VALIDITY ISSUES IN THE EVALUATION OF A MEASURE OF SCIENCE AND MATHEMATICS TEACHER KNOWLEDGE

by

Robert M. Talbot III

B.S., Indiana University, 1996

M.S., Indiana University, 2000

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

School of Education

2011

This thesis entitled:
Validity Issues in the Evaluation of a Measure of Science and Mathematics Teacher Knowledge
written by Robert M. Talbot III
has been approved for the School of Education

Dr. Derek Briggs (co-chair)_____

Dr. Valerie Otero (co-chair)_____

Dr. Erin Furtak_____

Dr. Steve Pollock_____

Dr. David Webb_____

Date _____

The final copy of this thesis has been examined by the signatories, and we
find that both the content and the form meet acceptable presentation standards
of scholarly work in the above mentioned discipline.

IRB Protocol # 0306.16

Talbot, Robert M. (Ph.D., Education)

Validity Issues in the Evaluation of a Measure of Science and Mathematics Teacher Knowledge

Thesis directed by Associate Professor Derek Briggs and Associate Professor Valerie Otero

This study investigates the reliability and validity of an instrument designed to measure science and mathematics teachers' *strategic knowledge*. Strategic knowledge is conceptualized as a construct that is related to pedagogical knowledge and is comprised of two dimensions: Flexible Application (FA) and Student Centered Instruction (SCI). The FA dimension describes how a science teacher invokes, applies and modifies her instructional repertoire in a given teaching context. The SCI dimension describes how a science teacher conceives of a given situation as an opportunity for active engagement with the students. The Flexible Application of Student-Centered Instruction (FASCI) survey instrument was designed to measure science teachers' strategic knowledge by eliciting open-ended responses to scenario-based items. This study addresses the following overarching question: What are some potential issues pertaining to the validity of measures of science and mathematics teacher knowledge? Using a validity argument framework, different sources of evidence are identified, collected, and evaluated to examine support for a set or propositions related to the intended score interpretation and instrument use: FASCI scores can be used to compare and distinguish the strategic knowledge of novice science and mathematics teachers in the evaluation of teacher education programs. Three separate but related studies are presented and discussed. These studies focus on the reliability of FASCI scores, the effect of adding specific science content to the scenario-based items, and the observation of strategic knowledge in teaching practice. Serious issues were found with the reliability of scores from the FASCI instrument. It was also found that adding science content to the scenario-based items has an effect on FASCI scores, but not for the reason hypothesized. Finally, it was found that more evidence is needed to make stronger claims about the relationship between FASCI scores and novice teachers'

practice. In concluding this work, a set of four recommendations are presented for others who are engaged in similar measure development efforts. These recommendations focus on the areas of construct definition, item design and development, rater recruitment and training, and the validation process.

This dissertation is dedicated to my wife, Catherine B. Talbot. Without her confidence in me, I could not have completed this marathon.

Acknowledgements

I would like to thank my advisers and dissertation co-chairs, Derek Briggs and Valerie Otero. Since taking my first course with Derek years ago, he has always challenged me to think critically about my own work, and has always been there to offer an alternative perspective on many a problem. Working with Derek has made me a better researcher, writer, and thinker, and for those things I am grateful. Valerie has provided me with many valuable opportunities during my doctoral career. Without her projects, connections, guidance, and motivations, this study would have never existed. Valerie has also helped me to see how research problems fit within a much bigger picture. As I begin my own career in academia, I realize how important this framing ability is, and I am grateful that I was able to observe her do that in many situations over the past few years.

My work would not have been possible without numerous other faculty mentors. Erin Furtak graciously allowed me to observe her secondary science methods course. From that experience I learned a lot about my own research problem and even more about secondary science teacher education. David Webb provided me with invaluable guidance on this and other research projects. Through a course that I took with him, he motivated me to develop a keen interest in teachers' assessment practices. Steve Pollock assisted me in the development of the physics-specific version of the survey used in my dissertation research. His knowledge of the structure of physics conceptual questions proved very valuable to my work.

My doctoral student colleagues have also provided me with incredible support and assistance throughout my studies. The LA-TEST research team helped me with collecting observation data, for which I am grateful. Specifically, Mike Ross assisted me in scoring survey responses and in working on other projects, and Kara Gray provided tremendous help with observation protocol data and was always willing to talk about analytical issues. There are two colleagues in particular that I cannot begin to

thank. Heidi Iverson has served as my confidant and sounding board for the past few years, and was always available to discuss research, teaching, and anything else that came to mind. She always pushed me to think carefully about my work. Heidi has a very strong repertoire of research skills, and often helped me to deal with some not so straightforward situations. Mark Lewis is a valuable colleague and an even better friend. He was always available to listen to an idea, help generate new ideas, and provide a critical eye for any piece of writing. From coursework to research to job search, Mark and I shared many experiences and many good times.

Finally, I could never have begun to engage in these studies without the support and guidance of my family. My mother, Sue, has always been my role model as an educator. An ardent supporter of public education, an experienced school teacher, and a respected university administrator, she has modeled for me the value of being educated. My father Bob, an artist and philosopher, has always been my most influential teacher. From an early age, dad has instilled in me a curiosity to ask questions about the world around me. My own daughters, Annabelle and Lilian, have provided me with a constant source of joy, inspiration, and motivation over the last twenty months. My wife Catherine has not only motivated me to work hard and succeed, but she has also been a true source of love and patience over the years. She has worked extremely hard so that I could pursue my goals, and has always provided a practical voice when I needed it the most. My “three girls” are the light of my life.

CONTENTS

TABLES

Chapter 2

Chapter 3

Chapter 4

Chapter 5

FIGURES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 1: Validity Issues in the Evaluation of a Measure of Science and Mathematics Teacher Knowledge

Introduction

As we strive to develop teacher education programs capable of preparing "highly qualified teachers" (United States Department of Education, 2007), we must be able to evaluate the effectiveness of these programs. Recent legislation in the state of Colorado mandates annual reporting on "the effectiveness of educator preparation programs" (Johnston & Merrifield, 2010). This reporting will include data from teachers who are graduates of a teacher preparation program and are in their first three years of practice (i.e., "novice" teachers). The legislation specifically calls for reporting on student academic achievement, educator placement, and educator mobility and retention. Though not specifically articulated in the legislation, measures of teacher knowledge are also an outcome of interest for this reporting.

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 which cannot be taken lightly. A teacher education program could be deemed ineffective based on such data, or at 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 both valid and reliable. 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). "Reliability" means the degree to which an instrument consistently measures that which it is intended to measure (cf., Traub, 1994). Both of these topics will be taken up and discussed in

detail in subsequent chapters. Lacking the characteristics or validity and reliability, 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, I present the case of one such instrument development and 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*. 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.

It is important to be able to measure science and mathematics teachers' strategic knowledge, as this knowledge is foundational to instructional and assessment practice. The characteristics of strategic knowledge (discussed below) are those of expertise in general, and are related to formative assessment practices specifically. I will further make this case in the next chapter. As no other instrument exists with which to measure science and mathematics teachers' strategic knowledge, it is imperative that this new instrument be subjected to the scrutiny of a validation effort.

Validation in Brief

Although there is a clear need for valid and reliable instrumentation which measures teacher quality (and in the case of this study, strategic knowledge), the process of investigating instrument validity is not a simple undertaking. Making a case for the validity of an instrument is complex, and

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 which are then evaluated—not the instrument itself. Based on the proposed score interpretation and instrument use, a set of propositions which 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 other ways than that defined by the proposed 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 which can be used to evaluate the effect of a teacher education program on novice teachers' knowledge. First and foremost 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? In the next chapter, I will describe why strategic knowledge is required of a highly qualified science or mathematics teacher, and how a teacher education program is expected to contribute to this knowledge base. This turns out to be one of the propositions that underlie the proposed use the instrument. Below, I briefly describe strategic knowledge and the instrument designed to measure it, before laying out in more detail the structure of the validity argument for the instrument.

Strategic Knowledge and the FASCI Instrument

The strategic knowledge a construct is comprised of two dimensions that are labeled, respectively, Flexible Application (FA) and Student Centered Instruction (SCI) (Briggs, Geil, Harlow, & Talbot, 2007). The relationship between the FA and SCI dimensions are presented here and will be explained in detail in the next chapter.

The FA dimension describes how a science or mathematics teacher invokes, applies and modifies her instructional repertoire in a given teaching context. At the most novice level in the FA dimension, a teacher has a very limited repertoire of strategies from which to draw, and with development she not only gains a larger repertoire of strategies, but she also gains 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, Smith, Baker, & Hattie, 2000; Hammerness et al., 2005).

The SCI dimension describes how a science or mathematics teacher conceives of a given situation as an opportunity for active engagement with the students, in order that she can identify the students' current understanding. At the lowest level, the teacher does not see the activity or scenario as an opportunity to elicit information from her students about their current level of understanding. At a high level of SCI, the teacher does 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, Verloop, & de Vos, 1998). In part, the teacher's "learner-centeredness" is what is being measured with the SCI dimension.

Measuring novice science and mathematics teacher's strategic knowledge is not so straight forward. Broadly, there are at least two ways to approach developing an instrument to measure strategic knowledge: (a) using instruments or protocols which yield direct measures of teaching practice based on observing teachers in the classroom, and (b) using instruments which 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. In this study, I focus on a more economical, efficient, and potentially less subjective approach to assessing strategic knowledge, through the scoring of responses to a scenario-based survey 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 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. 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, or 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 two groups 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, and those scores are used as the basis for comparing the strategic knowledge of novice science and mathematics teachers.

Proposed Score Interpretations and Use for the FASCI Instrument

As mentioned above, 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. In this section, I will describe these foundations and present the overall framework for this validity study.

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 for measuring levels of strategic knowledge (SK) among prospective teachers participating in the University of Colorado at Boulder (CU Boulder) Learning Assistant Program (LA Program; Otero, Finkelstein, McCray, & Pollock, 2006), who come from a variety of disciplines. 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 which might be used to evaluate each proposition, I use those categories set forth in the *Standards for Educational and Psychological Testing* (AERA, et al., 1999). These categories will be discussed in detail in the next chapter, and 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 which I collected to support them are shown in Figure 1.

Score Interpretation and Instrument Use: the Strategic Knowledge (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 teachers' SK	
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 observed in teaching practice	Comparison to observation protocol data [Evidence based on Relations to other Variables]
4. 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]
5. 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]

Figure 1. FASCI score interpretation, instrument use, supporting propositions, and sources of validity evidence

Propositions and Evidence Needed

In order to support the proposed score interpretation (comparing the strategic knowledge of novice science and mathematics teachers), I have identified five propositions that must be evaluated. These propositions guide the collection of evidence used in the validation effort. In this study, the propositions I have identified first focus on making a case that SK is important to measure (proposition one), and that it exists across all science and mathematics disciplines (proposition two). The argument supporting each of these propositions is a conceptual one, and depends on evidence based on test content. This evidence includes linking the FASCI instrument to previous research on strategic knowledge or related domains of knowledge. This evidence will be presented and discussed in chapters two and three. I then focus on the proposition that SK can be observed in teaching practice (proposition three). Evidence needed to evaluate this proposition comes from relations to other variables,

specifically comparing FASCI scores to those from an observation protocol. The methods for collecting this evidence and the evidence itself will be presented and discussed in chapters three and six.

Next is an evaluation of the proposition that SK can be measured reliably with a scenario-based survey (proposition four). 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. Further evidence needed to support a deeper investigation of score reliability comes from an analysis of the variance in observed scores that can be attributed to respondents, items, raters, and the interactions between these. These analyses will entail an application of Generalizability Theory (G Theory; Shavelson & Webb, 1991), which will be discussed in detail in subsequent chapters. The methods for collecting this data will be presented in chapter three, and the evidence and analysis will be presented and discussed in chapter four.

Because the FASCI instrument was designed to measure SK of science and mathematics teachers 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 five). 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. 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 the specific science or mathematics content of the respondent (e.g., physics) is able to access their SK differently than a content-neutral version. Evidence needed to support this proposition comes from test content, responses 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. The structure of these versions and the methods for collecting this evidence will be presented in chapter three, and the responses from each version will be presented and discussed in chapter five.

Research Questions

The above proposed score interpretations, related propositions, and associated validity evidence all contribute to addressing my specific research questions:

1. To what degree is the FASCI instrument valid for comparing and distinguishing the strategic knowledge of novice science and mathematics teachers?
2. What are some potential obstacles to developing valid and reliable measures of science and mathematics teachers' strategic knowledge?

The first of these questions will be evaluated by examining the validity argument (as outlined above) in its entirety. The second of these questions will be taken up in the final chapter. In that discussion, I will discuss the implications of this validation effort for conceptualizing, designing, and using instruments to measure science and mathematics teacher knowledge. In that discussion, I will present and discuss some implications and potential obstacles that should be considered when undertaking this type of work. I will also describe how these obstacles might be anticipated and dealt with in advance of their occurrence.

