Res Sci Educ
DOI 10.1007/s11165-016-9544-8

Scrutinizing a Survey-Based Measure of Science and Mathematics Teacher Knowledge: Relationship to Observations of Teaching Practice

Robert M. Talbot III [1]

Abstract There is a clear need for valid and reliable instrumentation that measures teacher knowledge. However, the process of investigating and making a case for instrument validity is not a simple undertaking; rather, it is a complex endeavor. This paper presents the empirical case of one aspect of such an instrument validation effort. The particular instrument under scrutiny was developed in order to determine the effect of a teacher education program on novice science and mathematics teachers' *strategic knowledge* (SK). The relationship between novice science and mathematics teachers' SK as measured by a survey and their SK as inferred from observations of practice using a widely used observation protocol is the subject of this paper. Moderate correlations between parts of the observation-based construct and the SK construct were observed. However, the main finding of this work is that the context in which the measurement is made (in situ observations vs. ex situ survey) is an essential factor in establishing the validity of the measurement itself.

Keywords Teacher knowledge · Survey measurement · Observation · Validity

Introduction

As we strive to develop teacher education programs capable of preparing "highly qualified teachers" (U.S. Department of Education 2002), we must be able to evaluate the effectiveness of these programs. While defining what it means to be a "highly qualified" teacher is itself a challenging endeavor, measuring it can also be equally challenging. Given the potential consequences of judgments resulting from uses of these measures, this is a challenge that cannot be taken lightly. A teacher education program could be deemed ineffective based on such data, or at

✉ Robert M. Talbot, III
robert.talbot@ucdenver.edu

[1] School of Education and Human Development, University of Colorado Denver, 1380 Lawrence St, PO Box 173364, Denver, CO 80217-3364, USA

Published online: 27 December 2016

least less effective than a competing program, and may lose its funding, accreditation, or enrollment. In other words, the stakes are conceivably high. In order to be able to make a strong case for the effect of a teacher education program on aspects of educator quality, the instrumentation from which these measures are derived must be valid. The term *validity* is used to denote the "degree to which evidence and theory support the interpretation of test scores entailed by the proposed test uses" (American Educational Research Association et al. 2014, p. 5). Lacking the characteristic of validity, uses of these measures in determining the effect of a preparation program on a teacher's qualifications could be unwarranted and may result in poor judgments being made.

Although there is a clear need for valid and reliable instrumentation that measures teacher knowledge (more specifically strategic knowledge), the process of investigating instrument validity is not a simple undertaking. Making a case for the validity of an instrument is complex; therefore, many things need to be considered. For example, one must begin by articulating the way in which scores resulting from the instrument will be interpreted and the intended use for the instrument. It is the interpretations of these scores that are then evaluated—not the instrument itself. Based on the proposed score interpretation and instrument use, a set of propositions that undergird that interpretation are then identified. These propositions frame and determine the types of evidence that need to be gathered in order to develop the larger validity argument. If the scores resulting from the instrument are to be used in ways that differ from the proposed definition of its score interpretation and instrument use, then this new interpretation must also be validated.

Given the complexity of such a validation effort, there are many potential obstacles to developing an instrument that can be used to evaluate the effect of a teacher education program on novice teachers' knowledge. Most important is the decision of what to measure. A foundational part of any score interpretation is that the score is of something that matters. For example, in the case of teacher education program evaluation, does the score represent an understanding, ability, or achievement level that matters for teaching and can be attributed to the program?

This paper presents the empirical case of one aspect of such an instrument validation effort. The particular instrument under scrutiny was developed in order to determine the effect of a teacher education program on novice science and mathematics teachers' *strategic knowledge* (SK). The relationship between novice science and mathematics teachers' SK, as *measured by a survey*, and their SK as *inferred from observations of practice*, is the subject of this paper and a central part of the validity argument for the survey-based measure.

Strategic Knowledge

The SK Construct

The strategic knowledge construct is comprised two dimensions that are labeled Flexible Application (FA) and Student-Centered Instruction (SCI) (Briggs et al. 2007). The FA and SCI dimensions are not conceptualized as orthogonal; rather, they are considered to be interrelated.

The FA dimension describes how science or mathematics teachers invoke, apply, and modify their instructional repertoire in a given teaching context. At the most novice level in the FA dimension, teachers have a very limited repertoire of strategies from which to draw, and with development they gain not only a larger repertoire of strategies but also both the ability to judge the appropriateness of various strategic approaches given the situational constraints and the ability to modify those strategies based on these constraints (e.g., Berliner 2001; Bond et al. 2000; Hammerness et al. 2005).

The SCI dimension describes how science or mathematics teachers conceive of a given situation as an opportunity for active engagement with the students in order to identify the students' current understanding. At the lowest level, teachers do not see the activity or scenario as an opportunity to elicit information from their students about their current level of understanding. At a high level of SCI, teachers see the activity as an opportunity to interact with the students in order to gauge their understanding and identify their needs (e.g., Van Driel et al. 1998). In part, teachers' "learner-centeredness" is what is being measured with the SCI dimension.

Measuring novice science and mathematics teachers' strategic knowledge is not so straightforward. Broadly, there are at least two ways to approach developing an instrument to measure strategic knowledge: (a) using instruments or protocols that yield direct measures of teaching practice based on observing teachers in the classroom and (b) using instruments that yield indirect measures based on what teachers *say* about their teaching practice, either in interviews or in response to survey prompts. Both direct classroom observations and teacher interviews can be costly, time-consuming, and subjective. This research program focuses on a potentially more economical, efficient, and less subjective approach to assessing strategic knowledge through the scoring of responses to a scenario-based survey instrument.

The FASCI Instrument

The Flexible Application of Student-Centered Instruction (FASCI) survey instrument was designed and developed to assess novice science and mathematics teachers' strategic knowledge. Briggs et al. (2007) hypothesized that teachers with high scores on the FASCI survey instrument could be characterized as being able to draw from a broad repertoire of teaching strategies and apply those strategies that are warranted by the given context (the FA dimension of strategic knowledge). In addition, these high-scoring teachers view instructional activities as an opportunity for students to be actively engaged in activities about the topic at hand so that the teacher can identify the student's level of understanding (the SCI dimension of strategic knowledge).

The scenario-based items on the FASCI, to which individuals respond in an open-ended fashion, all have a common form (see Table 1). In these items, a classroom scenario is presented that frames three prompts. The FASCI scenarios include a variety of classroom situations or events. Examples of these scenarios include students working in groups to discuss a conceptual problem, a teacher working an example problem on the board, or a teacher talking

Table 1 Example scenario-based FASCI item

Example FASCI item
For the question and scenarios that follow, please assume that you are teaching a high school course in physics, chemistry, biology, Earth science, or math to a class of 25–30 students.
1. Students are working in groups of four to discuss a conceptual question you provided them at the beginning of class.
(a) How might this activity facilitate student learning?
As the activity proceeds, one group gets frustrated and approaches you—they've come up with two solutions, but can't agree on which one is correct. You see that one solution is right, while the other is not.
(b) Describe both what would you do and what you would expect to happen as a result.
(c) If the approach you describe above in (b) didn't produce the result(s) you anticipated by the end of that class session, what would you do in the next class session?

one on one with a student. The first question prompt asks how respondents think the activity would facilitate student learning. A potential obstacle is then presented that further frames the scenario. For example, in the case of students working in groups to discuss a conceptual problem, the potential obstacle is that two groups cannot agree on the solution. In the second prompt, respondents are then asked what they would do in that situation, and finally, in the third prompt, respondents are asked what they would do next if their previously articulated approach did not produce the desired results. These open-ended responses are then scored by trained raters, and those scores are used as the basis for comparing the strategic knowledge of novice science and mathematics teachers.

