practices needed for future success. This demand for a range of both content knowledge and process skills has led to challenges to stakeholders in terms of implementation and assessment.

In response to these challenges, there has been a call, beginning with the National Research Council in 2011, for an indicator system to both establish the current level of STEM teaching and learning, and to track progress. While "a single indicator can rarely provide useful information about such a complex phenomenon as schooling," a system of indicators can be used "…to characterize the nature of a system through its components--how they are related and how they change over time" (Shavelson, McDonnell, & Oakes, 1991). "A monitoring and reporting system designed around these indicators would be unique in its focus on key aspects of teaching and learning and could enable education

leaders, researchers, and policy makers to better understand and improve national, state, and local STEM education for all students" (National Research Council, 2013, p. 3).

One indicator that could be developed using artifacts would be a measure of "the extent to which the instruction and learning activities students experience in a classroom cover content in a set of standards, are consistent with the performance-level expectations of those standards, and reflect the same conception of learning and instruction… capturing the enacted curriculum" (Means, Mislevy, Smith, Peters, & Gerard, 2016, p. 24). Therefore, use of an artifact measure in science as a component of an indicator system, to assess and monitor progress toward the goals articulated in the college and career readiness standards at either a local or more broad level, will advance both theory and practice.

The study will design an instrument and conduct an initial pilot of such an indicator system component that makes use of classroom artifacts as the primary source of evidence. This study tests the hypothesis that classroom artifacts, such as tests, homework, lab reports, etc., can provide streamlined access and meaningful evidence of alignment. Artifacts may provide insights beyond those available from other measures. Surveys can do a reasonable job of capturing content and skill coverage, but give little insight into the quality and depth of that coverage. Teacher observation protocols give valuable but incomplete data about teacher practice and student interaction, and require substantial resources to implement on a large scale. Therefore, classroom work (e.g. quizzes, class work) can be a complementary tool in understanding how curriculum is translated *by* teachers *for* students, and how students then respond to these classroom demands. If we are to gain insight into the extent to which reform initiatives are being

represented to students , including contextual factors such as curriculum, it is important to investigate classroom artifacts. However, as existing artifact protocols were not designed to function as measures of standard alignment, but instead to make inferences about individuals, a new protocol that provides information to stakeholders on alignment to emerging standards, including both content and practices, would be useful.

Classroom assignments call on domain-specific knowledge and skills and thus, the majority of research into classroom artifacts has made use of domain-specific protocols. Study three explores the development of an artifact indicator protocol for math. While the purpose is again to assess alignment to college and career readiness standards, these are different for math (e.g. the Common Core), and therefore the instrument development and pilot process will require attention to these domain differences.

Collectively, the studies represent an effort to better understand the potential use of artifacts to provide insight into the implementation of new STEM standards within classrooms and focus on the following research questions:

Synthesis Study

- *What are the important design characteristics of artifact study that have been considered in the literature?*

Indicator Protocol Studies

- *To what extent can the protocol be used to measure classroom practice articulated in college and career readiness STEM standards?*
- *To what extent is the protocol sensitive characteristics that may be of interest to stakeholders?*

Answers to these questions will be helpful to the artifact research community as well as stakeholders in STEM education as we move toward these key reforms in classroom instruction that includes both content and practices.

References

Allen, C. D., & Penuel, W. R. (2015). Studying teachers' sensemaking to investigate teachers' responses to professional development focused on new standards. *Journal of Teacher Education*, *66*(2), 136-149.

Ananiadou, K., & Claro, M. (2009). 21st century skills and competences for new millennium learners in OECD countries.

Borko, H., Stecher, B., Alonzo, A., Moncure, S., & McClam, S. (2005). Artifact Packages for Characterizing Classroom Practice: A Pilot Study. *Educational Assessment , 10* (2), 73-104.

Borko, H., Stecher, B., & Kuffner, K. (2007). *Using artifacts to characterize reform-oriented instruction: The Scoop Notebook and rating guide (CSE Technical Report 707).* LA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/UCLA.

Buckendahl, C. W., Plake, B. S., Impara, J. C., & Irwin, P. M. (2000). Alignment of Standardized Achievement Tests To State Content Standards: A Comparison of Publishers' and Teachers' Perspectives.

Campbell, Jay R., Kapinus, Barbara, and Beatty, Alexandra. "Interviewing Children About Their Literacy Experiences: Data from NAEP's Integrated Reading Performance Record (IRPR) at Grade 4." ETS, 1995.

Capps, D. K., Shemwell, J. T., & Young, A. M. (2016). Over reported and misunderstood? A study of teachers' reported enactment and knowledge of inquiry-based science teaching. *International Journal of Science Education*, *38*(6), 934-959.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of Observation Mode on Measures of Secondary Mathematics Teaching. *Educational and Psychological Measurement*, *73*(5), 757-783. DOI: 10.1177/0013164413486987

Clare, L., & Aschbacher, P. (2001). Exploring the Technical Quality of Using Assignments and Student Work as Indicators of Classroom Practice. *Educational Assessment* , 39-59.

Committee on the Evaluation Framework for Successful K-12 STEM Education; Board on Science Education; Board on Testing and Assessment; Division of Behavioral and Social Sciences and Education; National Research Council. (2013). *Monitoring Progress Toward Successful K-12 STEM Education: A Nation Advancing?* Washington, D.C.: The National Academies Press.

Gilson L. (2015) Lipsky's Street Level Bureaucracy. Chapter in Page E., Lodge M and Balla S (eds) Oxford Handbook of the Classics of Public Policy. Oxford: Oxford University Press

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and Implications. *Educational Measurement: Issues and Practice*, *13*(3), 5-16.

Le, Vi-Nhuan, Brian M. Stecher, J. R. Lockwood, Laura S. Hamilton, Abby Robyn, Valerie L. Williams, Gery W. Ryan, Kerri A. Kerr, Jose Felipe Martinez and Stephen P. Klein. Does Reform-Oriented Teaching Make a Difference? The Relationship Between Teaching Practices and Achievement in Mathematics and

Science. Santa Monica, CA: RAND Corporation, 2006. http://www.rand.org/pubs/research_briefs/RB9211.html.

LeMahieu, Paul G.; Gitomer, Drew H.and Eresh, Jo Anne T. “Portfolios in Large-Scale Assessment: Difficult But Not Impossible.” *Educational Measurement: Issues and Practice* 14, no. 3 (1995): 11–28. doi:10.1111/j.1745-3992.1995.tb00863.x.

Machi, E. (2009). Improving US Competitiveness with K-12 STEM Education and Training. Heritage Special Report. SR-57. A Report on the STEM Education and National Security Conference, October 21-23, 2008. *Heritage Foundation*.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research*, *79*(4), 1332-1361.

Matsumura, L., & Pascal, J. (2003). *Teachers' assignments and student work: Opening a window on classroom practice.* Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Matsumura, L., Garnier, H., Slater, S., & Boston, M. (2008). Toward Measuring Instructional Interactions “At-Scale”. *Educational Assessment* , 267–300.

Means, B., Mislevy, J., Smith, T., Peters, V., & Gerard, S. (2016). *Measuring the Monitoring Progress K-12 STEM Education Indicators: A Road Map* . Washington, D.C.: SRI Education.

Mitchell, K., Shkolnik, J., Song, M., Uekawa, K., Murphy, R., Garet, M., et al. (2005). *Rigor, Relevance, and Results: The Quality of Teacher Assignments and Student Work in New and Conventional High Schools.* Washington, D.C.: American Institutes for Research.

National Research Council. (2013). *Monitoring Progress Toward Successful K-12 STEM Education: A Nation Advancing?*. National Academies Press.

Newmann, F., Bryk, A., & Nagaoka, J. (2001). *Authentic Intellectual Work and Standardized Tests: Conflict or Coexistence?* Chicago: Consortium on Chicago School Research.

Nord, C., Roey, S., Perkins, R., Lyons, M., Lemanski, N., Brown, J., & Schuknecht, J. (2011). The Nation's Report Card [TM]: America's High School Graduates. Results of the 2009 NAEP High School Transcript Study. NCES 2011-462. *National Center for Education Statistics*.

Shavelson, R., McDonnell, L., & Oakes, J. (1991). What are educational indicators and indicator systems? *Practical Assessment, Research & Evaluation*.

Soland, J., Hamilton, L., & Stecher, B. (2013). *Measuring 21st Century Competencies: Guidance for Educators.* Global Cities Education Network.

Using Classroom Artifacts to Investigate STEM Instruction:

A Design Synthesis Study

Jeanette Joyce

Rutgers University

October 2017

Abstract

There has been an ongoing search for ways to capture STEM instruction in order to better understand teaching and learning. Previously, this attempt has relied chiefly on observations. However, study of classroom artifacts (e.g. homework, lesson plans, assessments, projects) provides evidence of STEM instructional practice through written description and materials that may provide additional, complementary insights. This study sets out to thematically analyze the work that has been done to date and to develop an overarching design framework for considering artifact measures as indicators of STEM teaching quality. All existing artifact studies, although they vary considerably in the particulars, can be described through a framework of design decisions: purpose, construct, sampling, contextual support, scoring, and validation. This study represents important movement toward cohesion in artifact study in STEM and it is hoped that it will lead to further standardization in future work, with clear reporting of protocols used and reliabilities attained, so that ongoing research in this promising area can advance understanding of instructional practice.

Introduction

There has been an ongoing search for ways to capture instruction in order to better understand STEM teaching and learning. To date, methods have focused on classroom observations (e.g. Danielson's Framework for Teaching), but the challenge remains: How can we characterize STEM instruction as given by a teacher to a classroom and to what extent can we leverage the insights gained to support improvements? Classroom artifacts, which can include both assigned tasks from teachers and the responding student work, have been used as one way of providing evidence about the nature of instruction available to students. Classroom artifacts can include any captured evidence of tasks set for students, and may include homework, classwork, assessments, and lesson plans. Inspired by early portfolio assessments in which student work was collected for assessment (Campbell, Kapinus, & Beatty, 1995; Koretz, Stecher, Klein, & McCaffrey, 1994; LeMahieu, Gitomer, & Eresh, 1995), these more recent attempts have developed protocols and methodologies that use artifacts as tangible traces of classroom instruction for assessment of both teachers and learners (Borko, Stecher, & Kuffner, 2007; Matsumura & Pascal, 2003).

Previously, the attempt to capture classroom instruction has relied chiefly on observations. While observations may more directly capture teacher and student real-time interactions (Gitomer & Bell, 2013), artifact study provides evidence of the instructional practice through written description and materials that may provide additional, complementary insights. With this potential to study classroom artifacts as a window into instructional practices, this paper sets out to thematically analyze the work

that has been done to date and to develop an overarching design framework for considering artifact measures as indicators of teaching quality.

This design framework is organized by critical features that characterize study of artifact measures. I present a brief overview of features here, and then will describe each in further detail. Artifact studies involve a complex set of decisions that support the *purpose(s)* of the artifact study, whether that is to evaluate teachers, or to improve instructional practice and/or to improve associated student outcomes. There is also the question of what *construct* underlies the measure. That is, what conceptualization of teaching or learning is intended to be measured using artifacts? The first design feature considered is sampling. While it would be ideal to examine every piece of evidence available, it is not feasible to examine and score the complete population of artifacts across an entire academic year (i.e., all work in all classrooms). For this reason, artifact studies must make *evidence-sampling* decisions about the artifacts themselves in terms of what to collect, when to collect it, and how much evidence is sufficient to understand the construct of interest. Additionally, decisions are made in terms of the needed *contextual support* in order to understand the artifact, and then to adapt or develop a *scoring* system or rubric used to convert the artifacts into analyzable data. Collectively, these decisions will inform inferences about teachers, schools, programs, or systems, and require both analytic methods to interpret the results and *validation* including psychometric evidence and correlational evidence from other educational measures.

All existing STEM artifact studies, although they vary considerably in the particulars, can be described through this framework of design decisions: purpose, construct, sampling, contextual support, scoring, and validation. However, to date, there

has not been a comprehensive review of artifact study in education in terms of a principled design analysis. As the use of classroom artifacts continues to expand in STEM educational research, there is use for a synthesis of work done to inform emerging design choices. It is important to consider the variations of existing artifact studies, and how these may be of use to the development of future work.

Purpose of the Artifact Study

It has been established that in well-designed research, purpose should inform methodology. As Stodolsky found in her seminal review of classroom observation research, "…the very close connection between purposes, goals, and methods must be explored…"(p.177, 1990) in order to interpret findings. This is also the case with the studies of classroom artifacts. These "frozen moments" of interaction have been used by researchers for the purpose of understanding and evaluating teacher expectations, implementation of intended instruction, and student learning through applications of assignment protocols.

In studies of summative assessment, the level of inference is also reflected in the purpose. In some cases, the purpose of the study is to evaluate instructional practice at the *teacher* level (e.g. Martinez, Borko, & Stecher, 2012). In such a study, the artifacts are used to make inferences about the quality of teaching within a given classroom, much in the same way that observation protocols have been previously applied. Other studies have been concerned with making inferences at the *school* level (e.g. Clare, Valdes, Pascal, & Steinberg, 2001), in some cases comparing artifacts as instruction across schools to make judgements about overall quality. A third group of summative studies is more concerned with making inferences at the *program* level, which may include using

artifacts to assess alignment to standards (e.g. Berry & Ellis, 2013). Artifacts have been also used for studies with formative purposes, as a way to direct changes in *teaching* (e.g. Borko, Stecher, & Kuffner, 2007) or to gain insight into *students*' understanding (e.g. Klenowski, 2011).

In each of these cases, the purpose for use of artifacts varies, and that dictates differences in the methodologies, in terms of sampling, contextual support, and scoring. For example, when using artifacts as part of a potentially high stakes evaluation such as making decisions about an individual teacher, the researchers seek to make the sample as comprehensive and representative as possible of the teacher's work. However, when the purpose is more formative, then the researchers do not attempt to capture as complete a picture of the totality of the teacher's instructional practice, but instead focus on the target instructional behavior that is under scrutiny. That is, if examining artifacts to improve addressing math misconceptions about fractions, it wouldn't be necessary to collect all math artifacts beyond that unit.

Construct

Artifact study most often centers on the construct of "intellectual demand." This takes into account both the rigor and relevance of tasks set for learners and emerges from the work of Anderson, Krathwohl, and Bloom (2001), reflecting the idea that while deep learning is not observable, certain types of tasks set for and questions asked of students are more likely to elicit indications of this type of complex cognitive activity. Artifacts can then be analyzed for the presence of different elements and the resulting student work can be analyzed for evidence that deep learning/critical thinking has taken place. In each study the construct of intellectual demand is revisited and may be viewed from a new

perspective. For example, in the 2010 study by Ruiz-Primo, Li, Tsai, and Schneider, the construct of intellectual demand is focused in the area of scientific inquiry, and artifacts are used to assess student's ability to make a claim, support it with evidence, and use reasoning to link the claim and evidence. This type of argumentation is one area of complex cognition that falls within the broader construct of intellectual demand.

Evidence Sampling

As mentioned earlier, it would be unwieldy to collect every artifact produced over a school year, even in terms of a single classroom. Furthermore, each artifact or instructional task generates another larger pool of student work. The sampling unit in an artifact study may range from a single assignment or assessment to an entire unit of work, with or without student work. This variation in sampling unit reflects the purpose of the study and impacts any inferences made. Therefore, it is important to consider what each of the studies has considered as a single observation for data collection and scoring purposes. I attempt to describe the series of decisions each of the studies had to undergo in order to determine the most appropriate sampling unit for its purpose. These decisions range from who selects the artifact, what type of artifact, when the artifacts are sampled, and how many artifacts are to be collected.

The first aspect of sampling to consider is who makes the selection of artifacts. In several studies, the teachers were asked to select artifacts, with minimal guidelines such as "typical, everyday assignments," "best work," or "challenging task" (e.g. Campbell, Kapinus, & Beatty, 1995). In these studies, the researchers left the selection to the teacher in order to gain insight into his or her perceptions of the aspect of the intellectual demand construct under investigation. In another set of studies, teachers or other data

collectors were asked to choose artifacts aligned with a stricter set of criteria, such as examples of science inquiry (e.g. Ruiz-Primo et al, 2010). Alternatively, all artifacts associated with a given unit or time period can be collected (e.g. Borko, Stecher, Alonzo, Moncure, & McClam, 2005) in order to more fully characterize instructional practice for that topic or time period. Finally, there can be a more random collection (e.g. Koh & Luke, 2009), without the teachers' input, which eliminates selection bias but also limits potential for inferences about the teacher's perceptions.

A second sampling decision involves the actual type of artifact collected. Studies to date have included assigned homework, classwork, projects, assessments, and lesson plans. In some studies, a mix of these was collected, while others chose to focus on a single category (e.g. Herman et al, 2005). When researchers were looking at instructional interactions across classrooms, schools, or programs, the samples tended to include multiple "angles" of practice (e.g. Newmann, Bryk, & Nagaoka, 2001). The type of artifacts collected represented the breadth of the inferences that could be validly made. The more general inference to be made in terms of overall practice, the more variety would be needed in type of artifact

Breadth of inference also influenced the frequency of the collection**.** In order to inform more generalized claims about instructional quality, a more comprehensive collection is needed. That is, to characterize a teacher's practice, artifacts would have to be collected in sufficient number across the year that the sample could be said to be adequately representative.

There is a further consideration of being able to make a claim about the stability of the artifact rating. Reliability in scoring a teacher's performance is improved when

multiple artifacts over multiple days are collected from the teacher. Previous studies indicated that at least three assignments would need to be collected from teachers to determine a "stable estimate of quality"(Clare et al, 2001), and that four samples made sense for generalizability (Matsumura, 2008), while Martinez et al (2012) found for their purposes, five days of artifact collection were needed. However, at a certain number of samples, these studies found that reliability estimates begin to asymptote, and the cost in time and effort of collecting more than five artifacts or five days of artifacts per classroom may outweigh any gain in reliability.

Another related issue has to do with timing of collection. Similar to the need for multiple samples of artifacts, researchers have found that artifact quality differed at different times of the school year (Clare; Joyce, Gitomer, & Iaconangelo) indicating that it may be important to sample across the academic year if the purpose of the study extends beyond what can be inferred from sampling a single unit of time.

Contextual Support

In order to fully understand and make inferences about the teacher-student exchange represented by the artifact, all studies found it necessary to collect additional information. This information sheds light into instructional goals, influences that affected the teacher's choice of assignment, and important factors that affected how the assignment is implemented, and may include direct teacher input and demographic information.

Researchers must decide, for the specific purpose of the study, how valuable and how feasible is collecting additional data. This could include direct teacher input, such as grading and feedback to the learner. Some studies sought demographic information at the

teacher, class, or school level in order to better characterize the setting of the task. Additionally, information may be elicited about the pressures that exist at school, district, and state levels, and may include curricular policy, school policy, availability of resources, and publisher bias.

Scoring

The next design decision point is the mechanism for scoring or rating of the artifacts. As the data used for analysis are the actual scores on the artifacts, this becomes critical to the study design. Many studies have developed a protocol that measures the targeted aspect of the construct of "intellectual demand," and then sought to establish validity and reliability of the developed measure, while others adapted existing protocols for their specific study. These protocols create dimensions that are domain specific. That is, a separate protocol is developed for ELA and math (e.g. Matsumura et al, 2006). An additional element of scoring is the selection and training of the raters themselves. Studies recruit raters, either using a convenience sample (e.g. graduate students) or seeking out those who have content area and teaching experience. Using the target protocol, studies develop anchor papers (assignments that serve as exemplars of specific scoring levels), and provide some amount of training ranging from a single session to multiple sessions, including re-alignment during the scoring period. The reliability of the raters affects the confidence with which the studies can make inferential claims.

Validation

In any performance-based assessment, there are design considerations that will impact the validity of the measure and its use. Confirmatory evidence from comparing artifact scores with observations and standard test scores is often reported. Additionally,

components such as reliability, comparability, and generalization are considered. Before a measure can be evaluated for validity, it would need to be consistent. Rater reliability is often addressed through percent agreement, but measures such as Cohen's Kappa and Intraclass Correlations provide evidence that raters' agreement is not due to chance and that there is a correlation in scores between raters. These measures provide some evidence that ratings can be consistent across settings, which is a necessary precursor to establishing validity. Comparability addresses the balance of standardization and flexibility in assessment design as the research moves across classrooms, schools, and subjects. This has not been well addressed in artifact study. However, if each study develops its own protocol with its own rating system, without consideration for the existing systems, it will be increasingly difficult to make inferences across studies about the type of work students are being asked to do beyond a very local level. This relates directly to generalizability which examines how well the measure of the construct in the specific research setting serves to evaluate the construct overall.

The purpose of this paper is to consider the state of classroom artifact study and to explore the variation in both the design of artifact protocols themselves, and the studies that then use these to gain insight into teaching and learning. As this study grew out of the literature review for a study on use of artifacts in the evaluation of STEM teaching (Joyce & Gitomer, 2017), the focus is on math and science artifact studies, although some of these studies included other subject areas as well.

Review of Three Key Artifact Studies

As I develop and discuss the framework, it may be helpful to look to three key studies as examples of work that has been done to date. Here, studies are briefly

described and considered in terms of the elements: purpose, construct, contextual support, scoring, and validation.

Classroom Artifacts as Measures of Teaching Quality

Purpose and Construct. The Classroom Artifacts as Measures of Teaching Quality Study (Joyce, Gitomer, and Iaconangelo, 2017) examined the level of intellectual demand (**construct**) of both typical and challenging math and ELA assignments, and the resulting student work in order to better understand the use of artifacts as a window into classroom interactions in order to assess teaching quality (**purpose**).

Sample. The study focused on 47 middle schools across 3 districts in one large metropolitan area, and collected data on 225 math and 225 ELA teachers. As part of a multi-disciplinary examination of measures of teaching quality (UTQ), teachers were asked to supply 6 assignments across the school year. Understanding that teachers give some assignments that are more routine and others that they think of as being more ambitious, the study asked teachers to submit both their typical assignments and those that they considered to be challenging. Teachers had latitude in determining what constituted "typical" and "challenging". While a *typical* assignment was described as "everyday work", a *challenging* one was described to teachers as "an assignment that gives you the best sense of how well your students are learning a subject or skill at their highest level. In this study, the **sample** was 6 artifacts, selected by the teacher using the criteria of 2 challenging and 4 typical from 6th to 8th grade classrooms. Work was collected in two separate visits, roughly categorized as fall and spring.

Contextual Support. Substantial context was available for this study, including 10 randomly selected samples of student work for each of the challenging tasks, which

were scored separately. Additionally, for each classroom, there was demographic information of prior achievement, percent eligible for Free-Reduced Price Lunch, percent Students with Disabilities, percent Gifted, and percent English Language Learners. Teachers did not provide any information beyond the rating of the student work as high, medium, and low, but information was available about teachers, including observation scores, test results of pedagogical and content knowledge, and value added measures.

Scoring. In order to evaluate the collected artifacts, the protocol from the Chicago Annenberg study was adapted. In 1996, Newmann, Bryk, and Lopez looked at classroom assignments in terms of authentic use of information to create new knowledge and to successfully communicate this knowledge. As a component of the Chicago Annenberg Research Project, Wenzel, Nagaoka, Morris, Billings, and Fendt developed an "Intellectual Demand Protocol" to rate separate dimensions of intellectual demand in assignments. The IDAP protocol considers intellectual demand in terms of 3 dimensions for math: communication, conceptual understanding, and real world connection. Teacher assignments and student work are each evaluated separately, and the range of scores may be different from dimension to dimension. A brief description of the math scales for assignments and student work are presented in Table 1.[1]

Validation. Dimensions were correlated to show internal consistency. Raters' scores were adjusted using the Multi Faceted Rasch Model to control for rater and scale severity and these adjusted scores were used for analyses. No comparisons were made between the math and ELA disciplines, but classrooms and teachers were compared based on the assumption that a teacher-determined artifact was indicative of the

[1] English language arts scales are available in Joyce, Gitomer, and Iaconangelo, 2017.

instruction delivered and any differences were related to teaching. Scores were correlated to teacher measures and variance was analyzed by demographic data.

TABLE 1
IDAP Math Assignment/Student Work Scale Dimensions
(*scale range*)

Scale	Teacher Assignment	Student Work
1	Written Communication *(1-3)*	Written Communication *(1-3)*
2	Conceptual Understanding *(1-4)*	Conceptual Understanding *(1-4)*
3	Relevant Context/ Real World Connection *(1-4)*	Reasoning *(1-4)*

SCOOP

Purpose and Construct. In a subsequent study, using a different approach, Borko, Stecher, and Kuffner (2007) developed the SCOOP protocol in order to use artifacts to characterize math and science classroom instruction. The study focused on the practical issue of having teachers look at classroom artifacts in a formative way, in order to aid in assessing both the students' understanding as well as the teachers' own process (**purpose**), and to explore the capacity of classroom artifacts as an indicator of reform-based instructional practices, such as cognitive depth (**construct).**

Sampling and Contextual Support. For this particular purpose, the data collection involved "scooping" up all instructional material and incorporating significant

reflective input from the teachers. Teachers were asked to provide lesson plans, handouts, coring rubrics, captured images of writing on the board or overheads, three samples of student work, homework, and projects for each task, rated high, medium, or low, and finally one "typical" assessment. They also provided answers to three different sets of reflective questions about context, lesson format, and strategies, before, immediately following, and at the conclusion of the week. One week's annotated artifacts, accompanied by teacher reflections and classroom photos were collected for evaluation.

In a follow-up study (Martinez, Borko, & Stecher, , 2012 used the SCOOP for making inferences about instructional practice in middle school science classrooms. Two field studies, one in California and one in Colorado, using in total 49 teachers from 25 schools, were conducted. Teachers were asked to complete SCOOP for typical classrooms

Scoring. As part of the pilot, a framework called SCOOP (Borko, Stecher, Alonzo, Moncure, & McClam, 2005) was developed with clearly defined steps for collecting, labeling, and reflecting on targeted student artifacts. Once collected, "...independent raters looked at 10 dimensions of the assignment [using] National Science Education Standards (NRC, 1996) and Principles and Standards for School Mathematics (NCTM, 2007) as a basis for identifying 10 dimensions of reform-oriented instructional practice in each content area,"(p.7). Each dimension was rated as high (5), medium (3), or low (1). A brief description of dimensions is presented in Table 2. Scores were based on artifacts, classroom observations, and teacher reflection. Dimension 6, "Discourse Community," was scored from classroom observation only and is not included here. In

the Martinez et al study, raters were again given anchor papers for the 1,3, and 5 score points, but were allowed to assign scores of 2 or 4 as well. The aggregated score was a holistic judgment on the part of the raters.

TABLE 2
SCOOP Math/Science Artifact Rating Dimensions

Dimension	Math	Science
1	Grouping (task was collaborative)	Grouping
2	Structure (task built on prior activities)	Structure
3	Multiple Representations	Hands on
4	Use of Math Tools	Use of Scientific Resources
5	Cognitive Depth*	Cognitive Depth
7	Explanation and Justification*	Explanation and Justification
8	Problem-solving*	Inquiry
9	Assessment	Assessment
10	Connections	Applications

Validation: Initially, the researchers found that the SCOOP system showed too much variability to be used for any high stakes teacher evaluation, but felt that the system represented a step toward a more complete description of instructional practice.

In the follow up study, Martinez, Borko, and Stecher (2012) focused specifically on validity and reliability issues. Scoop notebook inter-rater agreement ranged from 22% to

47% and the researchers described a need for clearer rating guidelines and further rater training. However, within one point agreement improved reliability percentages, and overall ratings showed stronger agreement than individual dimensions. Most variance between teachers reflected true differences between teachers but 50% of variance remained unexplained. The authors found that 3 raters and 5 days of data collection gave the best ratings. They also suggest condensing dimensions into instructional factors to improve measurement. For Science, it was suggested that an arithmetic average of factors or dimensions could give a better overall picture than a rater generated holistic score. This study collected teacher commentary on artifacts that was judged as helpful to raters on scoring, while photos, and before and after reflection were judged as less helpful.

Quality Assessment in Science

Purpose and Construct. In a similar study to the SCOOP notebook, Martinez, Borko, Stecher, Luskin, and Kloser (2012) conducted a pilot validation study of the Quality Assessment in Science notebook, "a portfolio-like instrument for measuring teacher assessment practices in middle school science classrooms" (p. 107) in order to gain insight into effective instruction in terms of quality assessment practice (**construct**). The purpose for this particular study was to develop a validation framework in order to support further applications of the designed instrument **(purpose).**

Sampling and Contextual Support. The team collected assessment artifacts from 42 8th grade science teachers twice yearly across a single state. These teachers were asked to collect assessment materials over a period of two instructional weeks (10 days) with accompanying annotations and reflections similar to the SCOOP study. The

students in these classes were asked to complete a survey and their end of the year test scores were collected. Additional information was collected in terms of classroom demographics similar to what was collected in the IDAP study.

Scoring. Eleven experienced 8th grade science teachers were recruited and trained and scored on nine dimensions with the tenth scoring dimension being a holistic rating. Dimensions are listed in Table 3. Each notebook was rated on each dimension from one (not present or realized) to five (fully present or realized), and then assigned an overall score based on the rater's general impression of the notebook's overall alignment with quality assessment practice.

TABLE 3
Quality Assessment in Science Notebook Dimensions

Dimension Description
Setting of clear learning goals
Frequency of assessment
Variety of assessment
Alignment of assessments to learning goals
Cognitive complexity
Scientific explanation/justification
Student involvement in self-assessment
Use of information for feedback to students
Use of information for instruction decisions

Validation. The authors found the QAS notebook to be a valid instrument in understanding teachers' assessment practices and to have predictive value in estimating

student achievement. They conducted generalizability and decision study analyses. Confirmatory evidence was sought through correlations and mean comparisons. Findings indicated that much of the variance in ratings remained unexplained by differences in teachers or raters, and that one dimension, "Alignment of Assessment and Learning Goals" had lower rater agreement than the others. Rating reliability was improved by adding a third rater. Moderate correlations with other assessment data provide evidence of validity.

