Embedding content into an instrument designed to measure novice science and mathematics teachers' strategic knowledge: A challenge for validity

Robert M. Talbot III
University of Colorado Denver, USA
Correspondence: robert.talbot@ucdenver.edu

Abstract
Strategic Knowledge (SK) is comprised of two dimensions: Flexible Application (FA) and Student Centered Instruction (SCI). The FA dimension describes how a teacher invokes, applies and modifies her instructional repertoire in a given teaching context. The SCI dimension describes how a teacher conceives of a given situation as an opportunity for active engagement with the students. The Flexible Application of Student-Centered Instruction (FASCI) instrument was designed to measure science and mathematics teachers' SK by eliciting open-ended responses to scenario-based items. This study focuses on the effect of embedding specific science content into these scenario-based items. It was found that adding content did have an effect on FASCI responses and scores. The difference appears to be due to the elicitation of construct irrelevant responses, rather than due to the accessing of a more sophisticated SK as was hypothesized. Implications for the difficulty of defining and measuring teacher knowledge constructs are discussed.

Keywords: Teacher knowledge, measurement, validity, reliability

1.1 Introduction

Defining science and mathematics teacher knowledge can be quite a challenge. Teacher knowledge is very complex, and is often thought of as some combination or synthesis of many different knowledge bases, some of them tacit (Verloop, Van Driel, & Meijer, 2001). Even if it can be defined, measuring it can be equally challenging. However, measures of teacher knowledge are increasingly being used to characterize outcomes of teacher education programs for accountability purposes (e.g., Johnston & Merrifield, 2010). Given the potential consequences of judgments resulting from uses of these measures, the challenges of defining and measuring teacher knowledge cannot be taken lightly. Therefore 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" (AERA, APA, & NCME, 1999, 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.

In this study, one aspect of such an instrument validation effort is presented and discussed. The particular instrument under scrutiny (the "Flexible Application of Student Centered Instruction" instrument, or FASCI) was developed in order to determine the effect of a teacher education program on novice science and mathematics teachers' *Strategic Knowledge* (SK). Strategic knowledge consists of how a teacher conceives of student engagement in the learning process, and what teaching strategies they apply in various teaching scenarios. As an example of strategic knowledge, when teaching a student about Newton's Third Law paired forces an expert teacher may consider the student's motivation

and prior ideas about the topic, and the contextual factors that bear on the teaching-learning interaction (time, resources available, etc.) before choosing to do a demonstration or engage the student in a Socratic dialog. In contrast, a content expert who is not an expert teacher may invoke the same teaching strategy (perhaps an explanation) regardless of the student or context.

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 who come from a variety of disciplines (e.g., chemistry, physics, mathematics, etc.). The term "content neutral" is used to imply that the situations presented in the FASCI scenarios are common to science and mathematics classrooms, but not to other disciplines (e.g., language arts). But this content neutrality may pose a threat to the validity of score interpretations if one believes that an instrument which is based in specific science or mathematics content is able to elicit a respondent's SK differently than a content-neutral version. The purpose of this study is to investigate if embedding specific science content into the items on the FASCI instrument has an effect on item responses and the resulting SK scores. Evidence needed to investigate this aspect of FASCI validity comes from comparing item responses and scores resulting from two versions of the FASCI instrument, the "content-neutral" version and one in which specific science content (in this case, physics) is embedded into the items.

1.2 The FASCI Instrument

The SK construct is comprised of two dimensions that are labeled Flexible Application (FA) and Student Centered Instruction (SCI) (Briggs, Geil, Harlow, & Talbot, 2007). The 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 which are warranted by the given context (the "Flexible Application" (FA) dimension of strategic knowledge). As well, 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 "Student-Centered Instruction" (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. These items are rather broadly contextualized: *"For the scenario that follows, please assume (unless it is otherwise specified) that you are teaching a high school course in physics, chemistry, biology, earth science or math to a class of 25-30 students."* It is this broad science and mathematics content characterization that is referred to as the "content neutral" character of the FASCI. An example FASCI scenario-based item is shown in Figure 1. In these items, a classroom scenario is presented which 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 one-on-one with a student. The first question prompt asks how the respondent thinks the activity would facilitate student learning. A potential obstacle is then presented which further frames the scenario. For example, in the case of students working in groups to discuss a conceptual problem, the potential obstacle is that the students in one group cannot agree on the solution. In the second prompt, the respondent is then asked what they would do in that

situation, and finally, in the third prompt the respondent is 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.

Example FASCI item

For the questions 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 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?

Figure 1. Example scenario introduction and scenario-based item on the FASCI

Most of the instruments reviewed which seek to measure some aspect of science and mathematics teacher knowledge do so in a less economical, more direct fashion: through observations of practice and follow-up interviews (cf. Talbot, 2011). This is what Kind (2009) refers to as an *in situ* approach. One exception is the Mathematical Knowledge for Teaching instrument (MKT; Hill, Schilling, & Ball, 2004), which does use a more indirect approach (what Kind refers to as the "prompt and probe" approach). But the MKT employs a multiple choice item design rather than an open-ended response format. The items on the MKT have a structure that presents hypothetical situations similar to the FASCI scenarios, but the MKT is much more focused on math teachers' subject matter knowledge, rather than strategic knowledge. Therefore the content-neutral, scenario-based, constructed-response structure of the FASCI items is somewhat unique.