Validity Argument Framework

Foundational to the structure of the validity argument for any instrument is an articulation of the proposed interpretation of scores resulting from that instrument and the use of the instrument. Once defined, specific propositions supporting the score interpretation and evidence needed to evaluate those propositions can then be outlined.

Scores on the FASCI instrument are interpreted such that the strategic knowledge of novice science and mathematics teachers can be compared and distinguished, both relatively (i.e., norm referenced) and absolutely (i.e., criterion referenced). This is the proposed score interpretation. The FASCI instrument was developed in order to evaluate the effect of a teacher education program on novice science and mathematics teachers' strategic knowledge. More specifically, it was designed to measure levels of SK among prospective teachers from a variety of disciplines participating in a learning assistant program (Talbot et al. 2015; Otero et al. 2006) at a large research university. This is the proposed instrument use.

In order to support the proposed score interpretation and guide the collection of evidence needed to build the validity argument for the FASCI instrument, a set of propositions must be outlined. In identifying sources of validity evidence that might be used to evaluate each proposition, the categories set forth in the *Standards for Educational and Psychological Testing* (American Educational Research et al. 2014) are used. These categories include: (a) evidence based on test content, (b) evidence based on response processes, (c) evidence based on internal structure, (d) evidence based on relations to other variables, and (e) evidence based on consequences of testing. These propositions and the associated evidence that was collected to support them are shown in Table 2.

Propositions and Evidence Needed

In order to support the proposed score interpretation, five propositions must be evaluated. These propositions guide the collection of evidence used in the validation effort. The first two propositions focus on making a case that SK is important to measure (Proposition 1) and that it exists across all science and mathematics disciplines (Proposition 2). The argument supporting each of these propositions is a conceptual one and depends on evidence based on test content. Proposition 3 asserts that SK can be measured reliably with a scenario-based survey. Evidence needed to evaluate this proposition comes from response processes, the internal structure of the instrument, and the test (instrument) content. This includes the scores of FASCI responses from three raters, analysis of rater-scoring agreement, and observed score reliabilities. Because the FASCI instrument was designed to measure the SK of science and mathematics teachers

Table 2 FASCI score interpretation, instrument use, supporting propositions, and sources of validity evidence

Propositions	Evidence
1. SK is one type of knowledge required to be a quality science or mathematics teacher.	Conceptual argument [evidence based on test content]
2. SK exists across all domains of science or mathematics teaching (e.g., biology, chemistry, physics, math, etc.).	Conceptual argument [evidence based on test content]
3. SK can be measured reliably with a scenario-based .survey	Survey responses, interviews, analysis of scoring and scores [evidence based on response processes, internal structure, test content]
4. SK score interpretations change when specific science content is added to the items.	Comparison of FASCI versions [evidence based on test content, response processes, internal structure]
5. SK can be observed in teaching practice.	Comparison to observation protocol data [evidence based on relations to other variables]

Score interpretation and instrument use: the SK of novice science and mathematics teachers can be compared and distinguished both relatively and absolutely in order to evaluate the effects of a teacher education program on novice science and mathematics teacher's SK

from a variety of disciplines (e.g., chemistry, physics, mathematics, etc.), it is important to evaluate the proposition that SK score interpretations change when specific content is added to the items on the FASCI instrument (Proposition 4). The FASCI instrument was purposefully designed to be content-neutral in order to be useful for measuring levels of SK among novice science and mathematics teachers. Evidence needed to support this proposition comes from test content, response processes, and the internal structure of two versions of the FASCI instrument, the content-neutral version and one in which specific science content (physics) is incorporated into the items.

Proposition 5 addresses the assertion that SK can be observed in teaching practice. Evidence needed to evaluate this proposition comes from relations to other variables, specifically comparing FASCI scores to those from an observation protocol. This paper will focus on the investigation of Proposition 5, accepting the conceptual propositions (1 and 2) as extant and leaving the discussion of Propositions 3 and 4 to other manuscripts (Talbot 2011). The collection and analysis of the evidence needed to evaluate the proposition that SK can be observed in teaching practice is discussed in the next section.

Methods

In order to support the proposed score interpretation for the FASCI instrument (SK scores of novice science, technology, engineering, and mathematics (STEM) teachers can be compared and distinguished), it is important that SK can be observed in teaching practice. This provides an important source of *convergent validity* evidence for SK scores. In the remainder of this paper, I evaluate this proposition by comparing SK scores from the FASCI instrument to scores from the Reformed Teaching Observation Protocol (RTOP; Sawada et al. 2002) for a sample of novice STEM teachers. The RTOP is the most commonly used holistic observation protocol in studies of secondary and post-secondary STEM education (Lund et al. 2015). Specifically, I compare FASCI scores for respondents to their RTOP factor scores, which are based on factor analyses that were conducted by the developers of the RTOP (discussed in detail below).

FASCI Pilot Testing and Scoring

Two pilot tests of the FASCI provided response data for the validity studies. In Pilot Test 1, a five-item version of the FASCI was administered to a sample of 63 respondents. These respondents included undergraduate learning assistants, university faculty, practicing K-12 teachers, and university graduate students. In Pilot Test 2, a six-item version of the FASCI was administered to a sample of 96 respondents consisting of pre-service and novice practicing secondary math and science teachers. The main difference between these two versions of the FASCI is that the first version had five scenario-based items and the second had six. Two of the items were common between versions. In other words, three items on Pilot Test 1 were replaced with four new items for Pilot Test 2.

The open-ended item responses from the FASCI are scored using a set of decision rules and scoring guides (see Tables 3 and 4). The initial set of these decision rules were the result of an iterative process involving the work of members of the FASCI development team (part of the larger learning assistant research team). Subsequently, a new scoring team further developed these scoring guides based on response data. In scoring a response with both the initial and new sets of scoring guides, the response to Prompt (a) of each scenario ("How might this activity facilitate student learning?") is used as the basis for assigning an SCI score for that scenario. In scoring SCI with these guides, scores of 0, 1, or 2 are given.

In assigning an FA score for each scenario, responses to item Prompts (b) ("Describe both what would you do and what you would expect to happen as a result.") and (c) ("If the approach you described above in [b] didn't produce the result[s] you anticipated by the end of that class session, what would you do in the next class session?") are used. The response to Prompt (b) served as a baseline for comparing the Prompt (c) response. In order to achieve a score of at least 1 (the middle level), respondents must give evidence that they would change or at least modify their teaching strategy when presented with the potential obstacle in each scenario. If they further specify the conditions or reasons that determine that shift or change in strategic approach, they achieve an FA score of 2 (the highest category) for that scenario.

Three raters were trained to score the open-ended responses from the FASCI. After substantial training, final pairwise rater agreement on the FA dimension ranged from 80% to 91% and Cohen's kappa ranged from 0.63 to 0.82. On the SCI dimension, pairwise rater agreement ranged from 76% to 88% and Cohen's kappa ranged from 0.40 to 0.57.