These three key studies help establish the framework for analysis of artifact research. Each had a specific purpose, and bounded a slightly different aspect of the intellectual demand construct. Sampling was comparable in that the work was collected from multiple teachers in middle school classrooms, but varied by the amount, type, contextual support, and frequency. Scoring either involved developing an instrument or adapting one for a new purpose. Finally, validation of the study and its instrument is addressed somewhat differently in all three.

Method

This review of existing protocols and artifact research uses a thematic synthesis approach (Petticrew & Roberts, 2006; Thomas &Harden, 2008). In thematic synthesis, key themes are identified, and studies are reviewed in order to fully develop the phenomena of interest. Studies were collected from available relevant peer-reviewed literature or equivalent[2], both in the US and outside, and authors were contacted if the protocol used for scoring was not available in the literature. The search was conducted between 2013-2015. The studies considered were conducted in the US, Australia, and

[2] Some studies were published as institutionally reviewed reports and the author's previous study is still in review for publication at this time but is included in analysis

Singapore and reported between 1995-2014. There was considerable difference noted in study criteria among the studies. The sole criterion for inclusion in the study was the use of classroom artifacts in math and/or science as a major data source. In total, 22 studies were investigated in order to better understand the purpose, sampling, and context of the study collection, as well as the protocol developed or used by the study for scoring. These are summarized in Table 4. The validity considerations of the study designs will also be examined and summarized. The overarching research question for this study is:

- *What are the important design characteristics of artifact study that have been considered in the literature.*

TABLE 4

STEM Artifact Studies

Study Protocol	Purpose	Sample	Frequency of Collection	Context	Scoring
1. Berry & Ellis (2012)	Summative	All	3x over course of study	Student work, Demographic, Observations	M-Scan (standards, SCOOP)
2. Borko et al (2005-7)	Formative, Student Understanding	All	1x over course of study, for a period of one week	Student work, Observations, Extended notations	SCOOP
3. Campbell et al (1995)	Summative	1-3 artifacts, Student/ teacher selected	1x over course of study	Student work Survey, Interviews	Study-designed
4. Castillo & Foley (2015)	Formative	All	1x over course of study	Demographic, Observations	Study-designed (NGSS)
5. Clare et al/ Matsumura et al (2001-8)	Summative	2-4 artifacts, typical and challenging, teacher selected,	2x over course of study	Student work, Demographics, Observations, Brief teacher comments, Interviews	Study-designed (IDAP)

TABLE 4

Artifact Studies (continued)

Study Protocol	Purpose	Sample	Frequency of Collection	Context	Scoring
6. Gentile et al (1995)	Summative	3 artifacts	1x over course of study	Student work, Demographics, Brief teacher comments, Student letter	Study-designed
7. Goldsmith & Seago (2013)	Formative	All	1x over course of study	Observations	Study-designed
8. Grant & Branch (2005)	Student Understanding	All	1x over course of study	Observations, Interviews	Study-designed
9. Herman et al (2005)	Summative	Year End Assessment	1x over course of study	None	Study-designed (Standards)

TABLE 4

Artifact Studies (continued)

Study Protocol	Purpose	Sample	Frequency of Collection	Context	Scoring
10. Joyce et al (2017)	Summative, Student Understanding	6 (4 typical/ 2 challenging), teacher selected	2x over course of study	Student work, Demographics, Observations	IDAP
11. Klenowski (2011)	Student Understanding	Assessment, Student Work	1x over course of study	None	Study-designed (Standards)
12. Koh & Luke (2009)	Summative	Random, 4 samples	Multiple times across the year	Student Work	Study-designed
13. Little et al (2003)	Formative	No information, Assignments and lesson plans	multiple times across the year	Student Work	Study-designed
14. Martinez et al (2012)	Summative	All Assessments	1x over course of study, for a period of 10 days	Student Work, Demographics, Annotations, Reflections, Surveys, Achievement	Study-designed (SCOOP)

TABLE 4

Artifact Studies (continued)

Study Protocol	Purpose	Sample	Frequency of Collection	Context	Scoring
15. Merritt et al (2010)	Formative	All	1x over course of study	Student work, Observations	M-Scan (SCOOP)
16. Mitchell et al (2005)	Summative	4 typical, 4 challenging assignments or assessments, teacher selected,	8x over study	Student work, Demographics, Observations, Achievement	Study-designed, (IDAP)
17. Morris & Hiebert (2011)	Formative	Lesson plans, teacher selected	1x over course of study	None	Study-designed
18. Newmann et al (2001)	Student Understanding	6 (4 typical/ 2 challenging), teacher selected,	Across the school year	Student work, Demographics, Achievement	IDAP
19. Ruiz-Primo et al (2002, 2010)	Formative/ Student Understanding	Lab notebooks	1x over course of study	Student work, Teacher feedback	Study-designed, (standards)
20. Shear et al. (2008)	Summative	1 assignment or assessment, teacher selected	8x over course of	Student work Demographics, Observations, Achievement	Study-designed (IDAP)

TABLE 4

Artifact Studies (continued)

Study Protocol	Purpose	Sample	Frequency of Collection	Context	Scoring
21. Silver et al. (2009)	Summative	2 assignments, 1 assessment, Randomly selected from best work portfolios (teacher selected)	Across the school year	None	Study - designed
22. University of Queensland (2002)	Summative	Longitudinal	across 3 years	Student work, Demographics, Observations, Interviews	IDAP

Results

Purpose

The purpose for studying artifacts was determined from the statements made in the rationale or purpose section of each study. These were classified as *summative*: in order to make a judgment of STEM teaching quality; *formative*: in order to make plans for improving STEM teaching practice; or *student understanding*: in order to make a judgment about the status of student mastery of STEM concepts. Overall, 12 of the 22 studies indicated that the purpose was summative, using language like "estimate of [instructional] quality" (Clare, Valdes, Pascal, & Steinberg, 2001, p. 4), "effective use of assessment by teachers as a critical component of quality instruction" (Martinez et al, 2012, p. 108), and "gauge the rigor of teacher assignments" (Mitchell et al, 2005, p. 2). Seven of the studies indicated a more formative intention, derived from phrases that included exploring "the extent to which teachers are using practices that are broadly endorsed in the reform literature" (Borko et al, 2005 p.77), "professional development…centered on exploration of classroom artifacts (Goldsmith & Seago, 2013, p. 1), and "…strive to improve [instructional] practice (Merritt et al, 2010, p.239). Finally, six of the studies were specifically to make an inference about student understanding as indicated in their purpose as using artifacts as "a natural strategy …to monitor students' progress" (Ruiz-Primo, Li, & Shavelson, 2002, p.2), linking high quality artifacts to "greater than average gains on [standardized testing]" (Newmann, Bryk, & Nagaoka, p. 2, 2001), and improving teachers' ability to "allow students to demonstrate their best work" (Klenowski, 2011, p. 12). This last study represents a dual purpose: both *formative* to change teaching practices and *student understanding* to make

judgments about students. As can seen by the summary of coding in Table 4, there was more than one study that was coded as having a dual purpose.

Within this framework, I return to the three exemplar studies. The Classroom Artifacts study had a *summative* purpose, to make inferences about teaching quality and the quality of resulting student work. The purpose led to design of a study that strove to be comprehensive and distributed across the year, and needed the corroboration of other validated measures of teaching quality. This is different from the purpose and design of the SCOOP study, which is *formative*, seeking to work with teachers to improve instruction. To achieve this goal, the design was comprehensive but narrowed in focus, and required more input from the teachers themselves in order to identify and target change. Finally, the QAS study adapts the SCOOP protocol to return to a more summative purpose, adding the investigation of alignment to standards that capture intellectual demand to its purpose. Here, a more distributed collection is again needed, but the component of extensive participation of the teacher remains. In all of the studies, there is some interest in *student understanding* through the analysis of student work, but the primary reason for including student work in these studies was to add context to understanding the teaching tasks and outcomes.

Construct

As discussed earlier, the underlying construct of teaching is related to the proposed conceptual lens of intellectual demand. This grows out of Archibald & Newmann (1988), who proposed a framework of cognitive complexity and social or personal relevance. Newmann, Bryk and Nagaoka (2001) describe authentic intellectual work as having three distinctive characteristics. First, it involves the *construction of*

knowledge, arguing that authentic work requires one to go beyond routine use of information and skills previously learned. Problem solvers must construct knowledge that involves "organizing, interpreting, evaluating, or synthesizing prior knowledge to solve new problems (p. 14)." The second characteristic of authentic intellectual work is *disciplined inquiry,* which involves the use of prior knowledge in a field, in-depth understanding, and elaborated communication. The final characterizing feature is *value beyond school*, the idea that work that people do authentically is intended to impact or influence others. While all studies considered here begin with this framing, the particular aspect of intellectual demand that is emphasized differs.

The construct was determined from the examination of the rationale or theoretical framework provided by the authors. The most prevalent characterization of the construct under scrutiny was "reform-based teaching." However, most papers stopped there, without further explanation of what the intent of the reforms was. From examination of the references cited, it can be inferred that these reforms are the ones discussed by Newmann et al above, with movement toward deeper, more authentic learning. By discipline, there were important differences. Science studies were more likely to cite inquiry-based tasks as critical to reforms, while math studies were more likely to mention problem-based instructional tasks.

In returning to our exemplar tasks, the Classroom Artifacts study clearly articulates a connection to Newmann et al's work and the scoring protocol is based on the three elements of intellectual demand described above. The SCOOP study follows the pattern previously stated, indicating that the underlying construct is focused on reform initiatives and, particularly, the value of inquiry-based learning. The QAS study looks

for evidence of cognitive demand through explanation and argumentation, both aspects of the intellectual demand construct viewed through the disciplinary specific lens of science.

Evidence Sampling

Four different sampling criteria were examined across the studies: number of artifacts sampled, type of artifacts sampled, frequency of sampling, and context sampled. Overall, it was found that more information is needed in this area, and that this is perhaps one area that could be somewhat standardized, which will be discussed in further detail.

Number of Artifacts. The number of artifacts collected varied widely across studies, ranging from all work produced in the classroom for a period of two weeks (e.g. Borko et al) to a single artifact (e.g. Shear et al). In most cases, for most purposes, all artifacts were collected for a designated period of time. However, that meant the number could vary from classroom to classroom within a study, as well as across studies.

Type of Artifacts. There were three studies (Herman, Klenowski, & Martinez et al) that focused solely on assessment type artifacts, while the rest collected a mix of assignments and assessments. Others have compared challenging to typical (e.g. Matsumura, Joyce), but there is a lack of empirical evidence to clarify any potentially significant differences by type of artifact collected.

Frequency of Sampling. The period of time over which the artifacts were collected also varied from a single visit to two full weeks of study, causing the number of artifacts to range widely. Another source of variation in total number of artifacts collected was the number of times artifacts were collected over a school year. Here, too, there was a range from a single visit to collection across the entire school year, with no clear association between articulated purpose and number of collection visits.

Context. In terms of the context collected along with the artifacts, while there was again no clear association with purpose, there were some interesting trends noted. In almost all cases, student work was collected to better interpret the assignment. However, there was no clear pattern as to how many samples of student work (ranging from 1 to 10) and as to whether these needed to be graded (A-F), rated (High-Medium-Low) or annotated. Another critical difference was whether the student work was scored separately against the construct (Joyce et al) or considered as part of the artifact package (Borko et al).

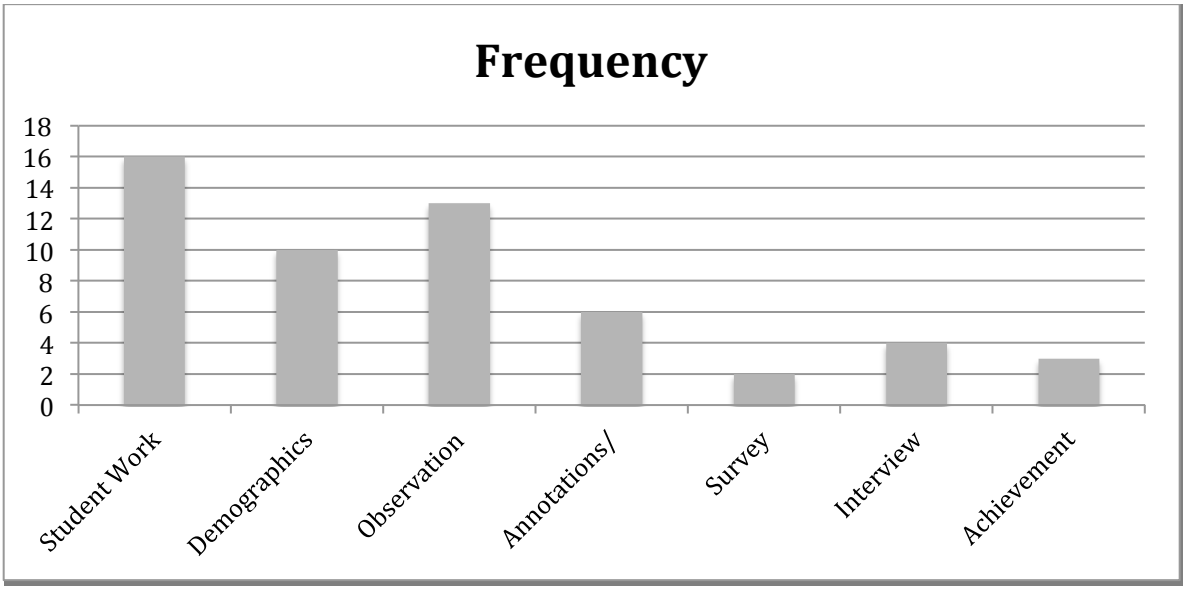


Figure 1. Summary by type of contextual data collected.

The second most frequently collected contextual data were observations of the teachers. These were most often used in order to validate the artifact ratings as a measure of teaching practice, and not to better interpret the artifacts. Alternatively, the frequent collection of demographic data does appear to add important context to the interpretation. That is, studies found it important to understand the make-up of the classroom in which the artifacts were situated in order to understand the task. Annotations, reflections and survey data were more often associated with formative tasks, in which it was deemed

important to gain access to the teacher's thinking. Finally, a few studies accessed student achievement data in order to validate the instrument through correlation. Frequency by type of context is summarized below in Figure 1.

In the three exemplar studies, we note key differences. For the Classroom Artifact study, there was a targeted sampling approach, across two sections attributed to the same teacher twice over the school year. The artifacts were classified by the teacher as either typical or challenging, and there was substantial contextual data collected. Ten samples of student work were randomly sampled and scored separately, and then used to support the claim that tasks with increased intellectual demand were associated with more in-depth student work. Teacher demographic information was provided but found not be associated with task ratings. Observations were conducted using multiple protocols. Although these observations were not specifically conducted for the lessons involving collected artifacts, the observation results were found to be aligned with findings of teaching quality based on artifacts, so, similar to student work, used as a confirmatory measure of validity. This trend was particularly strong for aspects of the observation protocols directly related to intellectual demand, such *as High cognitive demand, Use of representations/models, Student providing explanations,* and *Students making conjectures* (Joyce, Gitomer, & Iaconangelo, 2014). Extensive teacher annotations were not collected, which is not inconsistent with the summative purpose of the study. In this study, the artifacts were considered as a measure of teaching quality, and it is assumed that the extensive teacher annotations might have confounded the ratings of what the nature of the task. Aggregated achievement data was also used to examine potential differential access to demanding tasks based on prior performance. In sum, the

contextual information did not inform the artifact ratings but was used to determine whether significant differences by these criteria existed and to discuss possible implications of differing task quality for specific subgroups.

Both the SCOOP and QAS collected all artifacts of interest over a 10 day instructional period. For SCOOP, it was all material, and for QAS, assessment artifacts only. The collection was repeated twice over the academic year. Both included rated student work samples, with the SCOOP collecting three per artifact (hi-med-low) and the QAS collecting two per artifact (hi-low). No teacher classification (for example, "challenging") of the artifact was requested, but extensive teacher annotations were collected, and for the QAS, these were augmented by teacher and student surveys about assessment practices. The materials collected were chiefly used to understand teacher's perceptions about assessment, to encourage reflective practice, and to serve as an indication of any changes that occurred over the course of the study. Additionally, the QAS study collected student achievement data in order to explore the potential for artifacts to predict student scores.

Scoring

All studies either developed or adapted a protocol for scoring artifacts on the construct of interest. The majority of artifacts based their rating system on the Intellectual Demand of Artifacts Protocol developed by Wenzel (2002), the SCOOP protocol developed by Borko et al (2005), or National standards as existed at the time of the study. There are a few notable exceptions: the two studies using NAEP data (Campbell et al, 1995; Gentile et al, 1995) developed their own definitions and protocols based on the experts included in the research team, drawing from ETS research base.

Three studies (Grant & Branch, 2005; Little et al, 2003; Morris & Hiebert, 2011) developed unique protocols that were only descriptive coding, without use of quality ratings. Finally, two studies, Silver et al (2009) and Goldsmith & Seago (2013), developed unique protocols that invoked the same frameworks that underlie the IDAP, including Anderson and Krawthol's reworking of Bloom's taxonomy and other work in cognitive demand. That makes these study protocols adjacent to the IDAP in their development. Distribution of studies by study numbers from Table 4 is summarized in figure 2 below.

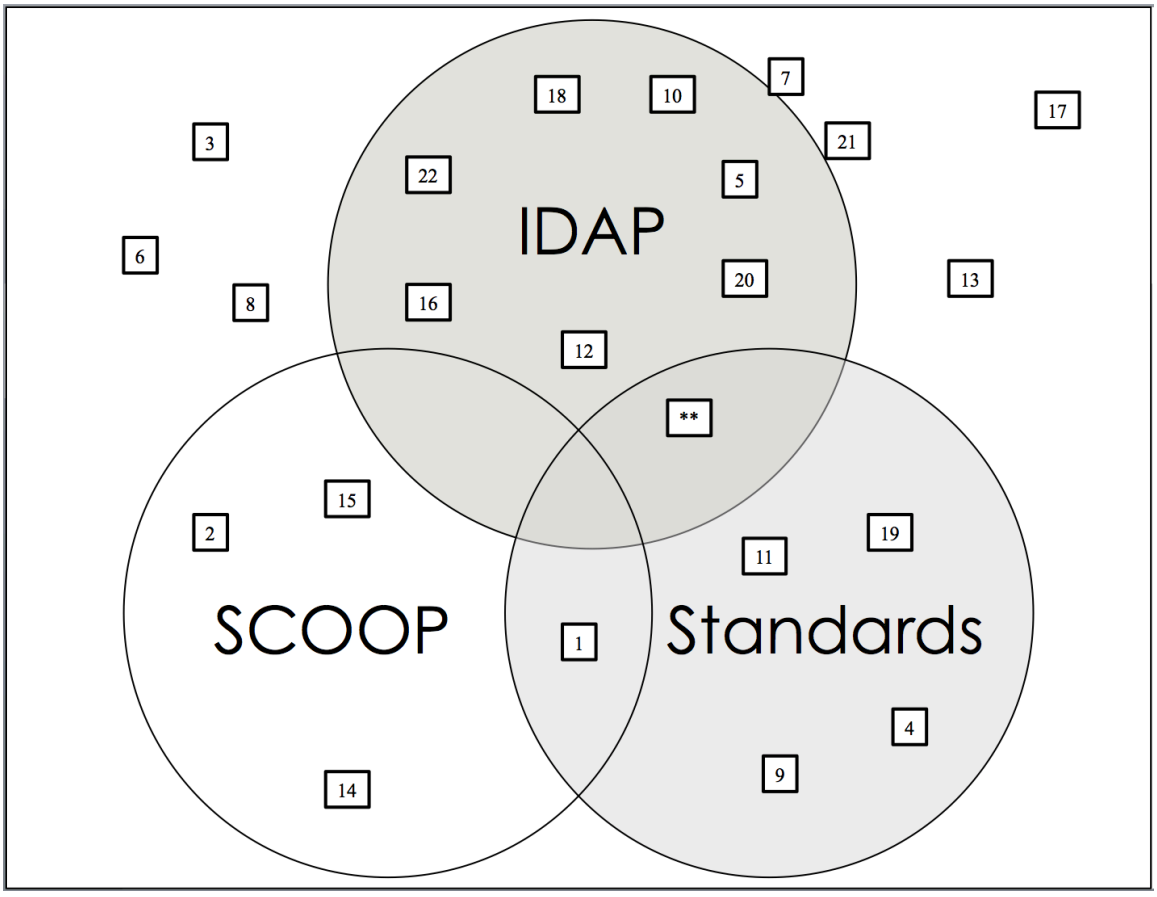


Figure 2. Scoring protocols adapted for study use. Numbers refer to the studies in Table 4. (Double dots (••) indicate current study from which this synthesis arose)

It is of interest that the three key studies focus on the two primary protocols: the IDAP and the SCOOP, which makes them useful as exemplars. The SCOOP protocol is to an extent a re-working of intellectual demand influenced by the NRC and NCTM standards, pre-cursors to the current Next Generation Science and Common Core State Standards. The QAS then follows the SCOOP, with some alterations that are appropriate to assessment practices.

Validation. All studies made some claim as to the validity of using the selected protocol for their purpose. However, not all studies provided evidence of underlying psychometric aspects of validity. These would be reported measures of *reliability, generalizability* and *comparability*. In terms of *reliability*, the majority of studies either used a single rater or reported "acceptable" reliability without empirical evidence. Eight studies did report percentage rater agreement, ranging from exact agreements as low as 50% (Borko et al, overall rating) to 92% (Ruiz-Primo et al, single dimension). However, the higher agreements were associated with dichotomous, descriptive ratings, such as presence or absence of a determined criterion, rather than more subjective ratings of a defined construct on an ordinal scale. Several studies set reliability goals at 65-70%, and did not allow raters to begin scoring until this benchmark had been met during training. Pertinent to overall reliability, studies also addressed re-alignment trainings as critical, as drift in ratings was noted over time.

There were even fewer mentions of *generalizability*. This is the assumption that the sample is representative enough of the population that an inference is justified, and can be expressed as a correlation or the variance explained by the artifact scores as opposed to random or rater effects (ANOVA). Generalizability with multiple raters has

been found to be as high as .98 (Herman, using 6 raters) or somewhat lower (.77) with 2 raters but multiple artifacts (Clare et al, using 4 artifacts).

The question of *comparability* remains unanswered with multiple protocols in existence. Can ratings from one study using one protocol be compared with findings from another study using another protocol? No metaanalysis exists to date that would shed light on that issue, but a current study is ongoing to compare rating the same set of artifacts with different protocols (Joyce, Zisk, and Gitomer, 2017).

An additional question exists in terms of extending a protocol's use or comparing findings across subject domains. That is, if a protocol yields a rating of 2 for a math artifact, how comparable is that with a rating of 2 on a science artifact using a related but not identical protocol? Most studies that covered multiple domains were careful to point out that while trends can be compared, scores should not.

Discussion and Conclusion

In the studies considered, it becomes clear that STEM artifact study is flexible enough to be useful for a variety of purposes. I found studies that were summative, formative, and based on student understanding, and several of the studies used the same base of data for multiple purposes (e.g. summative and student understanding in the Classroom Artifacts study). This flexibility is one of the greatest affordances of artifact study. However, as purpose guides design and analysis choices, it is important for further researchers to clearly define and articulate the purpose to which artifacts will be employed. Additionally, there is no evidence to date that artifacts have been used as an indicator of either a local or large-scale system, which would represent a significant

extension of purpose for artifacts. Further research is needed to clarify whether there is usefulness in engaging with artifact study for such a purpose.

Key design criteria that emerge from this study are that multiple artifacts from multiple times of year are essential to a stable rating, regardless of purpose. However, there is as yet no standardization in what defines "multiple." Studies on reliability have indicated that scoring 4 artifacts with 2 raters leads to a stable rating. Additionally, rater training is critical, with re-alignment, in order to make any purposeful inference from scores. This is particularly important when a high stakes decision may be made from the scores, such as characterizing the instruction by a particular teacher. Although many studies collected a multitude of contextual information that was useful to a certain extent in the study, the only context that was deemed essential to actually rating the artifact was student work, which varied from 3 to 10 samples. It seems that it is not possible to make rating judgments without this information about how the task is completed. Observation scores were also used in most studies, but as confirmatory evidence of validity rather than to influence scoring of the artifacts themselves.

An important finding in this study is that demographic information is frequently considered by researchers to be important to collect. As we work to better understand gaps in achievement and equity in educational opportunity, we will need further data as to the type of work that is being provided to different demographic subgroups, which requires information about the setting in which the artifacts are used. Also, the acknowledgement that the locus of control in setting tasks for students may rest beyond the teacher (in resources available, or departmental mandates) is critical before making inferences about a specific teacher's ability.

There appears to be some progress toward standardization of artifact study in terms of rating frameworks applied. Although studies refine scoring frameworks for their own specific interests, most researchers are able to draw on established frameworks to some extent, increasing *comparability* and potentially reliability of findings. At this point in time, general trends should be comparable across studies that have used different protocols That is, one study that finds a lack of intellectual demand could lend credibility to similar findings in another domain or using a different protocol. However, the purposes of the two studies should be aligned. Ongoing research as to comparability should contribute further to cohesion in the field of artifact study.

This study represents a first attempt to look across STEM artifact studies. A certain amount of inference was required in order to code for purpose, and the differences in reporting, for example correlation or ANOVA for generalizability, may have limited the usefulness of comparison. However, this study represents important movement toward cohesion in artifact study and it is hoped that it will lead to further standardization in future work, with clear reporting of protocols used, reliabilities attained, so that ongoing research in this promising area can advance understanding of instructional practice.

References

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Allyn & Bacon.

Archibald, D. A., & Newman, F. M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in secondary schools.* Washington, DC: National Association of Secondary School Principals.

Berry, III, R. Q., Rimm-Kaufman, S. E., Ottmar, E. M., Walkowiak, T. A., & Merritt, E. (2012). *The Mathematics Scan (M-Scan): A measure of standards-based mathematics teaching practices*. Unpublished, University of Virginia.

Borko, H., Stecher, B., & Kuffner, K. (2007). *Using artifacts to characterize reform-oriented instruction: The Scoop Notebook and rating guide (CSE Technical Report 707)*. LA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/UCLA.

Borko, H., Stecher, B., Alonzo, A., Moncure, S., & McClam, S. (2005). Artifact Packages for Characterizing Classroom Practice: A Pilot Study. *Educational Assessment , 10* (2), 73-104.

Campbell, Jay R., Kapinus, Barbara, and Beatty, Alexandra. "Interviewing Children About Their Literacy Experiences: Data from NAEP's Integrated Reading Performance Record (IRPR) at Grade 4." ETS, 1995.

Castillo, K., & Foley, B. (2015). Getting from worksheet science to the NGSS (Next Generation Science Standards): Facilitating collaborative data analysis with online tools. Presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Clare, L., & Aschbacher, P. (2001). Exploring the Technical Quality of Using Assignments and Student Work as Indicators of Classroom Practice. *Educational Assessment* , 39-59.

Clare, L., Valdes, R., Pascal, J., & Steinberg, J. (2001). *Teachers' Assignments as Indicators of Instructional Quality in Elementary Schools.* Los Angeles: CRESST.

Danielson, C. (2013). *2013 framework for teaching evaluation instrument*. Princeton, NJ: The Danielson Group. Retrieved from http://www.danielsongroup.org/framework/

Gentile, C. A., Martin-Rehrmann, J., & Kennedy, J.H (1995). *Windows into the classroom: NAEP's 1992 writing portfolio study*. Washington, DC: Office of Educational Research and Improvement, U.S. Dept. of Education.

Gitomer, D., & Bell, C. (2013). Evaluating Teaching and Teachers. In K. Geisinger, *APA Handbook of Testing and Assessment in Psychology: Volume 3 Testing and Assessment in School Psychology and Education* (pp. 415-444). American Psychological Association.

Goldsmith, L. T., & Seago, N. (2013). *Examining Mathematics Practice through Classroom Artifacts*. USA: Pearson.

Grant, M. M., & Branch, R. M. (2005). Project-based learning in a middle school: Tracing abilities through the artifacts of learning. *Journal of Research on*

Technology in Education, *38*(1), 65–98. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/15391523.2005.10782450

Herman, J. L., Webb, N. M., & Zuniga, S. A. (2005). *Measurement issues in the alignment of standards and assessments: A case study* (CSE Report No. 653). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation (CSE). Retrieved from http://www.cse.ucla.edu/products/reports/r653.pdf

Ingham, A. M. and Gilbert, J. K. (1991) The use of analogue models by students of chemistry at higher education level. *International Journal of Science Education*, 13, 193- 202.

Joyce, J., Gitomer, D.H., and Iaconangelo, C. *Assessment of Learning and Teaching Through Quality of Classroom Assignments* (European Association of Research on Learning and Instruction-SIG 1 Assessment. Madrid, Aug 2014).

Joyce, J**.,** Zisk, R.C., Gitomer, D.H. (2017, April). *Classroom Artifact Protocols in Assessment: The Role of Domain Specificity in Two Studies*. Roundtable to be presented at the annual meeting of the American Educational Research Association, San Antonio, Texas.

Klenowski, V. (2011). Assessment for learning in the accountability era: Queensland, Australia. *Studies in Educational Evaluation*, *37*(1), 78–83. http://doi.org/10.1016/j.stueduc.2011.03.003

Koh, K., & Luke, A. (2009). Authentic and conventional assessment in Singapore schools: an empirical study of teacher assignments and student work. *Assessment in Education: Principles, Policy, and Practice* , 291-318.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and Implications. *Educational Measurement: Issues and Practice*, *13*(3), 5-16.

LeMahieu, Paul G., Gitomer, Drew H., and Eresh, Jo Anne T. "Portfolios in Large-Scale Assessment: Difficult But Not Impossible." *Educational Measurement: Issues and Practice* 14, no. 3 (1995): 11–28. doi:10.1111/j.1745-3992.1995.tb00863.x.