This structure is not completely without precedence, however. For example, in many studies on teachers' pedagogical content knowledge (PCK), "critical events" are used as a prompt around which to discuss teacher actions (e.g., Hashweh, 1987; Shulman, 1986). Also, in the Mosaic II project, researchers at RAND developed what they call "vignette-based surveys" to measure science and mathematics teachers' reform-oriented practices (Le et al., 2004). Separate scenarios were created for mathematics and for science and present content specific to each domain. The scenarios are very specific for each grade level, subject, and the curriculum being taught by the teachers at each site. Responses are not open-ended; rather the teacher rates the likelihood (on a scale of 1-4) that they would engage is a particular practice, given the scenario. For example, after being presented with specific information about a math problem and how a particular group has gone about approaching the problem, the respondent is asked how likely they are that they will "ask the class if they can think of another way to solve the problem" (on a scale of 1-4, where 1 = very unlikely and 4 = very likely). Compared to these items, the scenario-based items on the FASCI are unique in the sense that they are not specific to a particular grade level, topic, and curriculum, and they require the respondent to construct an open-ended response.

The content-neutrality of the FASCI may be problematic from a validity standpoint, in that score interpretations could change if specific science or math content was added to the scenarios as in the Mosaic II project. In order to evaluate this potential problem, evidence based on an alternative instrument structure is needed—one which *embeds* specific content in the scenarios.

1.2.1 The Content Test

The "physics-FASCI" (or "p-FASCI") was developed based on the existing content-neutral FASCI (or "n-FASCI"). It was designed by placing the current FASCI scenario-based items within the context of specific physics topics: Newton's Third Law and Free Fall. These topics were chosen because of their ubiquity in general physics courses, and because they are "rich" topics in the sense that much research has been done on common student prior ideas/alternative conceptions related to these areas (e.g., Clement, 1982; Halloun & Hestenes, 1985; Viennot, 1979). The p-FASCI includes physics contextual information that precedes each set of scenario-based items. There are two of these item sets, the first focusing on the topic of Newton's Third Law and including three item scenarios, and the second focusing on Free Fall and including two item scenarios.

In the contextual information that precedes the scenarios on the p-FASCI, respondents are presented with the key concepts of each topic (in the form of learning objectives) and some problems/questions illustrating the topic. In this way, the FASCI scenarios have been made to be much more specific to a grade level, content, and curricular context, much like the scenarios in the Mosaic II project described above (though still requiring an open-ended response). Each of these elements is included for specific reasons. First, learning objectives related to the content are included to situate the physics content in the hypothetical course setting (a high school physics classroom). An example of one of these learning objectives is as follows: "Students should be able to apply Newton's Third Law in analyzing the forces that two objects in contact exert on each other when they accelerate together along a horizontal or vertical line, or the forces that two surfaces that slide across one another exert on each other." This piece of information tells the respondent exactly what they are trying to teach their students and therefore, what they (the teacher) need to know in terms of their own content knowledge. Second, example conceptual questions for each content piece are provided in order to further contextualize the scenario and in order to gather information about respondent content knowledge. These questions prompt for responses and can therefore be used to gauge the physics content knowledge of the respondents. These example questions were modified from the Force and Motion Conceptual Evaluation (FMCE; Thornton & Sokoloff, 1998). The n-FASCI also contained a set of the same 12 physics content items to which individuals responded at the end of the survey. These items were used as one way to characterize the physics expertise of respondents. An example of the scenario introduction and an item from the p-FASCI is shown in Figure 2.

p-FASCI

For the questions and scenarios that follow, please assume that you are teaching a high school course in physics to a class of 25-30 students. You have defined the following learning objectives for this class:

- Students should understand Newton's Third Law so that, for a given system, they can identify the force pairs and the objects on which the forces are exerted, and specify the magnitude and direction of each force.
- Students should be able to apply Newton's Third Law in analyzing the forces that two objects in contact exert on each other when they accelerate together along a horizontal or vertical line, or the forces that two surfaces that slide across one another exert on each other.

To assess your students' understanding of this content, you have given them the following conceptual questions:

The next set of questions refer to a large truck which breaks down out on the road and receives a push back to town by a small compact car. Pick one of the choices which correctly describes the forces between the car and the truck.

1) The car is pushing on the truck, but not hard enough to make the truck move…

Please respond to the following questions about your teaching:

2) Students are working in groups of four to discuss the conceptual questions about the car pushing the truck.

   a) How might this activity facilitate student learning?