Chapter Overviews

In the next chapter, I begin by discussing in more detail instrument validity and reliability. This discussion will rely heavily on the *Standards for Educational and Psychological Testing* (AERA, et al., 1999). I will also discuss some of the literature on science and mathematics teacher knowledge in order to provide evidence for the conceptual argument related to the first two propositions: that SK is required of science and mathematics teachers and that it exists across all domains within science and mathematics. In addition, I also review existing instruments which seek to measure (both directly and

indirectly) something related to science and mathematics teachers' strategic knowledge. In doing so, I will show that none measure SK as defined in this study, and that there is a lack of more economical instruments which seek to do this in an indirect fashion. In chapter three, I describe the methods for collecting evidence and data which will be used in evaluating each of the propositions. I also describe the methods used in analyzing this data. In chapter four I will describe in detail the FASCI response scoring process and score reliability. In chapter five, I present findings related to comparing scores between the two versions of the FASCI. In chapter six I will compare FASCI scores to scores resulting from an observation protocol for a sample of respondents. Finally, in chapter seven I discuss the overall validity argument, and evaluate each of the propositions in light of the evidence and findings. I also present and discuss some implications from this study and potential obstacles that could be encountered when undertaking such work, as well as recommendations for designing and developing instruments for measuring science and mathematics teacher knowledge. Finally, I conclude by discussing future directions for this and related research.

Chapter 2: Foundations of a Validity Argument for the FASCI Instrument

Introduction

Scores on the Flexible Application of Student-Centered Instruction (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). An example of the former would be determining if one respondent was able to be distinguished from another, while an example of the latter would be to see if a respondent's SK score is representative of a specific location on the SK continuum. The FASCI instrument was designed to evaluate the effects of a teacher education program on science and mathematics teachers' Strategic Knowledge. This proposed score interpretation and instrument use was presented in the previous chapter, and is supported by the set of five propositions also presented previously. Before beginning to examine each of those propositions and the evidence which will help to evaluate them, I will first discuss the concept of validity and the framework for structuring this validation effort. This discussion will rest heavily on the AERA *Standards for Educational and Psychological Testing*, but will also draw on some historical conceptions of instrument validity.

After discussing and presenting this framework, I will describe some existing instruments which are designed to measure science or math teacher knowledge and their associated validity evidence. I will show that there is a lack of instrumentation designed specifically to measure strategic knowledge (SK), and also very few which seek to measure science and mathematics teacher knowledge in an indirect fashion. Further, very few of these efforts explicitly discuss any validity evidence or validation studies.

Because SK is the characteristic of a respondent that the instrument is designed to measure, I refer to it as a *construct* (cf., AERA, et al., 1999; Wilson, 2005). Drawing on the literature on science and math teacher knowledge, I will provide support for the SK construct as being an important one to

identify and measure. This leads into a preliminary evaluation of the first two propositions: (1) SK is one type of knowledge required to be a quality science or mathematics teacher, and (2) SK exists across all domains of science and mathematics teaching. Each of these propositions will be evaluated based on previous research and theoretical evidence presented in this chapter.

Validity

The process of validation involves an evaluation of the argument for claims being made. More specifically, the validation effort rests on the evaluation of certain propositions which underlie the proposed interpretation of scores and uses of an instrument. In accordance with Kane (2006), I use the term "validation" to refer to an evaluation of the plausibility of the proposed interpretations and uses of an instrument, and "validity" to refer to the degree to which the evidence gathered supports or refutes the proposed interpretations and uses. In this section, I will discuss the concept of validity and its implications for my own validation effort.

Historically, validity has been conceptualized within one of three models or frameworks, or some combination thereof. These are the criterion, content, and construct models. I will briefly describe each of these before turning to a more contemporary conception of validity (and the one on which this study is based), that being the unified, argument-based approach.

The criterion model of validity is based on the concept that a test[1] is valid if scores on that test correlate with some other "objective measure" of the factor being measured, such as performance on some task (Angoff, 1988). The criterion model could be applied either concurrently or in a predictive fashion (Kane, 2006). In the former, the criterion score with which test scores are correlated is collected at the same (or at least near) time with the test scores. Predictive applications involve the correlation of test scores with some future performance (e.g., grade in a subsequent course of study). In the past,

[1] In the literature on validity theory and application, the term "test" is most often used. I keep with using that term in the context of this background discussion in order to faithfully preserve the intent of the literature cited. In this study, I use "test" as synonymous with "instrument."

predictive applications of the criterion model were widely used in testing efforts (e.g., in the armed services), while concurrent applications were more often used in making a case for the validity of a new instrument where an existing measure was the basis for the correlation (Angoff, 1988).

The content model of validity asks if test scores "based on a sample of performance in some area of activity [can serve] as an estimate of overall skill level in that activity"(Kane, 2006, p. 19). The observed performance (test score) can be considered an appropriate estimate of overall performance in the domain if "(a) the observed performances can be considered a representative sample from the domain, (b) the performances are evaluated appropriately and fairly, and (c) the sample is large enough to control sampling error" (Guion, 1977 as cited in Kane, 2006). Content validity is concerned with the representativeness of the tasks on the test and the ability to generalize the observed scores on that test to some estimate of ability within the content domain.

Construct validity considers the construct (the characteristic that the test is designed to measure) within a larger theory, which in turn is related to other theories in a hypothetico-deductive way. Networks link these theories to each other and to observations and/or scores which can serve as bases for making inferences about the existence of that construct in an individual. These networks of theories and inferences assume that the theory is fairly well-defined, but that it admittedly only approximates reality (Cronbach & Meehl, 1955). Construct validity has been further broken down into a substantive component, a structural component, and an external component (see Kane 2006 p.20 for a brief summary of this from Loevinger 1957). The construct model was originally proposed by Cronbach and Meehl as an alternative to the criterion and content models.

By the 1970's, researchers began advocating a unified approach to validation efforts. Messick (1989) was one of the first to outline a unified approach. Using the Construct model as a basis for this unified approach, he defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences*

and *actions* based on test scores or other modes of assessment" (Messick, 1989, p. 13, emphasis in original). One issue with this conception is that it does not provide much guidance for the validation effort. Because so much data and evidence could be considered relevant to making a case for the validity of a test, validation could end up being a lengthy, messy process.

Presenting the idea that test validation is an *evaluation*, Cronbach (1988) proposed the idea of a *validity argument*. He defined this argument as an evaluation of the proposed uses and interpretations of test scores. Describing the traditional trinity of validity conceptions (criterion, content, and construct) as "strands within a cable of validity argument," Cronbach emphasized the need to play devil's advocate in the development of a persuasive validity argument. The argument should not only seek to confirm, but also to falsify and contribute to revision—especially for a "young" instrument, such as that presented in this study.

A very approachable summary of this unified conception of validation and a guide for structuring validation efforts is presented in latest edition of the *Standards for Educational and Psychological Testing* (AERA, et al., 1999). In keeping with Cronbach's conception of the validity argument, the *Standards* define validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). Also emphasized is the idea that it is the score interpretations themselves that are evaluated in a validity argument—not the test itself. The implications of this idea are clear: if test scores are used or interpreted for a purpose other than the one being validated, then a new validity argument must be crafted.

As stated above, one potential complication with this concept of validity is that the validation process can become overwhelming. A vast amount of evidence could be brought to bear in supporting test use and score interpretation, and evaluation of that interpretation in light of that evidence could be complex. What is needed is a structure for guiding the validity argument, and for allocating resources during the development of such an argument.

The *Standards* provide such a structure. They begin by calling for an articulation of the proposed score interpretations and test use. The notion of a *construct* is central to this model—the proposed score interpretation is to be articulated in terms of the construct of measurement. Following the proposed use and interpretation is an explication of a set of propositions which support the proposed score interpretations. It is these propositions which provide the structure for the validity argument, as they guide the collection of evidence needed to build the argument. Again in keeping with Cronbach's conceptions, the *Standards* state that the identification of these propositions can be facilitated by playing devil's advocate, and considering alternative or rival hypotheses. For example, consider my second proposition presented in the previous chapter (and discussed further below): SK exists across all domains of science and mathematics (e.g., biology, chemistry, physics, math, etc.). This proposition comes about from the rival idea that strategic knowledge does not exist in some science and mathematics domains. Because the proposed *use* of the FASCI instrument is to evaluate the effects of a teacher education program on the SK of novice science and mathematics teachers' (who come from many disciplines), this proposition is foundational to that use.

Validity Evidence.

Once the specific propositions have been identified, they can be used to guide the collection of evidence needed to support the proposed score interpretations and instrument use. The *Standards* set forth five categories of evidence that can be collected to aid in evaluating each proposition. They are: (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. I will describe each of these in turn before re-visiting the structure of my validity argument as outlined in the previous chapter.

Evidence based on test content refers to the material contained within the test or instrument and the procedures for administering and scoring responses. This includes the items on the instrument

and the format of those items. It also includes an empirical analysis of the way that the content domain is being represented on the instrument and the appropriateness of that domain with respect to the proposed interpretation of test scores.

Evidence based on response processes provides information about how respondents are interpreting the items on a test and how they are formulating their responses to those items. An examination of this evidence can help to identify the degree to which there is a fit between the construct and the responses that individuals are formulating in response to the item prompts. This type of information can be gathered by asking respondents to think-aloud and discuss their responses. A detailed qualitative examination of the FASCI responses also contributes to this pool of evidence.

Evidence based on internal structure includes an examination of the relationship between the items on a test and how the items relate to the construct which the test is designed to measure. For example, item prompt a) in each scenario is designed to elicit information about the SCI dimension of SK. By examining the consistency with which this item prompt is measuring the SCI dimension (characterized by a reliability coefficient), I will be gathering information about the internal structure of the FASCI instrument.

Evidence based on relations to other variables provides an important source of validity evidence. This includes comparing scores from the instrument to scores from other instruments designed to measure the same or similar constructs (i.e., convergent evidence). It also includes comparison of scores to categorical variables of interest, such as group membership. This evidence contributes to an understanding of the degree to which these relationships are consistent with the construct and proposed score interpretation.

The last type of validity evidence presented in the *Standards* is *evidence based on the consequences of testing*. In discussing this type of evidence, the *Standards* state that it is important to distinguish consequential evidence which bears on instrument validity from that which bears on policy

decisions. For example, if differences in the SK between two groups of respondents are found from using the FASCI instrument and those differences do not result from validity issues such as construct underrepresentation or construct irrelevance, then this finding falls within the realm of a policy decision. If, however, the observed difference is due to a validity issue such as differential functioning for one group relative to the other, then the finding *is* a potential validity issue and should bear upon the validity argument.

Each of these sources of validity evidence will be collected and analyzed in the evaluation of the propositions in the validity argument. In the next chapter, I will discuss how this evidence will be collected and analyzed. The specific propositions are presented again below, and will be evaluated in chapter seven.

Structure of the FASCI Validity Argument.

As introduced in the previous chapter, the validity argument for the FASCI instrument begins with a statement of the proposed score interpretation and test use. A set of supporting propositions are then outlined, along with evidence that needs to be collected in order to evaluate those propositions. This structure is represented in Figure 1. This framework is presented again at this point in order to provide a basis for discussing the validation efforts of other instruments which seek to measure science and mathematics teacher knowledge.

Score Interpretation and Instrument Use: the Strategic Knowledge (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 teachers' Strategic Knowledge (SK)	
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 and mathematics teaching (e.g., biology, chemistry, physics, math, etc.)	Conceptual argument [Evidence Based on Test Content]
3. SK can be observed in teaching practice	Comparison to observation protocol data [Evidence based on Relations to other Variables]
4. 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]
5. 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]

Figure 1. FASCI score interpretation, instrument use, supporting propositions, and sources of validity evidence

In the next section, I will review some previous efforts to measure aspects of science and mathematics teacher knowledge that are similar to the conception of strategic knowledge underlying the FASCI instrument. In this review, any evidence gathered in making a case for the validity of each instrument will be presented and discussed. Following that, I will present and discuss the theoretical rationale for the Strategic Knowledge construct in detail before beginning to evaluate the first two propositions.

Some Other Instruments and their Validity Evidence

Many researchers have attempted to measure some aspect of science or mathematics teacher knowledge. The instruments developed are one of two types, or a combination thereof: direct observations of what teachers do in a classroom setting, and/or indirect observation of what teachers

say they would do in a classroom setting. These classifications are very similar to what Kind (2009) refers to as either direct observations (what she calls *in situ*) or indirect "prompt or probe" methods. She points out that the prompt and probe method has the advantage that it can be applicable across a wider range of contexts, but that the *in situ* method provides a "richer picture" of teacher knowledge. An example of this richer picture comes from the work of Loughran et al. (2001). Though Kind is specifically concerned with measures of science teachers' pedagogical content knowledge (PCK; Shulman, 1986), this distinction can be made for measures of other teacher knowledge constructs as well. For example, in my study I attempt to measure science and mathematics teachers' SK with the FASCI instrument (a "promote and probe" approach) and the Reformed Teaching Observation Protocol (an *"in situ"* approach). And while I would generally agree that a richer picture of this type of knowledge results from direct observations, I do think that a similar "rich picture" can be obtained through responses to complex scenario-based items (similar to performance tasks) such as those on the FASCI.