Little, J. W., Gearhart, M., Curry, M., & Kafka, J. (2003, November). Looking at student work for teacher learning, teacher community, and school reform. *Phi Delta Kappan*, *85*(3), 184–192. Retrieved from http://pdk.sagepub.com/content/85/3/184.full.pdf+html

Martínez, J. F., Borko, H., & Stecher, B. M. (2012). Measuring instructional practice in science using classroom artifacts: lessons learned from two validation studies. *Journal of Research in Science Teaching*, *49*(1), 38–67. http://doi.org/10.1002/tea.20447

Martínez, J. F., Borko, H., Stecher, B., Luskin, R., & Kloser, M. (2012). Measuring Classroom Assessment Practice Using Instructional Artifacts: A Validation Study of the QAS Notebook. *Educational Assessment*, *17*(2-3), 107–131. http://doi.org/10.1080/10627197.2012.715513

Matsumura, L., & Pascal, J. (2003). *Teachers' assignments and student work: Opening a window on classroom practice.* Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Merritt, E. G., Rimm-Kaufman, S. E., & Berry, III, R. Q. (2010). A reflection framework for teaching mathematics. *Teaching Children Mathematics*, *17*(4), 238–248.

Mitchell, K., Shkolnik, J., Song, M., Uekawa, K., Murphy, R., Garet, M., et al. (2005). *Rigor, Relevance, and Results: The Quality of Teacher Assignments and Student Work in New and Conventional High Schools.* Washington, D.C.: American Institutes for Research.

Morris, A. K., & Hiebert, J. (2011). Creating Shared Instructional Products: An Alternative Approach to Improving Teaching. *Educational Researcher*,*40*(1), 5–14. http://doi.org/10.3102/0013189X10393501

National Council of Teachers of Mathematics. (2007). Mathematics teaching today: Improving practice, improving student learning. Reston, VA: Author.

National Research Council (Ed.). (1996). *National science education standards*. National Academy Press.

Newmann, F., Bryk, A., & Nagaoka, J. (2001). *Authentic Intellectual Work and Standardized Tests: Conflict or Coexistence?* Chicago: Consortium on Chicago School Research.

Petticrew, M. & Roberts, H. (2006). *Systematic Reviews in the Social Sciences: A practical guide* Oxford: Blackwell Publishing.

Ruiz-Primo, M. A., Li, M., Tsai, S.-P., & Schneider, J. (2010). Testing one premise of scientific inquiry in science classrooms: Examining students' scientific explanations and student learning. *Journal of Research in Science Teaching*, n/a–n/a. http://doi.org/10.1002/tea.20356

Ruiz-Primo, M. A., Li, M., & Shavelson, R. J. (2002). *Looking into students' science notebooks: What do teachers do with them?* (CSE Technical Report No. 562). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation (CSE). Retrieved fromhttp://www.cse.ucla.edu/products/Reports/TR562.pdf

Shear, L., Means, B., Mitchell, K., House, A., Gorges, T., Joshi, A., et al. (2008). Contrasting Paths to Small-School Reform: Results of a 5-year Evaluation of the Bill & Melinda Gates Foundation's National High Schools Initiative. *Teachers College Record* , 1986-2039.

Silver, E. A., Mesa, V. M., Morris, K. A., Star, J. R., & Benken, B. M. (2009). Teaching Mathematics for Understanding: An Analysis of Lessons Submitted by Teachers Seeking NBPTS Certification. *American Educational Research Journal*, *46*(2), 501–531. http://doi.org/10.3102/0002831208326559.

The University of Queensland. (2001). *School reform longitudinal study: Final report (Volume 1)*. Brisbane, AU: School of Education, The University of Queensland.

Stodolsky, S. (1990). Classroom Observation. In J. Millman, & L. Darling-Hammond, *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 175-190). Newbury Park, CA: Sage.

Thomas, J. & Harden, A.(2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Med Res Meth*,**8:**45.

Wenzel, S., Nagaoka, J., Morris, L., Billings, S., & Fendt, C. (2002). *Documentation of the 1996-2002 Chicago Annenberg Research Project Strand in Intellectual Demand Exhibited in Assignments and Student Work.* Chicago: Consortium on Chicago School Research.

Using Classroom Artifacts to Track Enacted Science Reform:

The Artifact Indicator Protocol Study

Jeanette Joyce

October 2017

Abstract

There has been an ongoing call for science education reforms for nearly a century. The 21st century competencies described in the most recent phase of reform are represented in new standards (such as the Next Generation Science Standards) for developing global citizens. However, it is not enough to develop reforms through publication and legislation. It is important for stakeholders to understand how these policies are making their way into classrooms. In 2013, a NRC report called for a national indicator system that could be used to support the improvement of STEM education. This study explores how classroom artifacts could be used for such a purpose. Through literature synthesis and semi-structured interviews with eight experts in standards, artifacts, and large-scale data collection, a science-specific artifact measurement protocol was designed to serve as an indicator of both content coverage and practice alignment, The Artifact Indicator Protocol-Science (AIP-S) is designed to assess the quality of classroom assignments and assessments (artifacts) with respect to a set of dimensions that are aligned with new standards for science education such as those contained in college and career readiness standards. In order to gather empirical evidence for the soundness of the instrument, a study was conducted during the 2015-2016 academic year, with goals of feasibility of use and sensitivity of ratings to factors that were deemed likely to be of interest to stakeholders. Findings indicate that the instrument does hold promise as a tool for measuring alignment and potentially for self-study by a school, department, or district.

Background and Purpose

There has been an ongoing call for reforms in science education for nearly a century. Initially, these calls focused on expanding content, such as the National Science Foundation initiative to add Physics to High School curriculum in 1956 (Kaiser, 2002). These calls then eventually changed focus from what was taught to how science should be taught, with the growth of inquiry-based teaching and learning (Bybee. 1995). Other important shifts were from local benchmarks to more national standards that could then be adopted and adapted at the local level, and an increasing focus on building from elementary through secondary, rather than focusing solely on secondary education (Bybee, 1995). Most recently, continued low US achievement on measures such as Trends in International Mathematics and Science Study (TIMMS) and National Assessment of Educational Progress (NAEP) has spurred further reforms (Nord et al, 2011). The Next Generation Science Standards (NGSS) were developed and are in part a response to the Heritage Foundation report (2009) stating the future economic growth in the U.S. was dependent on improvement in science education.

The 21st century competencies described in this most recent phase of reform are represented in new standards (such as the Next Generation Science Standards) for developing global citizens. According to Ananiadou & Claro (2009) in their Organisation for Economic Co-operation and Development (OECD) report, "Developments in society and economy require that educational systems equip young people with new skills and competencies, which allow them to benefit from the emerging new forms of socialisation and to contribute actively to economic development under a system where the main asset is knowledge"(p. 5). In other words,

students who hope to participate in the future global economy need to not only master content but also the practices needed to critically understand the constant stream of emerging information and to potentially contribute to the knowledge base in science.

However, it is not enough to develop reforms through publication and legislation. What matters is how reform policies are interpreted by teachers and enacted in classrooms. This is in line with what Lipsky referred to as "street-level policy" (in Gibson, 2015), wherein ideas would be re-interpreted as they move from the halls of legislature to the halls of schools. Capps, Shemwell, and Young (2016) report that teachers can misunderstand new reforms and self-report that they are in compliance when tasks set for students are not truly aligned with standards. However, Bismack, Arias, Davis, and Palinscar (2014) found that, with support, teachers were able to incorporate new standards into classroom instruction. The challenge for stakeholders becomes how to elicit evidence of how standards represented in policies are actually being enacted in classrooms in order to offer support to teachers. Therefore, it becomes essential to have measures of how new reforms are reaching students and whether progress is being made toward reform goals. Such a set of measures would form an indicator system.

In 2013, a NRC report, *Monitoring Progress Toward Successful K-12 Education: A Nation Advancing?,* called for a national indicator system that could be used to support the improvement of STEM education. The report described 14 Indicators that were needed to guide improvement. Congress then directed the NSF to begin implementing a progress monitoring system for the indicators. In response, there is a call for development of new instruments to be used in an indicator system. "A monitoring and reporting system designed around these indicators would be unique in its focus on key aspects of

teaching and learning and could enable education leaders, researchers, and policy makers to better understand and improve national, state, and local STEM education for all students" (National Research Council, 2013, p. 3). The call is for an indicator system to describe the implementation of new college and career readiness standards into daily classroom tasks (Committee on the Evaluation Framework for Successful K-12 STEM Education; National Research Council, 2013; Means, Mislevy, Smith, Peters, & Gerard, 2016). One of these indicators that was identified as a priority was Indicator #5: *Classroom coverage of content and practices in NGSS*.

The indicator measures would serve a different purpose than what has been previously used in teaching quality. There currently exists a body of work describing various types of evaluation of instruction, including large-scale indicators like NAEP. These evaluations often make use of student achievement measures, observational measures, and survey measures. Each of these can make a useful contribution to understanding what is happening in classrooms and the extent to which reforms are implemented. Achievement measures can provide information on student mastery of content and practices. However, achievement measures can lag behind reform initiatives, particularly in historically "untested" subjects like science, and therefore may not assess reform-related curriculum effectively (Buckendahl, Plake, Impara, & Irwin, 2000; Martone & Sireci, 2009). Observations provide information on instructional exchanges between students and teachers, and allow for assessment of discourse. Finally, self-report through survey can give an indication of teachers' perception of their own practice. While these traditional methods can provide some insight into what is happening in classrooms, there is the potential for new measures that shed light, particularly in terms of

instruction around science practices.

This study explores how artifacts could be used for such a purpose. Classroom artifacts, such as labs, tests, and projects, have tremendous potential as one component of an indicator system, although they have as yet not been used for such a purpose. In an indicator study, the targeted inference would not focus on relative strengths of individuals, but rather, the extent to which certain skills and content foci are being addressed in practice across groups of classrooms (within a school, district, state, country). Previously, classroom artifacts have been used to make inferences at the teacher, student, or classroom level. This current study, although unique in its extension of the use of artifacts, draws from the existing body of research that has shown:

- Artifacts can serve as a window into classroom practices, interactions and enacted policies.(Borko, Stecher, & Kuffner, 2007; Matsumura & Pascal, 2003);
- Artifacts can be scored at an acceptable reliability level. (Borko et al, 2005; Clare & Aschbacher, 2001; Matsumura et al, 2008);
- Artifact study findings are similar to observation results (Joyce, Gitomer, & Iaconangelo, 2014);
- Artifact studies found assignments to be without the higher level cognitive demand associated with in-depth learning (Joyce, Gitomer, & Iaconangelo, 2014).

Again, these studies were making inferences about teachers, students, or classrooms, not as an indicator in a system. An indicator, as defined by the European Commission on Public Health, "…is a quantitative or qualitative measure of how close we are to achieving a set goal or policy outcome. They help us analyse and compare performance across population groups or geographic areas, and can be useful for

determining policy priorities"[3]. Indicators have previously relied heavily on surveys, and it has been suggested that artifacts may be too time and labor intensive to be incorporated into an indicator system. However, an indicator can function on multiple levels; there is, of course, need for a national indicator that shows how the nation as a whole is faring in its progress toward educational reform, but there is also a need to monitor systems on a finer grained level, for districts and schools themselves.

21st century competencies are represented in new standards, such as the Next Generation Science Standards, and incorporate not only content but also the development of practices needed for future success. This study undertakes the design of an indicator using classroom artifacts as a measure of "the extent to which the instruction and learning activities students experience in a classroom cover content in a set of standards, are consistent with the performance-level expectations of those standards, and reflect the same conception of learning and instruction… capturing the enacted curriculum" (Means, Mislevy, Smith, Peters, & Gerard, 2016, p. 24). It seeks to answer the question, "How would we know if educational practice is changing with the emergence of new standards?" and to explore the utility of using classroom artifacts to answer the call, beginning with the NRC in 2011, for an indicator system to both establish the current level of STEM teaching and learning, and to track progress.

This study tests the hypothesis that classroom artifacts can provide streamlined access to and meaningful evidence of enactment of new standards, particularly in respect to the practices articulated in the new standards. Artifacts may provide insights that are complementary to other indicators and serve as key evidence in understanding how

[3] See http://ec.europa.eu/health/communicable_diseases/indicators_en

curriculum is translated by teachers for students in terms of both content and practice. However, there are limitations. For example, a classroom artifact would not shed light on classroom discourse practices, but would be able to shed light on whether students were being tasked with formulating their own questions and planning an investigation. A protocol that identifies and codes the science practices that are captureable through classroom artifacts can serve as one indicator in the status of new standards' influence on classroom tasks. Since classroom assignments call on domain-specific knowledge and skills, the study presents a science-specific protocol that provides information to stakeholders on alignment to emerging US standards (NGSS), including both content and practices, and is related to measures included in TIMSS (2011) and discussed as global science literacy (Mayer and Tokuyama, 2002). The research questions include:

- *To what extent can the protocol be used to measure classroom practice articulated in science standards?*
- *To what extent is the protocol sensitive task characteristics that might be of interest to stakeholders?*

Methods

After the synthesis of available pertinent literature (Joyce, 2017) as well as semi-structured interviews with eight experts in standards, artifacts, and large-scale data collection, a science-specific artifact measurement protocol was designed (Figure 1). The protocol was designed to capture the kinds of understandings and practices embodied in the NGSS documents. Considerations included dimensions, scale ranges and scoring procedures.

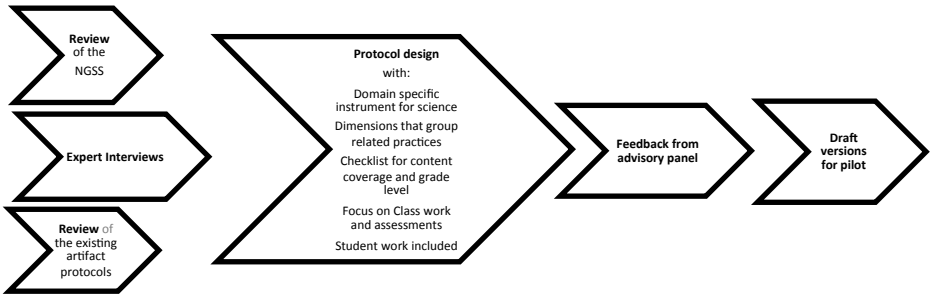


Figure 1. Summary of study methodology.

Instrument Design

This instrument development drew from both Mislevy's and Riconscente's work on evidence centered design (ECD) (2006) and the Rational Empirical Strategy of Test Construction(RESTC). Mislevy and Riconscente indicate that in any instrument design, the initial stages, or layers, must include domain analysis and modeling. In these initial stages, the researchers "gather substantive information about the domain of interest" and "express [the] assessment argument in narrative form" (2006, p. 67). The domain analysis and modeling was driven by the literature on artifact research as synthesized (Joyce, 2017), an in-depth review of the standards including the literature from which they emerged, and through expert interviews. Beginning in February 2015, I contacted experts in the areas of classroom artifact research, large-scale data collection and management, and science teaching. These experts were identified through the literature review as well as through "snowballing," in which interviewees were asked to provide names of other potential interviewees in their area of expertise. In total, eight semi-structured interviews were conducted. Points of convergence from the interviews and the

advisory panel are summarized in Table 1 and informed the development of the instrument.

Table 1
Points of Convergence from Interviews

General	Sampling	Scoring
•Artifacts can give important insight.	•Assessments and in-class work are more useful than homework and lesson plans.	• Raters should have teaching experience and extensive training.
•Not all standards lend themselves to artifact study.	•Multiple artifacts across the school year are needed.	•Dimensions should be limited and clearly defined.
	• Student work is critical.	•Separate content and practices.

The domain analysis and modeling then formed the theoretical justification needed for RESTC. This method, used by researchers such as Reinhart Pekrun in instrument development for motivation research, follows the belief that, in order to develop a good instrument, one needs this theoretical justification, followed by sound design process, and empirical back up of test validity and reliability. This provided the guidance and incentive for the pilot study of the instrument.

In designing the protocol, I considered not only what was critical in terms of the Next Generation Science Standards but also what is able to be seen in artifacts, as well as what may be trackable over time. Specifically, key considerations in the protocol design included attending to aspects of the standards for which artifacts provide evidence, accommodating the simultaneous independence and overlap of standards, accounting for cognitive and time demands on raters, accommodating likely variability in fidelity of

responses to artifact study instructions by participating teachers, and, finally, clarifying consideration of student work in determining ratings.

The next decision was how to create dimensions that were representative of new standards. It was decided to first separate content and practices. A content checklist was developed for science. This involved consideration of the range of the content (i.e. what general topics were present in the artifact) as well as the specific Disciplinary Core Ideas represented in the artifact, and sub-topics covered .

For practices, an effort was made to cluster the eight practices (listed in Table 2 below) into meaningful topics, rather than score each one separately. This was undertaken for ease of scoring and interpretation, necessary for an indicator system, as well as from a domain model that acknowledges that completely separating practices is somewhat artificial. For guidance, I turned to the NAP publication *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (2012) and the following diagram (see Figure 2). This publication proposes that a schema of three dimensions that "helps identify the function, significance, range, and diversity of practices embedded in the work of scientists and engineers." Although admittedly a simplification, the figure does identify "three overarching categories of practices and shows how they interact"(p. 46). Thus I grouped the eight practices into these dimensions. The protocol is described below and included in Appendix A.

Table 2
Next Generation Science Standards: Practices[4]

1. Asking questions (for science) and defining problems (for engineering)
2. Developing and using models
3. Planning and carrying out investigations
4. Analyzing and interpreting data
5. Using mathematics and computational thinking
6. Constructing explanations (for science) and designing solutions (for engineering)
7. Engaging in argument from evidence
8. Obtaining, evaluating, and communicating information

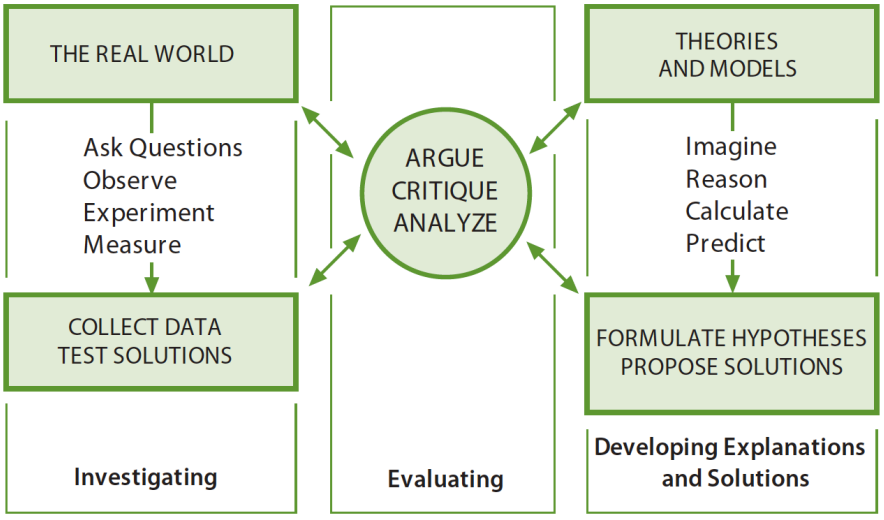


Figure 2. Graphic Depiction of NGSS from *A Framework* of *for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*

[4] (retrieved from http://www.nextgenscience.org/sites/default/files/Appendix%20F%20%20Science%20and%20Engineering%20Practices%20in%20the%20NGSS%20%20FINAL%20060513.pdf)

A decision was made in terms of developing a scale that could be used to describe an artifact in terms of the NGSS standards through the conceptualized dimensions. It was hypothesized that three levels of scoring (absent, partial, complete) would be too coarse grained, not allowing for refinement of levels of partial practice. It would be critical to describe levels of partial implementation during this early period of adoption and adaptation. Although wider ranges of scoring have been found to have higher reliability in surveys (Preston & Colman, 2000), it was felt that seven levels of scoring might be too burdensome for raters so instead a uniform four point score range was developed for each dimension. This range also avoids any tendency for raters to drift toward middle (Garland, 1991). The lowest level is set at zero (or absent) to increase interpretability of scoring and of analysis. Levels of scoring were as follows: 0 (absent), 1 (superficial), 2 (incomplete), 3 (present). The protocols and scoring guides were completed in August 2016. The pilot artifact protocol is described in detail below:

AIP-S. The Artifact Indicator Protocol-Science (AIP-S) is designed to assess the quality of classroom assignments and assessments (artifacts) with respect to a set of dimensions that are aligned with new standards for science education such as those contained in college and career readiness standards. For the purposes of the AIP-S, artifacts include both the assignments and assessments as assigned by the teacher and the student work that is associated with these assignments and assessments The AIP-S is designed to capture the extent to which students are asked to engage in science thinking and the extent to which **content** in terms of disciplinary core ideas and science **practices** are addressed across the science disciplines.

After the disciplinary core ideas are identified, the artifacts are rated on a four-point scale for the eight science practices that represent enactment of quality science thinking in instruction. These practices are organized into three dimensions, *Investigation* (INV), *Data Analysis and Evaluation* (DAE), and *Explanation, Argumentation, and Solution Design* (EASD), which are aligned with the NRC framework shown in Figure 2. That is, the Investigation dimension relates to the skills described in the left side of the graphic, including asking questions and collecting data. The dimensions of the AIP-S draw on available standards as well as other protocols exploring the intellectual demands manifest in classroom artifacts.

Specifically, the *Investigation* dimension focuses on the extent to which students are required to ask questions, observe, experiment, and measure data, connecting the real world to their conceptual understanding of science idea(s), and covers elements from NGSS practices: 1. *Asking questions (for science) and defining problems (for engineering); 3. Planning and carrying out investigations;* and *8. Obtaining… information* (Table 2). Artifacts that are high on this dimension ask students to develop their own investigations by drawing on conceptual understanding, to define real-world problems, and to obtain needed information in a systematic way from various resources including observations. Artifacts that ask students to complete tasks such as providing or selecting among definitions or carrying out a highly prescribed lab would be low on this dimension.

Specifically, a task such as filling in a provided graphic organizer from a lecture on photosynthesis or answering end of chapter recall questions from an assigned text would score zero(0). In tasks such as these, students are not asked to generate any

science-related questions, to plan or carry out an investigation, or to gather information independently. Instead, the teacher provides information through lecture or assigned texts. In order for a task to score one (1), there would have to be some beginning evidence of investigation or research. To continue with the example of photosynthesis, a task that asked students to complete a guided experiment comparing the growth of plants in the window with that of plants in the closet or independently take notes from the assigned text on photosynthesis would be scored as "surface practice." Here, students are asked to carry out prescribed investigations in which all questions and steps are specified and the outcome is known, or to gather information without synthesis from multiple sources and without consideration of credibility, accuracy, or bias. The next level of task alignment is "incomplete practice" and would score two (2). The students may be asked to carry out an experiment that gives evidence that plants need light to grow or to write a report about photosynthesis using at least two sources in addition to the class text. In these cases, the students are asked to plan and carry out systematic investigations or to gather, read, and synthesize information from multiple sources in order to answer *given* questions. In "complete practice," tasks that score three (3), would require students to formulate the question as well as plan and carry out the investigation within the bounds of classroom feasibility or to conduct synthesis research with attention to credibility and bias. In our example, this would be to design an experiment that investigates some factor of plant growth or to write a research paper with justification of choice of sources as reliable.

The *Data Analysis and Evaluation* dimension focuses on the extent to which students are asked to organize raw data including the identification of significant features

and patterns, use mathematics to represent relationships between variables, and take into account sources of error. It covers elements from practices: *4. Analyzing and interpreting data; 5. Using mathematics and computational thinking*; and *8. Evaluating, information* (Table 2). Artifacts that are high on this dimension ask students to display, analyze, interpret, and critique raw data or information using mathematical, computational, and statistical tools when appropriate. Scores at the low end of this dimension do not ask students to display, analyze, interpret, or critique raw data or information.

Similar to our examples for Investigation, guided lecture notes or chapter questions dealing with text would score as zero(0), as students are not asked to display, analyze, interpret, or critique raw data nor to engage in any mathematical, computational, or statistical thinking about data. In order to score as "surface practice," or one (1), the task for students would require them to display, analyze, or interpret simple patterns in given data representations. Asking the students to examine a table of plant growth and asking which grew the most is one example. "Incomplete practice," scoring two (2) describes a task in which the students are working with raw data and providing interpretations that involve mathematical concepts. An example would be monitoring and recording the growth of the class pea plants over the week and calculating the rate of growth for each. However, this is not "complete practice" because the students are not engaging with correlation or causation, or to consider sources of error. An example of complete practice, scoring three (3), would be similar to the one above, but with the questions that ask about the variation between plants or to make predictions for continued growth patterns, with consideration of error.

For *Explanation, Argumentation, and Solution Design*, the dimension focuses on the extent to which students are asked to analyze and/or represent situations and to develop and/or evaluate science arguments, explanations, and/or engineering solutions through written argumentation and/or development/revision of models, incorporating elements from practices: 2. *Developing and using models*: *6. Constructing explanations (for science) and designing solutions (for engineering); 7. Engaging in argument from evidence; and 8. ... communicating information* (Table 2). Artifacts that score high on this dimension require extended written communication to develop a science argument, explanation, or engineering solution description, with students engaging in theory and iterative model development as appropriate. Artifacts scoring low on this dimension do not require students to develop arguments, explanations, or engineering solutions or to engage in theory or model development.

We can continue with the theme of photosynthesis to examine the different score levels of this dimension. A word search of terms related to photosynthesis would score as zero (0), as students are not asked to construct explanations or design solutions, to interact with models, or to engage in any form of scientific argumentation about phenomena in the natural or designed worlds. A worksheet that asked students to label a given model of photosynthesis would score as one (1), "surface practice." While they are engaging with a scientific model, the students are not applying reasoning to represent phenomena in the natural or designed worlds. In order to score as "incomplete practice" at level two (2), the task must require the students to construct explanations, solutions, models, or arguments but with little evidence or evaluation of others' reasoning. An assignment or assessment that asks students to explain photosynthesis but not to support

their claims is an example. The same task that requires the students to explain or model photosynthesis using data from the class experiment would score as "complete" with a three (3).

Empirical Study

In order to gather empirical evidence for the soundness of the instrument, a study was conducted during the 2015-2016 academic year, with goals of feasibility of use and sensitivity of ratings to factors that were deemed likely to be of interest to stakeholders.

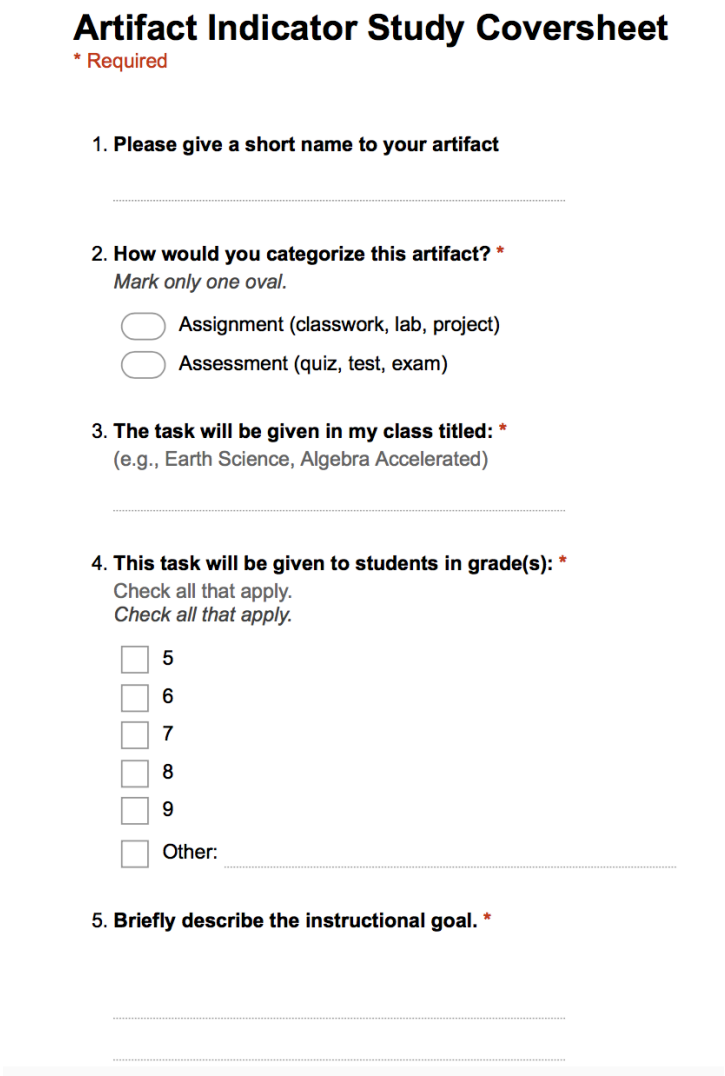
Data Sources and Sampling. The initial intent was to apply the protocols to artifacts collected from multiple districts, using the sampling guidance from the expert interviews (Table 1). To that purpose, IRB permissions were gained from ten large urban districts, allowing access to schools, but leaving participation to the discretion of the principal. In only two of the districts was there additional support from the central office. All in all, more than 300 schools were approached by mail and via phone calls, with positive responses from only five principals. Here, too, there was no offer of continued support, but rather only permission to contact teachers. Again, hundreds of teacher letters were sent out, which resulted in recruitment of three teachers. Even when budget and IRB documentation were amended to include a $200 stipend for what was to be 15 -30 minutes of additional work outside of regular classroom duties, no further participation was gained. The approach to recruitment was then revised to be more personal, pursuing connections with local schools and contacting colleagues for artifacts from their current or past research.

In this way, 25 teachers were recruited to each provide four artifacts (2 assessments and 2 assignments) from the current academic year which were added to pre-

standards artifacts gathered from previous artifact studies (SCOOP, QAS). Borko, Stecher, and Kuffner (2007) developed the SCOOP protocol in order to use artifacts to characterize math and science classroom instruction to aid in assessing both the students' understanding as well as the teachers' own process and to explore the capacity of classroom artifacts as an indicator of reform-based instructional practices**.** Subsequently, Martinez, Borko, Stecher, Luskin, and Kloser (2012) conducted a pilot validation study of the Quality Assessment in Science notebook in order to gain insight into effective instruction in terms of quality assessment practice. A current study with collaboratively designed science tasks across a small state also provided 40 artifacts, for a total closer to 200 artifacts, rather than the initial 500 planned. This revised sample included 115 artifacts collected from the 2015-16 academic year, across 3 states and 78 artifacts collected before 2011, in grades 5 through 9.