   As the activity proceeds, one group gets frustrated and approaches you—they cannot agree on the answers regarding the forces exerted by the car and truck on each other.

   b) Describe both what would you do and what you would expect to happen as a result.

   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?

Figure 2. Example scenario introduction and scenario-based item on the p-FASCI

As can be seen in Figure 2, the new p-FASCI items are essentially parallel, but not quite identical to, the scenario-based items on the n-FASCI. The same scenarios are used in both the p-FASCI and the n-FASCI, since the only variable manipulated was the specific science or mathematics content in the constraint which frames the items. The actual item prompts within each scenario are identical on both versions of the FASCI.

One might expect that score interpretations change for physics experts when physics content is added to the items due to the ability of the p-FASCI to access a "more sophisticated" knowledge base in these respondents, perhaps something akin to pedagogical content knowledge (PCK; Shulman, 1986). If score interpretations change because the instrument is accessing this more sophisticated knowledge base, then this could be a validity issue for the content-neutral version of the FASCI. If they change for some other reason, then it may be the case that this more sophisticated SK does not exist, or that the p-FASCI is not eliciting responses related to the SK construct.

1.3 Methods

1.3.1 Sample and Administration

The target population for this study consists of prospective science teachers in the teaching methods courses at institutions that are a part of The Physics Teacher Education Coalition

(PTEC; http://www.ptec.org). The PTEC institutions ($n \approx 175$) were chosen because they have a relatively large number of prospective physics teachers. Prospective science teachers are defined as those students who are in some kind of undergraduate science teaching course, whether they have committed to a career in teaching or not. A sample of prospective teachers from four PTEC sites was recruited to participate in this research (see Table 1). These sites were chosen because of their accessibility and due to the fact that at each, there was a teacher educator interested in the FASCI development. These prospective teachers were concentrating not only in physics, but also in different disciplines such as chemistry, biology, geology, and astronomy. In addition, participants at a fifth (non-PTEC) site (Central Research University) were asked to volunteer to take the surveys. The science methods course instructor at this institution expressed interest in having her students participate, though she did not require their participation. None of the respondents were offered incentives for participating.

At the beginning of the spring 2009 semester, the neutral and physics-FASCI surveys were administered to all available prospective teachers in science methods courses at the five participating sites (see Table 1). Students in each course section were randomly assigned to respond to either the n- or p-FASCI. There was no random assignment for students at Northeast Queen's University, because the course instructor wanted her students to respond only to the p-FASCI.

Table 1.
FASCI respondents by university and version

	Number Invited to Participate	Number of n-FASCI respondents	Number of p-FASCI respondents
Southeast Coastal University (SCU)	25	14	11
Northwest Pacific University (NPU)	4	2	2
Northeast Queen's University (NQU)	11	0	8
Western State University (WSU)	25	8	12
Central Research University (CRU)	13*	2	1

*Participation not required at Central Research University (a non-PTEC site)- students volunteered

In all administrations of the FASCI, demographic and academic background information was collected from participants. At the end of each version is a space for open-ended comments and feedback regarding participation. Each version of the FASCI was administered online using the QuestionPro web-based software (http://www.questionpro.com). Average completion time for respondents on the n-FASCI was 46 minutes, and for the p-FASCI it was 39 minutes. Responses from each version were downloaded, cleaned (e.g., blank response sets were deleted, formatting was corrected), and loaded into Google Docs for scoring through the use of forms. Results were exported to Microsoft Excel and SPSS for analysis.

1.3.2 Response Scoring

The open-ended item responses from each version of the FASCI are scored using a set of scoring guides. Abbreviated versions of these scoring guides are shown in Figures 3 and 4. The initial set of these scoring guides were the result of an iterative process involving the work of members of the FASCI development team. Subsequently, a new scoring team further developed these guides based on analyzing some response data. This new scoring team consisted of three practicing secondary science teachers who possess a high level of strategic knowledge. In scoring responses, 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. If the respondent conceives of the activity as an opportunity for interactive teaching and learning, then the response is given a score of at least 1. If they further specify a rationale for why they conceive of the situation as an interactive one, the response is given a score of 2 (see Figure 3). This scoring results in five unique SCI scores for each respondent on the FASCI, assuming that they responded to all item prompts.

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), a respondent has to 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 which determine that shift or change in strategic approach, they achieved an FA score of 2 (the highest category) for that scenario (see Figure 4). Again, five FA scores are possible for a respondent assuming that they responded to all item prompts.

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

Figure 3. SCI 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

Figure 4. FA scoring guide

For scoring purposes in this study, responses from both the neutral and physics versions of the FASCI were pooled together and raters were not told about the two different versions. In other words, they were blind to the FASCI version. None of the raters ever questioned the differences in responses or asked why some respondents cited specific physics content while others did not. For training purposes, a subset of responses from a separate administration of the n- and p-FASCI was used. This subset of responses was not used in quantitative analyses discussed below.