In identifying previous attempts to measure science and mathematics teacher knowledge, I first turned to the work of Taylor and Gess-Newsome (2007), who presented some examples of "Tools and Methods for Measuring PCK." For the past two decades, the PCK construct has received much attention in studies of science and mathematics teacher knowledge. Because of the prevalence of this construct in the science and mathematics teacher knowledge literature, and because of the relationship between the strategic knowledge construct and PCK (discussed below), I use the review of Taylor and Gess-Newsome as a starting point. The instruments included in their review (as well as two others that were not included in their review: the PRAXIS III (Dwyer, 1998) and the Mathematical Knowledge for Teaching measures (MKT; Rowan, Schilling, Ball, & Miller, 2001)) are summarized in Table 1 and discussed in more detail below. For each instrument, I discuss the types of validity evidence collected in the research and presented in publications or technical reports. With two exceptions (PRAXIS III and MKT) none of these efforts frames the evidence collected within a validity argument, or uses the term

"types of validity evidence" as in the *Standards for Educational and Psychological Testing* (AERA, et al., 1999). However, it should be noted that most of the publications and reports about these instruments were not structured as papers on instrument validity. Most were more general in nature, or focused on an application of the instrument and included only a brief section on validity.

Table 1.

Summary of some previous attempts to measure science and mathematics teacher knowledge constructs

Research Program	Type(s) of measures used	Common uses/examples	Types of Validity Evidence presented
Reformed Teaching Observation Protocol (RTOP; Piburn et al., 2000)	Observation protocol	As one component of program evaluation (e.g., Sawada, 2003)	Evidence based on test content, Evidence based on relations to other variables, Evidence based on internal structure
Local Systemic Change through Classroom Enhancement (LSC; Horizon Research, 1999a, 1999b)	Observation protocol, questionnaire	To evaluate LSC professional development programs	Evidence based on test content, Evidence based on internal structure
Expert Science Teaching Educational Evaluation Model (ESTEEM; Burry-Stock & Oxford, 1993)	Observation protocols, student outcome measures, teacher self-reports	Selected components used, such as the classroom observation rubric (e.g., Shin, Yager, Oh, & Lee, 2005)	Evidence based on test content, Evidence based on relations to other variables, Evidence based on internal structure
Science Case for Teacher Learning Project (Heller, Daehler, Shinohara, & Kaskowitz, 2004)	Interview rubric	Small-sample investigation of professional development program efficacy	Evidence based on test content, Evidence based on relations to other variables
Content Representation and Professional and Pedagogical Experience Repertoire (CoRe and PaP-eRs; Loughran, et al., 2001)	Interview data, observations, discussions	Development of a rich picture of teachers' content knowledge and teaching practice	Evidence based on test content (?)
Lee and Luft (2005)	Interview protocol and analysis rubric	Small sample investigation of practicing teachers' PCK (e.g., Lee, Brown, Luft, & Roehrig, 2007)	Evidence based on relations to other variables
PRAXIS III (Dwyer, 1998)	Observation protocol, interview protocol	Used in conjunction with PRAXIS II measures of subject-matter knowledge	Evidence based on test content, Evidence based on relations to other variables, Evidence based on internal structure
Mathematical Knowledge for Teaching (MKT; Rowan, et al., 2001)	Constrained (multiple choice) survey items	Large sample studies of teachers' knowledge of content and students	Evidence based on test content, Evidence based on relations to other variables, Evidence based on internal structure

Instruments without a Structured Validity Argument.

The instruments that I will review and discuss in this first group do not have structured validity arguments associated with them. Validity evidence is often provided in publications or in accessible technical reports, but sometimes in a piecemeal fashion which makes it difficult to evaluate the instrument validity with respect to any proposed score interpretation or use. I will describe each of these instruments and their validity evidence.

The Reformed Teaching Observation Protocol.

The Reformed teaching Observation Protocol (RTOP) is an observation protocol that is designed to measure "reformed" teaching. It is intended to be used by trained observers in order to rate the degree to which a teachers' practice is consistent with reform pedagogy. The instrument contains three categories (in addition to the background and context categories) to guide the observation of teaching and upon which the ratings are based. These categories are: (a) Lesson Design and Implementation; (b) Content (further subdivided into Propositional Knowledge and Procedural Knowledge); and (c) Classroom Culture (further subdivided into Communicative Interactions and Student/Teacher Relationships). The basis for defining "reformed teaching" and for these categories comes from the literature on constructivism, reform in science education (e.g., American Association for the Advancement of Science, 1990; National Research Council, 1996), and the National Council of teachers of Mathematics standards (2000). The discussion of these theoretical bases can be considered evidence based on test content. Although they do not cite Shulman or PCK as a basis for their instrument development, the researchers do refer to PCK when discussing their factor analysis of the RTOP items.

Evidence for the reliability of RTOP scores is provided in the form of correlations between observers' ratings. These correlations were generally very high, mostly above 0.90. Evidence for the validity of the RTOP is presented in three forms: 1) correlations between RTOP total score and subscale

scores, 2) correlation between student learning gains (as measured by a concept inventory, such as the Force Concept Inventory (Hestenes, Wells, & Swackhammer, 1992)) and their instructor's RTOP scores, and 3) factor analyses based on the observations from 153 classrooms. For the first two analyses (evidence based on relations to other variables), correlations were found to be high. The correlation between six instructors' RTOP scores and their students' FCI normalized gain scores for six physical science classrooms was 0.88. Similar comparisons were made between instructors' average RTOP scores and their students' normalized gain in physics (four classrooms, correlation = 0.97) and mathematics (six classrooms, correlation = 0.94 ("conceptual understanding") and 0.92 ("number sense")). The factor analyses (evidence based on internal structure) yielded interesting results. An initial principle component analysis indicated three unique factors. However, an examination of the item loadings on these three factors reveals that they are not consistent with the three broad design categories of the instrument (Lesson Design and Implementation, Content, and Classroom Culture). Instead, the evaluators 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 3 items load at 0.50 or greater, 2 of which also cross-load on factor 1). Further, when using a more common cut-off value for significance in factor loadings (0.30 rather than 0.50), the authors identified *five* factors rather than three, and two items which they did not group with any of these factors. They claimed that the four items which loaded on both factors 1 and 2 (out of the three identified) "define the meaning of *content pedagogical knowledge* operationally within the RTOP" (Piburn, et al., 2000, p. 22, emphasis in original). These four items come from all three observational categories, and focus on students' prior knowledge, students' ideas, "intellectual rigor," and encouraging students to entertain alternative solutions.

Though the RTOP researchers do provide some validity evidence for the instrument, it is not organized into an argument which could be interpreted with respect to specific uses and score

interpretations. Also, some of the evidence presented is difficult to interpret by itself, such as the correlations between instructor RTOP scores and their students' learning gains. The RTOP structure and analysis will be discussed in much more detail in chapter six.

The LSC Observation Protocol and Questionnaire.

The Local Systemic Change (LSC) Observation Protocol is intended to broadly characterize, through both constrained and open-ended items scored by an observer, the characteristics of an individual classroom "lesson" taught by a teacher of math or science content at the K-12 level. Specifically, it is designed to measure the "quality of an observed K-12 science or mathematics classroom lesson by examining the design, implementation, mathematics/science content, and culture of that lesson"(Horizon Research, 1999a). The associated LSC Science Teacher Questionnaire is designed to measure the "opinions, preparation, teaching practice, and the quality and impacts of professional development experiences" of 6-12 science teachers (Horizon Research, 1999b). It contains multiple subscales intended to measure the following constructs: attitudes toward reform-oriented teaching, pedagogical preparedness, content preparedness, use of traditional teaching practices, use of investigative teaching practices, use of practices that foster an investigative classroom culture, and perception of principal support. Both of these instruments (the observation protocol and the questionnaire) are intended to be used in concert with one another and were designed to contribute to the evaluation of the LSC Teacher Enhancement Program (a professional development program for practicing teachers). The LSC instruments result in a combination of direct (the observation protocol) and indirect (the questionnaire) measures.

Validity evidence for the LSC Observation protocol comes from evidence based on test content and evidence based on internal structure (Horizon Research, 2000). With respect to the former, the developers discuss the standards documents that the instrument was based on and the expert review of instrument drafts. They also discuss the rating scales and the reliability of ratings, which they state is of

"fundamental concern." Finally, the developers present measures of internal consistency (in the form of Cronbach's alpha) as evidence based on the internal structure of the instrument. Again, as with the RTOP, this evidence is not organized in a way that can be used to support any proposed interpretation of LSC Observation protocol scores or instrument uses. Also, in order to develop such an argument other evidence would need to be brought to bear. Missing is evidence based on relations to other variables, on response processes, and on consequences of testing.

ESTEEM.

Burry-Stock and Oxford set out to define characteristics of expert science teachers and develop a set of instruments to assess teachers with respect to these characteristics (ESTEEM; Burry-Stock & Oxford, 1993). The characteristics of expert science teachers that they define are based in part on the *Science Teaching Standards* (National Research Council, 1996) and the National Board of Professional Teaching Standards (Bond, et al., 2000). This discussion represents limited validity evidence based on test content. Some discussion of instrument administration and scoring could also be used as evidence based on test content, but does not seem to be brought to bear in any validity argument. The instruments themselves represent a sort of portfolio characterizing teaching expertise, and are to be considered together. They include: (a) a classroom observation rubric, (b) a student outcome assessment rubric, (c) a teacher self-report of frequency of practices, and (d) a teacher self-report of science grading practices. Two other instruments were planned (as of 1993, but no subsequent publication about them has been found): instructional design and reflective teaching practices. The authors recommend that the instruments be administered "over a several year period" in order to capture different stages of a teacher's career development from novice to expert. Subsequent uses of the ESTEEM to evaluate professional development programs have used select parts of the set of instruments, such as the classroom observation rubric (e.g., Shin, et al., 2005).

For the classroom observation rubric and the teacher practice inventory, evidence based on

internal structure is presented in the form of interrater reliability, score reliability coefficients and results from a principal components analysis. Some evidence based on relations to other variables is presented, in the form of correlational studies for expert teachers in their sample. As with the instruments discussed above, although the authors of ESTEEM present a variety of evidence which could be used in a validity argument, it is not presented within a single, coherent framework for judging the validity of the instruments.

Science Cases for Teaching Learning Project.

Central to the work of Heller, Daehler, Shinohara, and Kaskowitz (2004) is science teachers' subject matter knowledge, their knowledge of students' thinking, and their knowledge of pedagogical strategies. In their study, they set out to document how elementary teachers' content knowledge and PCK changed as they participated in a professional development program about electricity and magnetism. They developed a rubric for analyzing interview data from the teachers (n=18), and specifically focused on teachers' (a) perceptions of student difficulties understanding electric circuits; (b) instructional strategies for addressing those difficulties; (c) approaches to helping students understand what would happen if one of the bulbs were unscrewed in a parallel circuit; (d) interpretations of sample student responses to that problem; and (e) instructional strategies they would use to help those particular students (p. 1). Interview data was coded and scored, and conclusions were drawn about the professional development program efficacy.

Heller et al. present evidence based on test content in discussing the development of their rubric, but do not use this evidence in service of an explicit case for the validity of the rubric. They specifically discuss evidence for the validity of this rubric based on correlating coded interview scores with scores from a content test (evidence based on relations to other variables). They cite the observed strong, positive correlation as evidence of having "validated the content assessments." (p. 14). However they do not provide other evidence in support of the rubric's validity.

CoRe and PaP-eR.

The Content Representations (CoRe) and Pedagogical and Professional experience Repertoires (PaP-eR) work of Loughran, Mulhall, and Berry (2004) focuses on both teachers' content knowledge and their teaching practice. Both the CoRe and PaP-eR are made up of interview data, observations, and discussions between researchers and teachers. They focus on particular content, and multiple PaP-eRs are intended to be used in characterizing a teachers' PCK. This work certainly provides a rich picture of a teacher's content knowledge and PCK, but the data collection and analysis are intensely time-consuming and potentially subjective since rating each artifact is not part of the framework. Any rating (ordinal or otherwise) would be left up to the individual interpreting each artifact.

The authors write of PaP-eRs being "validated" by teachers and researchers, but do not discuss the types of evidence used in making validity judgments. Also discussed is the idea that the interpretation of CoRe and PaP-eRs is left to the reader, in that they can decide what aspects of this work are relevant to their own context. The authors cite this as being "enmeshed in an understanding of validity" (p. 382). Overall, it is unclear what types of evidence are provided in support of instrument validity, though it could be inferred that evidence based on test content is presented (in the form of theoretical backing). Again, lacking a coherent argument for the validity of the instrument, one cannot evaluate it with respect to proposed score interpretations or instrument use.

Lee and Luft Interview Protocol.

Another lesson-based interview protocol and associated rubric intended to measure science teachers' PCK was developed by Lee and Luft (2005). Using a case study approach with a small number of teachers (n = 4) the researchers characterized how experienced science teachers "revealed PCK throughout their teaching practices." In her related dissertation work, Lee (2005) describes this work as "meeting the standards for validity for naturalistic inquiry" (p. 57). She discusses triangulation of data sources and having conducted member checks on her interpretations (evidence based on relations to

other variables). No other information on the validity of the protocol or rubric is presented.