The classroom artifacts consisted of the assignment or assessment template (i.e., the blank form) as well as a selection of student work. One important point of divergence that had emerged from the expert interviews (Table 1) was whether there would be affordances in requesting "typical" or "challenging" work from the teachers. The question was posed whether, in an indicator study, one might be interested in the best the nation can produce (challenging) or what is pervasive in America's classrooms (typical). Based on the findings from the previous study (Joyce, Gitomer, and Iaconangelo, 2014) that there was only a slight difference in quality between typical and challenging math tasks as selected by teachers, and in the hope of capturing the high level of challenge embodied in new standards, it was decided by the study team, for the purposes of the pilot, to elicit challenging work.

The teachers completed a brief cover sheet online (see Figure 3 and Appendix C) to give some context to the work. The cover sheet asked for information about student demographics, lesson modification, and lesson origin. This data will be analyzed to better understand the context of the artifact itself. Teachers were instructed to provide one unit assessment and one challenging assignment for each of two rounds of data collection. They were also asked to provide six samples of student work, two each at high, medium, and low success levels. All artifacts were de-identified, coded for district, teacher, grade level, and time of year, and then scanned to a secure server. Some of these assignments were the result of a collaborative effort to develop assignments aligned with new standards in science and represent grades 5-8. The artifacts were collected in March and May 2016 from recruited schools.



Artifact Indicator Study Coversheet

* Required

1. **Please give a short name to your artifact**

2. **How would you categorize this artifact?** *
   Mark only one oval.
   - Assignment (classwork, lab, project)
   - Assessment (quiz, test, exam)

3. **The task will be given in my class titled:** *
   (e.g., Earth Science, Algebra Accelerated)

4. **This task will be given to students in grade(s):** *
   Check all that apply.
   Check all that apply.
   - 5
   - 6
   - 7
   - 8
   - 9
   - Other:

5. **Briefly describe the instructional goal.** *

Figure 3. Excerpt from artifact coversheet. Full coversheet available in Appendix C.

Additionally, artifacts were collected from an ongoing state reform initiative that is grounded in a competency-based educational approach designed to ensure that students have meaningful opportunities to achieve critical knowledge and skills. Other artifacts were selected from the SCOOP and QAS studies, which predate new standards and collected artifacts as a measure of teaching quality. In total, the 192 artifacts were comprised as follows: 90 collected from the recruitment effort, 25 from the current state reform project, 40 from SCOOP, and 37 from the QAS study.

Scoring. Raters were selected from a pool of 20 recruited from a call to Science Education graduate students and the State Association of Science teachers. Candidates were asked to describe their familiarity and experience with all three salient strands: NGSS, middle school science teaching, and use of a rubric for rating. Eight applicants were interviewed. The interview process included working with an unscored artifact, which candidates were asked to describe in terms of standards and teacher's perceived intent for the task. Following the interview process, three were selected, all of whom had extensive experience with the Next Generation Science Standards, and with teaching middle school science.

Training consisted of an in-person all-day session in which raters become familiar with the protocol and guiding questions before being asked to score anchor artifacts both with the study team and independently. Prior to training, scoring guidance materials were developed. These consisted of identifying critical components of the dimension and then creating focusing yes/no questions around these (see Appendix D). For example, a question for the *Investigation* dimension was "Does the artifact require the students to formulate their own questions?" In training, it was stressed that these questions were meant to facilitate focusing on the critical components, and not to translate into a score. That is, a certain number of "yeses" did not translate into a certain score. In fact, different dimensions had different numbers of questions. Raters did record their answers to each question, but the overall score per dimension required a more holistic decision related to the rubric. Also, key differences between score points were stressed in training, so that raters gained confidence in deciding between a 0/1, 1/2, and 2/3. All disputes were discussed and raters were directed to explain and record the justification

for their scoring in terms of the rubric. Raters then began scoring study artifacts first by content, and then one dimension at a time. It was possible to monitor online scoring in real time, and to note areas that needed further training. This training was accomplished through frequent re-alignment and troubleshooting check-ins, conducted via online video conferencing.

Additionally, timestamp information was collected in order to estimate time needed for scoring. Rater timestamp data was converted to duration of scoring per artifact per dimension. Averages were calculated per dimension after outliers (as defined by a duration that was more than twice as large as the mode) were removed. It was hypothesized that the excessively long durations indicated an interruption in scoring, and therefore did not accurately indicate the time to score. Rating took place between August and October 2016, with each artifact rated by two trained raters. Artifacts were randomly assigned to raters, and the order of scoring was randomized across dimensions.

Initially, artifacts are coded for content, focusing on the disciplinary core ideas categorized under Physical Sciences, Life Sciences, Earth and Space Sciences, and Engineering Design. Multiple content topics could be selected if present in the assignment or assessment, but raters were encouraged to choose a primary content area.

The blank assignment was scored first, and then student work samples (between five and ten available per artifact; work was rated high, med, low) would be used to determine if initial scoring is accurate. For example, an artifact that asked students to explain but still gave credit for shallow answers would have its score adjusted down while student work that showed students presenting models as evidence, even though there was no specific request to do so in the template, would be scored higher. Scoring

was conducted using online forms, with screen-captured versions included in Appendix D. Additionally, the forms included time stamp data for later analysis of the load to raters in scoring. All raters participated in a 30-45 minute exit interview after completing scoring. Interview questions are included in Appendix E.

Results and Analysis

Reliability

With any rater assessment based on a rubric, come concerns about the reliability of the scoring. Jonsson and Svingby (2007) in their review of research involving scoring rubrics indicated that "Ideally, an assessment should be independent of who does the scoring and the results similar no matter when and where the assessment is carried out, but this is hardly obtainable "(p.133). Furthermore, they used Stemler's 2004 criterion of 70% or greater for exact agreement and Stoddart, Abrams, Gasper, & Canaday's 2000 range of kappa values between .40 and .75 as "represent[ing] fair agreement beyond chance" (Jonsson and Svingby, p.133 , 2007). In previous artifact studies, reliability is reported as percent agreement, with scores ranging from moderately low (40% for overall artifact package in Martinez, Borko, & Stecher, 2012) to higher levels (86.4% for overall artifact in Clare & Aschbacher, 2001). Scale/dimension agreements have a similar wide range, with some reported as low as 22% (Martinez, Borko, & Stecher, 2012).

Table 3
Rater Reliability by Dimension

	% exact agreement	% adjacent	kappa	ICC
INV	70.5	94.4	.546***	.821***
DAE	59.7	94.2	.330***	.764***
EASD	62	92.7	.484***	.843***

For this study, rater agreement was described by percent exact and adjacent agreement, Cohen's kappa, and the intraclass correlations (ICC). ICC was run as oneway random, since all raters did not rate all artifacts, and is reported for means, as we are interested in the overall reliability of the scoring, and not the reliability of one particular rater. Results by dimension are summarized in Table 3.

While all dimensions had good agreement for adjacent scores, exact agreement for *Data Analysis and Evaluation* (DAE), and *Explanation, Argumentation, and Solution Design* (EASD) dimensions fell below the Stemler's 70% level, but are not out of line with those found in other artifact studies. The findings for kappa, while significantly different from chance agreement, also fell below the .40 cutoff for DAE, but above for INV and EASD. ICC indicated that raters' scores are significantly correlated for all dimensions.

Descriptive Analysis

Content. In an indicator system, the type of content covered (in this case, the Disciplinary Core Ideas or *DCI* as described by NGSS) would be of interest to stakeholders over a specified period of time or across a system. Therefore, it is important

to assess the instrument's ability to support content coding. Within our sample, there were artifacts across all four general topics, described earlier, as agreed upon by both raters. There were slightly more artifacts in the areas of Life Sciences and Physical Sciences than in Earth Sciences and Engineering design, even though Engineering units and classes were purposefully selected for inclusion in the sample. The distribution of general content is shown in Table 4 below. There were some artifacts that the raters were unable agree on classification even at the most general level, and one artifact that raters agreed had no true science content. In terms of change over time, which is one key function of an indicator measure, there were differences when the artifacts were separated by pre- and post-standards implementation.

Within the four general categories, there was a range of specific content coded. Most common themes in Physical Science were *Matter and its Interactions* and *Motion and Stability*, while *Energy* and *Waves and Applications* were found in very few artifacts. For Life Science, the predominant theme was *Molecules to Organisms*, found much more often than *Heredity, Evolution* or *Ecosystems*. For Earth Science, more artifacts were classified as pertaining to *Earth's Systems* than *Earth's Place in the Universe* and *Earth and Human Activity*. Finally, no artifacts in Engineering were associated with *Links Among Engineering, Technology, Science, and Society*. Scoring agreement decreased at this level, with raters failing to agree on 17% of the artifacts. At the most fine-grained coding of sub-topics, scoring agreement decreased further, with failure to agree on 23% of artifacts. Complete frequency information on content coding for Specific and Exact Coding, keyed to the protocol, can be found in Appendix F.

Table 4
Content Coding Frequencies by Percent

	Physical Science	Life Science	Earth Science	Engineering Design	No Agreement
Overall	36	32	16	10	6
Pre-Standard	62	10	17	4	7
Post-Standard	18	46	15	15	6

Practice Dimensions. For all dimensions, scores were given across all four rating points. However, as found in previous studies, the overall quality in terms of alignment to new standards found in artifacts tended to be at the lower end of the scoring scale, with a significant skew for *Investigation* and *Data Analysis and Evaluation*. Mean scores for all three practice dimensions are summarized in Table 5 and distributions are shown in Figures 4a-c. In terms of *Investigation*, most artifacts either did not ask students to answer any question using experimentation or research, or were heavily scaffolded in what raters often referred to as “cookbook labs”. These artifacts dictate step-by-step how and where information is obtained, and lead to only one possible outcome.

Table 5
Dimension Scores
(N=192)

Dimension	Mean	Standard Deviation
Investigation (INV)	.80	.81
Data Analysis & Evaluation (DAE)	.68	.81
Explanation, Argumentation, & Solution Design (EASD)	1.40	.97

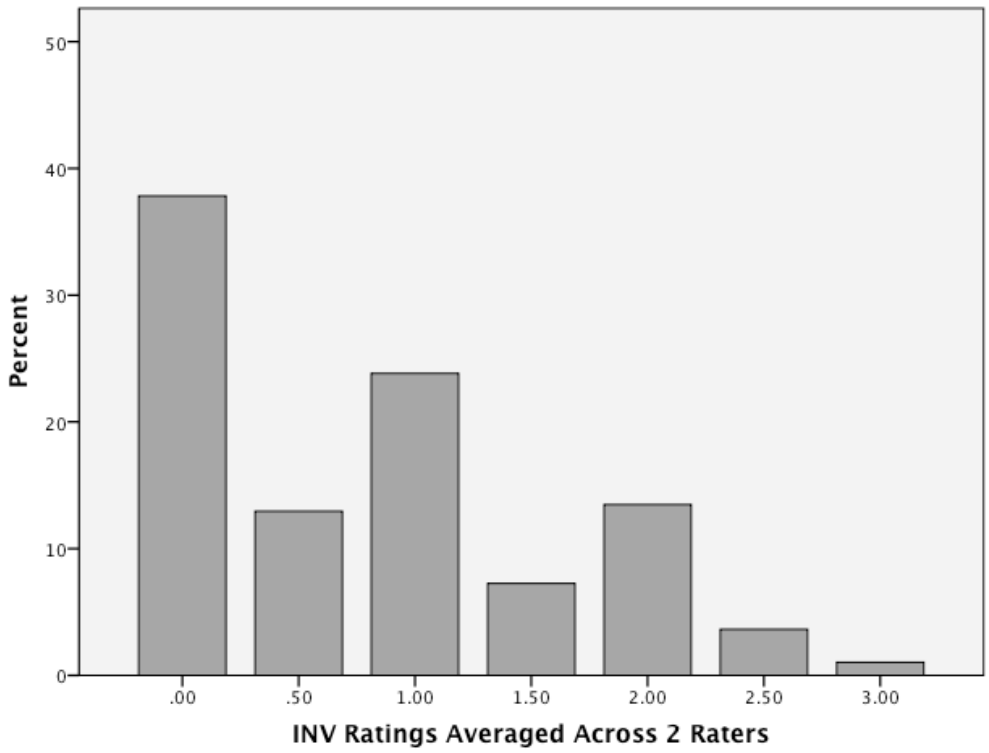


Figure 4a. Investigation

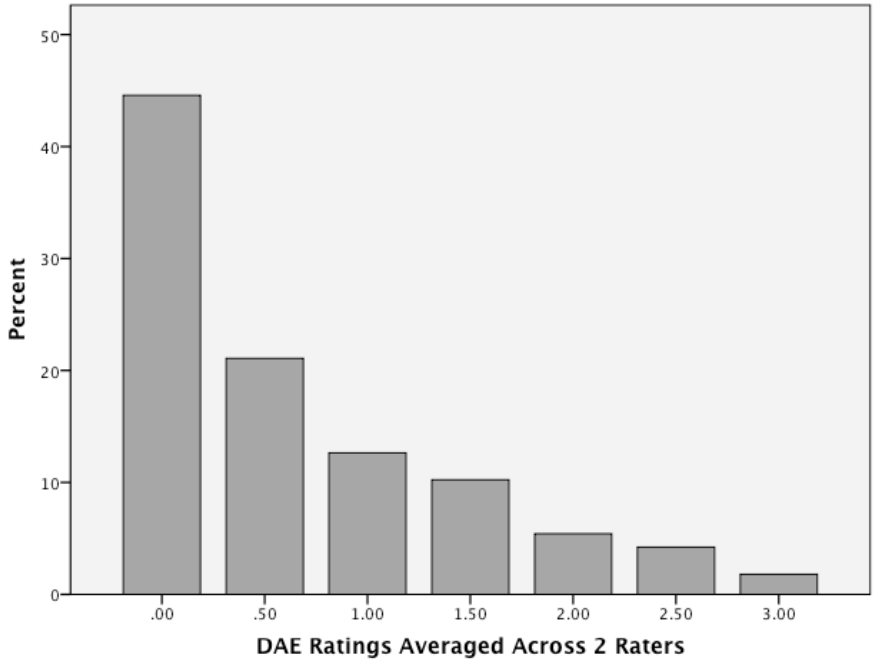


Figure 4b. Data Analysis and Evaluation

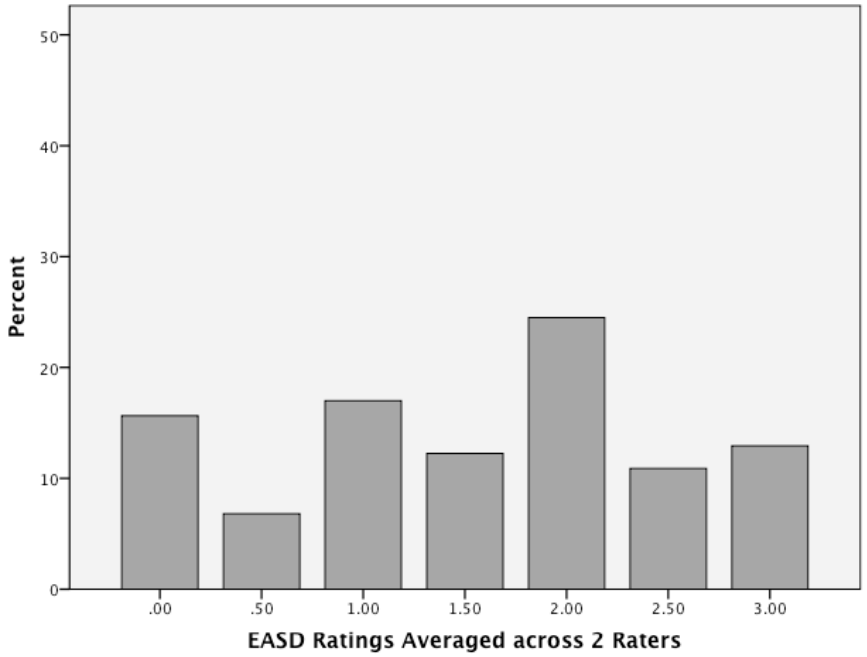


Figure 4c. Explanation, Argumentation, and Solution Design.
Figure 4a-c. Distributions of Averaged Ratings Across Dimensions.

The skewed pattern in *Investigation* was also found for *Data Analysis and Evaluation*. Again, scores covered the range of the scale, but in more than 40% of rated artifacts, students were not asked to display, analyze, interpret, or critique raw data. When the students were interacting with data, the artifacts tended to present data and ask for identification of simple patterns rather than asking students to engage in mathematical, computational, or statistical thinking or to consider sources of error.

The pattern for *Explanation, Argumentation and Solution Design* was somewhat different. While scores again were found at all points of the scale range, including almost 20% with no requirement for explanation, argumentation, or solution design, it does appear that students are to some extent being asked to explain, form an argument, or design an engineering solution. In order to meet the scoring criteria for "2" on this dimension, which accounted for more than 20% of the artifacts, the artifact must require the students to do one of the following:

- Students are asked to construct explanations and/or to design solutions with limited supporting evidence, principles, and/or theory. OR
- Students are asked to develop or describe models but are not asked to evaluate or revise models to explain, describe, test, and predict abstract phenomena and/or to design systems. OR
- Students are asked to construct an argument that supports or refutes claims for either explanations or solutions about the natural and/or designed world(s) using limited empirical evidence and scientific reasoning or agreed-upon design criteria when communicating scientific information.

Correlations and Aggregation

The dimensions were designed to measure different aspects of the practices associated with the NGSS. Although there was overlap, as the practices are inter-related, it is important that each dimension remains distinct in terms of exactly what is being described. In an effort to better understand the designed dimensions as related but not equivalent, correlations were run using the non-parametric Spearman's rho and are reported in Table 6. There is a moderate correlation between DAE and other dimensions, and a stronger correlation between INV and EASD.

Table 6
Correlations between Dimensions

	INV	DAE	EASD
INV	1	,403***	.623***
DAE		1	.395***
EASD			1

***p<.001

No formal aggregation was undertaken, as the dimensions are formulated to include different scientific practices that do not necessarily co-occur. However, artifacts were given an overall score that was equivalent to their highest score on any dimension in order to understand, in this sample, what were the levels of alignment and to what extent artifacts have uneven profiles. Results are shown in Figure 5.

$\overline{X}$ =1.57
sd = .92

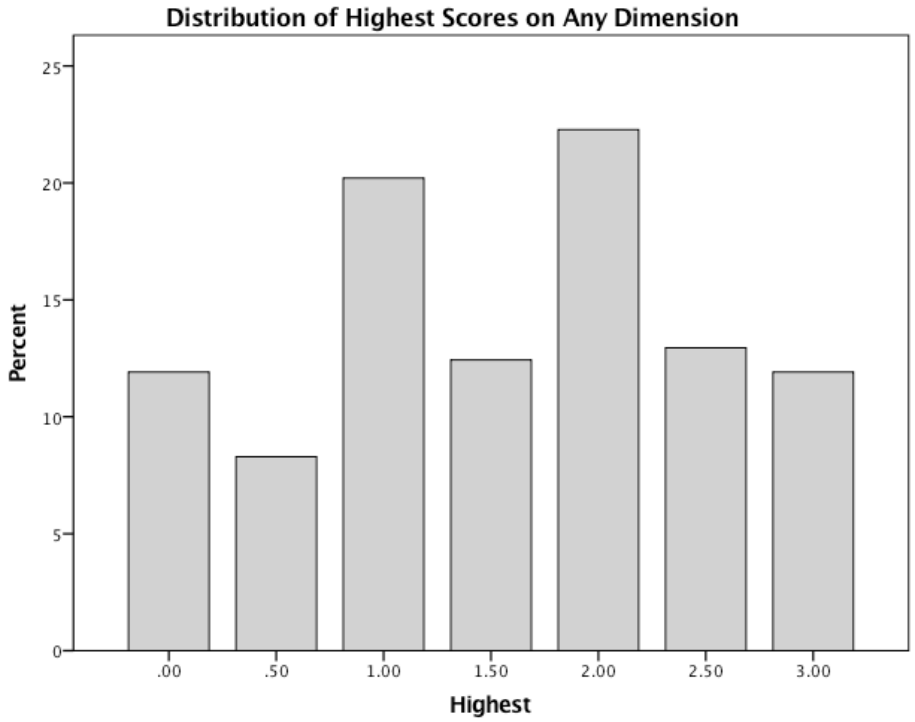


Figure 5. *Distribution of artifacts by highest score on any dimension.*

Analysis of Potential for Comparison

As an indicator measure, the protocols would be used to assess not only the extent to which content and practices were being implemented, but also to identify factors that may be associated with quality implementation. Although the convenience sample precludes inferences of this type, in this section, the potential of the protocol instrument is examined for such a purpose. Means were calculated by different criteria that had been collected with the artifacts either as provided by districts or elicited through the coversheet or short teacher survey (See Appendix G). Due to the inclusion of artifacts from other, older studies, not all information was available for all artifacts, and this often resulted in the number of artifacts included for these analyses as being smaller than the overall N of 192. However, many of the older studies did include teacher commentary

and annotation, and classifying information could be coded from these supplementary materials.

Although it is beyond the scope of this study to make inferences as to the alignment of subgroups of artifacts due to the selection bias in the sample, it is of interest to determine whether our instrument would be able to detect differences in a more random sample of the population. In order to investigate that potential, selected groupings from the descriptive section were submitted to the Wilcoxon–Mann–Whitney (WMW) 2-sample rank sum test. This is a nonparametric measure for skewed, small sample data and is appropriate for ordinal scores. It tests for equality of central tendency of the two unpaired distributions. First, all scores are ranked regardless of which sub-group the observation is from. The WMW then determines whether or not we can reject the null hypothesis that the two groups median ranking are the same in favor of the alternative hypothesis that one sub-group's median ranking is higher than the others. Each topic was tested by dimension to determine if it was significantly different from the group of artifacts not associated with that topic (e.g. Engineering v. not Engineering).

Task Characteristics. First, dimensions were subsetted by general content and then the grade level in which the task was assigned. The 16 artifacts in which the two raters disagreed as to the general Disciplinary Core Idea were not included in the analysis by topic. While in some cases, the classes represented mixed grades, the artifact was coded for the most predominant grade represented (e.g. more than 50% of the students were in this grade level as reported by the teacher).

There have been concerns about whether all science topics can be taught through the more inquiry-based type instruction emphasized in the new standards (Llewellyn,

2002; NRC, 2012). For this reason, it would be important to have an indicator measure that is sensitive to any differences by topic. Within our sample, Engineering Design and Solution tasks tended to be more aligned with standards than tasks in other domains, particularly in the *Investigation* and *Explanation, Argumentation and Solution Design* dimensions. This is in line with other work on the value of including engineering tasks in science education, which have found that "…engineering design experiences provide engaging experiences for students that help them develop science concept knowledge and higher order thinking skills such as analysis and synthesis" (Cantrell, Pekcan, Itani, & Velasquez-Bryant, 2006, p.307) and that "…achieving design challenges has the potential to afford exploration of issues important to understanding science concepts" (Hmelo, Holton, Kolodner, 2000, p. 252).

The interaction of grade level and quality of alignment is of interest due to previous findings that middle school students often experience science instruction as rote memorization through interaction with textbooks, while high school science tends to be more lab-based (Settlage & Meadows, 2002; Shaver et al., 2007). Results are represented in Tables 7 and 8 below. There appears to be a non-significant trend in this sample that the opportunity to work with data increases as does grade level. Overall, Grade 9 students, which can be the final year of middle or junior high school or the first year of high school, received more aligned work than students in earlier grades, although the practices included in this study are specifically designated for middle schoolers, with the exception of designing their own investigations, which appears in our sample to occur more often in Grade 6.

Table 7
Means by General Subject

	PHYS N=68	LIFE N=61	EARTH N=30	ENG N=18
INV	.54 (.64)	.84 (.82)	.60 (.76)	1.80*** (.62)
DAE	.62 (.73)	.84 (.91)	.42 (.66)	.89 (.92)
EASD	.98 (.89)	1.61 (.97)	1.32 (.83)	2.22*** (.77)

*** p<.001 from Wilcoxon test statistic that Engineering tasks differs from the combined set of other topics. p<.005 when Bonferroni correction for multiple comparisons applied.

Table 8
Means by Grade

	Grade 6 N=27	Grade 7 N=13	Grade 8 N=71	Grade 9 N=35
INV	1.04 (.80)	.69 (.88)	.75 (.75)	.97 (.99)
DAE	.33 (.44)	.58 (.67)	.63 (.78)	1.23*** (.93)
EASD	1.54 (.91)	1.42 (.70)	1.29 (.98)	1.86*** (.96)

*** p<.001 from Wilcoxon test statistic that 9th grade differs from the combined set of other grades. p<.005 when Bonferroni correction for multiple comparisons applied.

Next, means were grouped by type: whether the artifact was identified as an assignment or a summative assessment. As summarized in a recent European study of the importance of changing assessment, "…current assessment methods have a strong emphasis on knowledge recall and do not sufficiently capture the crucial skills…of key

competencies" (Finlayson, McLoughlin & McCabe, 2015, p. 227). Because there were two assignments and two assessments from the recruited teachers, as well as assignments and assessments from the same teacher in the older data, it was possible to do a matched comparison by teacher. In total, there were 61 pairs of assignment/assessments matched by teacher. Results are in the third column of Table 8. For all dimensions, assessments were found to be significantly less aligned with standards than assignments. Most notably, for *Explanation, Argumentation, and Solution Design*, there appear to be more opportunities for students to explain, argue, or propose a solution outside of a test or quiz. There is also a marked difference in terms of *Investigation*, which is not unforeseen, as students are unlikely to be asked to formulate their own questions on an assessment. These differences were supported by an examination of the artifacts themselves, which often involved rich tasks such as labs and projects classified as assignments, and then recall –based tests and quizzes, often given as machine scoreable bubble sheets.

Table 9
Means by Type

	Assignment (N=109)	Assessment (N=78)
INV	1.15 (.76)	.33*** (.60)
DAE	.74 (.88)	.62** (.72)
EASD	1.70 (.89)	1.03*** (.93)

***p<.001 **p<.01 from WMU test statistic, paired analysis (N=61)

Artifacts were then classified as being collaborative or individual work. Collaborative work has found to be more constructive and to lead to better learning outcomes (Chi & Wylie, 2014). An indicator instrument would need to be sensitive to such a difference. In order for an artifact from our sample to be identified as collaborative, the teacher needed to indicate that the work was completed by a pair or group, which could be either created by the teacher or self-managed by the students themselves. For INV and DAE, the score more than doubled when the work was collaborative and for EASD nearly doubled. Results are shown in Table 9.

Table 10
Means by Collaborative v. Individual Work

	Collaborative Work N=55	Individual Work N=120
INV	1.40 (.76)	.55*** (.71)
DAE	1.06 (.92)	.46*** (.66)
EASD	2.05 (.75)	1.17*** (.94)

***p<.001 from WMU test statistic

A similar pattern was noted when artifacts were sorted by length of time given to complete and is shown in Table 11. In the realm of project-based learning, which incorporates many of the same elements as the tasks emphasized by NGSS, working over an extended period of time is critical (Thomas, 2000). In our sample, when a task was designed to be completed over more than a single class period, the quality in terms of alignment to standards increased for all three dimensions.

Table 11
Means by Single v. Multiple Class Sessions

	Single Class N=111	Multiple Class N=73	
INV	.42 (.56)	1.36*** (.80)	-
DAE	.49 (.63)	.92*** (.96)	
EASD	1.06 (.90)	1.91*** (.83)	

p<.001 from WMU test statistic

Design Characteristics. Next, design practices were considered. Once tasks are identified as highly aligned, it will be of interest to stakeholders to understand the process that led to these tasks. For example, were they designed by a single teacher? pulled from certain textbook or internet resources? This section examines the reported task design practices in our sample. Artifacts were grouped by whether the teacher indicated that he or she had created the task or whether the task had been drawn from existing resources. These results are reported in Table 12 and no strong pattern is noted.

Another design practice considered was whether or not the teacher had worked singly or with collaborators to select, design, or adapt the task represented by the artifact. In view of the high alignment of *tasks that were designed **for** groups of students*, it is of interest to examine for similar trends in *tasks that were designed **by** groups of teachers*. These results are reported in Table 13 and show the instrument noting a weaker trend toward more aligned tasks when teachers worked together to determine the task.

Table 12
Means by Teacher Created v. Resourced

	Teacher Created N=54	Resourced N=88
INV	.87 (.94)	.89 (.77)
DAE	.66 (.90)	.76 (.82)
EASD	1.41 (.96)	1.64 (.91)

Table 13
Means by Individually v. Collaboratively Selected/Designed/Adapted

	Individually N=125	Collaboratively N=58
INV	.74 (.78)	.97 (.86)
DAE	.66 (.82)	.76 (.82)
EASD	1.26 (.95)	1.78** (.91)

**p<.01 from WMU test statistic

In acknowledgement of the increasing role the internet continues to play in all parts of education (Fermin & Koch, 1996), teachers were also asked whether they had

used internet resources in designing the artifact. These results are shown in Table 14 and no significant trend was found.

Table 14
Means by Internet Resource Used

	Internet (N=47)	Non-internet (N=91)
INV	1.06 (.78)	.76 (.79)
DAE	.85 (.87)	.61 (.81)
EASD	1.76 (.91)	1.45 (.90)

One proposed purpose of the instrument is to track change over time. To investigate its sensitivity to a hypothesized change in alignment between pre and post standards artifacts, scores were sorted into these two groups. Results are summarized in Table 15 and indicate that the instrument is detecting some difference between the two groups, particularly for the EASD dimension.