Overall rater agreement for the full response set (60 responses) is shown in Tables 2 and 3 for the FA and SCI dimensions respectively. On the FA dimension, agreement between pairs of raters ranged from 80%-90%. Cohen's kappa is also given for each pair of raters. This statistic is a bit more critical than percent agreement, in that it takes into account that rater agreement could have occurred by chance. On the FA dimension, kappa between pairs of raters ranged from 0.63 to 0.82. For the SCI dimension, agreement was not as high as that on the FA dimension. Percent agreement between pairs of raters ranged from 76%-88% and kappa ranged from 0.40 to 0.57.

Table 2.
Overall FA Rater Agreement

Rater Combination	Percent Agreement	Cohen's Kappa
r1-r2	83%	.68
r1-r3	80%	.63
r2-r3	91%	.82

Table 3.
Overall SCI Rater Agreement

Rater Combination	Percent Agreement	Cohen's Kappa
r1-r2	83%	.52
r1-r3	76%	.40
r2-r3	88%	.57

1.4 Analyses and Findings

1.4.1 Score Comparisons

Scores on each version of the FASCI were compared in two ways: 1) mean FA and SCI scores for each dimension by version (neutral or physics) that were averaged based on the number of items answered by each individual (i.e., based only on those items to which they responded), and 2) mean FA and SCI scores on each version for only those individuals who had complete response sets across both dimensions (i.e., they responded to all item prompts). The second analysis was conducted in order to check the sensitivity of the first score comparisons to the existence of missing data. In comparing results from the two analyses, they were found to be quite similar. However, the first method has the advantage of using all response sets and will therefore yield higher statistical power. Results from the first method are presented below.

Table 4 shows the mean scores for each dimension (FA and SCI) by version[1]. Scores on the p-FASCI were higher on FA but lower on SCI, and the difference was statistically significant for SCI. The magnitude of the difference for SCI was almost twice that of the difference for FA.

Table 4.
Mean FA and SCI scores averaged by number of responses (SD) by version

Dimension	n-FASCI sample size	n-FASCI mean score	p-FASCI sample size	p-FASCI mean score	Difference (p minus n)	p-value from t-test	Effect size (Cohen's d)	power
FA	26	0.57 (0.42)	34	0.69 (0.32)	0.12	0.19	0.33	0.35
SCI	26	0.78 (0.36)	34	0.58 (0.34)	-0.20	*0.03	-0.58	0.71

*significant at p<0.05

The differences in mean scores can also be expressed in terms of an effect size (in this case, as Cohen's d, see equation 1).

$$ES = \frac{(\bar{X}_{p-FASCI} - \bar{X}_{n-FASCI})}{SD_{pooled}} \quad (1)$$

These effect sizes are expressed as the difference between mean p-FASCI and n-FASCI scores, divided by the pooled standard deviation of these scores. A positive effect size represents a higher score on the p-FASCI; a negative effect size represents a higher score on the n-FASCI. Respondents who took the p-FASCI had SCI scores that were, on average, 0.58 standard deviations *lower* than the SCI scores of students who took the n-FASCI (an effect size of -0.58). This difference is quite large. The effect size for the mean FA score comparison is 0.33, which is a moderate effect size.

Statistical power was calculated *post hoc* for the tests of significance between mean score comparisons. These power estimates give the probability of rejecting the null hypothesis (in this case, that the mean scores between each version are the same) when it is false. For mean SCI scores between the n- and p-FASCI the statistical power is 0.71. This means that there is a 71% probability that a test of significance would reject the hypothesis that the n- and p-FASCI SCI scores are the same when in fact that is not true. For mean FA scores, statistical power is 0.35. The FA value for statistical power is quite low, indicating that the null hypothesis is not likely to be rejected if the mean scores between versions really are different. This finding should be interpreted with caution, as the *post hoc* calculation of

[1] In making this comparison, mean FA and SCI scores for incomplete response sets were averaged based on the number of scores in that set. For example, if a particular respondent answered only three FA items out of five possible, and the sum total of their FA scores was two, then their average FA score would be 0.66 (2/3), whereas a score for them based on all possible FA scores would be 0.40 (2/5). This averaging does not penalize the respondent for having missing response data. One shortcoming of this method is that it may bias the mean values if primarily easier or harder items were completed by the respondent. To investigate this, an analysis was conducted where missing scores were replaced with the mode score for that particular item and the resulting mean scores were compared based on this method. In this way, mean scores were not biased based on the difficulty of missing item response data. There was no difference in the mean scores; therefore one conclude that the averaged mean scores presented in Table 4 are not biased due to differences in the difficulty of items for which responses are missing.

statistical power in order to interpret results is somewhat controversial, unless the explanation is being used to inform the design of a future study (Howell, 2009).
The distribution of the mean FA and SCI scores on the n- and p-FASCI are shown in Figure 5. In these histograms, bin width is approximately equal to half of the standard deviation of each distribution of scores. In the right-hand panel of Figure 5, the lower SCI scores on the p-FASCI relative to the n-FASCI can be seen. In the left-hand panel, one can see that the mean FA scores are higher on the p-FASCI.