Instruments with a Structured Validity Argument.

Two instruments presented in Table 1 and discussed below are exceptions to those discussed above. The PRAXIS III and MKT instruments *do* present structured validity arguments which can be used to evaluate specific score interpretations and instrument uses. The structure of each of these arguments differs slightly from that which I propose in this study, but each stands as an exemplar. Below I discuss each of these instruments, their validity arguments, and how the work contributes to my study.

PRAXIS III.

Another instrument intended to measure one or more dimensions of pedagogical content knowledge is the Educational Testing Service's Praxis III (Dwyer, 1998). PRAXIS III is an observation and associated interview protocol meant to be used in conjunction with the PRAXIS II measures of subject matter knowledge (one for each discipline, e.g. biology, chemistry, physics, etc.). By itself, it is not intended to be a stand-alone measure of some aspect of pedagogical content knowledge, and it does not focus on any particular discipline (i.e., it is domain-general). PRAXIS III is intended to be used after the beginning teacher has had a fair amount of classroom experience, and emphasizes the application of subject matter and pedagogical knowledge in the context of the classroom. It consists of a variety of components: a class profile, instruction profile, pre-observation interview, classroom observation protocol, and post-observation interview. PRAXIS III administration is very time-consuming and observers and raters must first participate in lengthy training.

Dwyer (1998) begins her article on PRAXIS III by discussing a framework for validity, and cites Messick (1989) and the previous edition (1985) of the *Standards for Educational and Psychological Testing*. She states that although working within such a framework can be very challenging when

working with teacher assessments, it can contribute to producing technically sounds assessments. It should be noted that the form of this paper is very different from the other documents reviewed above. Dwyer's paper is on the technical and psychometric aspects of PRAXIS III, and accordingly places attention on issues of validity. She beings with an explicit statement of the proposed use of the instrument (teacher licensing), and then takes the reader through a detailed presentation and discussion of the test content, criteria, administration and scoring. She pays particular attention to discussion the observer's ratings, stating that the classroom context is central to interpreting teacher actions, and that the observer's judgments are the "cornerstone of defensibility of ratings" (p. 181). Although not working within the same framework that I propose to use (from the latest edition of the *Standards for Educational and Psychological Testing*), Dwyer's work is an example of a structured presentation of evidence for the validity of PRAXIS III, given the specific proposed use. In addition to this work providing a good example of a structured validity argument, it also points to at least one potential obstacle in undertaking this type of effort, that being rater judgments. These judgments are central to two of my propositions, those being the fourth (SK can be measured reliably with a scenario-based survey) and the fifth (SK can be measured dependably with a scenario-based survey).

MKT.

The Mathematical and Knowledge for Teaching (MKT) instrument was developed through a research project at the University of Michigan (Rowan, et al., 2001; Schilling & Hill, 2007). The MKT survey instrument employs constrained scenario-based items which (in its initial version) are intended to measure two constructs: mathematics teachers' content knowledge and their knowledge of content and students. These items present a situation from the teacher's point of view and then pose a question and give three or four answer choices. For example, one item shows three examples of how students have organized their work when multiplying large numbers, and then asks which method could be used to multiply any two whole numbers. Three answer choices are given with varying levels of

qualification for each of the students' methods (Hill, Schilling, & Ball, 2004). The items are particularly focused on teachers' subject matter knowledge, and do not aim to measure pedagogical knowledge separately from that content.

Schilling and Hill (2007) write explicitly about developing a validity argument for the MKT measures. These authors use Kane's approach (2001, 2004) to structuring a validity argument, which differs slightly from that which I have outlined based on the *Standards for Educational and Psychological Testing*. Kane's approach involves two types of argument: the interpretive argument and the validity argument. In the former, one outlines the assumptions and inferences that are central to the proposed score interpretations. In the latter, one evaluates these assumptions and inferences in light of the validity evidence that has been collected. Although very similar to the approach that I have outlined, Kane's approach places more emphasis on forming an interpretive argument before evaluating the proposed interpretations. Schilling and Hill further propose specific sub-types of assumptions in an effort to provide a more prescriptive framework than the one which Kane outlines. In a set of related papers in this special issue of the journal *Measurement: Interdisciplinary Research and Perspectives* (v 5, n 2-3), evidence for evaluating the validity of the MKT instrument is presented.

The authors note and discuss the discordance between validity theory and practice, noting that there are very few examples of validity arguments being developed. This has been echoed by others as well (e.g., Kane 2006). Their example of this type of approach supports the idea that validity is an ongoing process, and that developing a validity argument is a difficult undertaking.

Contributions of these efforts.

Each of these instruments and their associated validity evidence and arguments contributes in some way to my present work with the FASCI. Those instruments in the first group discussed point to the need for developing and presenting a coherent validity argument in which the proposed score interpretations and test uses are articulated. Also, they provide examples of the need for more and

varied evidence which can be brought to bear in examining instrument validity. Often, the existence of such evidence has to be inferred from these development efforts. Instead, it should be explicitly stated so that potential users of instruments and scores can make informed judgments.

The instruments in the second group (PRAXIS III and MKT) provide not only examples of how such a validity argument can be structured, but also indicate some of the potential obstacles to such efforts. For example, in each case one particular type of evidence (rater judgment for PRAXIS III and relation to other variables for MKT) seemed to be the lynchpin of the argument. Accordingly, the discussion of each of these pieces of evidence was comprehensive and came up again in the implications and discussion.

In the next section I describe the Strategic Knowledge (SK) construct. The discussion of this construct is meant to provide theoretical backing in support of the idea that SK is an important aspect of science and mathematics teacher knowledge. After this discussion, I will examine the first two propositions which support the proposed score interpretation.

The Strategic Knowledge Construct

The strategic knowledge construct is composed of the two dimensions of Flexible Application (FA) and Student Centered Instruction (SCI) which I described briefly in the first chapter. The dimensions are closely related to two characteristics of the construct of pedagogical content knowledge (PCK; Shulman, 1986): (a) teachers' representations of subject matter and strategies for teaching the subject matter (related to the FA dimension), and (b) teachers' knowledge of students' understanding of the subject matter (related to the SCI dimension). The latent variable underlying each of the dimensions of strategic knowledge (FA or SCI) can be described using what is known as a construct map (Wilson, 2005). A construct map describes the qualitatively distinct levels that are hypothesized to exist on the continuum of each latent variable. The construct maps for each of the FASCI dimensions are shown in Figures 2 (the FA construct map) and 3 (the SCI construct map).

Level	Respondent Characteristics
2	• The teacher has repertoire of strategies that can be used to facilitate student learning within a given class session. • If the teaching strategy comprised of these acts is not producing the desired result, sometimes it can be modified. • The teacher recognizes that the choice of a class activity and associated teaching strategy will depend upon variables specific to the classroom context.
1	• The teacher has a repertoire of strategies that can be used to facilitate student learning within a given class session. • If an activity based on a particular teaching strategy is not producing the desired result, the activity can be modified by selecting a different strategy.
0	• The teacher has a limited repertoire of strategies. • Once a particular activity has been selected for a class session, it is not easily modified with a different strategy.

Figure 2. Construct map for the Flexible Application (FA) dimension

The FA Dimension.

The FA dimension describes the strategic repertoire that a teacher possesses and how (at the highest level) she makes strategic decisions based on relevant contextual factors. A teacher who is at an expert level on the FA dimension (level "2") is an *adaptive expert* (as defined by Hatano & Inagaki, 1986) in that she has a large "adaptational repertoire" of teaching strategies like the experts in the study by Clermont, Borko, and Krajcik (1994). These strategies are based in the teachers' knowledge of representations of the subject matter, a characteristic of PCK and an essential component of the pedagogical reasoning process (Shulman, 1987). A teacher at this highest level is able to consider relevant constraints within their area, which is consistent with Chi's view of "expertise" (2006). At a slightly less sophisticated level on the FA dimension (level "1"), a teacher has a repertoire of instructional strategies but chooses a strategy without consideration of relevant contextual factors (such as student understanding). The teacher can adapt a teaching strategy as needed, but does not justify

her adaptation. At the lowest level on the FA dimension (level "0"), a teacher has a very limited repertoire of strategies to choose from and once she has decided on a strategic approach for a given context, she does not adapt or change it. Though this dimension is conceptualized as being continuous in nature, scorers could only come to acceptable levels of scoring agreement using a three level construct. Scorer training and results from scoring will be discussed in chapters three and four.

Level	Respondent Characteristics
2	• Discussion of interactive teaching which would be *observable* to the teacher or to an outside "other." • Discussion of a *rationale* for why they see this as an opportunity for interactive teaching and learning Teacher ⟷ Students and/or Students ⟷ Students
1	• Discussion of interactive teaching which would be *observable* to the teacher or to an outside "other." Teacher ⟷ Students and/or Students ⟷ Students
0	• No discussion of interactive teaching • Teacher primarily views classroom activities as ways to help students make sense of new ideas. Information goes from teacher to student. Teacher → Students

Figure 3. Construct map for the Student-Centered Instruction (SCI) dimension

The SCI Dimension.

The SCI dimension describes how a teacher views an activity as an opportunity for active engagement between herself and her students or among her students so that students' ideas are elicited and articulated. Eliciting and building on student prior knowledge is central to teaching for understanding (e.g., Bransford, Brown, & Cocking, 1999; Fosnot, 1996; Greeno, et al., 1996). At the expert level (level "2"), a teacher conceives of the situation as an opportunity for interaction between herself and the students, or between the students and each other. She also articulates a rationale for

why she conceives of the situation in this way. The articulation of this rationale demonstrates that the respondent is not merely providing a socially desirable response, but that she can also explain *why* the situation facilitates student learning. A teacher at the middle level (level "1") views the learning activity as a situation in which she and/or the students are interacting with each other. This level of thinking was identified in a study by Peterson and Treagust (1995). In this study, as teachers engaged in Shulman's pedagogical reasoning process, they became more student-centered. At a novice level on the SCI dimension (level "0") the teacher views the learning activity as a non-interactive place where she presents the material to her students without any adaptation or tailoring of the material and representations to her students' needs. Again, although the SCI dimension is conceptualized as being continuous in nature, in scoring moderation sessions we could only reach acceptable levels of scoring agreement using a three level construct. This will be discussed in detail in chapter three and four when presenting the results from scorer training.

Strategic Knowledge and Formative Assessment.

By eliciting and responding to student ideas, teachers can meet the diverse needs of their students (Hammer, 1996; McDermott, 1991; Minstrell, 1991; van Zee & Minstrell, 1997). When a teacher *responds* to the ideas of her students, and when this response is used to modify instruction, she is engaging in the process of formative assessment. This elicitation and response exists at the intersection of FA and SCI.

Formative assessment consists of stating the objectives, assessing, and giving feedback (Atkin, Black, & Coffey, 2001; Sadler, 1989). More specifically, assessment is formative when the information derived from the assessment informs instructional practices in order to meet needs of students (Black & Wiliam, 1998). The formative assessment process involves the teacher *responding* to students' conceptions by setting objectives, making instructional decisions, and providing feedback and relevant instruction. In seeking to measure teachers' strategic knowledge, I am partly interested in the extent to

which teachers' instruction involves formative assessment, or much more broadly, the extent to which teachers' instruction is student-centered. In a previous study, Otero and Nathan (2008) found that 61 pre-service elementary teachers commonly held one of four views about the role of a student's prior knowledge in instruction and in the formative assessment process. The authors did not make explicit claims about a hierarchical ordering of these four views on the basis of their sophistication. However, they did argue that a *flexible formative assessment* view would be the target of teacher education program. Further, they argued that this practice should be sensitive to contextual features within the classroom. Information about pre-service teacher knowledge from this study was initially used in the development of the SCI dimension of the FASCI instrument.

Expertise in Strategic Knowledge.

The strategic knowledge expert is one who has a large repertoire of teaching strategies to draw from. Not only do they have this repertoire, but they can apply the strategies conditionally based on the situation at hand. As described above, these are general characteristics of experts as identified by Hatano and Inagaki (1986), Chi (2006), and outlined in *How People Learn* (Bransford, et al., 1999). Further, the strategic knowledge expert conceives of a classroom situation as an opportunity for interaction between teacher and student or students and other students so that student ideas are elicited. They will also be able to articulate *why* they think the situation affords this interaction. This expertise is hypothesized based on the research literature and on empirical data. Responses from pilot testing of the FASCI instrument demonstrate the existence of this expertise. These exemplary or expert responses will be presented and examined in more detail when discussing response scoring in the next chapter.

In contrast to the expert, a more novice teacher (with respect to strategic knowledge) will possess a limited repertoire from which to draw, and will apply any strategies in a more haphazard

fashion. Neither will they be able to conceive of many classroom situations as opportunities for interaction. In short, they lack a student-centered, conditionalized strategic approach.

Strategic Knowledge as a Requisite for the Science or Mathematics Teacher

The first proposition that I presented in the previous chapter stated that strategic knowledge (SK) is one type of knowledge required to be a quality science or mathematics teacher. As such, it is an important construct to measure. Evidence supporting the importance of SK comes from linking the SK construct to previous research on science and mathematics teacher knowledge. Therefore evaluation of this proposition is based on the above description of the SK construct and on further support from the research literature.