Sample

The sample for comparing FASCI scores to observations of teaching practice consisted of 18 science and math teachers who were participants in an ongoing research program meant to assess the effectiveness of the Western State University (WSU) Learning Assistant (LA) program.

Table 3 FA scoring guide

Level	Modification of teaching approach	Discussion of contextual factors that bear on the modification of the teaching approach
2	YES	YES
1	YES	NO
0	NO	NO

Table 4 SCI scoring guide

Level	Discussion of interactive teaching	Discussion of a rationale for why they see this as an interactive situation
2	YES	YES
1	YES	NO
0	NO	NO

Seven of these teachers taught math, while the remaining 11 were science teachers. All were first-, second-, or third-year practicing teachers at the time of their FASCI participation (December 2008–January 2009). In these analyses, I compare the SK scores with the RTOP factor scores for these individuals. The RTOP instrument and the factor scores are described in detail below.

Each of these individuals responded to the FASCI and was observed at least two times during the spring semester of 2009 (two individuals were observed three times). These observations were conducted by five members of the WSU-LA research team who had established acceptable rater agreement on the RTOP prior to conducting the observations. The version of the FASCI to which they responded consisted of six scenario-based items (Pilot Test 2).

Observational Data

In addition to the FASCI scores for these individuals, their teaching episodes were scored with the RTOP at each observation. For these 18 respondents, there was no missing FASCI data and one missing RTOP observation. The RTOP instrument consists of 25 five-point Likert scale items[1] in three broad categories: lesson design and implementation, content, and classroom culture. The content category is further broken down into sections on propositional knowledge and procedural knowledge. The classroom culture category is further broken down into sections on communicative interactions and student/teacher relationships. Background information about the class and teacher are also noted on the first page of the protocol, and space is given to make notes about what occurs during the course of the observation. RTOP total scores are often used as the unit of analysis in research studies that use this instrument. A general rule of thumb is that aggregate RTOP scores above 50 (out of 100) are taken to indicate a reform orientation, while scores lower than that indicate a more traditional orientation (Piburn et al. 2000). While much broader than the SK construct, parts of the RTOP construct are related to FA or SCI. The relevant parts of the RTOP are represented by the factor scores used in the analysis below.

The RTOP was designed to measure reformed teaching in math and science. According to the developers, this construct is based in constructivism and the current reform movement in science and math education. In science education, the authors draw heavily on *Science for All Americans* from Project 2061 (Rutherford and Ahlgren 1991) and the *National Science Education Standards* (National Research Council 1996). From these, the RTOP developers highlight the importance of the standards for teachers of science. Specifically, the standards state that science teachers should promote investigations about nature, engage students actively in the process of learning science, and emphasize the importance of process rather than product. Also cited as foundational to the RTOP is the importance of moving students from concrete to abstract ideas and of working in collaborative environments (Piburn et al. 2000).

[1] The total possible score on the RTOP is 100 as the lowest category for rating on each item is zero.

From the mathematics education reform movement, the RTOP designers draw from *Principles and Standards for School Mathematics* (National Council of Teachers of Mathematics 2000). Specifically cited are the six principles and five generic standards. These principles are: (a) promotion of equity, (b) a vision for what is entailed in a curriculum, (c) a position on what knowledge is needed by mathematics teachers, (d) what it means to learn mathematics, (e) the importance of assessment, and (f) a promotion of the appropriate use of technology in teaching and learning mathematics. The standards cited focus on problem solving, reasoning and proof, communication, connections, representations, and having a vision of the classroom.

Using this framework as a guide, the RTOP developers drafted an observation protocol to be used in the evaluation of the Arizona Collaborative for the Excellence in the Preparation of Teachers (ACEPT) in 1989. The original version was designed for use in science classrooms and was revised into its present form (to be used in both science and math classrooms) after receiving input from mathematics educators. The 25 items divided into the three broad categories mentioned above are intended to "capture the full range of ACEPT reformed teaching" (Piburn et al. 2000, p. 9). These three broad categories and their sub-categories constitute the five subscales on the RTOP.

1. Lesson design and implementation
2. Content: Propositional pedagogic knowledge
3. Content: Procedural pedagogic knowledge
4. Classroom culture: Communicative interactions
5. Classroom culture: Student-teacher relationships

An analysis of the RTOP reliability and validity is presented in the reference manual and is based on 287 observations of 153 different classrooms that were conducted as part of a study comparing traditional and reformed teaching. Of particular relevance to this study are the results from the RTOP factor analysis that was conducted by the RTOP development team (Piburn et al. 2000). It is based on this analysis that I conceptualized the RTOP factor scores used for the sample in the present study. An initial principal component analysis conducted by the RTOP developers indicated three unique factors. However, the item loadings show that these three factors are not coincident with the three broad design categories of the instrument (lesson design and implementation, content, and classroom culture), as might be expected. The RTOP developers therefore identified and named three different factors: (a) inquiry orientation (onto which 20 of the 25 items load at 0.50 or greater), (b) content propositional knowledge (onto which five items load exclusively), and (c) collaboration (onto which three items load at 0.50 or greater, two of which also cross-load on Factor 1). Perhaps in response to the observed cross-loadings and the relatively small number of items loading onto one factor, the RTOP developers then used a cutoff value of 0.30 (rather than 0.50) for significance in factor loadings and subsequently identified five factors rather than three. These five factors seem to represent the most meaningful conceptual groupings of the items, and each of these factors was operationalized and described by the authors (Piburn et al. 2000). Therefore, it is these five factors (and their item groupings) that were identified by the RTOP developers and that I use as a basis for factor scores and for the comparisons with FA and SCI scores. The five RTOP factors are:

1. Inquiry orientation (items 3, 4, 11, 12, 13, 14, and 16). This is the same as Factor 1 identified in the initial principal component analysis. This factor is further described as "strongly suggestive of a pedagogy of inquiry."

2. Content propositional knowledge (items 6, 7, and 10). This is the same as Factor 2 identified in the initial principal component analysis. This factor is further described as "the scientific knowledge base contained in the lesson."
3. Content pedagogical knowledge (items 1, 5, 15, and 22). These items load on Factors 1 and 2. In discussing this factor, the authors relate it to pedagogical content knowledge (Shulman 1986).
4. Community of learners (items 2, 18, 20, 21, 24, and 25). These load onto Factor 1 and Factor 3 from the initial principal component analysis, not Factor 3 from this analysis. This factor is described as identifying the classroom as a collaborative place where the teacher acts as a resource person and a listener.
5. Reformed teaching (items 9, 17, and 19). These load on to all three factors from the initial principal component analysis. This factor describes a classroom that triggers divergent thinking where the teacher encourages student exploration.

The RTOP is strongly based in the literature of reformed math and science instruction. It also seems that the RTOP construct (reformed teaching) changed from conception to analysis based on the evolution of the instrument and on the validity evidence. Therefore, even though the RTOP and the FASCI are designed to accomplish a similar task (characterize science teachers' knowledge of practice), they have undergone different development processes and go about the task in different ways. However, because of the similarity between the RTOP and SK constructs, we should expect to see a positive correlation between the scores on each instrument.