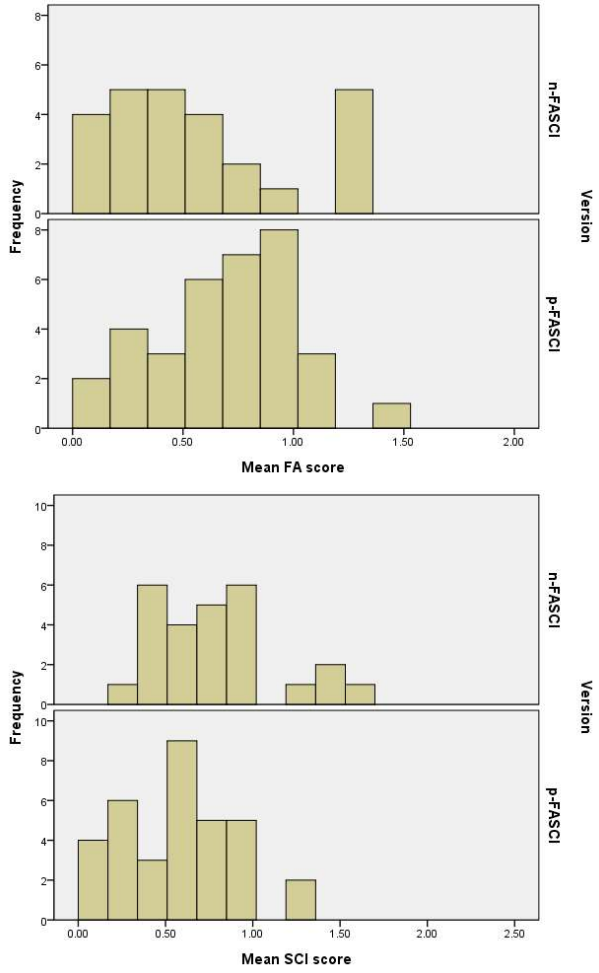

Figure 5. Distribution of mean FA and SCI scores averaged by number of responses for n- and p-FASCI

1.4.2 Score Comparisons by Physics Expertise

One might think that any differences in FA and SCI scores between the two versions of the FASCI instrument is related to the physics expertise of respondents in each group. Respondents were classified as either physics experts or novices based on each of three variables: 1) subject they plan on teaching (if physics, then they are an "expert"), 2) physics content knowledge score[2] (if greater than about 1 SD above the mean value, then they are an "expert"), and 3) number of physics courses taken (if greater than five then they are an "expert"). Comparisons between physics experts on each version based on these

[2] This measure is based on responses to the 12 physics content items (adapted from the FMCE) that were appended to the end of each version of the survey.

classifications are shown in Table 5. Note that these comparisons include the Northeast Queen's University sample, which accounts for the larger number of physics experts on the p-FASCI.

Table 5.
Comparison of FA and SCI scores between n- and p-FASCI (pre-test) for physics experts

Expertise Classification	Version	Number of Experts	FA mean (SD)	p-value	SCI mean (SD)	p-value
Subject plan on teaching (physics = expert)	n-FASCI	6	0.61 (0.32)	0.81	0.62 (0.38)	0.23
	p-FASCI	13	0.57 (0.30)		0.43 (0.27)	
Physics content knowledge (> .9 = expert)	n-FASCI	5	0.75 (0.52)	0.72	0.69 (0.20)	0.65
	p-FASCI	8	0.83 (0.25)		0.63 (0.23)	
Number of physics courses taken (> 5 = expert)	n-FASCI	3	0.71 (0.27)	0.99	0.71 (0.60)	0.22
	p-FASCI	10	0.71 (0.31)		0.40 (0.28)	

In these comparisons, FA scores are about the same or differ only slightly on each version and the differences are not statistically significant. The SCI scores are consistently higher on the n-FASCI, and the differences are not statistically significant for any of the expertise classifications. In summary then, it appears that there is no significant difference (statistical or practical) between FA scores or SCI scores for physics experts on each version of the FASCI.

The above comparison of scores between versions of the FASCI shows that for the aggregate sample (i.e., physics experts and non-experts) there is a significant difference in SCI scores between each version of the FASCI. Also, the magnitude of the difference in mean SCI scores is almost twice that of the differences in FA scores between versions. In order to gain further insight into the differences between scores on each version of the FASCI instrument, the next section addresses score reliabilities.

1.4.3 Score Reliability

The score reliabilities, observed score SDs, sample sizes and percent of missing response data are shown disaggregated by version of the FASCI in Table 6. These reliability estimates (reported as Cronbach's alpha[3]) are based on the average item and total scores across all three raters. Note that the reliability estimates are much lower for the SCI dimension than for the FA dimension. The reliability of SCI scores derived from the n-FASCI is slightly higher (by 0.05 to 0.09) than that found in previous pilot testing, and reliability of FA scores is

[3] Note that Cronbach's alpha can be thought of as an upper bound on score reliability, and as such is likely over-estimating reliability (Wainer & Thissen, 2001). In a separate study, I critically examine FASCI score reliability using a Generalizability Theory approach (G Theory; Brennan, 2001) in order to identify specific sources of measurement error.

within the range of that found in previous pilot testing (one of which had a sample size of 92).