As described above, the SK construct is based on the literature on expertise (e.g., Chi, 2006; Hatano & Inagaki, 1986) elicitation of student ideas (e.g., Bransford, et al., 1999; Fosnot, 1996; Greeno, et al., 1996), and formative assessment (e.g., Sadler, 1989). In particular, formative assessment is closely related to strategic knowledge. A teacher who sees the value of students' ideas and is able to elicit those ideas, and is then able to *conditionally* select a strategy from a repertoire of strategies is engaging in formative assessment, and is demonstrating expertise in strategic knowledge. I do not mean to imply that formative assessment and SK are one in the same, only that the two are related. Because formative assessment has been shown to be an effective instructional practice (cf., Black & Wiliam, 1998), it follows that SK is an important knowledge construct for teachers to possess so that they are able to enact these formative assessment practices.

Strategic Knowledge Exists across all Science and Mathematics Domains

The second proposition that I presented in the previous chapter stated that SK exists across all domains of science and mathematics teaching. In other words, to support the proposed score interpretation and instrument use, SK cannot be unique to physics teaching as compared to biology or math teaching. Support for this proposition follows in part from support for the first proposition. It

would be difficult to argue that being student-centered, possessing a repertoire of strategies, and being able to enact formative assessment practices are not important aspects of teaching within any science or mathematics domain.

A slightly different way to seek support for this proposition is to pose the following question: What makes the SK construct *unique* to science and mathematics teaching? The distinction for science and mathematics teachers lies in the types of strategies that are brought to bear and what constitutes a new strategic approach. For example, having students do a hands-on activity with some kind of manipulatives (as in a math classroom) or with lab equipment (as in a science classroom) are teaching approaches not often used in other disciplines. It follows then that what constitutes a change in strategic approach is also different. Moving from a grouping activity to a lab activity is something that often happens in a science and mathematics classroom and constitutes a change in strategic approach (which is characteristics of the middle and highest levels of the FA construct map). A framework for distinguishing these strategies in a science classroom (which can be applied to the math classroom as well) is provided by Treagust (2007). In summary then, it is the *strategic* part of SK that is unique to science and mathematics teachers.

This proposition is also related to proposition five: SK score interpretations change when specific science content is added to the survey items. Although the SK construct should exist across all science and mathematics domains in order to support the proposed score interpretation and instrument use, I originally hypothesized that the *measure* of SK would change when specific science or mathematics content is added to the items. For each teacher, one can assume that there is a more sophisticated way of thinking about the teaching *of their own discipline* as compared to teaching in general. This idea is based on Shulman's construct of PCK (1986). Therefore, an instrument designed to measure teacher knowledge will not be able to access this superior, more sophisticated way of thinking unless it is based within the respondent's discipline of specialty. Based on this assumption then, I would expect scores

from a content-specific (e.g., physics) version of the FASCI to be higher for those who are experts in that discipline as compared to what their scores would have been on the neutral version, because the content-specific version is accessing this more sophisticated knowledge base. If this is observed to be the case, then it could provide empirical evidence to help resolve the existence of PCK. This could also mean that the neutral version of the FASCI is not accessing the more sophisticated knowledge base, and is therefore a less valid measure of SK. This proposition and analysis will be discussed in detail in chapter five.

Discussion

Conceptions about instrument validity have changed quite a bit over the past half century. Correlation of a measure with some other measure or with performance relative to a criterion used to be the prevailing notion of what it meant for an instrument to be "valid." Content validity, or the representativeness of the tasks on the instrument to the actual tasks of the domain, also became a piece of the validity puzzle. If an instrument exhibited content validity, then a respondent's performance on that instrument could be used as an estimate of overall performance in that domain. Construct validity was originally proposed as an alternative to criterion and content validity. Cronbach and Meehl (1955) proposed use of the construct model when no "adequate" criterion existed for the construct of interest. The construct model considers the construct being measured within a larger theory or network of theories and observed attributes. If predictions based on those theories agree with observations, then that constitutes evidence for validity under the construct model. This conception of construct validity underlies a more contemporary conception of validity, that being the *validity argument*.

The validity argument structure that I use in this study is concordant with that presented in the latest edition of the *Standards for Educational and Psychological Testing* (AERA, et al., 1999). The *Standards* call for an articulation of the proposed test use and score interpretations, with the construct of interest being central to these statements. Following the proposed use and interpretation is an

explication of a set of propositions which support the proposed score interpretations. These propositions then provide the structure for the validity argument, and they guide the collection of evidence needed to build the argument. The propositions are then evaluated in terms of the evidence gathered.

In situating this study within other efforts to measure science or mathematics teacher knowledge, I found that few of these efforts have articulated a validity argument for their work. Two exceptions are the PRAXIS III (Dwyer, 1998) and MKT research program (Schilling & Hill, 2007). Especially in the case of the MKT research, a validity argument was structured and outlined and various studies were undertaken to evaluate the proposed inferences that are a part of that argument.

As a first step in developing a validity argument for the FASCI instrument, I have presented the proposed test use and score interpretations, as well as stated the propositions which support the score interpretation. In this chapter, I have also described the strategic knowledge construct which the FASCI is designed to measure, and have described what characteristics a strategic knowledge "expert" might have. Finally, I concluded this chapter with a discussion of the first two propositions in light of the theoretical backing and previous research evidence presented. In the next chapter, I will outline my methods for collecting the other evidence which supports the remaining propositions. I will also discuss the structure of the FASCI instrument in more detail, which will contribute to the evidence needed to evaluate proposition four: strategic knowledge can be measured reliably with a scenario-based instrument.

Chapter 3: Methods for Collecting Validity Evidence and Evaluating Propositions

Introduction

In the previous chapter I discussed the concept of test validity and presented the structure of the validity argument for the FASCI instrument. Central to that argument is the collection of evidence used to evaluate each of the propositions which underlie the proposed score interpretations. I also described each of the five categories of validity evidence presented in the *Standards for Educational and Psychological Testing* (AERA, et al., 1999), and discussed the theoretical bases of the FASCI construct. That description of the FASCI construct constitutes part of the evidence based on test content.

In this chapter I will describe the methods for collecting evidence based on internal structure, evidence based on response processes, and evidence based on relations to other variables. In describing each of these, I will also discuss how this evidence will be analyzed and used to evaluate each of the propositions. Building on the last chapter's presentation of other instruments designed to measure science or mathematics teacher knowledge and their associated validity arguments, I will first describe the internal structure of the FASCI instrument. I will then compare and contrast that with the structure of some of these other instruments. I will also discuss the associated propositions in this chapter.

The next two sections of this chapter serve as foundations for the next two chapters in this dissertation. My discussion of evidence based on response processes will include a detailed discussion of response scoring, scoring training, and analysis of those scores. I will also describe the methods for conducting think-aloud interviews with respondents. This lays the foundation for chapter four, in which I discuss the reliability of SK scores, and chapter five, in which I discuss the comparison of scores between two versions of the FASCI instrument.

My discussion of evidence based on relations to other variables serves as the foundation for chapter six. In this chapter, I will discuss the observations of practicing science and mathematics

teachers' strategic knowledge. Finally, I will discuss how each proposition can be evaluated at this point in the study, and what remains to be investigated in order to facilitate these evaluations.

Evidence based on Internal Structure

Evidence based on the internal structure of the FASCI instrument is needed in order to evaluate the following propositions: (4) strategic knowledge (SK) can be measured reliably with a scenario-based survey, and (5) SK score interpretations change when specific science content is added to the survey items. Specifically, the evidence needed to help evaluate these propositions comes from the structure and content of the FASCI survey items, and the consistency with which they measure science and mathematics teachers' SK. I will first describe the scenario-based structure of the FASCI, and then discuss the science and mathematics content-neutral character of these items. Next I will describe the way in which I collect evidence to evaluate this content-neutral character. Finally, I will discuss the concept of score reliability as a characterization of the consistency with which the survey items are measuring science and mathematics teachers' SK.

Instrument Structure.

As mentioned above, the FASCI instrument employs a scenario-based item design, which individuals respond to in an open-ended fashion. These items are rather broadly contextualized in terms of the content being taught: *"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 I refer to as the "content neutral" character of the FASCI. In other words, no specific science or mathematics content domain is specified in the scenarios. An example FASCI scenario-based item is shown in Figure 1.

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 n-FASCI

As discussed in the last chapter, most of the instruments reviewed which seek to measure some aspect of science and mathematics teacher knowledge do so in a more direct fashion, though observations of practice and follow-up interviews. An exception is the MKT, which does use a more indirect approach, but it employs a multiple choice item design rather than an open-ended response format. The items on the MKT do have a structure that presents hypothetical situations similar to the FASCI scenarios, but the MKT is much more focused on math teacher 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 without precedence, however. In many studies on teachers' 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 math teachers' reform-oriented practices. More specifically, these scenario-based measures were designed to measure "intent to engage in reformed approaches" (Le et al., 2004, p. 3). Separate scenarios were created for math 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). Between five and eight different rating questions are presented to the respondent in each scenario. 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.

However, this 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. This potential problem is what gives rise to proposition five in my validity argument for the FASCI: SK score interpretations change when specific science content is added to the survey items. In order to evaluate that proposition, evidence based on an alternative instrument structure is needed—one which *does* include specific content in the scenarios.

The Content Test.

I created the "physics-FASCI" (or "p-FASCI") based on the existing content-neutral FASCI (or "n-FASCI") and designed it by placing the current FASCI scenario-based items within the context of specific physics topics, namely Newton's Third Law and Free Fall. A comparison of scores between similar groups on the n- and p-FASCI discussed below) will provide further evidence needed to evaluate proposition five. The full n- and p-FASCI surveys can be seen in Appendix A. These particular physics 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). 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.

19) 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:

18) 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 I wanted to manipulate was the specific science or mathematics content in the constraint which frames the items. The actual item prompts within each scenario are the same on both versions of the FASCI. Having a parallel item design is important because it is the content constraint that is being tested in evaluating this proposition; therefore other aspects of the item need to be controlled.

The p-FASCI development was informed by three main activities: (a) discussion and meetings with other researchers in the School of Education and the Physics Department; (b) think-aloud

interviews with a practicing K-12 physics teacher and a pre-service physics teacher, both of whom had responded to the p-FASCI; and (c) a pilot test of an earlier version of the p-FASCI with a group of pre-service teachers at Northeast Queen's University, and subsequent conversations with their professor.

I would 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, as discussed in the last chapter. If score interpretations change for this reason, 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.

FASCI Administration.

Sample.

The target population for collecting this particular validity evidence (the "content test" described above) consists of all 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. I am defining prospective science teachers as those students who are in some kind of undergraduate science teaching course, whether they have committed to a career in teaching or not. My aim was to recruit a sample of prospective teachers from four PTEC sites to participate in this research: Western State University, Southeast Coastal University, Northwest Pacific University, and Northeast Queen's University. 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. Because the requirements for being a PTEC member institution are broadly conceived ("Institutions that are involved in preparing pre-service physics teachers are invited to join"), it is difficult to define the characteristics of

a representative institution. All four chosen sites are similar to each other in that they are active members of PTEC (e.g., a representative from each attended the PTEC conference in 2008), and three of the four sites have some form of a Learning Assistant (LA) program. At these sites, the full populations of prospective teachers in the science methods courses (which I expected to be about 20-25 students per course) were asked to participate. I expected these prospective teachers to be concentrating mainly in physics, but also in different science disciplines such as chemistry, biology, geology, and astronomy. I recruited participants through the course instructors and/or other teacher educators at these institutions. 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.

Administration.

At the beginning of the spring 2009 semester, I administered the neutral and physics-FASCI surveys to all available prospective teachers in science methods courses at the five participating sites (see Table 1). Students in each course section were assigned a randomly generated number and the course sections were divided into two equal halves based on the random numbers. This random number assignment was performed by the course instructors at Western State University, Southeast Coastal University, and Northwest Pacific University. I performed the random assignment for the few students participating from Central Research University, and there was no random assignment for students at Northeast Queen's University (discussed further below). Those students in the first half (lower half of random numbers) of each course section responded to the content-neutral FASCI while those students in the second half (upper half of random numbers) responded to the physics-FASCI. I also administered each version of the FASCI a second time at the end of the spring 2009 semester. However, attrition was high and there was a mix-up in version assignment at Southeast Coastal

University. For those reasons, scores from this administration are not used in analyses comparing scores between versions. Rather, these response sets were used for scorer training and qualitative analyses.

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- students volunteered

In all administrations of the FASCI, demographic and academic background information is collected from all participants. At the end of each version is a space for open-ended comments and feedback regarding participation. In previous pilot administrations of the FASCI, this space has proven valuable for information about respondents' perceived value of participating. In this particular study, respondents did comment in this space: 50% on the n-FASCI and 44% on the p-FASCI.

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 a Microsoft Access database for storage and scoring. The database was queried and results were imported into Microsoft Excel and SPSS for analysis. Response scoring will be discussed in detail below in the section of evidence based on response processes.