Comparing FA and SCI Scores to RTOP Factor Scores

Because the RTOP total score is representative of a very broad construct and not easily comparable to FA or SCI scores, I used scores on each of the five factors (determined by item groupings) discussed above in this comparison. My goal in these comparisons was to identify cases where teaching characterizations based on each instrument were consistent (i.e., rated similarly on both instruments) or inconsistent (i.e., rated dissimilarly). Once these cases were identified through the descriptive statistical analysis, I was able to identify representative cases and compare their FASCI scores and responses with notes from the observations.

In comparing FA and SCI scores to the five RTOP factor scores, I use the mean values for FA and SCI scores and the mean scores for each RTOP factor (averaged based on all items that comprise that factor, across all observations available for that individual).[2] Originally, the FASCI scores used were based on only my ratings of the responses. Subsequently, the newly trained raters scored these responses with similar agreement to that reported above in rater training (80–90 % agreement on the FA dimension and 75–87 % agreement on the SCI dimension). The raters' fully moderated scores were used as a basis for the present analysis. These new scores did not differ substantially from my original scores, but represent a more reliable set of scores. Correlations between these scores are shown in Table 5. In general, there is only one notable correlation between the mean FA score and any of the RTOP factor scores, that

[2] RTOP scores were averaged across all observations (either 1, 2, or 3, depending on the individual) because teachers were purposefully observed more than once in order to account for the possible effects of observing an atypical lesson.

Table 5 Correlations between the mean FA or SCI score and RTOP factor scores

	RTOP Factor 1: Inquiry orientation	RTOP Factor 2: Content propositional knowledge	RTOP Factor 3: Content pedagogical knowledge	RTOP Factor 4: Community of learners	RTOP Factor 5: Reformed teaching
FA	0.29	0.10	0.24	0.11	0.24
SCI	0.33	0.49*	0.38	0.36	0.35

$n = 18$ for all correlations

*$p < 0.05$

being RTOP Factor 1 (inquiry orientation). However, in all cases, there are stronger correlations between the mean SCI score and RTOP factor scores than there are between the mean FA and RTOP factor scores. Note that for all correlations, the sample size is small ($n = 18$).

These correlations suggest focusing on the relationship between individuals' SCI scores and their scores on all RTOP factors and the relationship between their FA scores and scores on RTOP Factor 1. In order to identify cases that do not fit these trends, I examine scatter plots from each of these relationships (Figs. 1, 2, 3, 4, 5, and 6).

Although each of these scatter plots shows the general trend in relationship between each of the variables compared, they are more useful in identifying both outliers (cases that appear to be far outside the rest of the population) and consistently rated cases. For example, in four of these plots, one particular respondent (highlighted with a square shape in the plots of mean SCI

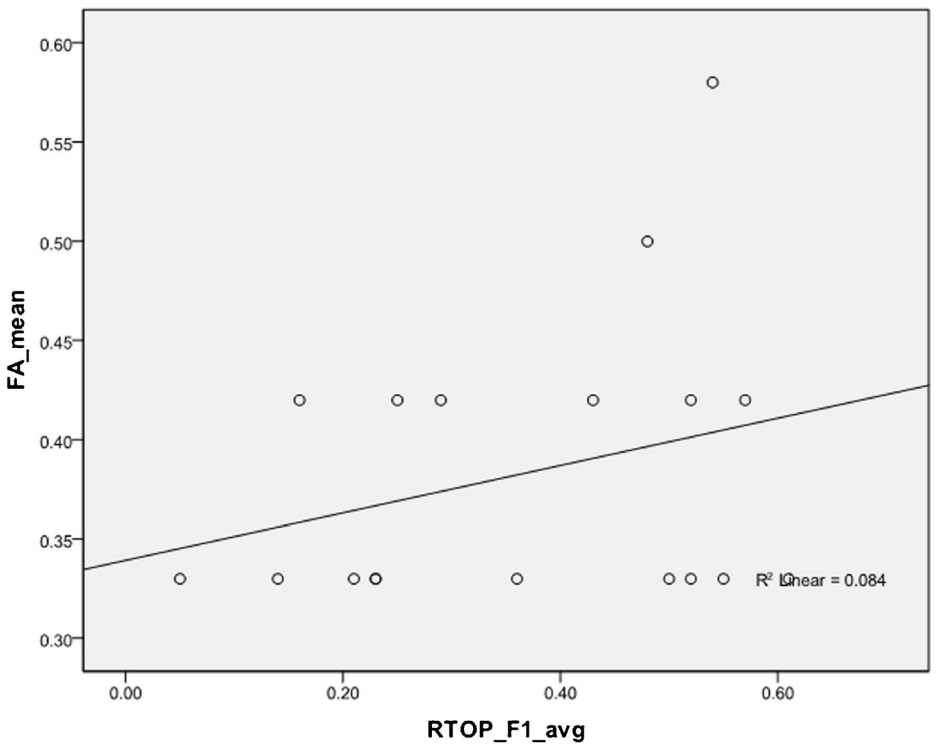

Fig. 1 Mean FA score vs. RTOP Factor 1 (inquiry orientation) score

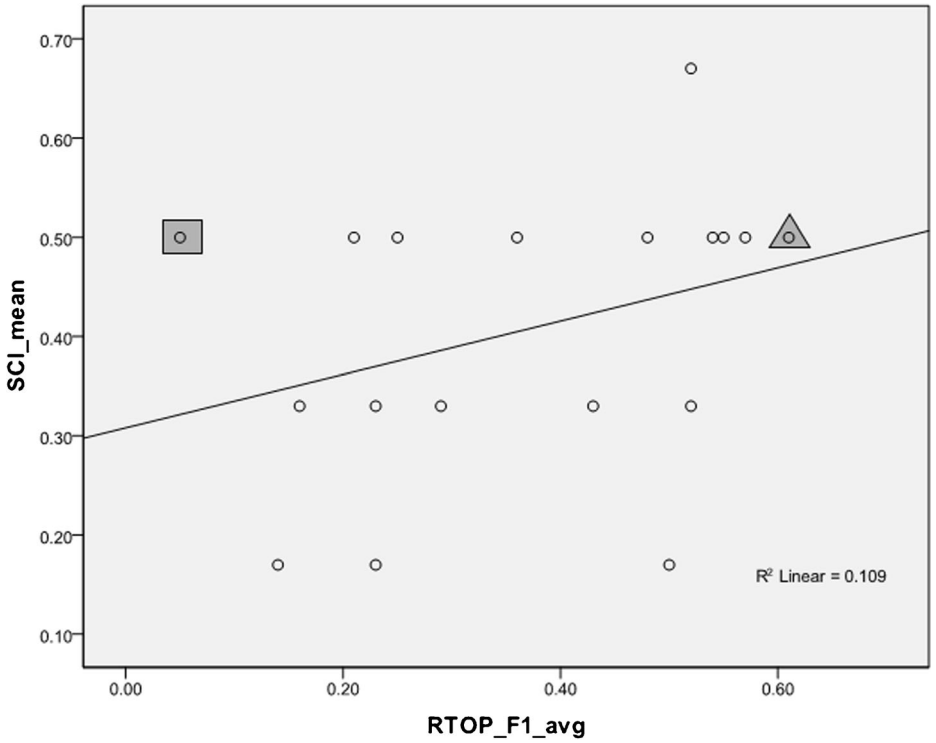

Fig. 2 Mean SCI score vs. RTOP Factor 1 (inquiry orientation) score

score vs. RTOP Factors 1, 3, 4, and 5) has high SCI scores and low RTOP factor scores and clearly exists as an outlier. Because this individual (George) was characterized differently based on his FASCI responses and observations of his teaching, he represents an interesting case to examine qualitatively and will be discussed below. In addition, one of the consistent cases (Ellie) has high SCI scores and RTOP factor scores and is highlighted with a triangular shape in the plots. I will also discuss her case in detail below.