Table 6.
Reliability estimates and overall percentage of missing data for each dimension by version

Version	Sample size	FA alpha	SD of Observed FA score	% FA missing	SCI alpha	SD of Observed SCI score	% SCI missing
n-FASCI	26	0.69	0.42	7.7	0.51	0.36	3.8
p-FASCI	34	0.45	0.32	32.4	0.41	0.34	20.6

As seen in Table 6, p-FASCI score reliability is lower than that for the n-FASCI on both dimensions. Two findings discussed below help to explain this large difference: 1) a larger percentage of missing response data on the p-FASCI and 2) the elicitation of construct irrelevant responses on the p-FASCI.

The disproportionately high number of incomplete response sets and percentage of missing data on the p-FASCI deserves scrutiny, given that each sample was similar on the characteristics surveyed. There are two plausible explanations for this missingness. First, some respondents found the physics content frustrating (especially if they are self described as not being "physics people"). The second explanation applies specifically to the Northeast Queen's University sample, for which there was a very large percentage of incomplete response sets (38%). Respondents from this university were mostly characterized as physics experts, but were frustrated with the scenario-based items of the FASCI in general, not with the content. In personal communications with their instructor, I became aware of the fact that students often stopped responding because they found the teaching scenarios presented to be discordant with their own beliefs about teaching and learning. For example, when some of these respondents encountered the FASCI scenario which begins with the statement "You have just finished giving a presentation", they became frustrated because they did not believe a presentation could facilitate student learning (example response to item 3, prompt a): *"I have no clue how it would exactly as this is a very vague statement. I don't like the idea of presenting"* ID 3491548).

It is reasonable to think that the high percentage of missing data on the p-FASCI has an effect on score reliability for that version of the instrument. Replacing this missing data with mean item or person scores is one way to deal with the missing data and may yield more robust estimates of reliability for the p-FASCI (cf., Downey & King, 1998)[4]. However, the observation that the p-FASCI has so much missing data relative to the n-FASCI calls into question the validity of the p-FASCI itself. If respondents are not completing the survey out of frustration or because of some other factor, then it might be difficult to ever obtain complete response sets with this version of the instrument.

The very low SCI score reliability, taken together with the finding of statistically significant differences in mean SCI scores between versions, calls for a deeper investigation of the SCI item responses.

[4] I replaced missing data on the p-FASCI with mean scores for each rater-item combination and computed reliability estimates again. Doing this increases both FA and SCI score reliability only slightly.

1.4.4 Qualitative Analyses of SCI Item Responses

To further investigate the significant difference in SCI scores between the n- and p-FASCI, the prompt a) item responses from each version were examined. While qualitatively examining these responses, it was found that many p-FASCI respondents discussed the content provided in the scenario, rather than discussing the students. This finding is interpreted as a potential source of construct-irrelevant variance (Messick, 1994). In other words, item responses to prompt a) on the p-FASCI include both information about to target construct (strategic knowledge) and information about some other construct. This could be biasing subjective judgments about these responses and therefore distorting response scoring.

Table 7 shows a comparison of the prompt a) item responses on the n- and p-FASCI. Note that this table does not include 100% of prompt a) responses on each version, as some responses discussed something other than students or content.

Table 7.

Percentage of responses to prompt a) on each version of the FASCI coded as discussing students or content

Version	Percentage of Responses Discussing Students	Percentage of Responses Discussing Content
n-FASCI	89%	1.6%
p-FASCI	73%	16%

This examination of prompt a) responses on the n-FASCI revealed that the respondents were predominately discussing students—89% of n-FASCI responses focused on the students. However, the p-FASCI prompt a) responses looked different. Fewer of the responses discussed the students (73%), whereas a larger proportion of responses were coded as discussing the physics content (16%). This indicates that prompt a) on the p-FASCI is eliciting targeted responses *less consistently*. As discussed above, SCI score reliability was lower on the p-FASCI than that on the n-FASCI, which corresponds with the observation that 16% of prompt a) responses on the p-FASCI were coded as discussing the content. Some example prompt a) responses can help to illustrate this finding.

Below is an example of a response to prompt a) of item 2 on the n-FASCI ("You are working out an example problem up on the board. How might this activity facilitate student learning?") which was coded as discussing students (ID 3449137):

> *it would facilitate student learning by having the visual and audio aspect of teaching to the students. I would allow questions and I would ask students what would be the next step for me in the problem to get closer to the answer.*

This response was assigned an SCI score of 1 because the respondent discusses interacting with the students ("…allow questions…" and "…ask students what would be the next step…"). It appears that this individual conceives of the teaching situation as an opportunity for the students to have an active role.