Score Reliability.

An examination of how consistently the items (from which the scores are derived) are measuring each latent trait is central to evaluating the internal structure of the FASCI. The reliability of scores on each dimension of the FASCI can be used to characterize this consistency. In examining score reliability, I will discuss reliability from previous pilot testing of the FASCI, as well as that from the "content test" administrations described above.

According to classical test theory, an individual's observed score consists of their true score and some error component. If we were able to repeatedly administer a particular set of items to the individual (assuming that at each subsequent administration they did not have the experience of the previous- a sort of "reset" between administrations), then we would expect a distribution of observed scores around his or her "true score", where a true score is defined as the average of all possible observed scores. The "error" associated with an individual's observed score represents the deviation from the true score, and can be negative or positive, large or small, and has the same variance and standard deviation as the observed score for that individual. The relationship between observed score, true score, and error under classical test theory is shown in equation 1.

$$x_p = \tau_p + e_p \quad (1)$$

When applying this equation to the context of the present study, x_p can be interpreted as the observed score for an individual on one dimension and version of the FASCI, τ_p represents their true score on that dimension, and e_p represents measurement error.

Score reliability is estimated using a reliability coefficient, which is generally defined as the ratio of true score variance to observed score variance. Observed score variance can be further expressed as the sum of true score variance and error score variance. This coefficient ranges from zero to one, and is near zero when observed score variance is mostly due to error score variance and near one when

observed score variance is mostly due to differences in true scores. A high reliability coefficient indicates a measure with relatively high precision (Traub, 1994).

In this research, one way I estimate score reliability is to use Cronbach's alpha. Cronbach's alpha (expressed in equation 2) is often referred to as a measure of internal consistency, and can be thought of as the lower bound on reliability for a set of items (Wainer & Thissen, 2001, p. 33).

$$\propto = \frac{I}{I-1}\left(1 - \frac{\sum_{i=1}^{I} \sigma_{y_i}^2}{\sigma_x^2}\right) \quad (2)$$

In equation 2, σ_x^2 is the variance in observed total scores (which are represented by x_p in equation 1 above), and $\sum_{i=1}^{I} \sigma_{y_i}^2$ is the sum of the variances in observed item scores across all individual items (*i*). The term $\frac{I}{I-1}$ takes into account the total number of survey items on the survey (*I*) and is used to give weight to a survey with more items (i.e., as *I* increases, the multiplicative $\frac{I}{I-1}$ factor also increases, making $\propto$ larger).

Score reliability can be considered a necessary condition for validity, and is therefore generally important to investigate. Depending on the proposed instrument use and score interpretations, differing degrees of precision in measurement or reliability can be required for making a validity argument. More specifically, proposition four asserts that SK can be measured reliably with a survey-based instrument. With respect to proposition five about the stability of score interpretations when content is added to the items, comparing score reliabilities between versions of the FASCI is of interest. It should be noted that while high score reliability is desirable for precision in characterizing an individual's level of strategic knowledge (an absolute or criterion-referenced interpretation) it is not as critical for making norm-referenced comparisons. Having relatively low score reliability does not necessarily preclude a measure from being used for these relative comparisons but it does make it more difficult to rule out chance as the reason for observed differences. Both uses are intended for the FASCI; therefore the relative and absolute reliability of scores is of interest and will be discussed.

A Deeper Investigation of Score Reliability.

Cronbach's alpha is an estimate of *relative* score reliability, but it is one that can often over-estimate reliability. This is because it takes into account only one source of measurement error, represented by the composite error term in equation 1. When used in this research context, it is essentially ignoring any error due to the raters, and is only considering error due to the items (in terms of item-level variance, see equation 2). A more critical approach to evaluating score reliability would be able take into account the potential error due to raters, and would also provide a characterization of score reliability for *absolute* decisions in addition to *relative* ones. Recall that both are intended uses of the FASCI instrument.

Respondent performance on the FASCI can be thought of as a *sample* of their performance drawn from a universe of potential items and scored by a universe of potential raters. Conceiving of FASCI scores in this way allows for a more fine-grained approach to analyzing error variance and score reliability. Specifically, this conception allows me to examine multiple components of error variance (e.g., items, raters, and their interactions) rather than considering one composite "error", as in the classical test theory conceptualization of observed score. When respondent performance is observed to vary substantially from one item to another, or when scored by one rater or another, then measurement error can be thought of as being due to variability in sampling (Shavelson, Baxter, & Gao, 1993). In this conceptualization, the items are a potential source of measurement error due to the difficulty of the items for the respondent. The raters are also a source of potential measurement error due to differences in the severity of the raters.

A framework exists for conceiving of respondent performance as a sample drawn from some universe, and therefore for examining the reliability of FASCI scores in this deeper way: Generalizability Theory (G Theory; Brennan, 2001). Using this framework will allows one to generate estimates of the error variance attributable to items and raters, as well as interactions between these. Foundational to

this approach is an examination of how much of the variability in observed scores can be attributed items, raters, and the interactions of these (the *facets of measurement*) with each other and with persons (the *object of measurement*). Further, one can investigate how these sources of "error" variance could be minimized if the number of items and/or raters were to be increased. There are certainly other facets of measurement that could be examined, for example the testing occasion, but items and raters are of interest given the FASCI structure and intended uses. And because the FASCI is taken in a more controlled environment (computer-based) as compared to observations collected during teaching episodes, the occasion facet in a sense being controlled in this study.

Using a G Theory approach, estimates of both relative and absolute reliability for different combinations of items and raters can be generated. In G Theory, the *generalizability coefficient* (sometimes referred to as the "G coefficient" or as ρ^2) is analogous to a reliability coefficient in classical test theory, such as Cronbach's alpha. The G coefficient is a characterization of *relative* error, and as such is of interest when comparative decisions are being made, such as identifying the top performers within a sample. Cronbach's alpha is much the same—a characterization of relative error. Because of this similarity, a comparison can be made between Cronbach's alpha and the G coefficient for FASCI scores. The *dependability coefficient* (Φ) is the other reliability estimate generated in a G Theory study, and is used in characterizing *absolute* (criterion-referenced) score reliability. By using G Theory, one is able to generate both of these estimates of score reliability and examine how they could change if the number of items and/or raters was varied. In this way, the G Theory approach represents a deeper examination of score reliability than that based on Cronbach's alpha—one that is consistent with the intended relative and absolute uses of FASCI scores. In the next chapter, I will discuss the G Theory framework in much more detail and present and discuss these reliability estimates.

Evidence based on Response Processes

In addition to the evidence based on the internal structure of the FASCI discussed above, evidence based on response processes is needed to evaluate propositions four and five: (4) SK can be measured reliably with a scenario-based survey, and (6) SK score interpretations change when specific science content is added to the survey items. Evidence based on response processes provides information about how respondents are interpreting the items on a test and how they are formulating their responses to those items. Although response scoring and scorer training are often included in discussion of evidence based on test content (cf., AERA, et al., 1999), I will discuss it in this section on response processes. The reason I choose to include it here is because it is closely related to evidence from think-aloud interviews, which clearly falls under response processes.

Response Scoring.

The open-ended item responses from each version of the FASCI are scored using a set of decision rules and scoring guides (see Appendix B). 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. This further development is discussed in detail below. In scoring 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. This scoring results in five unique SCI scores for each respondent on the FASCI, assuming that they responded to all item prompts. In scoring SCI with the initial set of guides, scores of 0 or 1 are given. Some discussion of interactive teaching and learning has to be present in the response for it to be assigned a score of 1, else it is assigned a score of 0.

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. Responses which exemplify each scoring category can be seen in Figure B1 (Appendix B). Again, five FA scores are possible for a respondent assuming that they responded to all item prompts. FA scoring is essentially the same in the initial and new sets of scoring guides. Only minor clarifications were made.

New Rater Training.

The FASCI development team originally created the strategic knowledge construct and scoring guides as described above. During pilot testing of the FASCI and initial analyses in the present validity studies, rater agreement was found to be rather low when responses were scored with the scoring guides discussed above. For example in this validity study, on a random sample of 20% of responses agreement on the FA dimension between a second rater and myself was 66% on the neutral-FASCI and 76% on the physics-FASCI. For that same random sample on the SCI dimension, agreement was 71% on the neutral-FASCI and 83% on the physics-FASCI. This low agreement in scoring is one of the reasons that a deeper examination of score reliability is warranted—one that takes into account the potential error due to raters. In order to address this relatively low rater agreement and to see if there were other patterns in the response data not picked up in the previous scoring, a new scoring team was assembled and trained. This new scoring team consisted of three practicing secondary science teachers who possess a high level of strategic knowledge. That is, they were chosen because they have a large repertoire of strategic approaches, can apply them conditionally, and are student-centered in nature.

These new raters were trained over a series of three face-to-face meetings and engaged in three independent practice scoring tasks. In this training, responses from both the neutral and physics versions of the FASCI were pooled together and scorers were not told about the two different versions. In other words, they were blind to the FASCI version. Interestingly, 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 the end of semester administration of the n- and p-FASCI (described above) was used. This subset of responses was not used in quantitative analyses to compare group means, due to attrition and random assignment mix-ups from the pre-semester administration to the post. I purposefully selected response sets for training that were either easy or hard for the previous rater and I to agree on. For example, in the first training meeting, we discussed and scored together three response sets that were easy for the previous rater and I to agree upon, and two that were hard for us to agree upon. Throughout the training phase, I used both easy and hard to agree upon response sets.

In our first training meeting, I introduced the overall purpose of this work, the FASCI instrument, and briefly described the proposed score interpretation and instrument use. I then showed an example FASCI scenario-based item and an example response. Without showing the initial set of scoring guides (developed previously), we discussed the characteristic of the response with respect to my brief descriptions of FA and SCI. I then introduced the scoring guides and we scored the response together, agreeing on the same scores that it had been given in previous scoring. We did this for one other "easy" response set before scoring another "easy" response set individually. During our discussion of scoring this response set, there was disagreement on the SCI score. All of the new raters scored this particular response as SCI = 1, but they thought that it was a "better 1" than the previous response, which also was scored SCI = 1. This led to a lengthy discussion about what makes a response a "better 1" than a "lesser 1." One of the new raters suggested the idea that the "better 1" response included some statement of

rationale for why they envisioned this particular activity as an opportunity for interactive teaching. For example, a response which states that "students could ask questions" would be scored as SCI = 1, and is not as sophisticated as a similar response which states "students could ask questions which would help them articulate their ideas" which includes a rationale and therefore would be scored as SCI = 2. We incorporated this idea into a new highest level for SCI, SCI = 2 (see scoring guide in Appendix B).

After this first face-to-face training meeting, raters were given five response sets to score individually before our next face-to-face meeting. Rater agreement on this first independent scoring task is summarized in Tables 2 and 3. SCI score agreement between pairs of raters was quite good, but FA score agreement was not.

Table 2.

SCI rater agreement on first independent scoring task, five response sets

	Rater 1	Rater 2	Rater 3
Rater 1	1	80%	80%
Rater 2		1	100%
Rater 3			1

Table 3.

FA rater agreement on first independent scoring task, five response sets

	Rater 1	Rater 2	Rater 3
Rater 1	1	10%	20%
Rater 2		1	30%
Rater 3			1

In the second face-to-face training meeting, we concentrated on discussing the FA dimension. The main point of disagreement on scoring this dimension surrounded the definition of context which bore on strategic choice. We discussed and clarified this point, and practiced together with five more response sets in that meeting (two "easy" to score and three "hard" to score). Once again, the raters

were tasked with scoring five response sets individually. Rater agreement on this independent scoring task is shown in Tables 4 and 5.

Table 4.

SCI rater agreement on second independent scoring task, five response sets

	Rater 1	Rater 2	Rater 3
Rater 1	1	80%	80%
Rater 2		1	80%
Rater 3			1

Table 5.

FA rater agreement on second independent scoring task, five response sets

	Rater 1	Rater 2	Rater 3
Rater 1	1	80%	80%
Rater 2		1	80%
Rater 3			1

While rater agreement was indeed better on FA, it was no better than before on SCI (and in once case, between raters 1 and 3, it was worse). Therefore we had one more face-to-face meeting in which were scored five more example response sets together (all "hard" to score), and further clarified some of the language in the scoring guides and out interpretation of that language. Once again, raters were given five response sets to score independently. Results from this last round of independent scoring were quite encouraging, and are summarized in Tables 6 and 7.

Table 6.

SCI rater agreement on third independent scoring task, five response sets

	Rater 1	Rater 2	Rater 3
Rater 1	1	100%	100%
Rater 2		1	100%
Rater 3			1

Table 7.

FA rater agreement on third independent scoring task, five response sets

	Rater 1	Rater 2	Rater 3
Rater 1	1	100%	80%
Rater 2		1	80%
Rater 3			1

After this third round of independent scoring, the raters were given access to all 60 response sets from the pre-semester administration of the n- and p-versions of the FASCI. Again, raters were blind to survey version. Scoring was accomplished using spreadsheets and forms in Google Docs. Results from this scoring will be discussed in the next chapter.

Analysis of Scores.