To further examine the relationships between individuals' characterizations based on FASCI responses and those based on the RTOP, I also compute cross-tabulations of categorical scores on each measure. I conducted this analysis because correlations based on a small sample size can be sensitive to outliers, which clearly exist as observed in the plots. Average FA, SCI, and RTOP factor scores were binned into discrete categories: 0, 1, or 2 for FA and 0 or 1 for SCI (corresponding to the rating levels), and 0 (never occurred), 1 (low), or 2 (high) for the RTOP factor scores. Because the RTOP training manual specifically states that a rating of 0 corresponds to "never observed or occurred," I chose to isolate that from a low categorization and make it a distinct category. Average factor scores (on a scale of 0 to 1) greater than 0 but less than or equal to 0.50 were binned into the low category (1), and those greater than 0.50 were binned into the high category (2). Note that there were no average RTOP factor scores of 0 nor were there any high FA category scores for any of the individuals in the sample.

In examining these cross-tabulations, I systematically identified those individuals who were most frequently characterized inconsistently based on comparing the two measures (e.g., as low on the FASCI but high on the RTOP factor, or vice versa) and those who were most

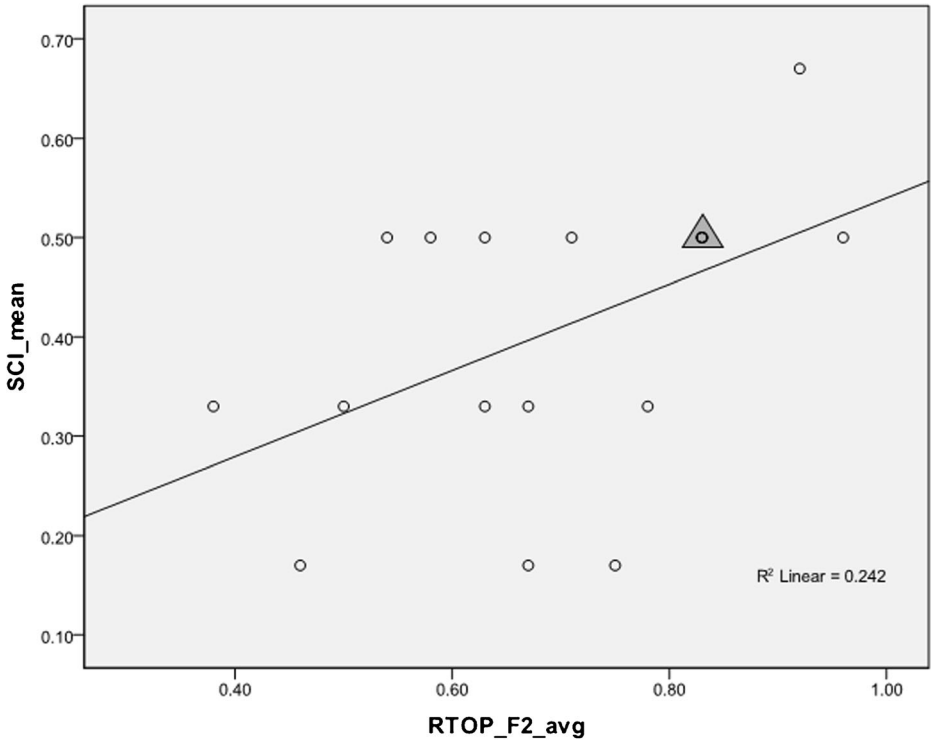

Fig. 3 Mean SCI score vs. RTOP Factor 2 (content propositional knowledge) score

frequently characterized consistently (e.g., low or high on both FASCI and RTOP). Because there were no high FA categorizations, for this comparison, I considered the middle level for FA (1) to be high. The pattern of these consistent/inconsistent characterizations is shown in Table 6. For each individual, there are 10 category comparisons: FA with each of the five RTOP factors and SCI with each of the five RTOP factors. Therefore, the number of comparisons in each row sums to 10. The inconsistent characterization columns (low:high and high:low) are shaded in Table 6.

Cases Identified for Further Analysis

Individuals of particular interest for qualitative analysis are listed by their pseudonym in Table 7. Jason and Ellie were chosen because they always scored high on the FASCI and on each of the RTOP factors. Laura was chosen because she was consistently low on both measures. James was chosen because he predominately scored low on the FASCI but rated high on the RTOP, although for two comparisons he was low on both measures. Finally, George (the individual identified as an outlier in the scatter plots) was chosen because of the inconsistent pattern of his characterizations (predominately either low on both measures or high on the FASCI and low on the RTOP). The FA, SCI, and RTOP factor scores (expressed in standard units) for each of the cases identified are shown in Table 7. In analyzing each of these cases, I present commonalities and differences in the rating comparisons identified above.

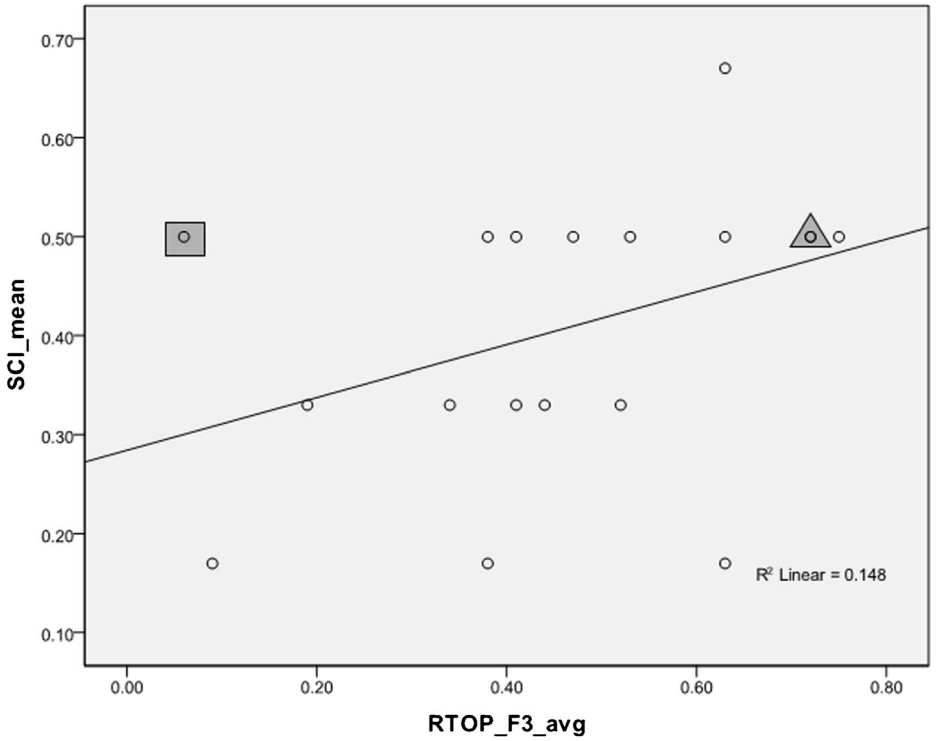

Fig. 4 Mean SCI score vs. RTOP Factor 3 (content pedagogical knowledge) score

Consistent Case Analyses

Laura scored consistently low on both dimensions of the FASCI and on the RTOP factors. Her mean scores were below the group mean values, far below in some cases (e.g., SCI and RTOP Factor 3). She was in her first year of teaching, taught ninth grade math, and was observed three times: twice in mid-February and once in early April. During two of the observations, student desks were grouped into threes, and during the third observation student desks were in rows facing the front of the class. In each class that was observed, Laura had a 5-minute warm-up task for the students at the beginning of class. Students worked on these tasks individually, and then Laura had volunteers work the warm-up problem on the board in front of the class. During two of the three observations, students in the class took a quiz (individually), and during one class they prepared for the upcoming State Student Assessment Program test.