Table 15
Means by Pre v. Post Standards

	Pre N=76	Post N=115	WMU Test Statistic	
INV	.54 (.64)	.99 (.86)	3.625***	
DAE	.53 (.68)	.78 (.88)	1.820	-
EASD	.87 (.84)	1.75 (.90)	6.083***	

***p<.001

Demographic Characteristics. Next, means were grouped by the limited teacher level and demographic information available, as an indicator measure may be used by stakeholders to understand equity of access to aligned assignments for all students. For the 25 teachers who were recruited specifically for the artifact study, detailed information on experience, education, and classroom demographic information was available, and will be described as a case study in a subsequent paper. However, for the larger sample, there is limited information, and artifacts can only be categorized by presence or absence of Students with Disabilities (SwD) in the classroom, presence or absence of English Language Learners (ELL) in the classroom, and high/med/low prior achievement levels (ACH). Again, the hypothesis here based on prior research is that we would expect differences to exist and the intent of this sub-grouping is to see if the designed instrument

is sensitive to such differences. Results are summarized in Tables 16-18 below. There appears to be a very small trend in the predicted directions for inclusive classrooms on the dimension of *Investigation*. That is, the means for artifacts given in classrooms that include students with disabilities or in classes labeled as remedial are slightly lower than means for their counterparts. The artifacts for the remedial classes were also less aligned for the EASD. No differences were noted for artifacts from classrooms with English language learners. However, the sample and identification of class characteristics was somewhat problematic, and the results may not be interpretable. It may be that the sample is too flawed or that the instrument is not sensitive to demographic differences rather than that there is little difference between these target sub-groups.

Table 16
Means by Inclusive Classrooms

	Inclusive N=61	Non-Inclusive N=92
INV	.70** (.75)	.92 (.85)
DAE	.54 (.75)	.78 (.84)
EASD	1.40 (.84)	1.59 (1.00)

**p<.01 from WMU test statistic

Table 17
Means by Classrooms with English Language Learners

	Presence of ELL N=39	Non-Presence of ELL N=147
INV	.65 (.61)	.84 (.86)
DAE	.44 (.64)	.75 (.85)
EASD	1.19 (1.00)	1.45 (.95)

Table 18
Means by Prior Achievement Level of Class

	High (described as Honors or Accel) N=24	Med (no designation) N=122	Low (described as remedial or below proficient) N=18
INV	.79 (.85)	.93 (.83)	.37** (.40)
DAE	.65 (.81)	.74 (.85)	.72 (.73)
EASD	1.40 (1.07)	1.61 (.92)	.95** (.86)

**p<.01 from WMU test statistic comparing Low with not Low
p<.05 when Bonferroni correction is applied

Analysis of Feasibility

Because this study involves a feasibility component, rater scoring timestamp data was converted to duration of scoring per artifact per dimension. Results are shown in Table 19. Raters were asked to identify *Disciplinary Core Ideas* at the General and Specific levels, and to provide justification for their choices while scoring *Investigation*, so that dimension was therefore more time consuming. One rater worked at a slower average pace than the other two, which could indicate that the Average Across Raters is a high estimate of the time per dimension. Also, rating time decreased as raters scored further, indicating a learning curve. For example, the time for rating EASD for Rater 1 averaged 2 minutes and 57 seconds for the first 10 artifacts scored but 1 minute and 58 seconds for the last 10 artifacts scored.

Table 19
Time to score (in minutes)

	Rater 1	Rater 2	Rater 3	Average Across Raters
INV	4.90	6.85	7.90	6.51
DAE	4.22	8.43	3.82	5.43
EASD	2.78	6.20	3.48	4.23

Discussion

Drawing from Mislevy's and Riconscente's work on Evidence-Centered Design (2006) as well as the Rational Empirical Strategy of Test Construction, the Artifact Indicator Protocol for Science content and practice was designed, based on the Next

Generation Science Standards, design criteria extracted from a thematic synthesis of existing artifact protocols, and interviews with experts. This following questions were addressed in the results and are discussed here:

- *To what extent can the protocol be used to measure classroom practice articulated in science standards?*

The Artifact Indicator Protocol for Science was able to describe classroom practice in terms of the NGSS. There are several components to consider in order to fully address this question. First, we determined that raters were able to utilize the protocol. The protocol was given to raters with expertise in middle school science and deep familiarity with NGSS. Training was conducted and then raters were asked to score 193 artifacts, one dimension at a time. Raters indicated during exit interviews that the provided rubrics, guiding questions, examples, and anchor papers made scoring manageable. Timestamp data indicated that raters' perceptions of scoring duration per artifact and actual duration was similar, and was approximately 5 minutes or less for *Data Analysis and Evaluation* (DAE) and *Explanation, Argumentation, and Solution Design* (EASD), and slightly longer for *Investigatio*n (INV) including *Disciplinary Core Ideas* (DCI). One rater, who was older and less comfortable with the technology involved in scoring (e.g. google forms, dropbox) took more time to score, which indicates that selection of raters is a critical consideration.

Next, we must consider the reliability of the ratings. Rater reliability measures indicated that although there was less than optimal exact agreement, raters were at the same end of the range for artifacts more than 90% of the time. In the exit interviews, raters most often articulated difficulties deciding between rating an artifact as 2 (partial

practice) and 3 (complete practice). Their suggestions in terms of revision or clarification of wording will be incorporated in the revision of the protocol and training materials.

Once we have established that we can feasibly reach a stable rating, we can examine whether the instrument is able to capture evidence of the construct under scrutiny: the content and practices of the NGSS. First, we consider content coding. Here the convenience sample had a marked impact on the findings. Six of the 25 teachers who were willing to participate taught 9th grade biology, which increased the number of artifacts classified as Life Sciences. Further, at the finer grained levels, rater agreement decreased. More specific guidelines need to be developed if scoring content is of continued interest, and content may need to be scored on a separate form to lessen the burden to raters of scoring both content and a practice dimension in a single sitting. Consensus meetings to discuss disputed content may be of use before revising this portion of the protocol.

In terms of practices, there was more success in identifying meaningful differences among artifacts. The dimensional practice scoring did populate all 4 score points, although data distributions for INV and DAE were highly skewed at the low end of the scoring range. This is not unexpected, as NGSS has either not yet been implemented in all schools sampled, or is at the very early stages, so that high numbers of artifacts that lack alignment or are only aligned superficially (i.e. say "investigate" but really mean "follow directions") are not surprising. A more interesting preliminary finding is that for this sample, there were almost half of the sampled artifacts showing partial or complete alignment with the EASD dimension. Work as early as *Benchmarks for Scientific Literacy* (AAAS, 1993) in terms of Claims-Evidence-Reasoning as the basis

for inquiry-based learning seems to be already making inroads in the science classes sampled. This may be what teachers mean when they respond to new standards by saying "But I am already doing that."

- *To what extent is the protocol potentially sensitive to assignment characteristics that are of interest to stakeholders?*

The purpose of an indicator measure is to provide valuable information to stakeholders on factors of interest. To this purpose, data were collected through a digital coversheet, a teacher survey, and attached annotations. Means and standard deviations were compared by these criteria and it appeared that the protocol was able to detect some differences in the sample. Many of these were significant using the WMW statistic. Others did not show a difference. Three of particular interest are inclusion of SwD and ELL, as well as FRLP as an indicator of socio-economic status. Questions of equity in terms of opportunity to learn often use these measures to identify groups who may be underserved. However, in the scoring of the artifacts collected, none of these showed means that were different enough to merit further statistical analysis. This may result from a number of factors. First, the extreme skew and lack of variance in data for our sample in INV and DAE may make it difficult for the instrument to detect differences. Also, ELL, which can in some way serve as an indicator of minority enrollment, did show some differences, but overall the lack of detailed data for many of the artifacts which could not be directly linked to a particular classroom may have obscured any significance.

Yet, when further exploratory analysis was done between two districts within the same state with very different student profiles (District A: wealthy suburban with 26%

minority enrollment, N=51 and District B: economically disadvantaged urban with 99% minority enrollment, N=34), the instrument did detect a significant difference using WMU analysis ($p<.04$) for EASD scores. In terms of explaining the significance of the finding for EASD in this case study district comparison, it may be related to the Anyon's classic study (1980) of differential tasks set for learners in different socio-economic classes. For students who are in a "working class" district, there may be little incentive to teach them to explain, argue or design solutions, as these are characteristics most often associated with white collar jobs.

Challenges and Limitations

The difficulties in recruitment that led to a different sample than initially desired were discussed earlier. It may be that concern about high-stakes evaluation or increased workloads contributed to low participation of teachers. The difficulty with recruitment, creating a convenience sample, will limit any inferences made from the study findings. That is, determining whether the instrument was able to detect differences in artifacts, is for a more site-specific sample than originally anticipated, and for a group of teachers whose characteristics may be unusual because they were willing to participate in the study.

There is also some initial indication that there are substantial challenges to use of classroom artifacts as part of a national indicator system, as the collection process was labor-intensive. Frequent correspondence and in-person follow-up were required to gain teacher compliance, and uploading of artifacts using existing technology is still somewhat cumbersome. Menial tasks, such as removing staples and feeding pages into a portable scanner, can be time-consuming and limit the feasibility of scaling up. However, one

tool that holds promise of streamlining the process is the sharing through personal cloud storage services, which two of the teachers used for sharing artifacts. While this is virtually effortless, there are some security issues that must be considered in terms of privacy.

It may be that artifacts' greatest affordances as an indicator are at a more local level. It appears that this instrument could be used a baseline for identifying and setting a school's or district's goals in terms of furthering alignment and then be used to monitor change over time. All raters indicated in interviews that the protocol could be useful for self-study or as a reference for teachers in choosing, adapting, or designing classroom tasks.

Conclusion

The Artifact Indicator Protocol-S is a protocol that was designed to measure content and practice alignment to NGSS standards of classroom tasks as represented in classroom artifacts. Quality in terms of intellectual demand, deep thinking, and challenge were included in the formulation of the three dimensions: *Investigation, Data Analysis and Evaluation, and Explanation, Argumentation, and Solution Design*. The protocols were than used to score a small convenience sample of Science tasks across 5^{th} to 9^{th} grade classes across several districts. Findings indicated that the instrument does hold promise as a tool for measuring alignment and for self-study by a school, department, or district. However, there are concerns for the potential of scaling up to a state or national level due to the arduous tasks still associated with collecting and scoring artifacts. Further research is needed in the potential affordances of technology to streamline the process to a more scaleable level. For example, could there be an algorithm written for preliminary

machine scoring of artifacts? Although a recent review by Shermis (2015) found that, to date, machine scoring of short constructed responses fail to agree with human raters, even the ability to sort out "0"s would significantly reduce the load on human raters. The argument could be made, however, that artifacts may capture information about science practices more objectively than self-report. For this reason, classroom artifacts are worthy of further consideration as part of an indicator system.

Additionally, the protocol shows some promise in terms of monitoring opportunities to learn for identified sub-groups. With sufficient and specific data about the makeup of classrooms, one could potentially apply the rubric to artifacts to ensure that high quality, aligned tasks are available to all learners. Using classroom artifacts and having a protocol available to gauge and track implementation of the reforms embodied in the Next Generation Science Standards would be of value to stakeholders, and could support the instructional practices that lead to a science-literate citizenry.

References

American Association for the Advancement of Science. (1993). Benchmarks for scientific literacy. Washington, DC.

Ananiadou, K., & Claro, M. (2009). 21st century skills and competences for new millennium learners in OECD countries.

Anyon, J. (1980). Social class and the hidden curriculum of work. *Journal of education*, 67-92.

Borko, H., Stecher, B., & Kuffner, K. (2007). *Using artifacts to characterize reform-oriented instruction: The Scoop Notebook and rating guide (CSE Technical Report 707)*. LA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/UCLA.

Borko, H., Stecher, B., Alonzo, A., Moncure, S., & McClam, S. (2005). Artifact Packages for Characterizing Classroom Practice: A Pilot Study. *Educational Assessment , 10* (2), 73-104.

Cantrell, P., Pekcan, G., Itani, A., & Velasquez-Bryant, N. (2006). The effects of engineering modules on student learning in middle school science classrooms. *Journal of Engineering Education*, *95*(4), 301-308.

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49, 219-243.

Clare, L., & Aschbacher, P. (2001). Exploring the technical quality of using assignments and student work as indicators of classroom practice. *Educational Assessment*, *7*(1), 39–59. doi: 10.1207/S15326977EA0701_5

Finlayson, O., McLoughlin, E., & McCabe, D. (2015). Strategies for the Assessment of Inquiry Learning in Science (SAILS) A European Project in Science Teacher Education. *New Perspectives in Science Education 4th Edition Proceedings*, p. 225-229.

Garland, R. (1991). The mid-point on a rating scale: Is it desirable. *Marketing bulletin,* 2(1), 66-70.

Hmelo, C. E., Holton, D. L., & Kolodner, J. L. (2000). Designing to learn about complex systems. *The Journal of the Learning Sciences*, 9(3), 247-298.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review*, *2*(2), 130-144.

Joyce, J. & Gitomer, D.H. A Thematic Synthesis of Artifact Research in Teaching Quality. (in preparation).

Joyce, J., Gitomer, D.H., and Iaconangelo, C. *Assessment of Learning and Teaching Through Quality of Classroom Assignments* (European Association of Research on Learning and Instruction-SIG 1 Assessment. Madrid, Aug 2014).

Llewellyn, D. 2002. *Inquire within, implementing inquiry based science standards*. Thousand Oaks, CA: Corwin Press.

Martínez, J. F., Borko, H., Stecher, B., Luskin, R., & Kloser, M. (2012). Measuring Classroom Assessment Practice Using Instructional Artifacts: A Validation Study of the QAS Notebook. *Educational Assessment*, *17*(2-3), 107–131. http://doi.org/10.1080/10627197.2012.715513

Matsumura, L. C., Garnier, H. E., Slater, S. C., & Boston, M. D. (2008). Toward measuring instructional interactions "at-scale". *Educational Assessment*, *13*(4), 267–300. doi:10.1080/10627190802602541

Matsumura, L., & Pascal, J. (2003). *Teachers' assignments and student work: Opening a window on classroom practice.* Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Mayer, V. J., & Tokuyama, A. (2002). Evolution of global science literacy as a curriculum construct. In *Global science literacy* (pp. 3-24). Springer Netherlands.

Means, B., Mislevy, J., Smith, T., Peters, V., & Gerard, S. (2016). *Measuring the Monitoring Progress K-12 STEM Education Indicators: A Road Map* . Washington, D.C.: SRI Education.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design. *Handbook of test development*, 61-90.

Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 Assessment Frameworks*. International Association for the Evaluation of Educational Achievement. Herengracht 487, Amsterdam, 1017 BT, The Netherlands.

National Research Council of the National Academies. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas.* Washington, DC: The National Academies Press.

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary educational psychology*, *36*(1), 36-48.

Preston, C and A. Coleman. "Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences." acta psychologica (2000): 1-15.

Serim, F., & Koch, M. (1996). *NetLearning: Why Teachers Use the Internet*. Songline Studios, Inc. and O'Reilly & Associates, Inc., 101 Morris St., Sebastopol, CA 95472.

Settlage, J., & Meadows, L. (2002). Standards based reform and its unintended consequences: Implications for science education within America's urban schools. Journal of Research in Science Teaching, 39, 114–127.

Shaver, A., Cuevas, P., Lee, O., & Avalos, M. (2007). Teachers perceptions of policy influences on science instruction with culturally and linguistically diverse elementary students. Journal of Research in Science Teaching, 44, 725–746

Shermis, M. D. (2015). Contrasting state-of-the-art in the machine scoring of short-form constructed responses. Educational Assessment, 20(1), 46-65.

Thomas, J. W. (2000). *A review of research on project-based learning*. Retrieved from: http://www.ri.net/middletown/mef/linksresources/documents/researchreviewPBL_070226.pdf.

Appendix A

Disciplinary Core Ideas Dimension of the Next Generation Science Standards (NGSS)

Physical Sciences	**Life Sciences**	**Earth and Space Sciences**	**Engineering Design**
• Matter and its interactions 1. structure and properties 2. chemical reactions 3. definitions of energy • Motion and stability 1. forces and motion 2. types of interactions • Energy 1. definitions of energy 2. conservation of energy and transfer 3. energy/force relationship • Waves and applications 1. wave properties 2. electromagnetic radiation 3. information technologies and instrumentation	• Molecules to organisms 1. structure and function 2. growth and development 3. organization for matter and energy flow in organisms • Ecosystems 1. interdependent relationships 2. cycle of matter and energy transfer 3. ecosystem dynamics, functioning, and resilience 4. biodiversity and humans • Heredity 1. growth and development of organisms 2. inheritance of traits 3. variation of traits • Evolution 1. evidence of common ancestry and diversity 2. natural selection 3. adaptation	• Earth's place in the Universe 1. stars 2. solar system 3. Earth's history • Earth's systems 1. Earth's history 2. Earth's material 3. Earth's water 4. plate tectonics • Earth and human activity 1. natural resources 2. natural hazards 3. human impact 4. global climate change	• Engineering design 1. defining and delimiting an engineering problem 2. developing possible solutions 3. optimizing the design solution • Links Among Engineering, Technology, Science, and Society

Appendix B

Scoring Practices

Artifact Indicator Protocol – Science: Investigation

Justification for this dimension comes from:
http://www.nextgenscience.org/sites/default/files/Appendix%20F%20%20Science%20and%20Engineering%20Practices%20in%20the%20NGSS%20-%20FINAL%20060513.pdf

Investigation: This dimension focuses on the extent to which students are required to ask questions, observe, experiment, and measure data, connecting the real world to their conceptual understanding of science idea(s). Artifacts that are high on this dimension ask students to develop their own investigations by drawing on conceptual understanding, to define real-world problems, and to obtain needed information in a systematic way from various resources including observations. Artifacts that ask students to complete tasks such as providing or selecting among definitions or carrying out a highly prescribed lab would be low on this dimension.			
0 – Absent	**1 – Surface Practice**	**2 – Incomplete Practice**	**3 – Developed Practice**
Students are not asked to generate any science-related questions, to plan or carry out an investigation, or to gather information independently. The teacher provides information through lecture or assigned texts.	Students are asked to carry out prescribed investigations in which all questions and steps are specified. Students may be asked to identify variables and to collect and analyze data, but there is no independent planning component. Students are asked to gather information from classroom texts and/or handouts without synthesizing information from multiple sources or evaluating the credibility, accuracy, or possible bias of information used.	Students are provided with the investigation questions and are asked to plan and carry out systematic investigations in order to answer the given questions. Students may be asked to identify variables and to collect and analyze data. Students are asked to gather, read, and synthesize information from multiple sources but are not asked to evaluate the credibility, accuracy, or possible bias of each publication and method used.	Students are asked to formulate questions and to plan and carry out an investigation that can be realistically accomplished, that identifies dependent and independent variables, and that produces data in a systematic way. Students are asked to gather, read, and synthesize information from multiple appropriate sources and to assess the credibility, accuracy, and possible bias of each source.

Artifact Indicator Protocol – Science: Data Analysis and Evaluation

Justification for this dimension comes from:
http://www.nextgenscience.org/sites/default/files/Appendix%20F%20%20Science%20and%20Engineering%20Practices%20in%20the%20NGSS%20-%20FINAL%20060513.pdf

Data Analysis and Evaluation: This dimension focuses on the extent to which students are asked to organize raw data including the identification of significant features and patterns, use mathematics to represent relationships between variables, and take into account sources of error. Artifacts that are high on this dimension ask students to display, analyze, interpret, and critique raw data or information using mathematical, computational, and statistical tools when appropriate. Scores at the low end of this dimension do not ask students to display, analyze, interpret, or critique raw data or information.			
0 – Absent	**1 – Surface Practice**	**2 – Incomplete Practice**	**3 – Developed Practice**
Students are not asked to display, analyze, interpret, or critique raw data.	Students are asked to display, analyze, interpret, or critique raw data by identifying simple patterns in given data representations.	Students are asked to display, analyze, and critique raw data and are asked to provide interpretations that address mathematical concepts such as similarities and differences but not causation or correlation.	Students are asked to display, analyze, interpret, and critique raw data quantitatively in order to determine patterns such as similarities and differences, causation, or correlation.
Students are not asked to engage in any mathematical, computational, or statistical thinking about data, or to consider sources of error.	Students are asked to apply given formulae but are not asked to engage in mathematical, computational, or statistical thinking or to consider sources of error.	Students are asked to represent data using mathematical or computational thinking in order to describe patterns in data but are not asked to apply mathematical or statistical concepts to support explanations or arguments or to consider sources of error.	Students are asked to use mathematical representations or computational methods to describe patterns in large data sets (possibly drawn from big data) and to use mathematical concepts to support explanations or arguments. There may be some consideration of sources of error.

Artifact Indicator Protocol – Science: Explanation/Argumentation and Solution Design

Justification for this dimension comes from:
http://www.nextgenscience.org/sites/default/files/Appendix%20F%20%20Science%20and%20Engineering%20Practices%20in%20the%20NGSS%20-%20FINAL%20060513.pdf

Explanation/Argumentation and Solution Design: This dimension focuses on the extent to which students are asked to analyze and/or represent situations and to develop and/or evaluate science arguments, explanations, and/or engineering solutions through written argumentation and/or development/revision of models. Artifacts that score high on this dimension require extended written communication to develop a science argument, explanation, or engineering solution description, with students engaging in theory and iterative model development as appropriate. Artifacts scoring low on this dimension do not require students to develop arguments, explanations, or engineering solutions or to engage in theory or model development.			
0 – Absent	**1 – Surface Practice**	**2 – Incomplete Practice**	**3 – Developed Practice**
Students are not asked to construct explanations or to design solutions.	Students are asked to construct explanations and/or to design solutions without specific supporting evidence, principles, and/or theory.	Students are asked to construct explanations and/or to design solutions with limited supporting evidence, principles, and/or theory.	Students are asked to construct explanations and/or to design solutions and provide strong evidence, scientific ideas, principles, and/or theory.
Students are not asked to interact with models.	Students are asked to identify or label parts of models, but they do not develop or revise models for reasoning about science or engineering concepts.	Students are asked to develop or describe models but are not asked to evaluate or revise models to explain, describe, test, and predict abstract phenomena and/or to design systems.	Students are asked to develop, evaluate, and revise models to explain, test, and predict abstract phenomena and/or to design systems.
Students are not asked to engage in any form of scientific argumentation about phenomena in the natural and/or designed world(s).	Students are asked to make a claim(s) but are not asked to construct an argument about phenomena in the natural and/or designed world(s) that uses empirical evidence, scientific reasoning, or design criteria when communicating scientific information.	Students are asked to construct an argument that supports or refutes claims for either explanations or solutions about the natural and/or designed world(s) using *limited* empirical evidence and scientific reasoning or agreed-upon design criteria when communicating scientific information.	Students are asked to construct an argument that supports or refutes claims for either explanations or solutions about the natural and/or designed world(s) using *strong* empirical evidence and scientific reasoning or agreed-upon design criteria when communicating scientific information.

Appendix C

Artifact Indicator Study Coversheet

* Required

1. **Please give a short name to your artifact**

2. **How would you categorize this artifact?** * *Mark only one oval.*
Assignment (classwork, lab, project) Assessment (quiz, test, exam)

3. **The task will be given in my class titled:** * (e.g., Earth Science, Algebra Accelerated)

4. **This task will be given to students in grade(s):** * Check all that apply
Check all that apply. 5 6 7 8 9 Other:

5. **Briefly describe the instructional goal.** *

6. **In creating this artifact, which of the following did you use?** * Check all that apply. *Check all that apply.* commercially published curricular material district resources department resources within my school collaboration with colleagues Internet resources Other:

7. **If you checked internet source above, please list the url here:**

8. **Students will be working on this task** * *Mark only one oval.*
individually. with a partner of their own choosing. in a group of their choosing. in a teacher created group. with a teacher assigned partner. unspecified

9. **How much time will be provided to complete the task?** * *Mark only one oval.* Part of a class period All of a class period 2-3 class periods This is a long-term task, which will be worked on both inside and outside of class.

10. **How many samples of student work are attached?** *

11. **Was this assignment given in the same format to all students?**
Mark only one oval. yes no

Appendix D: Artifact Scoring Sheets

Artifact Scoring Sheet- Sci Investigation

* Required

Coder Initials *

Your answer

Artifact ID *

Your answer

Disciplinary Core Ideas-General *

Choose

DCI-Specific *

Your answer

DCI justification *

Your answer

Investigation Scoring Elements *

	YES	NO	UNCLEAR
Does the artifact require the student to carry out an investigation or gather information from source(s) in order to answer a research question?	○	○	○
Does the artifact require the student to plan an investigation or obtain information from multiple sources?	○	○	○
Does the artifact require consideration of the credibility of information sources or data?	○	○	○
Does the artifact require students to develop their own questions?	○	○	○

Investigation Score *

Choose

Notes

Your answer

SUBMIT

Page 1 of 1

Artifact Scoring Sheet- Sci DAE

* Required

Coder Initials *

Your answer

Artifact ID *

Your answer

Data Analysis and Evaluation Scoring Elements *

	YES	NO	UNCLEAR
Does the artifact require the student to organize and display raw data graphically or to make use of a graphical display of big data?	○	○	○
Does the artifact require mathematical thinking to detect patterns?	○	○	○
Does the artifact require the student analyze and interpret data in terms of causal or correlational relationships?	○	○	○
Does the artifact require the student to apply concepts of statistics to characterize data?	○	○	○
Does the artifact require the student to consider sources of error?	○	○	○
Does the artifact incorporate the use of computational thinking, or use of digital tools?	○	○	○

Data Analysis and Evaluation Score *

Choose

Artifact Scoring Sheet- Sci EASD

* Required

Coder Initials *

Your answer

Artifact ID *

Your answer

Explanation/Argumentation/Solution Design Scoring Elements *

	YES	NO	UNCLEAR
Is the student asked to construct an argument, explanation or design solution?	○	○	○
Is the student asked to provide strong supporting evidence?	○	○	○
Is the student asked to create a model of phenomena in the natural or designed worlds?	○	○	○
Is the student asked to evaluate or revise a model of phenomena in the natural or designed worlds?	○	○	○

Explanation/Argumentation/Solution Design Score *

Choose

Notes

Your answer

SUBMIT

Page 1 of 1

Appendix E

Exit Interview Questions

Exit Interview-Science Raters

Overall Protocol

-How did you approach the scoring task? Can you talk us through your procedure in scoring?
-What role did student work play in your scoring?
-Which practice or component of a practice was more difficult to see in the classroom artifacts?
-Were there rubrics or elements of rubrics that were particularly challenging or stood out in any way? If so, in what ways?
-Are there materials you wished you had in scoring?

Investigation and Content Scoring

-About how long did it take you to an artifact score for this dimension?
-What were the challenges in scoring specific content?
-What were the challenges in scoring this practice?

Data Analysis and Evaluation

-About how long did it take you to score an artifact for this dimension?
-What were the challenges in scoring this practice?

Explanation, Argumentation, and Solution Design

-About how long did it take you to an artifact score for this dimension?
-What were the challenges in scoring this practice?

Final Questions
-In your opinion, would access to these protocols be helpful to teachers in planning or assessing instructional materials?

-What else do you think we need to know overall in order to improve the usability of the protocols?

Appendix F

Frequencies of Specific and Exact Content

Keyed to Protocol:

Specific Content

ECO=Ecosystem
ED=Engineering Design
EH=Earth and Human Activity
ES=Earth's Systems
EU=Earth's Place in the Universe
EVO=Evolution
HER=Heredity
MOL=Molecules to Organisms

Exact Content

Numbers indicate subtopic on protocol

Specific Content	Frequency	Percent
ECO	6	4
ED	19	12
EH	7	4
ENERGY	2	1
ES	14	9
EU	6	4
EVO	5	3
HER	9	5
MATTER	26	16
MOL	27	16
MOTION	12	7
NOAGREEMENT	27	16
NONE	3	2
WAVE	1	1

Exact Content	Frequency	Percent
ECO 1-4	5	3
ED1	1	1
ED1-3	4	2
ED2	1	1
ED2,3	5	3
ED3	3	2
EH3	4	2
EH4	2	1
ENERGY1	1	1
ENERGY3	1	1
ES1	1	1
ES1,4	1	1
ES2	2	1
ES3	8	5
ES4	2	1
EU1	3	2
EU1-3	1	1
EU2	2	1
EVO1-3	1	1
EVO1	1	1
EVO2	1	1
EVO2,3	2	1
HER2	3	2
HER2,3	4	2
HER3	1	1
MATTER	1	1
MATTER1	11	7
MATTER1,2	5	3
MATTER1,3	1	1
MATTER2	6	4
MOL1	18	11
MOL1,3	3	2
MOL3	3	2
MOT1	13	8
NOAGREEMENT	39	24
NONE	3	2
WAVE1	1	1

Appendix G
Teacher Survey

There have been recent changes in different curriculum standards for math and science. The questions below ask for your input on how these changes may have affected your teaching.
Individual answers will be kept confidential and shared only with the study team, not with your school or district.
In the last 3 years, have you participated in any professional development courses or programs concerning math and science standards?
_____ Yes No
If so, please describe below:

Title	Duration	Delivery model (on-line/short course/seminar)	Content covered	Offered by

_____
Please select one answer for each of the following questions about college and career readiness standards.

Strongly Disagree/ Disagree/ Neither Agree nor Disagree/ Agree/ Strongly Agree

I am very familiar with college and career readiness standards in my subject area.

My district has changed its curriculum to align with college and career readiness standards.

My district has provided support for teaching to college and career readiness standards.

My school has provided support with college and career readiness standards.

My classroom teaching has changed to align with college and career readiness standards.

_____
We consider assignments to be the instructional tasks that teacher present to students. Please select one answer for each of the following questions about assignments:

Never/ Occasionally/ Fairly Many Times/ Very Often/ Always

I have control over what classroom assignments I give my students.

I develop my own assignments.

I develop assignments with colleagues in my school.

I use assignments that are given to me by my school or district.

I use assignments that have been developed by publishers or professional curriculum developers.

I choose assignments from Internet sources.

Please complete the following section to help us understand your teaching background. Some of the questions ask about previous courses and academic work; please answer to the best of your recollection. Please note: the information here will be used to understand the background of teachers overall. No evaluations are being made.

Please list the courses you are currently teaching.Name of class (e.g. Earth Science, Algebra Accel, Math 5)

What was your undergraduate major?

Do you have a graduate degree?
Yes
No

If you have a graduate degree, what degree did you receive? (e.g. MS in Biology, MA in Math Ed)

How many years of teaching experience do you have?

How many years of teaching do you have in your current subject?

How many years of teaching do you have in your current subject at your current grade level?

Which certifications or endorsements do you hold?