In chapter five I compare the mean scores for each dimension on the neutral and physics FASCI in order to evaluate proposition five (SK score interpretations change when specific science content is added to the survey items). This analysis begins with two approaches to calculating mean scores: (a) averaged based on the number of items that were completed by a respondent, and (b) calculated using only complete response sets. I conduct the second approach as a check of the sensitivity of the first approach to the existence of missing data. In each of these mean score comparisons, I use independent sample t-tests to determine if the differences between FA and SCI scores on each version are statistically significant. I also express the difference in scores on each dimension and version in terms of effect size units, and calculate the statistical power of the tests of significance for each of these differences. Effect size is calculated as Cohen's d (equation 3).

$$ES = \frac{\overline{Y_p} - \overline{Y_n}}{SD_p} \quad (3)$$

In this equation, Effect Size (*ES*) is calculated by taking the difference between the mean p-FASCI ($\overline{Y_p}$) and n-FASCI scores ($\overline{Y}_n$) for the dimension of interest, and dividing it by the pooled standard deviation (SD_p) of those scores.

Statistical power estimates give the probability of rejecting the null hypothesis (in this case, that the mean scores between each version are the same) if it is, in fact, false. A higher value for statistical power can be thought of as a more confident rejection of the null hypothesis. For all power calculations, alpha (the probability of falsely accepting the alternative hypothesis when the null hypothesis is true) was set at 0.05. For the statistical power calculations, I used the software G*Power (Faul, Erdfelder, Buchner, & Lang, 2009). These calculations were performed *post hoc* and are based on Effect Size and sample size. However, using an *a priori* approach, I can get some sense of the sample size required in order to achieve a specific value for power. For example, given a moderate effect size of 0.50, in order to achieve a power value of 0.80 (a strong value), I would need sample sizes of 51 on both versions of the FASCI. Given that my samples are smaller (26 on the n-FASCI and 34 on the p-FASCI), it is reasonable to believe that the statistical power will be low for score comparisons unless the effect sizes for difference in scores between versions are very large.

One should keep in mind that a finding of statistical significance when comparing n- and p-FASCI mean scores would *support* the proposition as stated: that SK score interpretations change when specific science content is added to the items. In addition, observing qualitative differences in the item responses between versions may also support this proposition. However, the potential cause of any observed differences will need to be examined in order to evaluate the hypothesized differences.

As is evident in the above *a priori* statistical power calculations, power can be increased by using larger samples. However, it can also increase if the reliability of the measure used increases. In fact, the effect of change in reliability on statistical power can in some cases be the same as the effect of changing the sample size (Haertel, 2006). Of course, this depends on there being a difference between

the groups being compared (i.e., the null hypothesis is rejected). Score reliability therefore also has a central role in the comparison of these scores.

In addition to comparing mean scores, I also compare item difficulties (as p-values) on each dimension and version of the FASCI at each administration. The p-value (not to be confused with a p-value in statistical tests of significance) is simply the proportion correct on an item. A comparison of p-values between the same items on different versions of the FASCI can give a sense of how difficult or easy that particular item is for respondents on each version, and when coupled with a qualitative examination of item responses can provide insight into how the inclusion of content in the FASCI items affects the response process, if at all.

In comparing scores between each version of the FASCI, I also examine the relationship between these scores and some physics expertise variables. On each version of the FASCI, respondents were presented with a question about what subject they plan on teaching (e.g., biology, chemistry, physics, etc.). They were also asked how many physics courses they had taken since being in university. Finally, all individuals were prompted to respond to the 12 physics content questions (embedded in the p-FASCI and appended to the end of the n-FASCI). These content questions were scored and each respondent was assigned a physics content knowledge score. Each of these three pieces of information (subject they plan on teaching, number of physics courses taken, and physics content knowledge score) are used to characterize a respondent's physics expertise.

Response Data from Previous Pilot Testing.

Two previous pilot tests of the FASCI provide other response data which adds to the evidence based on response processes. In pilot test one, the same version of the FASCI that was used in the n-FASCI/p-FASCI score comparison part of this study (see Appendix A) was administered to a sample of 63 respondents. These respondents included undergraduate Learning Assistants (LAs), University Faculty, practicing K-12 teachers, and university graduate students. In pilot test two, the version of the n-FASCI

used was the same one as that used in comparing FASCI and observation protocol scores (discussed below). 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 one were replaced with four new items for pilot test two. In the second pilot test, the FASCI was administered to a sample of 96 respondents from a population similar to that of pilot test one.

Response data from these pilot tests was scored using the initial set of scoring rules created by the FASCI development team. The reliability of these scores will be discussed in the next chapter, and will contribute to an evaluation of proposition four (the reliability of SK scores).

Think-Aloud Interviews.

In order to collect further evidence base on response processes, I conducted think-aloud interviews. These interviews were conducted following the first and second administrations of both versions of the FASCI with a subsample of six participants from Western State University (WSU), three who took the n-FASCI and three who took the p-FASCI. All WSU respondents were asked to volunteer to be interviewed after they responded to the first administration of the n- and p-FASCI. A compensation of $20 was offered for participating in an interview lasting approximately 30-45 minutes. Six students in the WSU sub-sample indicated (by noting in the comments field of each survey) that they would be interested in participating in an interview, and I contacted each of them personally. Each interview was audio recorded and transcribed for analysis. The average length of the first six interviews was 28 minutes. All six participants also agreed to (and did) engage in a second think-aloud interview after they completed the post-semester survey, for which they were also paid $20. These interviews averaged 34 minutes in length. A description of the interview participants is provided in Table 8.

Table 8.

Description of Think-Aloud Interview Participants (frequency counts or mean/SD)

Variable		n-FASCI	p-FASCI
Number of respondents (pre-test ID)		3 (3475220, 3479452, 3473229)	3 (3473183, 3473454, 3474297)
Group	Learning Assistants		1
	non-Learning Assistants	3	2
Age		23 (1.3)	23 (1.6)
Gender	Female	2	1
	Male	1	2
Want to teach	Astronomy		1
	Biology	2	
	Chemistry	1	1
	Geology		1
Years teaching experience		2.3 (0.7)	2 (0.8)
Number of physics courses		3 (0.0)	4 (1.2)
Physics Content knowledge		0.72 (0.64)	0.72 (0.54)

Conducting the Think-Aloud Interviews.

During these interviews, I asked participants about the entire experience of having responded to the survey, including what they thought it was designed to measure, how well they thought it did that, and what was difficult and easy about the survey. Most importantly, I wanted to have them explain to me their thought processes as they were formulating their written responses, and to see if the way they described their thinking was reflected in what they had written. I chose to structure this part of the interview in a way that would focus the participants on discussing their thought processes so that I could gain insight about their response processes and assign scores for them on a particular item based on their comments. Had I left the interview format more open, participants may have discussed any item of their choosing or may not have described they thought processes in a way that provided scoreable comments.

I made this purpose (identifying their thought processes) explicit to the respondents as an introduction to explaining to them what I meant by "think-aloud." I then asked each participant to read their responses to a particular item and to talk me through their thought processes when they were constructing their response. In prompting them to discuss their thought processes, I used various prompts and probes, such as "Tell me what you were thinking about when writing this responses" and "Can you explain your thinking on that to me?" I used the same think-aloud prompts in both the n- and p-FASCI interviews.

The interview protocol that I used differed from first to second administration interviews in one main way: in the second interview, I told the respondents exactly what the survey was designed to measure, and asked them for feedback on how well they thought it accomplished this. I also asked them to self-rate their own ability, based on the FA and SCI scoring guides presented to them. It should be noted that these were the initial scoring guides, not the revised ones that resulted from the new scorer training. The interview protocols for first and second administrations can be seen in Appendix C.

Analyzing the Interview Data.

I conducted two types of analyses in examining the interview data: (a) assigning FA and SCI scores based on each participant's discussion of their thought processes, and (b) assigning specific codes to interviewee comments. With respect to the first of these, I specifically focus on respondent's comments about their thought processes when they were constructing their written responses. I then assigned FA and SCI scores based on how they discussed their response to the particular item, using the same *initial* scoring guides as those used in assigning scores for written responses (see Appendix B). My aim was to derive scores from the interviews that I could compare to their scores based on their written responses.

I also assigned specific codes to comments in each interview using an *a priori* framework (see Appendix D). I developed this framework based on what I *expected* the participants might discuss, such

as content and context-dependence of their responses. I expected these to be salient codes based on my analysis of pilot test responses to the FASCI scenarios and based on the FASCI scoring guides. For example, conditioning strategic choices in response to prompt c ("What would you do next if that didn't work?") is a necessary component of a high-category FA response. Therefore, I want to specifically identify if and how respondents discussed such conditioning when describing their response processes to me. I created the code "context-dependence" to capture comments which were identified as such. An example response assigned this code comes from an n-FASCI respondent (ID 3626337):

> *"...not knowing anything about this particular student, it's hard for me to say specifically why I would do something one way with them if I don't know them."*

I chose not to use an entirely emergent coding scheme since I started with a definite framework about what I expected to see in the data. However, I remained open to other patterns in the data, and did develop new codes which emerged as relevant during the analysis process. For example, I had not expected to code responses to prompt a) ("How might this activity facilitate student learning") as discussing either the content *or* the students, but this distinction emerged as a very important one with respect to the score reliabilities. Therefore, this emergent code became a part of the coding framework. I compare the results from this particular coding to the score reliabilities in chapter four.

Two interviews (17% of the total) were also coded by a second (trained and experienced) interview rater, and the agreement between my coding and theirs was 84%. That is, 84% of time we assigned the same codes to passages. Once the interviews were coded, I searched each set of codes for patterns within and across interviews, both from first and second administrations. I first examined the percentage of each interview (by time) that was assigned various codes, and then examined the passages within each interview that were coded similarly to search for patterns. These coding frequencies and patterns were used to inform the comparisons of scores based on item responses with those from interview comments.

Evidence based on Relations to other Variables

In order to evaluate proposition three (SK can be observed in teaching practice), evidence based on relations to other variables is needed. More specifically, I describe collecting evidence from observations of teachers' practice.

Observing SK in Practice.

I conducted observations of teaching for a separate sample of respondents from Western State University. These individuals were not a part of the neutral and physics-FASCI sample in the study described above. The observations were conducted by myself and other members of the Learning Assistant (LA) program research team. This sample consisted of 18 science and math teachers who were participants in the ongoing LA program research, meant to assess the effectiveness of the WSU LA program. 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). These teachers were observed using the Reformed Teaching Observation Protocol (RTOP; Sawada et al., 2002), discussed in the previous chapter.

FASCI Participation.

All of these participants responded to the neutral FASCI, which was being used in the larger LA research program. The version of the n-FASCI to which they responded was that used in Pilot Test 2 (six items; discussed above), whereas the version used in the n- and p-FASCI study described above was that used in Pilot Test 1 (five items). In these analyses I compare the teachers' FA and SCI scores with their RTOP scores.

Conducting Observations of Practice.

Each of these individuals was observed at least two times during the spring semester of 2009. Each observation was followed by an interview, and their teaching practice was scored used the RTOP

(Appendix E; Sawada, et al., 2002). The RTOP structure and analysis will be discussed in more detail in chapter five when presenting this data. During the observations, extensive field notes were taken by the observer (either myself or other members of the LA research team) and noted on the RTOP forms in the space provided. These notes proved to be important in characterizing teaching practice.

Comparing FASCI scores and RTOP scores.

I compared the FA and SCI scores for these respondents with their RTOP scores. I compare scores by examining three types of data: (a) correlations, (b) scatterplots, and (c) cross-tabulations of score categories (e.g., low, medium, or high). The goal of these comparisons is to identify cases which were rated consistently or inconsistently based on each measure. I then examine representative cases in detail, also using evidence from notes taken during their observations. These analyses and the related findings will be discussed in chapter five.

Discussion

Evidence based on response processes comes from response scoring, qualitative analysis of responses, and think-aloud interviews. In chapter four, I will discuss these things when presenting evidence needed to evaluate proposition four (SK can be measured reliably with a scenario-based survey). In that chapter I will further discuss the results from the rater training and item scoring. I will also examine the reliability of those scores. Also in chapter four, I will present and discuss the estimates of reliability based on different combinations of items and raters.

The evidence based on internal structure that I collected consists of information about the scenario-based items on the FASCI, the content specificity (or lack thereof) of those items, and the reliability of scores resulting from the FASCI. In chapter five, I will discuss results from the "content test" experiment described above. In that experiment, two versions of the FASCI (the content-neutral and physics versions) were administered at random to a population of prospective science teachers. In order to investigate proposition five (SK score interpretations change when specific science content is

added to the survey items), I will compare scores from each version of the instrument. I will also discuss respondent comments from some of the think-aloud interviews that were conducted with a subsample of survey respondents.

Evidence based on relations to other variables is used to evaluate propositions three (SK can be observed in teaching practice). In chapter six, I present and discuss evidence from observations of teaching practice and compare those to FASCI scores for a sample of practicing teachers.