Laura's FASCI responses were very brief, often characterized by one- or two-word responses and simple phrases. For example, three of her responses to Prompt (a) ("How might this activity facilitate student learning?") were "technology," "guided learning," and "higher level of thinking." She did not fully explicate how she conceived of each scenario in terms of facilitating student learning. However, her characterization based on her FASCI responses was consistent with that based on her RTOP scores. She scored well below the group mean on all RTOP factors. An examination of her item scores for RTOP Factor 4 (community of learners) shows that they were very low. Notes from observations confirm that she communicated very little with the students during her teaching and did not elicit their ideas at all. An examination of Laura's FASCI Prompt (b) and (c) responses shows that her low FA score was due to the

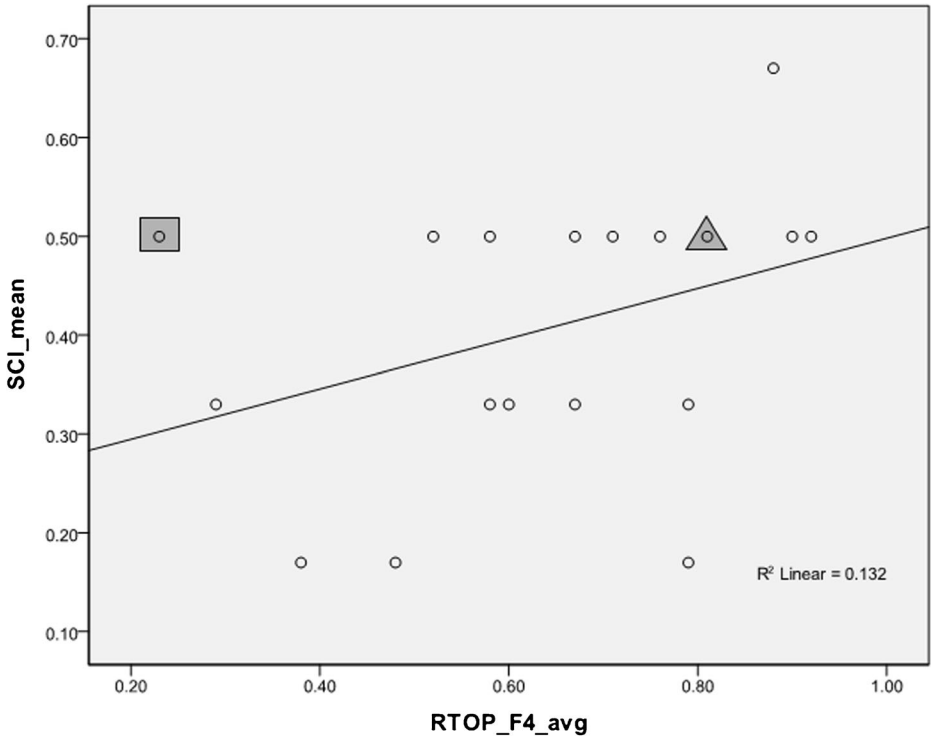

Fig. 5 Mean SCI score vs. RTOP Factor 4 (community of learners) score

fact that she repeated the same strategy whenever faced with a potential obstacle. For example, she often wrote that she would "go over another similar problem." In summary, Laura's FASCI and RTOP characterizations were consistently low based on quantitative and qualitative comparisons. I inferred that she did not appear to be very student centered, nor did she have a very large repertoire of strategies from which to draw upon. Her case provides convergent evidence for the validity of FASCI score interpretation.

Ellie scored consistently high on all FASCI-RTOP comparisons, and her mean scores were above the group mean in all but one category (RTOP Factor 2). She was a first-year teacher who taught math to 10th, 11th, and 12th grade students. She was observed once in early March and twice in April. During each observation, Ellie had the students working in groups on assignments, worksheets, or conceptual questions (e.g., "Come up with a definition of asymptote."). In two of the three observations, it is clear that she interacted quite a bit with each of the groups, asking questions such as "What do you think?" and "Do you agree?" These interactions are typified by an instance where she took time to talk to a group that did not want to work together and presented an alternative for them in which they worked independently but discussed with each other before writing down their final answers. I also noted that there was a high degree of student–student talk and interaction in her class. In addition to group work, Ellie also used presentation, explanation, and discussion strategies in her teaching.

Ellie's FASCI responses were consistent with these observations. She mentioned students "reasoning through their opinions and defending or rejecting them." She often invoked questioning strategies in response to Prompt (b), further confirmation of her conception of student involvement in her class. In addition to questioning, she also cited the use of