Using Classroom Artifacts to Assess Enacted Math Standards:

The Artifact Indicator Protocol Study

Jeanette Joyce

October 2017

Abstract

There has been a call for math education reforms over the past century. The 21st century competencies described in the most recent phase of reform are represented in new standards (such as the Common Core State Standards for Math or CCSS-M). However, it is not enough to develop reforms through publication and legislation. It is important for stakeholders to understand how these policies are making their way into classrooms. In 2013, a NRC report called for a national indicator system that could be used to support the improvement of STEM education. This study explores how classroom artifacts could be used for such a purpose. Through literature synthesis and semi-structured interviews with eight experts in standards, artifacts, and large-scale data collection, a math-specific artifact measurement protocol was designed to serve as an indicator of both content coverage and practice alignment, The Artifact Indicator Protocol-Math is designed to assess the quality of classroom assignments and assessments (artifacts) with respect to a set of dimensions that are aligned with new standards for math education, such as those contained in college and career readiness standards. In order to gather empirical evidence for the soundness of the instrument, a study was conducted during the 2015-2016 academic year, with goals of feasibility of use and sensitivity of ratings to factors that were deemed likely to be of interest to stakeholders. Findings indicate that the developed protocol may be helpful in identifying the extent to which teachers are setting tasks for students that are aligned with CCSS-M.

Background and Purpose

There has been a call for reforms in math education for the past century. These calls have fluctuated between "back to basics" movements and progressivist movements for discovery-based pedagogy. One such movement emerged with the 1989 National Council of Teachers of Mathematics (NCTM) publication of standards which sought to "create a coherent vision of what it means to be mathematically literate both in a world that relies on calculators and computers to carry out mathematical procedures, and in a world where mathematics is rapidly growing and is extensively being applied to diverse fields," and to "create a set of standards to guide the revision of the school mathematics curriculum and its associated evaluation toward this vision." (p. 1). In 2000, the NCTM released a revised document, *Principles and Standards for School Mathematics,* an attempt to integrate both progressive pedagogy and basic algorithmic proficiency.

Most recently, continued low US achievement on measures such as Trends in International Mathematics and Science Study (TIMMS) and National Assessment of Educational Progress (NAEP) has spurred further reforms (Nord et al, 2011). The Common Core State Standards in Math (CCSS-M) were developed and are in part a response to the Carnegie Foundation report (2009) stating the future economic growth in the U.S. was dependent on improvement in math education.

The 21st century competencies described in this most recent phase of reform are represented in new standards (such as the Common Core State Standards for Math) in order to develop global citizens. According to Ananiadou & Claro (2009) in their Organisation of Economic Co-operation and Development (OECD) report, "Developments in society and economy require that educational systems equip young

people with new skills and competencies, which allow them to benefit from the emerging new forms of socialisation and to contribute actively to economic development under a system where the main asset is knowledge"(p. 5). In this case, the report is mandating a level of mathematical proficiency that will be essential for students' future contributions.

However, it is not enough to develop reforms through publication and legislation. What matters is how reform policies are interpreted by teachers and enacted in classrooms. This is in line with what Lipsky referred to as "street-level policy" (in Gibson, 2015), wherein ideas would be re-interpreted as they move from the halls of legislature to the halls of schools. Capps, Shemwell, and Young (2016) report that teachers can misunderstand new reforms and self-report that they are in compliance when tasks set for students are not truly aligned with standards. However, Bismack, Arias, Davis, and Palinscar (2014) found that, with support, teachers were able to incorporate new standards into classroom instruction. The challenge for stakeholders becomes how to elicit evidence of how standards represented in policies are actually being enacted in classrooms. Therefore, it becomes essential to have measures of how new reforms are reaching students and whether progress is being made toward reform goals. Such a set of measures would form an indicator system.

In 2013, a NRC report, *Monitoring Progress Toward Successful K-12 Education: A Nation Advancing?,* called for a national indicator system that could be used to support the improvement of STEM education. The report described 14 Indicators that were needed to guide improvement. Congress then directed the NSF to begin implementing a progress monitoring system for the indicators. In response, there is a call for development of new instruments to be used in an indicator system. "A monitoring and reporting

system designed around these indicators would be unique in its focus on key aspects of teaching and learning and could enable education leaders, researchers, and policy makers to better understand and improve national, state, and local STEM education for all students" (National Research Council, 2013, p. 3). The call is for an indicator system to describe the implementation of new college and career readiness standards into daily classroom tasks (Committee on the Evaluation Framework for Successful K-12 STEM Education; National Research Council, 2013; Means, Mislevy, Smith, Peters, & Gerard, 2016). One of these indicators that was identified as a priority was Indicator #5: *Classroom coverage of content and practices in CCSS*.

The indicator measures would serve a different purpose than that which has been previously used in teaching quality. There currently exists a body of work describing various types of evaluation of instruction, including large-scale indicators like NAEP. These evaluations often make use of student achievement measures, observational measures, and survey measures. Each of these can make a useful contribution to understanding what is happening in classrooms and the extent to which reforms are being implemented. Achievement measures can provide information on student mastery of content and practices. However, achievement measures can lag behind reform initiatives, and therefore may not assess reform-related curriculum effectively (Buckendahl, Plake, Impara, & Irwin, 2000; Martone & Sireci, 2009). Observations provide information on instructional exchanges between students and teachers, and allow for assessment of discourse. Finally, self-report through survey can be an indication of teachers' perception of their own practice. While these traditional methods can provide some insight into what is happening in classrooms, there is the potential for new measures that

shed light, particularly in terms of instruction around mathematical practices.

As new standards, such as the Common Core, strive to represent 21st century competencies, there is a need to investigate methods to develop a scalable system of indicators. These indicators could then provide insight for stakeholders into the *range* and *quality* of alignment with emerging college and career readiness standards. In response to this challenge, there is a call for development of new instruments that could be used in such an indicator system. "A monitoring and reporting system designed around these indicators would be unique in its focus on key aspects of teaching and learning and could enable education leaders, researchers, and policy makers to better understand and improve national, state, and local STEM education for all students" (National Research Council, 2013, p. 3).

This study will make use of classroom artifacts as the primary source of evidence of implementation of mathematical standards for inclusion in an indicator system. Classroom artifacts are an important source of evidence of the quality of classroom interactions and can provide information that is not as accessible through other measures (Joyce, Gitomer, & Iaconangelo, 2014). Classroom artifacts, which include both assigned tasks from teachers and student work, are very useful in providing clear evidence about the nature of expectations that are held for students in terms of new standards, content, reasoning, and communication.

In order to meet this new challenge of assessing mathematical standards alignment, this study proposes the design of an artifact-based indicator measure that can be used as evidence of implementation of instructional practices, as well as content, embodied in new college and career readiness standards. The second phase of the study

involves piloting key components of the system to assess feasibility for use as either part of a large-scale survey or as a tool for states and districts to engage in self-study, and for researchers to analyze STEM teaching.

Classroom artifacts, such as in-class assignments, tests, and projects, hold unrealized potential as one aspect of an indicator system. In an indicator study, the targeted inference would not focus on relative strengths of individuals, as was the case in earlier studies. Previously, classroom artifacts have been used to make inferences at the teacher, student, or classroom level. Prior research provides the following foundational assertions:

- Artifacts can provide insight into classroom practices and interactions and, as such, give evidence of enacted policies (Borko, Stecher, & Kuffner, 2007; Matsumura & Pascal, 2003).
- Artifact study findings are confirmatory with observation results (Joyce, Gitomer, & Iaconangelo, 2014).
- Artifact studies found assignments to lack the deeper learning associated with high level cognitive demand (Joyce, Gitomer, & Iaconangelo, 2014).

This study seeks to extend previous work to use of classroom artifacts to use as an indicator within a system. One definition of an indicator, given by the European Commission on Public Health, "…is a quantitative or qualitative measure of how close we are to achieving a set goal or policy outcome. They help us analyse and compare performance across population groups or geographic areas, and can be useful for determining policy priorities"[5]. Indicators have historically utilized data from surveys,

[5] Retrieved from http://ec.europa.eu/health/indicators/policy/index_en.htm

rather than the more time and labor intensive methods such as observations. The current study will assess the feasibility of artifacts to be incorporated into an indicator system.

I undertake the design of an indicator using classroom artifacts as a measure of "the extent to which the instruction and learning activities students experience in a classroom cover content in a set of standards, are consistent with the performance-level expectations of those standards, and reflect the same conception of learning and instruction… capturing the enacted curriculum" (Means, Mislevy, Smith, Peters, & Gerard, 2016, p. 24). Specifically, the study examines the potential for classroom work to serve as an indicator of both the current level and measured progress for the implementation of math content and practices described in new standards.

This study tests the hypothesis that classroom artifacts may capture evidence that other indicators are not equipped to provide and shed light in understanding how teachers are interpreting standards-based curriculum for their students. Clearly, there are limitations. For example, a classroom artifact would not necessarily provide insight into appropriate use of physical math tools or how fluently a student is able to access correct procedures. However, a protocol that identifies and codes practices that are captureable through classroom artifacts, such as the use of strategies as math tools or the accuracy with which students select and apply math procedures, could be useful in assessing the status of new standards' influence on classroom instruction. The study presents a math-specific protocol that provides information to stakeholders on alignment to emerging standards (Common Core Math, 2011), including both content and practices, and is related to measures included in TIMSS (Mullis et al, 2009). The research questions include:

- *To what extent can the protocol be used to measure classroom practice articulated in math standards?*
- *To what extent is the protocol sensitive to characteristics of instruction that may be of interest to stakeholders?*

Methods

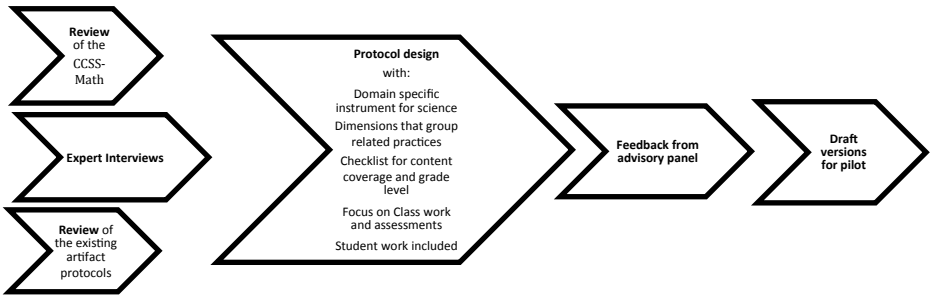


Figure 1. Summary of study methodology.

Initially, currently available artifact literature was synthesized in order to better understand design criteria (Joyce, 2017). Next, semi-structured interviews were conducted with eight experts in standards, artifacts, and large-scale data collection, before undertaking the design of a math-specific artifact measurement (Figure 1). The information from the synthesis study and the interviews were then refined into design criteria that were incorporated into the domain-specific rating instrument. The protocol was designed to capture the kinds of understandings and practices embodied in the Common Core Math documentation. In particular, dimensions, scale ranges and scoring procedures were considered.

Instrument Design

The instrument development drew from both Mislevy's and Riconscente's work on evidence centered design (ECD) (2006) and the Rational Empirical Strategy of Test Construction(RESTC). Mislevy and Riconscente indicate that in any instrument design, the initial stages, or layers, must include domain analysis and modeling. In these initial stages, the researchers "gather substantive information about the domain of interest" and "express [the] assessment argument in narrative form" (2006, p. 67). The domain analysis and modeling was driven by the literature on artifact research as synthesized (Joyce, 2017), through an in-depth review of the standards including the literature from which they emerged, and through expert interviews. I used the Rational Empirical Strategy of Test Construction to guide instrument development. This method, as used by researchers such as Reinhart Pekrun in instrument development for motivation research, specifies that, valid instrument development involves theoretical justification, evidenced design process, and empirical indication of test validity and reliability. The theoretical base was drawn from the literature on protocol development, and extended through the expert interviews.

Experts in the areas of classroom artifact research, large-scale data collection and management, and math teaching were interviewed beginning in February 2015. In total, eight semi-structured interviews were conducted. Points of convergence from the interviews and the advisory panel are summarized in Table 1, and informed the development of the instrument. This follows the line taken by Denner, Salzman, & Bangert (2001) in their work developing an instrument to assess Teacher Work Samples by defining "indicators of the standards that our professional community agreed provided

the evidence of performance one would look for to evaluate whether or not the targeted standards were met"(p.289).

The next decision was how to create dimensions that were representative of Common Core standards. In designing the protocol, I considered not only what was critical in terms of the standards but also what is able to be seen in artifacts, as well as what may be trackable over time. Specifically, key considerations in the protocol design included attending to aspects of the standards for which artifacts provide evidence, accommodating the simultaneous independence and overlap of standards, accounting for cognitive and time demands on raters, accommodating likely variability in fidelity of responses to artifact study instructions by participating teachers, and, finally, clarifying consideration of student work in determining ratings.

Table 1
Points of Convergence from Interviews

General	Sampling	Scoring
•Artifacts can give important insight.	•Assessments and in-class work are more useful than homework and lesson plans.	• Raters should have teaching experience and extensive training.
•Not all standards lend themselves to artifact study.	•Multiple artifacts across the school year are needed.	•Dimensions should be limited and clearly defined.
	• Student work is critical.	•Separate content and practices.

It was decided to first separate content and practices. A content checklist by grade level was developed. This involved consideration of what the broad content was (e.g. functions, expressions) as well as what the CCSS grade level was associated with

the specific task. For the purposes of this study and the potential for an indicator system, it was determined that information was needed about what students are actually being asked to do within their current classrooms, whether it is "on grade level" or not. For this reason, dimensions would contain grade level information, but also look at the content and practices as they are being incorporated into classrooms. While other studies (Polikoff, 2015; Schmidt & Houang, 2014) are currently examining the math content in available published materials and its alignment to standards, this study will lend valuable insight into how these resources are informing the tasks that students are actually being asked to do.

For practices, an effort was made to cluster the eight practices listed in Table 2 below into meaningful topics for ease of scoring and interpretation, rather than to score individual practices. This was done as an acknowledgment that completely separating practices is somewhat artificial, and that not all practices can be found through artifact study. These dimensions drew from the NAP publication *Adding It Up: Helping Children Learn Mathematics* (2001). This publication identifies the five strands of mathematical proficiency and emphasizes how these "are not independent; they represent different aspects of a complex whole. The most important observation we make here…, is that the five strands are interwoven and interdependent in the development of proficiency in mathematics"(p. 116). The strands are represented graphically in Figure 2. Thus, the choice was made to group practices roughly by these strands into four dimensions, with some overlap, rather than to tease them apart and evaluate the math practices one by one. The current protocol is described below and included in the appendix.

Table 2

Common Core State Standards: Math Practices[6]

1. Make sense of problems and persevere in solving them.
2. Reason abstractly and quantitatively.
3. Construct viable arguments and critique the reasoning of others.
4. Model with mathematics.
5. Use appropriate tools strategically.
6. Attend to precision.
7. Look for and make use of structure.
8. Look for and express regularity in repeated reasoning.

A decision was made in terms of developing a scale that could be used to describe an artifact in terms of the CCSS through the conceptualized dimensions. It was hypothesized that three levels of scoring (absent, partial, complete) would be too coarse grained, not allowing for refinement of levels of partial practice. In this early stage of standard implementation, it is critical to have information as to the extent of partial execution. Although wider ranges of scoring have been found to have higher reliability in surveys (Preston & Colman, 2000), it was felt that seven levels of scoring might be too burdensome for raters so instead a uniform four point score range was developed for each dimension, similar to that used by Denner et al (2001). This range also avoids any tendency for raters to drift toward middle (Garland, 1991). The lowest level is set at zero (or absent) to increase interpretability of scoring and of analysis. Levels of scoring are

[6] (retrieved from http://www.corestandards.org/Math/Practice/)

described below but generally follow 0 (absent), 1 (superficial), 2 (incomplete), 3 (present). The protocols and scoring guides were completed in August 2016. The pilot artifact protocol is described in detail below:

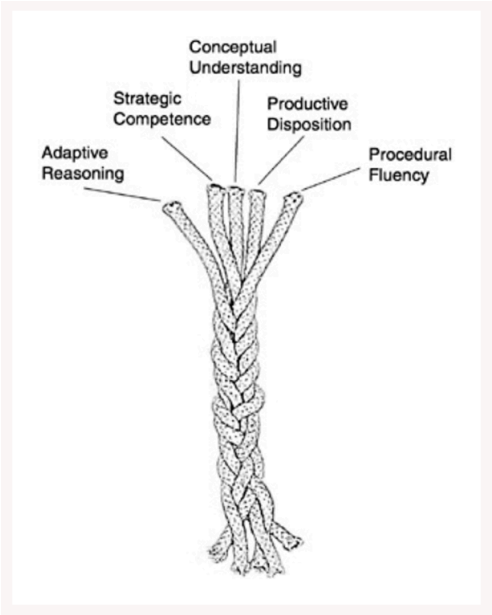


Figure 2. *A graphic representation of the 5 strands of mathematical proficiency. Retrieved from Adding It Up: Helping Children Learn Mathematics (2001).*

AIP-M. The Artifact Indicator Protocol for Math (AIP-M) is designed to assess the quality of classroom assignments and assessments (artifacts) that are aligned with new college and career readiness standards for mathematics education such as those contained in the CCSS framework or other state-specific standards that focus on similar sets of expectations for students. Artifacts include both the assignments and assessments as set by the teacher and the student work that is associated with these assignments and assessments. The AIP-M is designed to capture the extent to which students are asked to demonstrate mathematical proficiency in terms of both **content** and **practices**.

In the content dimension, there is scoring by topic and grade level. The current focus is on middle school math, a crucial transition period between mastering

computation and developing the abilities needed for the pursuit of higher level math. Topics include:

- Operations & Algebraic Thinking,
- Numbers & Operations in Base Ten,
- The Number System,
- Fractions,
- Measurement & Data,
- Geometry,
- Expressions & Equations,
- Ratios & Proportional Relationships,
- Functions, and
- Statistics & Probability.

For practice dimensions, the AIP-M draws on other protocols exploring the intellectual demands manifest in classroom artifacts. We code for four broad dimensions of mathematical thinking. Specifically, the AIP-M encapsulates the extent to which the content is addressed in terms of practices associated with *Conceptual Understanding* (CU), *Procedural Skills and Fluency* (PSF), *Application and Relevance* (AR), and *Argumentation and Communication* (ACom).

The *Conceptual Understanding* dimension focuses on the extent to which students are asked to provide explicit evidence of their conceptual understanding of mathematical idea(s) and contains elements of the following practices: *1. Make sense of problems and persevere in solving them*; *2. Reason abstractly and quantitatively;*

and *7. Look for and make use of structure* (Table 2). Artifacts that are high on this dimension ask students to provide evidence of their understanding of mathematics concepts as well as relationships among concepts; to represent their thinking about these concepts; to see connections with other ideas in mathematics; and to apply concept(s) to solve problems. Artifacts that ask students to do such things as routine solving of equations would be low on this dimension.

Specifically, for a score of zero (0), the task does not ask students to apply or provide explicit evidence of any conceptual understanding. This type of work gives a rote example of a procedure, and the student is asked to replicate this in subsequent problems. An example of this would be an assignment that asks students simply to find the volume of a cylinder with $r = 2$, $h = 4$ or to identify the slope in the line $y = -2x + 3$. For a task that shows "surface projects," at a score level of one (1), students are asked to apply concept(s) but are not asked to provide explicit evidence of their understanding. This type of work indicates which concept to access and does not require explanation or justification. To continue with our previous examples, this task would present a cylinder with measures given and then ask the student to find the volume of this cylinder or to calculate the slope of the line that contains the points (-2, 4) and (2, -4). As we move to "incomplete practice," scoring two (2), students are asked to apply concept(s) and provide explicit evidence of their understanding, but evidence is developed with scaffolding. In our cylinder example, the students would be asked to answer a question such as:

> Two cylinders are both 8 inches high. One has a volume of 32Pi cu in. What is its radius? The other has a volume of 128Pi cu in. What is its radius? Explain how the volume of a cylinder changes with change in radius.

or to approach a slope problem such as:

> Two lines have a slope of -2. One passes through the origin, and one passes through the point (0, 2). What is another point on each of these lines? Graph each, and explain how lines with the same slope are related.

The final level of scoring, "complete practice" or three(3) is for tasks that require Students are asked to both apply and provide explicit evidence of their understanding independently, such as explaining how the volume of a cylinder changes as the diameter of the cylinder changes, using examples, words, or drawings to explain the relationship. Alternatively, with the slope example, the students would be asked to explain how knowing the slope helps define the relationship between the lines, again, using examples, words, or drawings to demonstrate.

The *Procedural Skills and Fluency* dimension focuses on the extent to which students are asked to make sense of problems and persevere in solving them by using appropriate strategies, while attending to precision and conditions of use. This dimension considers elements of the following practices: *1. Make sense of problems and persevere in solving them*; *5. Use appropriate tools strategically; 6. Attend to precision;* and *8. Look for and express regularity in repeated reasoning* (Table 2). Artifacts that are high on this dimension ask students to understand procedures and conditions of use, to interpret procedural outcomes, to use multiple procedures in a coordinated manner, to describe these procedures and the solution path, and to check work. Artifacts that ask

students to do such things as routine solving of problems using a single, given procedure would be low on this dimension.

For example, at the level scoring zero, students are given a procedure to follow in order to solve problems. No sense-making is expected. Students are not asked to make a selection of a procedure to use. Tasks receiving a score of 0 do not ask students to choose procedures. This type of work gives an example of the procedure, and the student is asked to replicate this in subsequent problems. For example, an assignment might ask students to use the distance formula to find the length of a segment with endpoints of (2, 3) and (-7, 9). At the level of "surface practice," or one (1), the task requires students to select and implement single procedures and to describe a solution path OR to select, implement, and coordinate multiple procedures with no solution path. In this case, the exact procedure is not given and the student may be asked to complete tasks such as determining the length of the line segment with endpoints of (2, 3) and (-7, 9). As we shift to "incomplete practice" or score of two (2), students are asked, with scaffolding, to select, implement, and coordinate multiple procedures and to describe their solution path in order to make sense of problems. This type of work involves multiple steps, formulae, or calculations. An example would be:

> John needs to visit a store that is 30 miles north and 40 miles east of his home via the N-S highway and the E-W highway (speed limit, 45 mph) or via the direct local road (speed limit, 30 mph). Which is the faster route? Explain why. First, you will need to determine the length of the local road and then the time of travel for each.

Finally, "complete practice" or score of three(3), includes tasks that ask students to select, implement, and coordinate multiple procedures and to describe their solution path independently in order to make sense of problems. A specific example would be:

> A car and a bus set out at 2 p.m. from the same point, headed in the same direction. The average speed of the car is 30 mph—slower than twice the speed of the bus. In two hours, the car is 20 miles ahead of the bus. Find the speed of the car. Explicit checking of work may be required in order to attend to precision.

It is important to note that solution paths can be described in multiple ways. For example, this item would be a 3 if the student were asked to create a table or other representation of the solution path.

These first two dimensions are directly related to the single strands of mathematical proficiency (Figure 2). The final two, *Application and Relevance* and *Argumentation and Communication*, are my attempt to capture the kinds of mathematical thinking that are reflected in the standards for mathematical practice and overlap with all strands in Figure 2, particularly covering strategic competence, adaptive reasoning, and productive disposition. The *Argumentation and Communication* dimension is an attempt to capture the request for students "…to formulate, represent, and solve mathematical problems" and to demonstrate the "…capacity for logical thought, reflection, explanation, and justification", while the *Application and Relevance* dimension seeks to indicate the extent to which students are asked to develop a "habitual inclination to see mathematics as sensible, useful, and worthwhile…" (NRC, 2001, p. 117).

The *Argumentation and Communication* dimension focuses on the extent to which students are asked to develop and communicate mathematical conjectures and arguments, and contain elements of practices *1. Make sense of problems and persevere in solving them; 2. Reason abstractly and quantitatively; 3. Construct viable arguments and critique the reasoning of others;* and *4. Model with mathematics* (Table 2). Artifacts that score high on this dimension ask students to make mathematical conjectures, to develop a mathematical argument, and to communicate their thinking coherently to others, as well

as to evaluate the arguments and communication of others. Artifacts that ask students to perform tasks such as routine solving of equations without extended writing would be low on this dimension.

At the scoring level of zero (0), students are not asked to construct an argument. Writing does not extend beyond providing a simple mathematical or verbal solution. This type of work sets a task that does not require the student to go beyond selecting or calculating an answer. For example, students match equations with the mathematical property illustrated. At the level of "surface practice," with a score of one (1), **s**tudents are asked to use given assumptions or definitions to support a provided argument. Writing or representation is highly structured and/or constrained, such as, an assignment in which students are asked to choose the property that is illustrated by an equation and to explain their choice. As we move to "incomplete practice," score of two (2), students are asked to develop their own argument or to evaluate the arguments of others, but writing is not elaborated in a logical progression through use of counterexamples, alternatives, and/or representations. For example, a question that gives an irregularly shaded region and asks, if someone looked at the figure and determined that the probability of landing on the shaded area was ¾, would you agree? Why or why not? In the "complete practice" level with a score of three (3), students are asked to develop their own complete argument or to evaluate the arguments of others and are asked to develop a complete or elaborated communication. Artifacts at this level ask students to demonstrate reasoning through the use of counterexamples, alternatives, representations, and/or elaborated writing in a logical progression. In order for the example above were to be scored at this level, the student would be asked to describe the

error as well as to redraw the figure so that the answer was correct and compare the two figures.

The *Application and Relevance* dimension focuses on the extent to which students are asked to make sense of real world problems or to model real world situations using mathematical representations and reasoning, referring to elements in practices *1. Make sense of problems and persevere in solving them; 2. Reason abstractly and quantitatively;* and *4. Model with mathematics* (Table 2). Artifacts that are high on this dimension ask students to apply mathematics concepts to real world problems or to work with real world data. Artifacts that ask students to perform tasks such as routine solving of equations without real world context or with real world context that is superficial to the solution, would be low on this dimension.

For a score of zero (0), students are not asked to solve real-world problems or to work with real-world data. Tasks have no context. This type of work presents a problem that requires only procedural execution, such as to find (4 + 5 + 6 + 7 + 7 + 8)/6. As we move to level one (1), or "surface practice," students are asked to solve problems with a real-world context that is superficial and does not add critical information. Students may be asked to use a data set, such as a table that shows number of rock concerts attended next to names with the request to find the average number of concerts attended. Here, the task receives a score of 1 because it uses context only as a veneer. With a score of two (2) or "incomplete practice," students are asked to solve problems with a simplified real-world context but are not asked to model real-world situations. Students may be asked to use data but are not asked to address real-world problems. Tasks receiving a score of 2 present problems that are only academic in nature. such as asking student to interview

each other about the type, number, and ages of pets they have. There is no real world purpose behind the task. Finally, at level three (3) or "complete practice," students are asked to address real-world problems or to model real-world situations. Students are asked to use data from real-world contexts. Tasks receiving a score of 3 present problems that are relevant to their daily lives. For example:

> We want to add to our classroom storage cubes to hold backpacks and lunch boxes. We need to know how big to make the cubes. It is okay if some students cannot fit all of their belongings into the cube, but most of the students need to be able to store their stuff.
>
> Find out how much space each student will need, and share your data with the class. What is the storage cube size that makes the most sense? Explain why.

Together, these four domains articulate a range of ways that students engage in and demonstrate mathematical knowledge, reasoning and strategic competence.

Empirical Study

In order to gather empirical evidence for the soundness of the instrument, a pilot study was conducted during the 2015-2016 academic year with goals of feasibility of use and sensitivity of ratings to factors that were deemed likely to be of interest to stakeholders.

Data Sources and Sampling. The initial intent was to apply the protocols to artifacts collected from multiple districts, using the sampling guidance from the expert interviews (Table 1). To that purpose, IRB permissions were gained from ten large urban districts, allowing access to schools, but leaving participation to the discretion of the principal. In only two of the districts was there additional support from the central office. All in all, more than 300 schools were approached by mail and via phone calls, with positive responses from only five principals. Here, too, there was no offer of continued support,

but rather only permission to contact teachers. Again, hundreds of teacher letters were sent out, which resulted in recruitment of three teachers. Even when budget and IRB documentation were amended to include a $200 stipend for what was to be 15 -30 minutes of additional work outside of regular classroom duties, no further participation was gained. The approach to recruitment was then revised to be more personal, pursuing connections with local schools and contacting colleagues for artifacts from their current or past research.

The resulting sample included 79 artifacts collected from the 2015-16 academic year, and 83 artifacts from previous artifact studies (SCOOP, UTQ), collected before 2011, for a total of 162 unique artifacts in grades 6 through 9, across 5 states. Borko, Stecher, and Kuffner (2007) conducted the SCOOP study in order to use artifacts to characterize math and science classroom instruction to aid in assessing both the students' understanding as well as the teachers' own process and to explore the capacity of classroom artifacts as an indicator of reform-based instructional practices**.** The UTQ project, as described in Joyce, Gitomer, and Iaconangelo (2014), used artifacts in math as one of several attempts to attend to characteristics of teaching practice.

The classroom artifacts consisted of the assignment or assessment template (i.e., the blank form) as well as a selection of student work. 25% of these had only the template with no associated student work. The remaining artifacts had 1 to 13 samples of student work, with 2/3 of these in the 1-4 sample range.

In developing the sample, one important point of divergence considered emerged from expert interviews (Table 1): whether there would be affordances in requesting "typical" or "challenging" work from the teachers. The question was posed whether, in

an indicator study, one might be interested in the best the nation can produce (challenging) or what is pervasive in America's classrooms (typical). Based on the findings from the previous study (Joyce, Gitomer, and Iaconangelo, 2014) that there was only a slight difference in quality between typical and challenging math tasks as selected by teachers, and in the hope of capturing the high level of challenge embodied in new standards, it was decided by the study team, for the purposes of the pilot, to elicit challenging work.

Artifacts were categorized as either assignments or assessments. All artifacts were de-identified, coded for the state of origin, grade level, and date assigned, and then scanned to a secure server. Most of the current artifacts were the result of an ongoing state reform initiative that is grounded in a competency-based educational approach designed to ensure that students have meaningful opportunities to achieve critical knowledge and skills aligned with new standards in math. These artifacts were collected in March and May 2016, while others were selected from the previous studies (UTQ and SCOOP), which predate new standards.