A different response to the same prompt is presented below, this one from a comparable individual at the same university who responded to the p-FASCI. Item 2 on the p-FASCI reads "On the board, you are drawing free body diagrams of the car and the truck." This individual's response to prompt a) ("How might this activity facilitate student learning?"; ID 3458893):

Free-body diagrams facilitate students learning because it demonstrates all the forces that have been taken into consideration and makes it easier to spot for problems.

This second respondent focuses on the physics content rather than on the students in the classroom. This response was assigned an SCI score of 0, as she did not discuss anything about students taking an active role in the scenario. These two very similar respondents (at least in terms of the characteristics surveyed) approached the same scenario and prompt very differently on the two versions of the FASCI. A similar comparison can be made across other paired responses to prompt a) from other scenarios. For example, consider prompt a) responses to item 1 which has consistently been the easiest SCI item in previous administrations of the n-FASCI. On the n-FASCI, this item reads "Students are working in groups of four on a conceptual question you gave them at the beginning of class" and on the p-FASCI, it reads "Students are working in groups of four to discuss the conceptual questions about the car pushing the truck." In the present study, many p-FASCI respondents scored 0 on this item. Examples from two p-FASCI respondents:

(ID 3476461):

They might better understand force diagrams and statics by recognizing the balance of forces.

(ID 3482375):

It shows real world examples of various forces involved in acceleration/deceleration.

It seems that the physics content presented in the p-FASCI is eliciting from the respondents a discussion of the content as well as the students. In other words, item prompt a) (which targets SCI) is performing differently on the two versions: on the n-FASCI, it is eliciting responses about the students, but when the content is embedded in the scenario (as in the p-FASCI), item prompt a) is eliciting something different, and not the "more sophisticated" SK that was hypothesized. Respondents are not discussing students as much; rather they are discussing the content more. This discussion of content can be interpreted as being irrelevant to the target construct (strategic knowledge) and can therefore be considered to be a source of construct-irrelevant variance. As mentioned above, this could have biased scoring judgments and is in turn reflected in the observed differences between SCI scores and score reliabilities on each version of the FASCI.

1.5 Discussion

Can novice science and mathematics teachers' Strategic Knowledge (SK) be measured in a *valid* way with the FASCI instrument? This question drives a larger set of validity studies of the FASCI instrument. This particular study focuses on investigating the effect of embedding specific science content into the content-neutral FASCI scenarios. It was hypothesized that a "more sophisticated" SK exists that is based in a teachers' content area of expertise, and that this knowledge base could be accessed by a content-specific version of the FASCI instrument. However, this was not found to be the case.

Other challenges to establishing the validity of the FASCI instrument have been identified (notably, score reliability and observing SK in practice) but the inclusion of specific science content into the scenarios poses a particularly interesting challenge to FASCI validity. The SK construct was conceptualized as being specific to the domains of science and mathematics in that the teaching scenarios presented and strategies most often used are not common to other domains (e.g., language arts). However, the SK construct was

operationalized in a content-neutral way, not specifying any particular science or mathematics content in the construct or associated survey items.

SK is also conceptualized as being related to pedagogical knowledge (PK) which, in turn, is one component of pedagogical content knowledge (PCK; Shulman, 1986). Though PCK can be thought of as being some integration or synthesis of many different knowledge bases (cf. Grossman, 1990; Shulman, 1987), most researchers agree on PK and subject matter knowledge (SMK) as being central to PCK (Gess-Newsome, 1999; van Driel, Verloop, & de Vos, 1998). However, it remains unclear how PK and SMK are related in comprising PCK (Kind, 2009), as little empirical evidence exists in order to make a case for a well-defined relationship between these constructs. The uncertainty in the relationships between SK, PK, SMK, and PCK are illustrated in Figure 6.