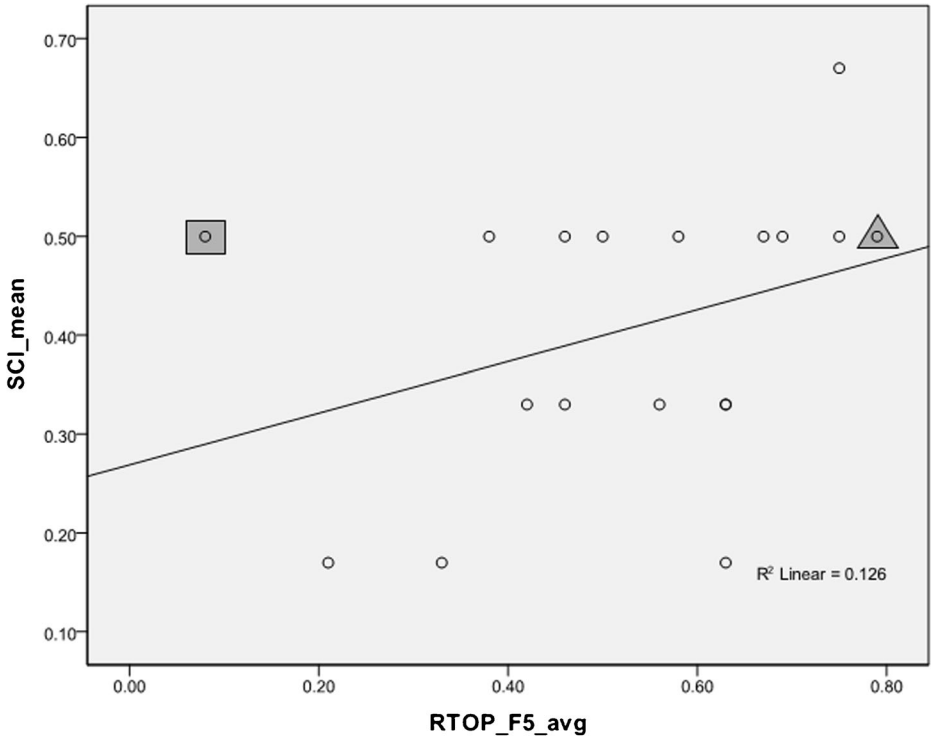

Fig. 6 Mean SCI score vs. RTOP Factor 5 (reformed teaching) score

presentation, explanation, and visual representations in her FASCI responses. Though she did not cite the contextual dependence of her strategic choices, it is clear that she had a repertoire of strategies from which to draw (based on her FASCI responses and RTOP notes). For example, in response to the FASCI item about having made a mistake when working a problem on the board, Ellie first wrote that she would have the students find the mistake. When prompted for what she would do next if that approach did not work (Prompt (c)), she wrote that she would have the students estimate a reasonable answer to the problem. Further evidence of her repertoire of strategies comes from one observation of her teaching in which Ellie was observed to use questioning, modeling, and explanation strategies all in a span of 15 minutes. Based on these teaching descriptions and item responses, I inferred that Ellie was student-centered in her thinking and that she had a repertoire of strategies to draw upon. Again, this case provides convergent evidence for the validity of the FASCI.

Jason, like Ellie, was also consistently high on all FASCI–RTOP factor comparisons. Of the cases identified for analysis, he had the highest mean FA score (0.58) and the highest mean score on three of the RTOP factors (2, 3, and 4). Jason was a first-year teacher at the time and taught life science to seventh grade students. He was observed twice, once in February and once in April. In each of the observed classes, students began with a warm-up activity related to the day's topic and shared their work before Jason proceeded with any formal presentation of the material. A discussion of science-related current events also took place each day. It is apparent from the notes that students' ideas were elicited and valued, and there was a high degree of student talk during each class. For example, Jason would often pose a question and then have the students discuss it in pairs before sharing with the class. In one classroom observation, this

Table 6 Comparison of FA and SCI category rating to RTOP factor category rating, all individuals in the sample

	[FA/SCI Categorization] : [RTOP Factor Categorization]			
Individual	[Low] : [Low]	[Low] : [High]	[High] : [Low]	[High] : [High]
1	2	3	2	3
2	3	2	3	2
3	.	.	6	4
4	.	5	.	5
5	.	.	2	8
6	3	2	3	2
7	8	2	.	.
8	2	3	2	3
Jason	.	.	.	10
10	.	5	.	5
George	4	1	4	1
12	.	5	.	5
Ellie	.	.	.	10
James	2	8	.	.
Laura	10	.	.	.
16	.	5	.	5
17	5	.	5	.
18	8	2	.	.

strategy was observed three times. It also appeared that each time Jason asked for a volunteer to share an idea with the class, there were many student responses. Jason posed many divergent questions to his students (e.g., "Science is global. What do you think is meant by that?"). He used multiple strategies to facilitate student discussion, such as individual work time, think–pair–share activities, clickers, and whole-class discussions. All of this explains why Jason had the highest mean score on RTOP Factor 4 (community of learners) among the group.

Jason's FASCI responses confirm and support these classroom observations. His high mean FA score is due to the fact that he invoked multiple strategies in response to the FASCI scenarios and sometimes cited the contextual dependence of his strategic choices (e.g., "dependent upon time…" and "if the students thought it made sense…"). His frequent choice of using questioning strategies is also evidence of his desire to hear students' ideas and to

Table 7 FA, SCI, and RTOP factor scores (standard units) for identified cases

Case	FASCI:RTOP	FA	SCI	RTOP F1	RTOP F2	RTOP F3	RTOP F4	RTOP F5
Jason	High:high	2.72	0.65	0.98	0.84	1.26	0.82	0.73
Ellie	High:high	0.51	0.65	1.15	−0.71	0.82	0.58	0.83
George	Mixed	−0.74	0.65	−1.82	−0.40	−1.95	−2.02	−2.33
James	Low:high	−0.74	−1.66	0.75	−0.15	0.82	0.73	0.52
Laura	Low:low	−0.74	−1.66	−1.31	−1.45	−1.81	−1.29	−1.65

engage them in the lesson, which is evidence of his student-centeredness. In his Prompt (a) responses, Jason mentioned the importance of having students "verbalize their thoughts and convey them to others." He also mentioned having them do this in pairs, which is consistent with what was observed in his classroom.

In summary for these consistent cases, there is a strong agreement in characterization based on the RTOP and FASCI. In each case, specific strategic choices and student-centered dispositions can be seen in the observation notes and in the FASCI responses. However, note that each of these individuals represents an extreme case; Laura rates very low on the constructs, and Ellie and Jason both rate very high. Although each of these cases seems to support the validity for FASCI score interpretations, none of them could be considered average based on comparing their mean scores to those of other teachers in this sample.

Inconsistent Case Analyses

James was in his second year teaching ninth grade math at an urban high school when he was observed. He was observed three times, once in late January, once in April, and once in early May. His classroom was equipped with a Promethean projection system that he used each day for formal presentation. James class consisted of about 16 Hispanic students, about two thirds of which were female. His students used the AVID (Advancement Via Individual Determination 2015) notebook structure. He began class with a warm-up activity projected onto the front board that students worked on individually. James would generally circulate around the room and help students as they worked on this activity. He then presented the material for the day before giving them individual or small-group work time to complete a related homework assignment. James gave each student in his class individual attention at some point during the class period. For example, after formal presentation, James would walk around to each student and talk with them about their work. He spoke with them individually rather than addressing the group in which the student was working. He encouraged them to participate in the work and in answering questions during whole-class activities, though only a few students ever volunteered to answer questions during class. There was not much student talk during the classes and very little talk between students (about the topic at hand).

All of James' inconsistent FASCI–RTOP factor comparisons come from having a low rating on FA or SCI and a high rating on the RTOP (refer to Table 6). His mean FA score was relatively low (0.33) and his mean SCI score was very low (0.17). In his responses to Prompt (a) on the FASCI, James only once mentions students interacting with each other in the teaching scenarios. All of his other comments were about students working through something or thinking about something individually. This seems somewhat consistent with what was observed in his classes, but what the FASCI did not detect is the individual attention James gave to each student during class. In part, this led James to achieve higher than average scores on most RTOP factors. In general, his ratings on items within the classroom culture category were higher than average, indicating that James had interactive relationships with his students. In the observations, it was evident that James wanted to involve every student and did this on a one-on-one basis. None of his survey responses indicated this type of student-teacher interaction.

One possible reason for these differences is the uniqueness of James' teaching situation relative to the other teachers who were observed. Perhaps the FASCI scenarios were different enough from James' classroom environment that his constructed responses were not framed in his actual practice. In other words, what he wrote on the FASCI could have been completely hypothetical in his mind and not related to what happens in his classroom. If this were the case,

then the FASCI could be contextually limited in the sense that the teaching scenarios are being interpreted by some respondents as assuming a common set of conditions or constraints that do not exist across all classroom environments. Another possible reason for the difference is the amount of James' teaching experience relative to the other teachers in the sample. He was in his second year of teaching at the time and had been a learning assistant as well, meaning that he has had substantially more teaching experience than the teachers discussed above.