Scoring. Raters were selected from a pool of nine recruited applicants from a call to Math Education graduate students and the State Association of Math teachers. Six applicants were interviewed, including commenting on an unscored artifact in terms of the CCSS for math, and then three were selected, all of whom had experience with the CCSS and with teaching middle school math.

Training consisted of an in-person all-day session in which raters become familiar with the protocol and guiding questions before scoring anchor artifacts both with the study team and independently. Prior to training, scoring guidance materials were

developed. These consisted of identifying critical components of the dimension and then creating focusing yes/no questions around these (see Appendix B). For example, a question for the *Conceptual Understanding* dimension was "Does the artifact require the students to complete the task independent of scaffolding?" In training, it was stressed that these questions were meant to facilitate focusing on the critical components, and not to translate into a score. That is, a certain number of "yeses" did not translate into a certain score. In fact, different dimensions had different numbers of questions. Raters did record their answers to each question, but the overall score per dimension required a more holistic decision related to the rubric. Key differences between score points were stressed, so that raters felt confident in deciding between a 0/1, 1/2, and 2/3. All disputes were discussed and raters were directed to explain their scoring in terms of the rubric.

Following training, one rater was replaced because of an inability to conform to the rubric in the protocol. Another applicant was then trained and then all three raters began scoring with frequent re-alignment and troubleshooting check-ins conducted via online video conferencing. It was possible to monitor online scoring in real time, and to note areas that needed further training. Additionally, timestamp information was collected in order to estimate time needed for scoring. Rater timestamp data was converted to duration of scoring per artifact per dimension. Averages were calculated per dimension after outliers (as defined by a duration that was more than twice as large as the mode) were removed. It was hypothesized that the excessively long durations indicated an interruption in scoring, and therefore did not accurately indicate the time to score. Between August and October 2016, each artifact was rated by two trained raters.

For content, raters were given the quick reference guide included in Appendix A as well as the complete Common Core content guidelines found here: http://www.corestandards.org/Math/. Using these references, raters were asked to choose the primary topic area, noting other topics covered, and then identify the predominant grade level associated with the task under that topic. If more than one grade level was equally represented, raters were asked to note the higher grade level, with the understanding that lower level skills are often needed to approach higher level tasks.

Scoring begins with the artifact itself and then considers whether accompanying student work indicates that the score is accurate or needs to be adjusted. That is, raters are asked to consider the task in the blank template, make a judgment, and then seek confirmatory evidence for their interpretation from sampled student work. If the samples indicate that the students are answering an implicit expectation, such as show your work, use examples, or create a representation, even though there is no implicit ask in the template, then the artifact score may be increased. For example, an artifact that asked students to explain but still gave credit for shallow answers would have its score adjusted down, while student work that showed students presenting representations without explicitly being asked would be adjusted upward. Scoring forms are included in Appendix B. All raters participated in a 30-45 minute exit interview after completing scoring. Interview questions are included in Appendix C.

Results and Analysis

Reliability

With any rater assessment based on a rubric come concerns about the reliability of the scoring. Jonsson and Svingby (2007) in their review of research involving scoring rubrics indicated that "Ideally, an assessment should be independent of who does the scoring and the results similar no matter when and where the assessment is carried out, but this is hardly obtainable "(p.133), acknowledging that high percentages of agreement are not common in this type of rating. They used Stemler's 2004 criterion of 70% or greater for exact agreement, and Stoddart, Abrams, Gasper, & Canaday's 2000 range of kappa values between .40 and .75 as "represent[ing] fair agreement beyond chance" (Jonsson and Svingby, p.133 , 2007). In previous artifact studies, reliability is reported as percent agreement, with scores ranging from moderately low (40% for overall artifact package in Martinez, Borko, & Stecher, 2012) to higher levels (86.4% for overall artifact in Clare & Aschbacher, 2001). Scale/dimension agreements have a similar wide range, with some reported as low as 22% (Martinez, Borko, & Stecher, 2012).

For this study, rater agreement was described by percent exact and adjacent agreement, Cohen's kappa, and the intraclass correlations (ICC). ICC was run as oneway random, since all raters did not rate all artifacts, and is reported for means, as we are interested in the overall reliability of the scoring, and not the reliability of one particular rater. Results by dimension are summarized in Table 3 below.

While the *Conceptual Understanding* (CU)*, Application & Relevance* (AR), and *Argumentation and Communication* (ACom) dimensions had agreement above 90% for adjacent scores, exact agreement for all dimensions fell below the Stemler's 70% level,

but are not out of line with those found in other artifact studies. Of particular note, is the *Procedural Skills & Fluency* (PSF) dimension, in which rater agreement was below 90% for adjacent scores and below 50% for exact agreement. As it was possible to monitor scoring real-time through a cloud based scoring form, several re-alignment meetings were scheduled. Alignment improved following each of these, but drift was noted, indicating that this dimension may need to be revisited.

Kappa, while significantly different from chance agreement for all dimensions, fell below Stoddart et al's .40 cutoff for PSF and ACom, but above for CU and AR. ICC indicated that raters' scores are significantly correlated for all dimensions. This is an indication that there can be confidence that another pool of raters would score artifacts similarly to the chosen raters.

Table 3
Reliability

	% exact agreement	% adjacent	kappa	ICC
CU	63%	92%	.436***	.712***
PSF	48%	87%	.254***	.639***
AR	63%	95%	.476***	.820***
ACom	58%	93%	.384***	.824***

Descriptive Analysis

Content. As part of an indicator system, the designed measure would need to provide information to stake holders in terms of topics covered across a specified period of time or across a system. Although the artifacts collected were derived from a convenience

sample, they did cover six of the ten topic areas described in the Common Core State Standards for middle school grades: *Expressions and Equations*; *Functions*; *Geometry*; *Measurement; Ratios and Proportions*; and *Statistics and Probability*. Most artifacts were coded as *Geometry* or *Expressions and Equations*. The raters were unable to reach agreement on content topic in 20% of the artifacts. However, the distribution of agreed upon general content is shown in Table 4 below.

Table 4
Content Coding Frequencies by Percent

Expressions & Equations	Functions	Geometry	Measurement & Data	Ratios& Proportional Relationships	Statistics & Probability
29	13	33	5	14	6

It may also be of interest to stakeholders to understand the grade level of the artifacts. In this study, grade level was coded in two ways. First, information was collected from the teacher as to grade level of the class assigned the task. This was not available for some of the older artifacts, but 99 artifacts were assigned a grade level. In addition, the raters assigned a grade level based on the content alignment with Common Core grade levels. When both raters agreed on the grade level, it was included in Table 5 below. However, raters had difficulty reaching agreement on the grade levels. In training, raters often had trouble deciding on a single grade level that represented the entire artifact. These experienced math teachers indicated that often lower level skills are present in a task in order to build to targeted higher level skills. They were instructed to score at the highest predominant grade level, but disagreement remained at 39%, higher than the content topic disagreement at 20%. The 64 assignments that had both a reported

grade level and an agreed upon scored grade level were analyzed to better understand whether there was alignment of the two levels. In 47% of these cases, the two scores were aligned. In 34% of the cases, the raters scored the task above grade level, and in 19% of the cases, the raters' score fell below grade level.

Table 5
Grade Level Coding by Percent

	Below 6th	6th	7th	8th	Above 8th
Reported	1	36	20	33	10
Rated	7	17	25	38	13

Practice Dimensions. Distributions were calculated for all four dimensions that encompass the Common Core Math practices. Means and standard deviations are shown in Table 6 and distributions are graphically represented in Figure 3. There are comparability caveats between dimensions. Although each dimension uses the same four point scoring range, each represents a different interweaving of practices. Therefore, a 2 on *Conceptual Understanding* cannot be equated with a 2 on *Argumentation & Communication*. Consequently, we cannot interpret the mean on one dimension in relation to the other dimensions. Only trends can be compared.

For *Conceptual Understanding*, the majority of the artifacts were rated as Superficial (1) or Incomplete Practice (2). These include tasks that ask students to apply conceptual knowledge but either do not ask them to provide explicit evidence of their understanding or give the students substantial scaffolding in order to elicit this evidence. For example, a student may be asked to decide which rectangular prism has the greater volume, but not to explain how she determined the answer, or the student may be told to

compare the heights, widths, and depths of two rectangular prisms, and then to compare the volumes and to explain why one would hold more than the other. In this second example, the student is guided toward the answer with a certain amount of "handholding." For this dimension, only one artifact was scored as absent of *Conceptual Understanding*, indicating that tasks set for students in our sample required some application of mathematical knowledge or that our raters/instrument were unable to identify a lack of conceptual understanding.

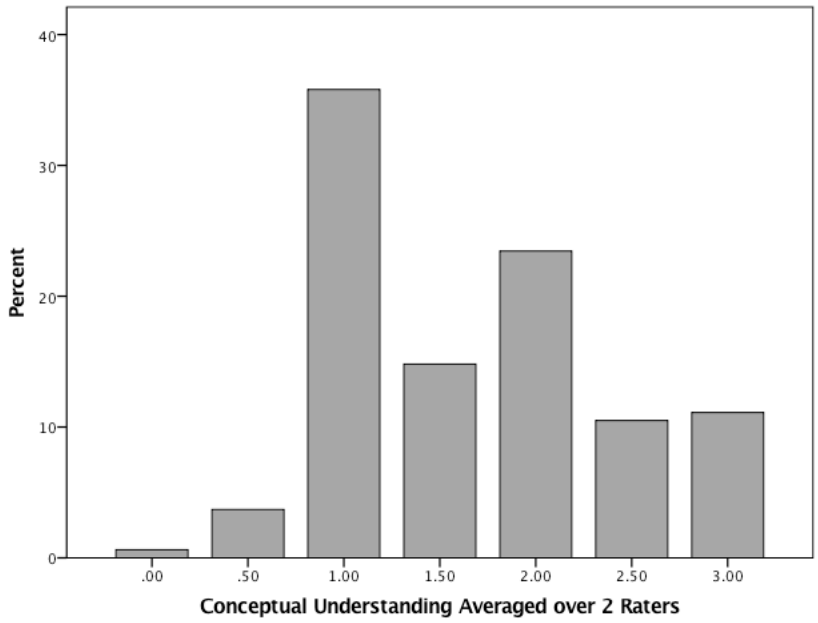
Table 6
Dimension Scores
(N=162)

Dimension	Mean	Standard Deviation
CU	1.66	.73
PSF	1.35	.81
AR	1.02	.85
ACom	.95	1.01

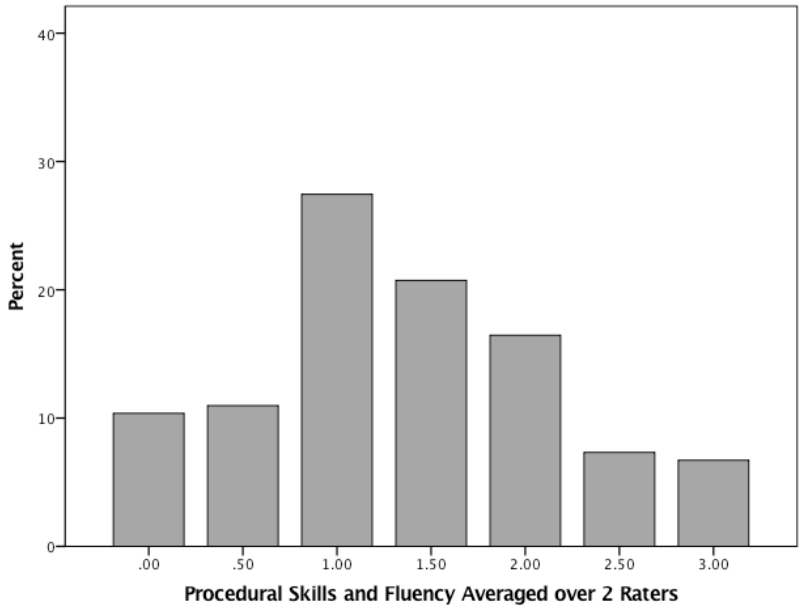
For *Procedural Skills and Fluency*, there is a similar pattern, with the majority of the artifacts either asking students to determine and use single procedures or heavily scaffolding student's use of multiple procedures. On this dimension, however, 10% of the artifacts were scored at 0 (absent). In these artifacts, students are given the procedure to follow and are not asked to make sense of the problem independently. It is also worth noting that this dimension had the lowest rater reliability. Raters reported difficulties interpreting what scaffolding entailed in contrast with having the structure of the artifact build complexity. That is, the difference between a task that tells the student, "First, do

this, then use that answer to find the next" versus one that asks students to perform tasks that could then be built upon to answer further questions without being directed to do so. Also, in re-alignment meetings, raters discussed the difference between coordinated multiple procedures (rated at 2) versus a sequence of single procedures contained in the same artifact (rated at 1).

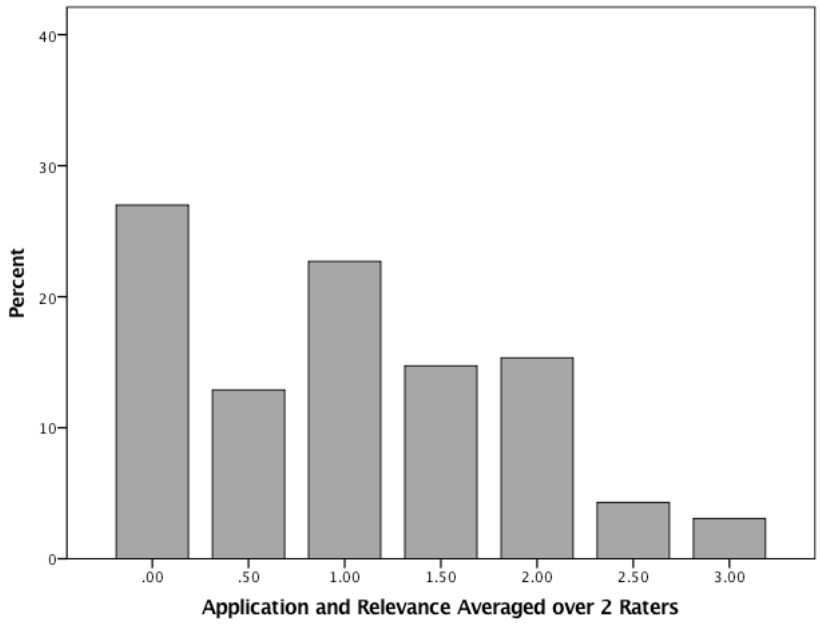
There was a higher percentage of artifacts scored at 0 for both the *Application and Relevance* (27%) and *Argumentation and Communication* (42%), indicating that within our sample there were few tasks that asked students to perform real world tasks using math or to explain or argue their thinking.



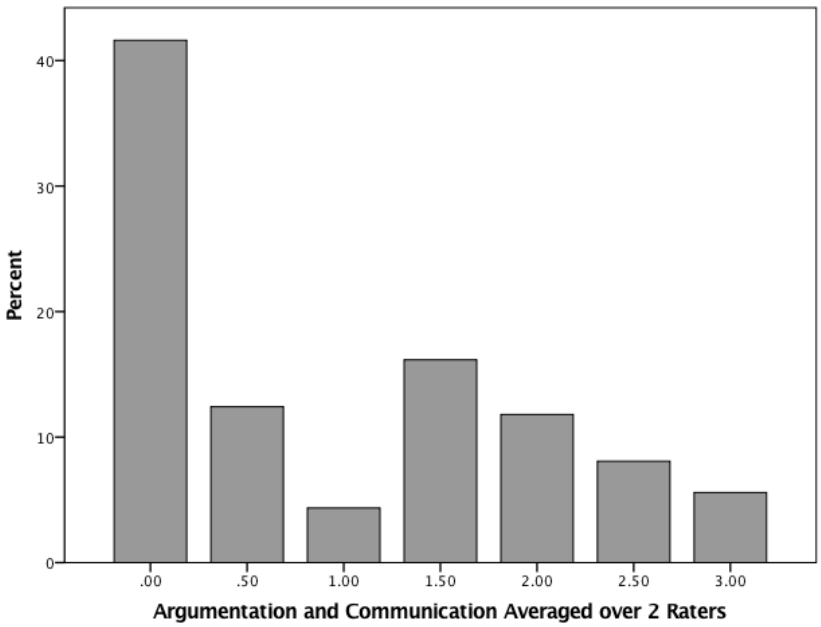
3a. *Conceptual Understanding*.



3b. *Procedural Skills and Fluency*



3c. *Application and Relevance*.



3d. *Argumentation and Communication*.

Figure 3a-d. *Distributions of Averaged Ratings Across Dimensions.*

Correlations between Dimensions

The dimensions were designed to measure different aspects of the practices associated with the CCSS-Math. Although there was overlap, as the practices are inter-related, it is important that each dimension remains distinct in terms of exactly what is being described. In an effort to better understand the designed dimensions as related but not equivalent, correlations were run using the non-parametric Spearman's rho and are reported in Table 7. There are moderate correlations among dimensions, with a strong correlation between *Conceptual Understanding* and *Argumentation & Communication*, indicating that the dimensions are related but not identical, and, specifically, that there is a strong connection between comprehension of a mathematical idea and the ability to express that idea either linguistically or through quantitative representation.

Table 7

Correlations between Dimensions

	CU	PSF	AR	ACom
CU	1	.597***	.539***	.726***
PSF		1	.425***	.584***
AR			1	.560***
ACom				1

Analysis of Potential for Comparison

As an indicator measure, the protocols would be used to assess not only the extent to which content and practices were being implemented, but also to identify factors that may be associated with quality implementation. Although the convenience sample precludes inferences of this type, in this section, the potential of the protocol instrument is examined for such a purpose.

Contextual information was collected from annotations associated with artifacts. In order to investigate that potential, descriptive sub-groupings were submitted to the Wilcoxon–Mann–Whitney (WMW) 2-sample rank sum test. This is a nonparametric measure for skewed, small sample data and is appropriate for ordinal scores. It tests for equality of central tendency of the two unpaired distributions. First, all scores are ranked regardless of which sub-group the observation is from. The WMW then determines whether or not we can reject the null hypothesis that the two groups median ranking are the same in favor of the alternative hypothesis that one sub-group's median ranking is higher than the others. Due to the lack of background information from some of the older

artifacts, and the grouped nature of some of the newer artifacts, the sample size for these subgroupings was often different from the total 162 artifacts, and often quite small.

First, contextual content focus results were analyzed to understand if the instrument noted any differences in ratings by topic and grade. A recent report from California (McLaughlin, Glaab, & Carrasco, 2014) documents concerns from both administrative and educational stakeholders in that state about the feasibility and fidelity of implementing Common Core Standards in math, in terms of teacher knowledge and the shift in grade level curriculum (i.e., a topic that was once 7^{th} grade being shifted to 8^{th} grade). This report states that students may "miss out" on topics as teachers get up to speed in areas of mathematical concepts that they may not have deep knowledge of or by being in a grade that didn't cover the topic prior to implementation and then being in a grade that no longer covers it under the new system. An indicator system would need to be able to provide stakeholders with information pertaining to what is happening by topic and by grade in order to identify any areas that require curricular attention.

To that purpose, means and standard deviations were calculated by content topic (Table 8). Each topic was tested by dimension to determine if it was significantly different from the group of artifacts not associated with that topic (e.g. Geometry v. not Geometry). The instrument was invariate across topics, with the possible exception of Measurement and Data, which was significantly lower in the dimension of *Procedural Skills and Fluency*, and Ratio and Proportional Reasoning, which was significantly higher in the dimension of *Application and Relevance*. In terms of means by grade (Table 9), each grade was tested against the remaining grades (e.g. Grade 6 v. not Grade 6). The instrument was also invariate with the possible exception of Grade 7, in which the test

detected a significant difference in the dimension of *Application and Relevance* when compared to other grades.

Table 8
Means by Mathematical Topic

	Exp Equat	Func	Geometry	Measure & Data	Ratios& Prop Rel	Stat & Prob
	N=37	N=17	N=42	N=6	N=19	N=8
CU	1.42 (.65)	1.68 (.77)	1.70 (.73)	1.67 (1.17)	1.76 (.87)	1.69 (.84)
PSF	1.23 (.67)	1.18 (.68)	1.44 (.87)	.58* (.49)	1.42 (.82)	1.06 (.78)
AR	.86 (.88)	.68 (.86)	.90 (.81)	1.08 (.58)	1.37* (.70)	1.38 (.58)
ACom	.81 (1.08)	1.15 (1.23)	.88 (.97)	.58 (.74)	.64 (.98)	1.21 (.81)

*p<.05 from WMU test statistic. Not significant when Bonferroni correction for multiple comparisons is applied

Table 9
Means by Scored Grade

	Below 6th N=7	Grade 6 N=17	Grade 7 N=25	Grade 8 N=37	Above 8th N=13
CU	1.36 (.38)	1.56 (.73)	1.84 (.76)	1.58 (.69)	1.54 (.83)
PSF	.79 (.39)	1.35 (.96)	1.48 (.67)	1.27 (.77)	1.42 (.73)
AR	1.21 (.64)	1.15 (.81)	1.26** (.90)	.70 (.77)	.54 (.85)
ACom	.71 (.76)	.76 (.92)	1.07 (1.09)	.86 (1.00)	.77 (1.03)

**p<.01 from WMU test statistic. p<.05 when Bonferroni correction for multiple comparisons is applied.

One proposed purpose of the instrument is to track change over time. To investigate its sensitivity to a hypothesized change in alignment between pre and post standards artifacts, scores were sorted into these two groups. Results are summarized in Table 10 and indicate that the instrument is detecting some difference between the two groups for the *Procedural Skills & Fluency* and *Argumentation & Communication* dimensions.

Table 10
Means by Pre v. Post Standards

	Pre N=84	Post N=79
CU	1.57 (.72)	1.76 (.72)
PSF	1.14 (.77)	1.59*** (.79)
AR	1.00 (.85)	1.04 (.85)
ACom	.78 (.89)	1.14* (1.10)

***p<.001 *p<.05 by WMU test statistic

Next, means were grouped by type: whether the artifact was identified as an assignment or a summative assessment. As summarized in a recent European study of the importance of changing assessment in STEM, "…current assessment methods have a strong emphasis on knowledge recall and do not sufficiently capture the crucial skills…of key competencies" (Finlayson, McLoughlin & McCabe, 2015, p. 227). The instrument

detected some differences in the *Application and Relevance* dimension (Table 11) between assignments and assessments, with assignments described more often as moving beyond asking students to solve problems with a real-world context that is superficial and does not add critical information, to contexts that were more real-world and might involve making use of a relevant data set.

Table 11
Means by Type

	Assignment (N=84)	Assessment (N=79)
CU	1.72 (.78)	1.60 (.67)
PSF	1.38 (.94)	1.33 (.64)
AR	1.25 (.92)	.77*** (.68)
ACom	1.06 (1.03)	.84 (.98)

***p<.001 by WMU test statistic

Artifacts were then classified as being collaborative or individual work and as having been assigned for a single class or over multiple class periods. Collaborative work has been found to be more constructive and to lead to better learning outcomes (Chi & Wylie, 2014). In our sample, there were no differences detected for any dimension when comparing collaborative or group work with individual tasks (Table 12). It is notable that in our sample of 160 tasks, only 13 were identified by teachers as group or collaborative work, with the vast majority of tasks set for individual students. A similar pattern is

noted when comparing tasks to be completed in a single class session and those assigned over multiple class periods (Table 13,). Although research finds that working over an extended period of time is critical (Thomas, 2000), only 16 artifacts in the study were to be completed in more than a single class period.

Table 12
Means by Collaborative v. Individual Work

	Collaborative Work N= 13	Individual Work N= 147
CU	1.62 (.62)	1.66 (.74)
PSF	1.50 (.94)	1.35 (.80)
AR	1.31 (.95)	.99 (.84)
ACom	1.15 (1.09)	.93 (1.00)

Next, design practices were considered. Once tasks are identified as highly aligned, it will be of interest to stakeholders to understand the process that led to these tasks. For example, were they designed by a single teacher? pulled from certain textbook or internet resources? This section examines the reported task design practices in our sample.

Therefore, final sub-grouping criteria involved contextual design factors, such as whether the teacher independently created the artifact or drew on the work of others (including published materials), whether or not the internet was used in creating the task, or whether the artifact was designed in collaboration with colleagues. Again, the

instrument detected no significant differences for the first two (Tables 14 and 15), but there were also very few artifacts identified as drawn from the internet or independently created. Our instrument did detect a difference between the 21 artifacts that teachers created collaboratively and the other tasks (Table 16). Sixteen of these twenty one artifacts were not only created by groups of teachers in the same grade but also emerged from a multiple session workshop in which the teachers were given training and support in alignment with new standards. This subset of artifacts was rated significantly higher than the rest of the artifacts for all four dimensions.

Table 13
Means by Single v. Multiple Class Sessions

	Single Class N= 139	Multiple Class N= 16
CU	1.62 (.70)	1.72 (.88)
PSF	1.36 (.78)	1.06 (1.05)
AR	.97 (.83)	1.13 (.87)
ACom	.92 (1.01)	.87 (.74)

Table 14
Means by Teacher Created v. Resourced

	Teacher Created N= 26	Resourced N= 116
CU	1.62 (.71)	1.66 (.72)
PSF	1.38 (.78)	1.20 (.91)
AR	.92 (.78)	1.19 (.87)
ACom	.91 (1.07)	.94 (.99)

Table 15
Means by Internet Resource Used

	Internet (N=8)	Non-internet (N=147)
CU	1.88 (.88)	1.65 (.72)
PSF	1.56 (1.27)	1.33 (.78)
AR	.50 (.80)	1.03 (.85)
ACom	1.19 (1.19)	.90 (.98)

Table 16
Means by Individually v. Collaboratively Selected/Designed/Adapted

	Individually N=128	Collaborative N=21
CU	1.51 (.67)	2.14*** (.64)
PSF	1.19 (.76)	2.19*** (.72)
AR	.87 (.77)	1.81*** (.73)
ACom	.73 (.91)	1.93*** (.89)

***p<.001 from WMU test statistic

Analysis of Feasibility

As a component of an indicator system, the AIP-M would require evidence of feasibility in terms of implementation. One consideration is time to score. To this end, rater scoring timestamp data was converted to duration of scoring per artifact per dimension. Results are shown in Table 17. Raters were asked to identify topic and grade level, and to provide justification for their choices while scoring Conceptual Understanding, so that dimension was therefore more time consuming. One rater worked at a slower average pace than the other two, which could indicate that the Average Across Raters is a high estimate of the time per dimension. Also, rating time decreased as raters scored further, indicating a learning curve. For example, the time for rating ACom for Rater 1 averaged 4 minutes and 6 seconds for the first 10 artifacts scored but 1 minute and 45 seconds for the last 10 artifacts scored.

Table 17
Time to score (in minutes)

	Rater 1	Rater 2	Rater 3	Average Across Raters
CU*	5.08	3.28	7.45	5.18
PSF	3.40	2.82	5.07	3.73
AR	1.75	1.62	4.07	2.41
ACom	2.10	1.37	5.12	2.78

*includes topic and grade scoring

Discussion

Drawing from Mislevy's and Riconscente's work on Evidence-Centered Design (2006) as well as the Rational Empirical Strategy of Test Construction, the Artifact Indicator Protocol for Math content and practice was designed, based on the Common Core Math Standards, design criteria extracted from a thematic synthesis of existing artifact protocols, and interviews with experts. This following questions was addressed in the results and is discussed here:

- *To what extent can the protocol be used to measure classroom practice articulated in math standards?*

The Artifact Indicator Protocol for Science was able to describe classroom practice in terms of the CCSS-Math. There are several components to consider in order to fully address this question. First, we determined that raters were able to utilize the protocol. The protocol was given to raters with expertise in middle school math and the

CCSS in math. Training was conducted and then raters were asked to score 162 artifacts, one dimension at a time. Raters indicated during exit interviews that the provided rubrics, guiding questions, examples, and anchor papers made scoring manageable. Timestamp data indicated that raters' perceptions of scoring duration per artifact and actual duration were similar, and that scoring was less than 5 minutes per artifact for all dimensions except for *Conceptual Understanding. Conceptual Understanding*, which included determining topic and grade level, was scored in just over five minutes. One rater, who was less comfortable with the technology involved in scoring (e.g. google forms, dropbox) took more time to score.

Next, we must consider the reliability of the ratings. Rater reliability measures indicated that although there was less than optimal exact agreement, raters were at the same end of the range for artifacts more than 90% of the time for all dimensions with the exception of *Procedural Skills and Fluency*. Here raters sometimes differed by more than one point. Raters described challenges in scoring this dimension in terms of operationalization of scaffolding and multiple procedures. Further exploration of reliability for this dimension showed that there were differences by score point, and a decrease in reliability when scoring assessments than when scoring assignments. The PSF dimension has been revisited, reworked, and is currently being piloted in a case study. This should provide further information about whether there was a sizeable flaw in the original design or whether this dimension is particularly difficult to assess in artifacts. Overall, however, the null hypothesis that raters agree by chance was rejected for all dimensions.

Once we have established that we can feasibly reach a stable rating, we can examine whether the instrument is able to capture evidence of the construct under scrutiny: the content and practices of the CCSS-Math. Content and grade level scoring proved more problematic than anticipated. Raters struggled to choose a single description of the artifact as a whole. This is less surprising for grade levels, in which lower level scores may be nested within a higher level task, or work with lower level scores may be extended with one or two instances of challenge. However, it was not anticipated that raters would disagree on the topic most representative of an artifact. This was more marked with assessments than with assignments. It may be that with a test or quiz, the teacher is checking mastery of multiple concepts and that a single descriptor does not make sense. In a recent study of physics artifacts (Zisk, R., Etkina, E. & Gitomer, D.H., 2015), assessments were scored item by item, rather than holistically, and that method should potentially be examined in further study. There is also the possibility that math educators lack deep content knowledge in some areas and that makes them less likely to identify these topics.