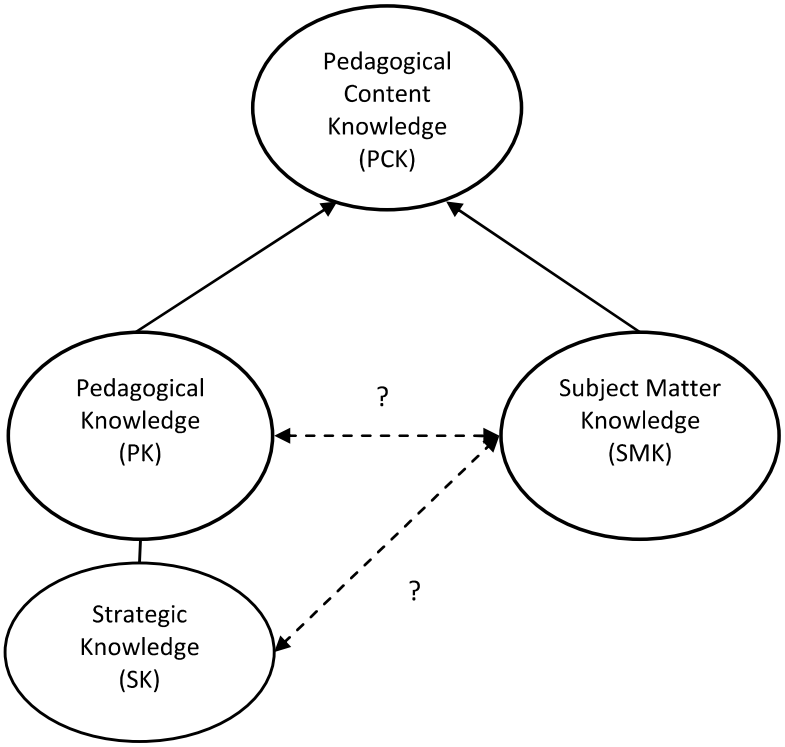

Figure 6. The uncertainty in relationships between teacher knowledge constructs

While one might hypothesize (based on the construct of PCK) that an instrument with specific science content embedded would do a better job in accessing an individual's SK, this was not found to be the case in the present study. The inclusion of specific content elicited responses that were not relevant to the target domain (SK). As a result, scores were lower for the p-FASCI respondents, and item responses on the p-FASCI referenced the content more often than the students. This biasing of scores due to construct-irrelevant variance (Messick, 1994) in p-FASCI responses presents a major challenge for the validity of the p-FASCI.

The findings from this study highlight the difficulties inherent in measuring teacher knowledge and in defining the relationships between teacher knowledge constructs. One implication of these findings is that the finer-grained components of teacher knowledge (e.g., SK, PK, SMK, etc.) can likely be measured more reliably and in a more valid way than some combination or amalgam of these knowledge bases (e.g., SK + SMK, PCK, etc.). Defining these knowledge bases is in itself quite a challenge, and operationalizing them in a

way that can lead to a valid measure is even more challenging. But these challenges must be pursued if we are to truly understand science and mathematics teacher knowledge.

1.6 References

AERA, APA, & NCME. (1999). Validity. In Aera (Ed.), *Standards for Educational and Psychological Testing* (pp. 9-24): AERA.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

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

Clement, J. (1982). Students' Preconceptions in Introductory Mechanics. *American Journal of Physics, 50*(1), 66-71.

Downey, R., & King, C. (1998). Missing data in Likert ratings: A comparison of replacement methods. *Journal of General Psychology*, 175-191.

Gess-Newsome, J. (1999). Pedagogical Content Knowledge: An Introduction and Orientation. In J. Gess-Newsome & N. G. Lederman (Eds.), *Examining Pedagogical Content Knowledge: The Construct and its Implications for Science Education* (pp. 3-17). Dordrecht: Kluwer.

Grossman, P. (1990). *The Making of a Teacher: Teacher Knowledge & Teacher Education*. New York: Teachers College Press.

Halloun, J. A., & Hestenes, D. (1985). Common sense concepts about motion. *American Journal of Physics, 53*, 1056-1065.

Hashweh, M. Z. (1987). Effects of subject-matter knowledge in the teaching of biology and physics. *Teaching and Teacher Education, 3*(2), 109.

Hill, H., Schilling, S., & Ball, D. (2004). Developing Measures of teachers' measures mathematics knowledge for teaching. *Elementary School Journal, 105*(1), 11-30.

Howell, D. C. (2009). *Statistical Methods for Psychology* (7th ed.): Wadsworth Publishing.

Johnston, M., & Merrifield, M. (2010). *Colorado Senate Bill 10-036: Program Results for Teacher Preparation*.

Kind, V. (2009). Pedagogical content knowledge in science education: perspectives and potential for progress. *Studies in Science Education*, 169-204.

Le, V.-N., Stecher, B., Hamilton, L., Ryan, G., Williams, V., Robyn, A., & Alonzo, A. (2004). Vignette-Based Surveys and the Mosaic II Project: RAND.

Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher, 23*(2), 13-23.

Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher, 15*(2), 4-14.

Shulman, L. S. (1987). Knowledge and Teaching: Foundations of the New Reform. *Harvard Educational Review, 57*(1), 1-22.

Talbot, R. M. (2011). *Validity Issues in the Evaluation of a Measure of Science and Mathematics Teacher Knowledge.* (PhD dissertation), University of Colorado, Boulder. Retrieved from http://goo.gl/OPPFz

Thornton, R. K., & Sokoloff, D. R. (1998). Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation *American Journal of Physics 66* (4), 228-351.

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.

Verloop, N., Van Driel, J., & Meijer, P. (2001). Teacher knowledge and the knowledge base of teaching. *International Journal of Educational Research, 35*(5), 441-461.

Viennot, L. (1979). Spontaneous Reasoning in Elementary Dynamics. *European Journal of Science Education, 1*(2), 205-221.

Wainer, H., & Thissen, D. (2001). True Score Theory: The Traditional Method. In D. Thissen & H. Wainer (Eds.), *Test Scoring* (pp. 23-72). Mahwah, NJ: Lawrence Erlbaum Associates.