George only rated highly on RTOP Factor 2 (content propositional knowledge) and rated very low on the other RTOP factors. His SCI score was high (0.50) and his FA score was low (0.33). He was in his first year of teaching science in an ethnically diverse high school classroom (about 45 % Hispanic, 5 % African American, and 50 % Caucasian) and was observed three times. He started each class period with a question of the day on the board (e.g., "Who is Rocky the Rock Cycle?"), which he had students write down in their notebooks. In most cases, George then began class by presenting the content, after which he had the students work on some task either individually or in groups. Based on the observation notes, George often had to address off-task behaviors and activities (e.g., taking away an iPod, kids hitting each other, and off-task conversations). Notes from each of the three observations also indicate that George's class was very content-focused. There were few observations of student talk or teacher elicitation of students' ideas, but many notes about definitions (of an igneous rock, for example) and observations of students working from the textbook or on worksheets. He employed mostly lecture or explanation, followed by individual student work (worksheets, students filling in diagrams from information in their book, and individual student writing assignments). Very little student talk was noted, nor were students' ideas ever observed to be the focus of the lesson or class activities.

Based on the observation data, the biggest inconsistency in George's characterization was appears to be related to the SCI dimension. Although he scored relatively high on SCI, George's classroom practices did not look very student-centered, which was reflected in his RTOP factor scores (especially RTOP Factors 1, 3, 4, and 5, which are inquiry orientation, content pedagogical knowledge, community of learners, and reformed teaching). It appears that although although George discussed students' active engagement in the learning process in his FASCI responses, his practice did not reflect this conception. With respect to the FA dimension, there is not as much discrepancy: George's relatively low mean FA score (0.33) was consistent with his very low RTOP Factor 1, 3, and 5 scores (those that correlated most highly with the mean FA score). In each of his observations, off-task behavior and disciplinary issues were observed. Among the observed sample of teachers, this was unique to George's teaching setting.

Discussion

In comparing the consistent and inconsistent cases with respect to FASCI and RTOP ratings, one important distinction arises: teaching context. In the cases of Laura, Ellie, and Jason (consistent cases), nothing was noted in the observations that seemed unique when compared to the rather general contextualization of the FASCI scenarios. In the inconsistent cases (James and George), observations did indicate a somewhat unique teaching context when compared to the FASCI scenarios. In the case of James, his classroom environment was characterized by trying to actively engage his students who seemed very reluctant to participate. Although he gave each student individual attention (and therefore scored relatively high on RTOP factors), his SCI responses did not reflect this teaching practice. As stated above, the generic framing of

the FASCI scenarios may have been so different from his classroom environment that they were not consistent with his classroom experiences. In the case of George, his classroom was characterized by off-task and behavior issues during his content-heavy presentations of the material. Though he scored high on SCI, his conceptions about student involvement were not reflected in his practice, perhaps due in part to these classroom management issues. In his case, the FASCI scenarios may have been hypothetical situations that were unattainable in the teaching and learning context that he and his students shared. Interestingly, George also had relatively low FA scores and did not cite these classroom management issues as relevant contextual factors that affected his strategic choices. If this difference between actual classroom context and FASCI teaching context is really a difference that matters, then perhaps the FASCI scenarios are contextually limiting.

Conclusions

The context of measurement matters in characterizing teacher knowledge. Survey-based measures need to have flexibly contextualized prompts that still yield reliable measures. But that is a true challenge—too much latitude in a prompt may lead to increased construct-irrelevant variance in responses and result in low score reliability (Messick 1989, 1995). Indeed, in the development of the FASCI, both highly constrained contexts and more open contexts were tested for the scenario-based items, but finding a balance to achieve highly informative yet reliably scorable responses was difficult. The work presented in this paper highlights this tension and suggests that different teaching contexts might call for different measures of teacher knowledge. Future work could test this hypothesis by developing two contextually different versions of an instrument for measuring teacher knowledge and administering both versions to a sample of teachers whose practice aligns with these two contexts. Inferences resulting from these measures could then be compared to inferences drawn from observations of practice, which would either support or refute the findings presented in this study.

If measuring a teacher's knowledge was as simple as using a meter stick, none of this would constitute a contribution to the field. However, any dimension of teaching knowledge is far more complex than the basic dimension of length and, accordingly, our tools for measurement are not as simple as those used for measuring length. However, whether simple or complex, attempting to measure something requires one to develop a deeper understanding of the object of measurement and of the context within which that measurement is being made.

References

Advancement Via Individual Determination. (2015). *About AVID*. http://www.avid.org/about.ashx. Accessed 14 Jan 2015.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research, 35*, 463–482.

Bond, L., Smith, T., Baker, W. K., & Hattie, J. A. (2000). *A distinction that matters—why national teacher certification makes a difference*. Greensboro, NC: National Board for Professional Teaching Standards.

Briggs, D., Geil, K., Harlow, D., & Talbot, R. M. (2007). *Measuring the pedagogical sophistication of Math and Science Teachers using Scenario-based Items*. Paper presented at the American Educational Research Association Annual Meeting.

Hammerness, K., Darling-Hammond, L., Bransford, J., Berliner, D. C., Cochran-Smith, M., McDonald, M., & Zeichner, K. M. (2005). How teachers learn and develop. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: what teachers should learn and be able to do* (pp. 358–389). San Francisco: Jossey-Bass.

Lund, T. J., Pilarz, M., Velasco, J. B., Chakraverty, D., Rosploch, K., Undersander, M., & Stains, M. (2015). The best of both worlds: building on the COPUS and RTOP observation protocols to easily and reliably measure various levels of reformed instructional practice. *CBE Life Sciences Education*, *14*(2), ar18.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational and psychological measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

National Research Council. (1996). *National Science Education Standards: observe, interact, change, learn*. Washington, DC: National Academies Press.

Otero, V., Finkelstein, N., McCray, R., & Pollock, S. J. (2006). Who is responsible for preparing science teachers? *Science, 313*(5786), 445–446.

Piburn, M. D., Sawada, D., Turley, J., Falconer, K., Benford, R., Bloom, I., & Judson, E. (2000). *Reformed teaching observation protocol* (*RTOP*) *reference manual*. Tempe, AZ: Arizona Collaborative for Excellence in the Preparation of Teachers, Arizona State University.

Rutherford, F. J., & Ahlgren, A. (1991). *Science for all Americans*. Oxford: Oxford University Press.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: the reformed teaching observation protocol. *School Science and Mathematics, 102*(6), 245–253.

Shulman, L. S. (1986). Those who understand: knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14.

Talbot, R. M. (2011). *Embedding content into an instrument designed to measure Novice Science and Mathematics Teachers' Strategic Knowledge: A challenge for validity*. Phalaborwa, South Africa: International Conference on Mathematics, Science, and Technology Education.

Talbot, R. M., Hartley, L., Marzetta, K., & Wee, B. (2015). Transforming undergraduate science education with learning assistants: student satisfaction in large enrollment courses. *Journal of College Science Teaching, 44*(5), 24–30.

U.S. Department of Education, Office of Postsecondary Education, Office of Policy Planning and Innovation (2002). *Meeting the highly qualified teachers challenge: the secretary's annual report on teacher quality*. Washington, DC

Van Driel, J. H., Verloop, N., & de Vos, W. (1998). Developing science teachers' pedagogical content knowledge. *Journal of Research in Science Teaching, 35*(6), 673–695.