Scoring with the protocol did populate all points of the scale for all math practice dimensions, although the zero or absent rating was not often used for the *Conceptual Understanding* dimension, and there was noticeable skew toward the lower end of the scale in the scoring for the *Argumentation and Communication* dimension. These findings indicate that the developed protocol may be helpful in identifying the extent to which teachers are setting tasks for students that are aligned with Common Core State Standards, and particularly, whether teachers are asking students to "talk about math" through written communication or mathematical modeling, or whether they are still being

asked to solve problems without formulating an argument (Burns, 2004; Whitin& Whitin, 2000).

In light of the findings, the Artifact Indicator Protocol for Math is in further revision, following Dwyer's recommendation (1994) that protocol designs be subject to "iterative reviews and revisions" (p.144).

- *To what extent is the protocol sensitive to characteristics of instruction that may be of interest to stakeholders?*

The purpose of an indicator measure is to provide valuable information to stakeholders on factors of interest. Means and standard deviations were compared by several criteria and it appeared that the protocol was able to detect some differences in the sample. Some of these were significant using the WMW statistic. Others did not show a difference. Artifacts were identified by several instructional characteristics, including topic, grade, type, collaboratively completed, time allowed for completion, teacher creation, collaboration in task design, use of internet resources, and creation pre- or post- standards. The study-designed protocol was invariate to most characteristics with the exception of collaboratively designed. This does not clearly indicate, however, that teachers working together automatically create more highly aligned tasks, because the teachers in our sample were given more support than just the opportunity to work together. It does provide some evidence that working together with guidance, teachers are able to develop tasks that align with the high demand of the Common Core Standards, and that the designed instrument is able to detect these types of differences.

Given the low reliability for describing tasks and the convenience sampling, the findings that certain topics were more aligned on *Procedural Skills and Fluency*, and on

Application and Relevance post-standards in comparison with pre-standards are not interpretable. However, the instrument is sensitive to some shift here, and could be useful in further study to determine whether students are encountering more aligned tasks in certain math areas.

The instrument also was sensitive to some characteristic in the dimension of *Application and Relevance* by grade. Again, while some of this may be accounted for by the biased sample and, rater disagreement, this does provide evidence that the instrument may be able to detect differences by instructional features. A planned second pilot, following the revision to the instrument, may clarify the instrument's usefulness further.

Although the characteristics of length of time, collaborative work, and teacher resources did not provide differing outcomes with this protocol, there is important information to consider, although with the limits of the convenience sample in mind. Fewer than 10% of the collected artifacts allowed students to work over multiple sessions or together with classmates. Given the literature that describes these types of tasks as potentially enriching (Chi & Wylie, 2014; Thomas, 2000), more investigation is needed into the role of this characteristic in tasks that reflect the CCSS-math.

The finding that assessments within our study are less likely to have real-world context through the *Application and Relevance* dimensions is in line with research that shows teachers tend to use tests and quizzes to assess mastery of more procedural skills (Finlayson, McLoughlin & McCabe, 2015). In order to address this mismatch between changing instruction and less dynamic assessment, there are two paths to be considered. One method is to redesign existing assessments, which is currently being investigated in a case study framework. Another is to extend the definition of what constitutes an

assessment, as done in a related science study by Martinez et al (2012) in which a variety of assessments were collected.

The difference found between artifacts collected before standards and those collected after the Common Core State Standards were introduced provides evidence that the instrument could be of use to stakeholders attempting to track progress over time. Our sample showed a shift in terms of *Argumentation and Communication.* In a larger, more balanced sample, this could reflect calls for "talking about math" that have increased in recent years.

The lack of examples of teacher designed or internet resourced artifacts in the sample supports findings from the previous study (Joyce, Gitomer, & Iaconangelo, 2014) that the types of tasks set for students rely heavily on what is available from publishers. Findings from studies about the alignment of current math materials (Polikoff, 2015; Schmidt & Houang, 2014) may provide important insights about what students are experiencing and lead to pressures on publishers to produce more aligned materials.

Challenges and Limitations

The initial intent of the pilot was to apply the protocols to artifacts collected from multiple districts. However, despite IRB approvals, recruitment was unsuccessful. Even when budget and IRB documentation were amended to include a $200 stipend for what was to be 15 -30 minutes of additional work outside of regular classroom duties, no further participation was gained. Once the recruitment approach was revised to include personal connections and older artifacts previously collected, 77 artifacts (assessments and assignments) were collected from the current academic year, which were added to pre-standards artifacts gathered from previous artifact studies (SCOOP, UTQ). Therefore,

total artifacts collected were closer to 150 artifacts, rather than the initial 500 planned, and the sample was less representative.

Concern about high-stakes evaluation or increased workloads may have contributed to low enrollment of teachers. The difficulty with recruitment, creating a convenience sample, will limit any inferences made from the study findings. That is, determining whether the instrument was able to detect differences in artifacts, is for a more site specific sample than originally anticipated, and for a pair of teachers whose characteristics may be unusual because they were willing to participate in the study.

There is also some initial indication that there are substantial challenges to use of classroom artifacts as part of a national indicator system, as the collection process was labor-intensive. Frequent correspondence and in-person follow-up were required to gain teacher compliance, and uploading of artifacts using existing technology is still somewhat cumbersome. Menial tasks, such as removing staples and feeding pages into a portable scanner, can be time-consuming and limit the feasibility of scaling up. However, one tool that holds promise of streamlining the process is the sharing of google drive folders, which teachers may use for sharing artifacts. While this is virtually effortless, there are some security issues that must be considered in terms of privacy. It may be that artifacts' greatest affordances as an indicator are at a more local level, as discussed earlier.

Conclusion

The AIP-M is a protocol that was designed to measure content and practice alignment to Common Core State Standards of classroom tasks found in artifacts. Quality in terms of intellectual demand, deep thinking, and challenge as represented in the math practices were included in the formulation of the four dimensions: *Conceptual*

Understanding, Procedural Skills and Fluency, Application and Relevance, and Argumentation and Communication. The protocols were than used to score a small convenience sample of math tasks across 6th to 9th grade classes across several districts. Findings indicated that the instrument was invariate in terms of most characteristics studied, and that further revision is merited for the *Procedural Skills and Fluency* dimension to improve rater understanding and reliability.

However, the instrument does hold promise as a tool for self-study by a school, department, or district, as it did capture shifts in alignment on all dimensions between older, pre-standard artifacts and current artifacts, as well as the 21 artifacts that resulted from a collaborative workshop on the new standards.

There are concerns for the potential of scaling up to a state or national level due to the arduous tasks still associated with collecting and scoring artifacts. Further research is needed in the potential affordances of technology to streamline the process to a more scaleable level. For example, could there be an algorithm written for preliminary machine scoring of artifacts? Even the ability to sort out "0"s would significantly reduce the load on human raters. Thus, the argument could be made, that math artifacts capture rich information about the quality and alignment of tasks set for students, and are worthy of further consideration as part of an indicator system.

References

Borko, H., Stecher, B., & Kuffner, K. (2007). *Using artifacts to characterize reform-oriented instruction: The Scoop Notebook and rating guide (CSE Technical Report 707)*. LA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/UCLA.

Borko, H., Stecher, B., Alonzo, A., Moncure, S., & McClam, S. (2005). Artifact Packages for Characterizing Classroom Practice: A Pilot Study. *Educational Assessment , 10* (2), 73-104.

Burns, M. (2004). Writing in math. *Educational Leadership*, *62*(2), 30-33.

Clare, L., & Aschbacher, P. (2001). Exploring the technical quality of using assignments and student work as indicators of classroom practice. *Educational Assessment*, *7*(1), 39–59. doi: 10.1207/S15326977EA0701_5

Common Core State Standards Initiative. (2011). *Common core state standards for mathematics*.

Denner, P. R., Salzman, S. A., & Bangert, A. W. (2001). Linking teacher assessment to student performance: A benchmarking, generalizability, and validity study of the use of teacher work samples. *Journal of Personnel Evaluation in Education*, *15*(4), 287-307.

Dwyer, C. A. (1994). Criteria for performance-based teacher assessments: Validity, standards, and issues. *Journal of personnel evaluation in education*, *8*(2), 135-150.

Finlayson, O., McLoughlin, E., & McCabe, D. (2015). Strategies for the Assessment of Inquiry Learning in Science (SAILS) A European Project in Science Teacher Education. *New Perspectives in Science Education 4th Edition Proceedings*, p225-229.

Garland, R. (1991). The mid-point on a rating scale: Is it desirable. *Marketing bulletin,* 2(1), 66-70.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review*, *2*(2), 130-144.

Joyce, J. A Thematic Synthesis of Artifact Research in Teaching Quality. (in preparation).

Joyce, J., Gitomer, D.H., and Iaconangelo, C. *Assessment of Learning and Teaching Through Quality of Classroom Assignments* (European Association of Research on Learning and Instruction-SIG 1 Assessment. Madrid, Aug 2014).

Klein, D. (2003). A brief history of American K-12 mathematics education in the 20th century. *Mathematical cognition*, 175-259.

Martínez, J. F., Borko, H., & Stecher, B. M. (2012). Measuring instructional practice in science using classroom artifacts: Lessons learned from two validation studies. *Journal of Research in Science Teaching*, *49*(1), 38-67.

Martínez, J. F., Borko, H., Stecher, B., Luskin, R., & Kloser, M. (2012). Measuring Classroom Assessment Practice Using Instructional Artifacts: A Validation Study of the QAS Notebook. *Educational Assessment*, *17*(2-3), 107–131. http://doi.org/10.1080/10627197.2012.715513

Matsumura, L. C., Garnier, H. E., Slater, S. C., & Boston, M. D. (2008). Toward measuring instructional interactions "at-scale". *Educational Assessment*, *13*(4), 267–300. doi:10.1080/10627190802602541

Matsumura, L., & Pascal, J. (2003). *Teachers' assignments and student work: Opening a window on classroom practice.* Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

McLaughlin, M., Glaab, L., & Carrasco, I. H. (2014). Implementing Common Core State Standards in California: A Report from the Field. *Policy Analysis for California Education, PACE*.

Means, B., Mislevy, J., Smith, T., Peters, V., & Gerard, S. (2016). *Measuring the Monitoring Progress K-12 STEM Education Indicators: A Road Map* . Washington, D.C.: SRI Education.

Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 Assessment Frameworks*. International Association for the Evaluation of Educational Achievement. Herengracht 487, Amsterdam, 1017 BT, The Netherlands.

National Council of Teachers of Mathematics Commission on Standards for School Mathematics. (1989). Curriculum and evaluation standards for school mathematics. Reston VA: The Council. http://www.standards.nctm.org/index.htm

National Research Council. (2001). *Adding It Up: Helping Children Learn Mathematics.* Washington, DC: National Academy Press.

National Research Council. (2013). *Monitoring progress toward successful K-12 STEM education: A nation advancing.*

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary educational psychology*, *36*(1), 36-48.

Polikoff, M. S. (2015). How Well Aligned Are Textbooks to the Common Core Standards in Mathematics?. *American Educational Research Journal*, 0002831215584435.

Preston, C and A. Coleman. "Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences." acta psychologica (2000): 1-15.

Schmidt, W. H., & Houang, R. T. (2012). Curricular coherence and the common core state standards for mathematics. *Educational Researcher*, *41*(8), 29

Whitin, P., & Whitin, D. J. (2000). *Math Is Language Too: Talking and Writing in the Mathematics Classroom*. National Council of Teachers of English, 1111 W. Kenyon Road, Urbana, IL 61801-1096

Zisk, R., Etkina, E., & Gitomer, D. (2015). *Developing a Protocol to Assess Instructional Artifacts in Physics.* American Association of Physics Teachers Summer Meeting. College Park.

Appendix A: AIP-M

Table 1

Mathematics Content Domains and Standards

Domain	Below Grade 6	Grade 6	Grade 7	Grade 8	Above Grade 8
Operations & Algebraic Thinking	• Write and interpret numerical expressions *using simple expressions with numbers*. • Analyze patterns and relationships *using two given rules, and graph the ordered pairs*.				
Numbers & Operations in Base Ten	• Understand the place value system. • Perform operations with multi-digit whole numbers and with decimals to hundredths.				
The Number System		• Apply and extend previous understandings of multiplication and division to divide fractions by fractions. • Compute fluently with multi-digit numbers and find common factors and multiples. • Apply and extend previous understandings of numbers to the system of rational numbers.	• Apply and extend previous understandings of operations with fractions.	• Know that there are numbers that are not rational, and approximate them by rational numbers.	
Numbers & Operations - Fractions	• Use equivalent fractions as a strategy to add and subtract fractions. • Apply and extend previous				

multiplication and division.

Table 1

Mathematics Content Domains and Standards (continued)

Domain	Below Grade 6	Grade 6	Grade 7	Grade 8	Above Grade 8
Measurement & Data	• Convert like measurement units within a given measurement system. • Represent and interpret data. • Geometric measurement: understand concepts of volume.				
Geometry	• Graph points on the coordinate plane to solve real-world and mathematical problems. • Classify two-dimensional figures into categories based on their properties.	• Solve real-world and mathematical problems involving area, surface area, and volume.	• Draw, construct, and describe geometrical figures and describe the relationships between them. • Solve real-life and mathematical problems involving angle measure, area, surface area, and volume.	• Understand congruence and similarity using physical models, transparencies, or geometry software. • Understand and apply the Pythagorean Theorem. • Solve real-world and mathematical problems involving volume of cylinders, cones, and spheres.	
Expressions & Equations		• Apply and extend previous understandings of arithmetic to algebraic expressions. • Reason about and solve one-variable equations and inequalities. • Represent and analyze quantitative relationships between dependent and independent variables.	• Use properties of operations to generate equivalent expressions. • Solve real-life and mathematical problems using numerical and algebraic expressions and equations.	• Work with radicals and integer exponents. • Understand the connections between proportional relationships, lines, and linear equations. • Analyze and solve linear equations and pairs of simultaneous linear equations.	

Table 1

Mathematics Content Domains and Standards (continued)

Domain	Below Grade 6	Grade 6	Grade 7	Grade 8	Above Grade 8
Ratios & Proportional Relationships		• Understand ratio concepts and use ratio reasoning to solve problems.	• Analyze proportional relationships and use them to solve real-world and mathematical problems.		
Functions				• Define, evaluate, and compare functions. • Use functions to model relationships between quantities.	
Statistics & Probability		• Develop understanding of statistical variability. • Summarize and describe distributions.	• Use random sampling to draw inferences about a population. • Draw informal comparative inferences about two populations. • Investigate chance processes and develop, use, and evaluate probability models.	• Investigate patterns of association in bivariate data.	• Calculate expected values and use them to solve problems. • Use probability to evaluate outcomes of decisions.

Conceptual Understanding: This dimension focuses on the extent to which students are asked to provide explicit evidence of their conceptual understanding of mathematical idea(s). Artifacts that are high on this dimension ask students to provide evidence of their understanding of mathematics concepts as well as relationships among concepts; to represent their thinking about these concepts; to see connections with other ideas in mathematics; and to apply concept(s) to solve problems. Artifacts that ask students to do such things as routine solving of equations would be low on this dimension.			
0 - Absent	**1-Surface Practice**	**2-Incomplete Practice**	**3-Developed Practice**
Students are not asked to apply or provide explicit evidence of any conceptual understanding.	Students are asked to apply concept(s) but are not asked to provide explicit evidence of their understanding.	Students are asked to apply concept(s) and provide explicit evidence of their understanding, but evidence is developed with scaffolding	Students are asked to both apply and provide explicit evidence of their understanding independently.

Justification for this dimension comes from : (http://www.nap.edu/read/9822/chapter/6#118)

Procedural Skills and Fluency: This dimension focuses on the extent to which students are asked to make sense of problems and persevere in solving them by using appropriate strategies while attending to precision and conditions of use. Artifacts that are high on this dimension ask students to understand procedures and conditions of use, to interpret procedural outcomes, to use multiple procedures in a coordinated manner, to describe these procedures and the solution path, and to check work. Artifacts that ask students to do such things as routine solving of problems using a single, given procedure would be low on this dimension.			
0 - Absent	**1-Surface Practice**	**2-Incomplete Practice**	**3-Developed Practice**
Students are given a procedure to follow in order to solve problems. No sense-making is expected. A task receiving a score of 0 does not ask students to choose procedures.	Students are required to select and implement single procedures and to describe a solution path OR to select, implement and coordinate multiple procedures with no solution path in order to make sense of problems. Explicit checking of work may be required in order to attend to precision..	Students are required to select, implement, and coordinate multiple procedures and to describe their solution path, with scaffolding, in order to make sense of problems Explicit checking of work may be required in order to attend to precision.	Students are required to select, implement, and coordinate multiple procedures and to describe their solution path independently in order to make sense of problems.. Explicit checking of work may be required in order to attend to precision..

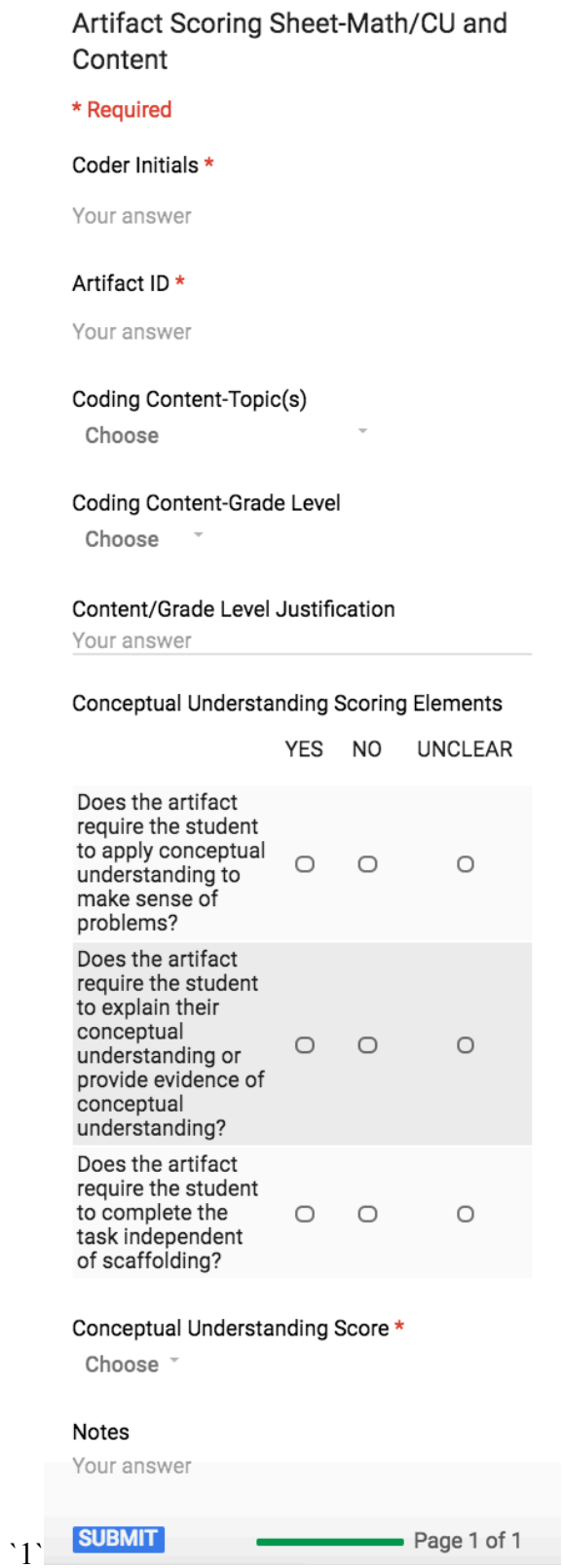
Justification for this dimension comes from : (http://www.nctm.org/Standards-and-Positions/Position-Statements/Procedural-Fluency-in-Mathematics/)

Application and Relevance: This dimension focuses on the extent to which students are asked to make sense of real world problems or to model real world situations using mathematical representations and reasoning. Artifacts that are high on this dimension ask students to apply mathematics concepts to real world problems or to work with real world data. Artifacts that ask students to do such things as routine solving of equations without real world context or with real world context that is superficial to the solution would be low on this dimension.			
0 - Absent	**1-Surface Practice**	**2-Incomplete Practice**	**3-Developed Practice**
Students are **not asked to solve real world problems**	Students are asked to solve problems with a real world **context that is superficial and does not add critical information**. Problems may require the presence or creation of a data set. Tasks receiving a score of 1 use context only as a veneer.	Students are asked to solve problems with a **simplified real world context and data set** but are not asked to address real world problems or to model real world situations. Tasks receiving a score of 2 solve problems that are only academic in nature.	Students are asked to address real world problems or model real world situations using data from real world contexts.

Argumentation and Communication: This dimension focuses on the extent to which students are asked to develop and communicate mathematical conjectures and arguments. Artifacts that score high on this dimension ask students to make mathematical conjectures, to develop a mathematical argument, and to communicate their thinking coherently to others, as well as to evaluate the arguments and communication of others. Artifacts that ask students to do such things as routine solving of equations without extended writing would be low on this dimension.			
0 - Absent	**1-Surface Practice**	**2-Incomplete Practice**	**3-Developed Practice**
Students are not asked to construct an argument. Writing does not extend beyond providing a simple mathematical or verbal solution	Students are asked to use given assumptions or definitions to support a provided argument. Writing or representation is structured.	Students are asked to develop their own argument or to evaluate the argument of others but are not asked to develop a complete or elaborated communication. Artifacts at this level do not ask students to demonstrate reasoning through the use of counterexamples, alternatives, representations and/or elaborated writing in a logical progression.	Students are asked to develop their own complete argument or to evaluate the argument of others and are asked to develop a complete or elaborated communication. Artifacts at this level ask students to demonstrate reasoning through the use of counterexamples, alternatives, representations and/or elaborated writing in a logical progression.

Justification for this dimension comes from: http://www.nctm.org/Standards-and-Positions/Principles-and-Standards/Process/

Appendix B: Scoring Sheets



Artifact Scoring Sheet-Math/CU and Content

* Required

Coder Initials *

Your answer

Artifact ID *

Your answer

Coding Content-Topic(s)

Choose

Coding Content-Grade Level

Choose

Content/Grade Level Justification

Your answer

Conceptual Understanding Scoring Elements

	YES	NO	UNCLEAR
Does the artifact require the student to apply conceptual understanding to make sense of problems?	○	○	○
Does the artifact require the student to explain their conceptual understanding or provide evidence of conceptual understanding?	○	○	○
Does the artifact require the student to complete the task independent of scaffolding?	○	○	○

Conceptual Understanding Score *

Choose

Notes

Your answer

SUBMIT Page 1 of 1

`1`

Artifact Scoring Sheet-Math/PSF

* Required

Coder Initials *

Your answer

Artifact ID *

Your answer

Procedural Skills and Fluency Elements *

	YES	NO	UNCLEAR
Does the artifact require the student to select and implement procedure(s)?	○	○	○
Does the artifact require the student to coordinate multiple procedures?	○	○	○
Does the artifact require the student to describe the solution path?	○	○	○
Does the artifact require the student to complete the task independent of scaffolding?	○	○	○
Does the artifact require the student to attend to precision by checking his/her work?	○	○	○

Procedural Skills and Fluency Score *

Choose

Notes

Artifact Scoring Sheet-Math/AR

* Required

Coder Initials *

Your answer

Artifact ID *

Your answer

Application and Relevance Scoring Elements *

	YES	NO	UNCLEAR
Does the artifact provide a real-world context or use a real-world data set?	○	○	○
Is the context critical to make sense of the problem, or does the context require use of a real-world data set?	○	○	○
Does the artifact provide a real-world context or data set that is relevant to the student?	○	○	○
Does the artifact require the student to mathematically model real-world situations?	○	○	○

Application and Relevance Score *

Choose

Notes

Your answer

Artifact Scoring Sheet-Math/ACom

* Required

Coder Initials *

Your answer

Artifact ID *

Your answer

Argumentation and Communication Scoring Elements *

	YES	NO	UNCLEAR
Does the artifact require the student to construct an argument or to evaluate the arguments of others?	○	○	○
Does the artifact require the student to support an argument with evidence drawn from theory and/or examples?	○	○	○
Is the student asked to write a complete and elaborated argument including elements such as examples, counterexamples, alternative explanations, and representations?	○	○	○

Argumentation and Communication Score *

Choose

Notes

Appendix C: Rater Exit Interview Questions

Overall
How did you approach the scoring task? Can you talk us through your procedure in scoring?
What role did student work play in your scoring?
Which practice or component of a practice was more difficult to see in the classroom artifacts?
Which rubric or element of a rubric was difficult to learn?
What differences did you note, if any, between scoring assessments such as quizzes and tests, and scoring assignments?

Content and Conceptual Understanding
What were some of the challenges in scoring content and grade level of math assignments and assessments?
What were the challenges in scoring this practice?
About how long did it take you to score an artifact for this dimension?

Procedure Skills and Fluency
What were the challenges in scoring this practice?
About how long did it take you to score an artifact for this dimension?
How could we go about describing scaffolding to make scoring easier for this dimension?

Application and Relevance
What were the challenges in scoring this practice?
About how long did it take you to score an artifact for this dimension?

Argument and Communication
What were the challenges in scoring this practice?
About how long did it take you to score an artifact for this dimension?

Final Questions
In your opinion, would access to these protocols be helpful to teachers in planning or assessing instructional materials?
What else do you think we need to know overall in order to improve the usability of the protocols?

Conclusion

This research investigates the potential of classroom artifacts in providing evidence that helps to answer questions about progress toward educational reform goals, and to give support to teachers and administrators in their quest to improve alignment of tasks set for students. First, it provided a thematic synthesis of existing STEM artifact studies and how these have been designed in order to establish a foundational resource for artifact study. Then, the research extends the to-date uses of artifact study to a new use, as a component of an indicator system that has the capacity to describe the alignment of classroom work and assessments with the current college and career readiness standards. Two domain specific studies were conducted, one in science and one in math, that sought to develop and pilot protocols that could be used to describe alignment of tasks given to students with the reforms embodied in the NGSS and CCSS-Math.

In the Synthesis Study, it becomes clear that STEM artifact study is flexible enough to be useful for a variety of purposes. Several key decision points were identified in terms of sampling, scoring, and reliability. The study finds that multiple artifacts from multiple times of year are essential to a stable rating, regardless of purpose, and that some sample of student work is critical to understanding the nature of the instructional task. Additionally, demographic information is frequently considered by researchers to be important to collect. This is critical to understanding the situated nature of instruction. However, there is, to date, no agreement on the extent that constitutes necessary or sufficient. There appears to be some progress toward standardization of artifact study in terms of rating frameworks applied, with three types of scoring rubrics emerging (IDAP, SCOOP, Standards). This should lead to increased comparability within disciplinary

domains. Ongoing research as to comparability across studies and with differing protocols should contribute further to cohesion in the field of artifact study. A certain amount of inference was required in order to code for purpose, and the differences in reporting may have limited the usefulness of comparison. However, this study does represent important movement toward cohesion in artifact study and it is hoped that it will lead to further standardization in future work, with clear reporting of protocols used, reliabilities attained, so that ongoing research in this promising area can advance understanding of instructional practice.

In the second study, The Artifact Indicator Protocol-S was designed to measure content and practice alignment to NGSS standards of classroom tasks as represented in classroom artifacts, and was able to describe classroom coverage of content and practice in terms of the NGSS. Raters were able to utilize the protocol within acceptable reliability parameters. There was success in identifying meaningful differences among artifacts. Additionally, the protocol shows some promise in terms of monitoring opportunities to learn for identified sub-groups. Using classroom artifacts and having a protocol available to gauge and track implementation of the reforms embodied in the Next Generation Science Standards would be of value to stakeholders, and could support the instructional practices that lead to a science-literate citizenry.

A third related, but distinct study was designed to measure content and practice alignment to Common Core State Standards of classroom tasks found in math artifacts. While raters were able to utilize the protocol to describe alignment for artifacts, there were some concerns about agreement on topic, grade, and the designed dimension of *Procedural Skills and Fluency*. Further information is needed about whether there is a

flaw in the original design or whether the PSF dimension is particularly difficult to assess in artifacts. Additionally, further exploration of the difficulties in reaching consensus on topic and grade is needed. Exit interviews were conducted with raters, and further analysis is ongoing. The purpose of an indicator measure is to provide valuable information to stakeholders on factors of interest. Means and standard deviations were compared by several criteria and it appeared that the AIP-M was able to detect some differences in the sample. Findings of significant differences between collaboratively designed artifacts and more individually selected tasks does provide some evidence that working together with guidance, teachers are able to develop tasks that align with the high demand of the Common Core Standards, and that the designed instrument is able to detect these types of differences. Additionally, the lack of examples of teacher designed or internet resourced artifacts in the sample supports findings from previous studies that the types of math tasks set for students rely heavily on what is available from publishers. Overall, math artifacts capture rich information about the quality and alignment of tasks set for students, and are worthy of further consideration as part of an indicator system.

However, due to recruitment issues in both studies, determining whether the instruments were able to detect differences in artifacts is for a more site-specific sample than originally anticipated. There is also some concern that there are substantial challenges to use of classroom artifacts as part of a national indicator system, as the collection process was labor-intensive. Further research is needed with a broader sample, and that explores to a greater extent the affordances of new technologies to streamline the collection and scoring process.

Another potential use for the standards-aligned protocol, although beyond the scope of the current studies, would be in a more local context to plan curriculum and professional development. Further research is needed to examine whether the same instrument, with some locally desirable modifications, could be potentially be used for this formative assessment through self-study. Additionally, a collection of exemplar artifacts collected in the two completed studies could be developed and provided to teachers as a self-guided professional development tool. There is potential in the instruments developed in to be used for such a purpose, which could be explored through interaction with teachers in a case study setting. Such a study could utilize a Professional Learning Community type structure - teachers and academic administrators, and a research team working together:

- to develop a language and tools to evaluate classroom assignments and assessments
- to judge quality of work students are being asked to do with a goal of identification and improvement.

The completed research here would support such an extension of purpose.

The Artifact studies here further our understanding of how to answer the question "How would we know if educational practice was changing with the emergence of new STEM standards?" through investigation of classroom artifacts. Beginning with a comprehensive synthesis of work with STEM artifacts to date, and then moving beyond to explore the possibility of designing and implementing disciplinary specific protocols for middle school tasks, allows us to ask, "Is this what we want students to do? Are there

ways we might change what we ask of students?" Through the use of the Artifact Indicator Protocols in Science and Math, we can begin to find the answers.