In the next three chapters, I shift from organizing the discussion around types of validity evidence to organizing around the specific propositions presented in support of the proposed score interpretation. In chapter four, I will discuss proposition four (SK can be measured reliably with a scenario-based survey). In chapter five, I will discuss proposition five (SK score interpretations change when specific science content is added to the survey items). In chapter six, I will discuss proposition three (SK can be observed in teaching practice).

Chapter 4: How Reliably can Strategic Knowledge be Measured?

Introduction

In my validity argument structure, I presented a set of propositions which are central to the proposed instrument use: to evaluate the effects of a teacher education program on novice science and mathematics teachers' strategic knowledge (SK). This proposed use involves making both norm-referenced decisions (e.g., did a particular novice science or mathematics teacher achieve a higher SK score than another novice science or mathematics teacher?), and criterion-referenced decisions (e.g., did a particular novice science or mathematics teacher achieve a certain level on the SK construct?). Score reliability is central to both of these situations, as precision in measurement is necessary in order to make such decisions.

Score reliability is often thought of in terms of Cronbach's alpha, which is a characterization of *relative* measurement error, and is of interest when making norm-referenced decisions such as that in the first situation posed above. As discussed in the last chapter, Cronbach's alpha considers only one composite source of measurement error and does not take into account a potentially important source of error in this study (that due to raters), therefore it over-estimates score reliability. But this classical test theory conception (represented by Cronbach's alpha) is not the only way to think about score reliability. By taking into account other *facets of measurement* such as the raters, and interactions between these facets, reliability can be investigated more deeply. When considering the items, raters, and their interactions as potential sources of error, one can examine not only the reliability for making *relative* (norm-referenced) decisions, but also that for making *absolute* (or criterion-referenced) decisions. In this chapter, I will discuss proposition four (SK can be measured reliably with a scenario-based survey) by investigating reliability in three ways: 1) rater agreement in scoring, 2) the classical test theory conception of score reliability, and 3) score reliability conceptualized within a Generalizability Theory (G Theory; Brennan, 2001) framework.

New Response Scoring

One approach that instrument developers often take to examining score reliability is to assess the agreement of raters in scoring open-ended responses or observable actions. For example, in chapter two I described the development of the Reformed Teaching Observation Protocol (RTOP; Sawada, et al., 2002) in which this type of approach was used. As mentioned in the previous chapter, checks of agreement between my FASCI scores and those of another trained rater showed relatively low agreement. The recruitment and training of three different raters was undertaken in response to this relatively low agreement. These new raters were blind to FASCI version (neutral or physics) and for scoring purposes the responses from each version were pooled together. None of these new raters inquired about the differences in item responses with respect to some referencing physics content while others did not. In training, I only gave them access to the neutral version of the FASCI instrument.

During training, the raters added a level to the SCI scoring guide and worked to further clarify and define some of the scoring guide language. Once they became familiar with the task at hand, they required very little prompting from me in making these changes. These raters talked about how they would need to come to agreement in scoring without me pushing them to do so. Further, they saw the need for better definition on the scoring guides so that they could agree on scores in training, and made the necessary changes. The resulting FA and SCI scoring guides are shown in Figures 1 and 2 respectively, with the new or clarified language *italicized*. Very little change was made to the FA scoring guide—most of the work on this dimension was in having a discussion about what constituted a new strategy and an appropriate contextual factor, and in further defining some of the wording on the FA construct map. As discussed in chapter three, level two on the SCI scoring guide was added to accommodate the new idea that a high-level response should include some *rationale* for why the situation was seen as an opportunity of interactive teaching and learning. The modified FA and SCI construct maps can be seen in Appendix B.

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 1. New FA 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*

Figure 2. New SCI scoring guide

On the last independent scoring task in training, rater agreement was quite good on both the FA and SCI dimensions (see Tables 6 and 7, chapter three). Overall rater agreement for the full response set (60 responses) is shown in Tables 1 and 2 for the FA and SCI dimensions respectively. Rater agreement by individual item can be seen in Appendix B. While not quite as good as in training, rater agreement was still quite a bit higher than in previous efforts. 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 1.

Overall FA Rater Agreement

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

Table 2.

Overall SCI Rater Agreement

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

All three of these new raters were practicing secondary science teachers who were thought to exhibit expertise in SK. Because they were similar to each other in this characteristic (more so than the scoring teams in prior FASCI scoring), and because of their initiative in taking the lead in coming to agreement in scoring, their scoring agreement was better than in previous scoring efforts. Although this new scoring exhibited higher levels of agreement than past efforts, this information alone is not enough to support the reliability of scores on the FASCI. A first step towards a deeper examination of score reliability comes from computing Cronbach's alpha for these scores.

Cronbach's alpha of SK Scores

I first estimated score reliability by calculating Cronbach's alpha for each dimension and version of the FASCI. In this case any measurement error cannot be attributed to scoring by the raters. This is because in classical test theory, the only source of measurement error is, in theory, due to the differences in scores that would be observed if the same respondent were to be surveyed over and over again repeatedly. There is no specific "rater component" of measurement error. This is in contrast to

the approach taken in the analyses discussed below, where observed score variance is decomposed into multiple facets of the survey process (e.g., items, raters) that contribute to measurement error.

In this classical test theory approach to examining score reliability, the value of the reliability coefficient (the ratio of true score variance to observed score variance) will be higher if the variability of the error component of observed scores is lower than the variability of the true score component (see equation 1). In this equation, σ_τ^2 represents the variance in true score for an individual on one dimension and version of the FASCI, σ_x^2 represents the variance in observed score on that dimension, and σ_e^2 is the variance in the measurement error. This relationship shows that a decrease in error variance (σ_e^2) will increase score reliability.

$$reliability = \frac{\sigma_\tau^2}{\sigma_x^2} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_e^2} \quad (1)$$

I estimate score reliability for both n- and p-FASCI pooled together and compare this information to results from previous pilot testing of the FASCI. I do this in order to see if score reliabilities resulting from this study are consistent with those from previous pilot tests, one of which had a significantly larger sample size from a similar population (see Table 3). For this study, I found that score reliability for each dimension is within the range of those found in previous pilot testing of the FASCI. Score reliability, sample size, and variability in observed scores (expressed in terms of standard deviations) from this study and from previous pilot tests are shown in Table 3. FA score reliability for this study was 0.62, and in the two previous pilot tests it was found to be 0.69 and 0.43. SCI score reliability was 0.51, and in the two previous pilot tests it was found to be 0.42 and 0.46. The SD of observed FA and SCI scores for the aggregate (n-FASCI + p-FASCI) samples were lower than those observed in previous pilot testing. This is due to the fact the scores in this study are based on three raters, while in previous pilot studies scores were based on only one rater. By averaging a respondent's score over three raters, there are likely to be fewer outliers (extremely high or low scores) that result

from particularly harsh or easy scoring. This serves to decrease the spread of the observed scores. Note that there is a significant amount of missing data. This has an effect on score reliability, and on comparisons of scores between versions (discussed further in the next chapter).

Table 3.

Reliability estimates (Cronbach's Alpha) and overall percentage of missing data for each dimension, this study and previous pilot testing

		Sample size	FA alpha	SD of Observed FA score	% FA missing	SCI alpha	SD of Observed SCI score	% SCI missing
Present study	n+p-FASCI pooled	60	0.62	0.37	21.7	0.51	0.36	13.3
	n-FASCI	26	0.69	0.42	7.7	0.51	0.35	3.8
	p-FASCI	34	0.45	0.32	32.4	0.41	0.34	20.6
[¥]FASCI Pilot test 1	n-FASCI	63	0.69	2.05	6.5	0.42	1.13	5.5
[¥]FASCI Pilot test 2	n-FASCI	96	0.43	1.73	2.4	0.46	1.24	1.7

[¥]Results from these pilot tests of the n-FASCI are reported in Briggs, Talbot, and Otero (in progress)

According to Traub (1994), reliability estimates of 0.80 or higher are routinely achieved for objectively scored tests, while those for subjectively scored tests (such as the FASCI) can be lower. This target value of 0.80 is similar to that achieved in the vignette-based surveys of mathematics teacher practice described in the previous chapter (Le, et al., 2004), where estimates of internal consistency were observed to be between 0.80 and 0.86. However, that survey was not open-ended, rather the respondents ranked the likelihood that they would engage in a particular behavior. Therefore it is reasonable to think that the lower end of that range (0.80) is a good target for an open-ended scenario-based instrument such as the FASCI. Given this benchmark, the score reliability observed for the FA and SCI dimensions is not very high, indicating a lack of internal consistency in this measure of SK. Note that reliability for the p-FASCI is lower than that for the n-FASCI, but both n- and p-FASCI score reliability were within the range of that observed in previous pilot tests for the FA dimension, and close to that

observed in previous pilot testing for the SCI dimension. I will discuss the differences between n- and p-FASCI score reliability and the existence of missing data in the next chapter.

As mentioned above, these estimates do not take into account any potential error due to the raters and therefore likely over-estimate score reliability. Because of the previously observed low rater agreement, it is reasonable to think that there would be some error due to the raters. What is needed is a framework for decomposing the composite error variance of observed scores into separate components.

A Deeper Investigation of the Reliability of SK Scores

In this deeper investigation of score reliability, I am interested in: 1) assessing how much of the variability in observed scores can be attributed to the items, raters, and interactions between respondents and these facets, and 2) how these sources of ”error” variance could be reduced if the number of items and/or raters were to be increased. The first of these points is addressed by estimating error variances based on the observed scores, and will be discussed directly below. The second of these points involves estimating how score reliability and the standard error of measurement would change if the number of items and/or raters associated with FASCI scoring was changed.

These analyses are accomplished by applying Generalizability Theory (G Theory; Brennan, 2001). In a G Theory analysis, multiple sources of error variance for single observations (e.g., a person’s score on a single item scored by a single rater) can be estimated. This is done as a part of the generalizability study (or “G Study”). In subsequent decision study (or “D Study”) one then investigates the variance estimates associated with *mean* scores from some sample (e.g., of persons across a set of items and raters). Mean scores are the focus of the D Study as it is these scores that are the basis for any decisions made about a person. The D Study therefore helps to inform decisions about the optimal configurations for maximizing score reliability related to the proposed score interpretation and instrument use.

Sources of possible error variance are referred to as *facets of measurement* in G Theory. It is assumed that measurements are made from a *universe of admissible observations*, which are observations that are seen as interchangeable. Each facet is considered to have its own universe. For example, in this study the facets of interest are items and raters. Each of these is conceptualized as being from an infinite universe of admissible items or raters. In other words, in theory at least, there are an infinite number of possible item scenarios from which those on the survey are drawn from, and an infinite number of potential raters of responses to those items. It is plausible to think that there are a very large number of potential teaching scenarios which could be included in a FASCI item, given the complexity of teaching. It is also plausible to think that there exists a very large pool of potential raters of FASCI responses. Conceiving of observed scores as being a *sample* of respondent performance drawn from a universe of potential item scenarios and rated by a universe of potential raters allows one to think of measurement error as being due to sampling variability. Accordingly, we would expect sampling variability to decrease as sample size increases. However, if all of the items in the universe are of similar quality, then scores resulting from a sample of items drawn from that universe will not differ greatly from scores resulting from any other sample drawn from the same universe. The same would hold true for raters. Ideally, we would want the bulk of observed score variability to be attributable to the *object of measurement* (persons, denoted by "*p*"). The facets of measurement that are not the object of measurement are therefore thought of as sources of error (in this case, items, denoted by "*i*" and raters, denoted by "*r*"), and accordingly we would want to minimize the role that the variance associated with each of those facets plays in FASCI score interpretations.

G Theory represents a more nuanced way of thinking about score reliability than the classical test theory conception discussed above (cf., Thompson, 2003). Rather than considering an observed score as being composed of a true score and an error score component (see equation 1), in a G Theory analysis one can consider the error attributable to multiple components (the facets). Because a G

Theory analysis simultaneously estimates variance components for the object of measurement (e.g., persons) and multiple measurement facets (e.g., items and raters), the concept of observed score represented in equation 1 is somewhat different. Rather than thinking of true score and error score components of observed scores, we now think of *universe scores* and their variance components due to each facet of measurement.

The item and rater facets are of interest because one would like to be able to generalize from one set of items or raters to a different (larger or smaller) set of each. Therefore in this study, the *universe of generalization* consists of different specified item/rater combinations. In making *absolute* decisions (i.e., whether or not a person achieves some specific SK score), each of these facets is a potential source of error in generalization. Items vary in difficulty and quality, so a respondent's score on one set of items may not be the same as that on another set. Raters vary in their judgments when scoring—some are harsher than others, for example. Each of these potential variances can be estimated in a G Theory analysis, and each are of particular interest when making absolute decisions. This is one of the proposed uses of the FASCI instrument as described above.

As stated above, in this study I consider items and raters (both conceptualized as being drawn from an infinite "universe" of each) as potential sources of error variability in *universe scores*. This study therefore has two facets of measurement. Further, the interactions between these facets and the object of measurement (persons), and the facets and each other, are other potential sources of variability. Because I would accept within the universe of admissible observations the response of any person (*p)* to any item *(i)* scored by any rater *(r)*, then the G Study design is said to be *fully crossed*. This is represented by the notation *p x i x r* (read "p cross i cross r", or "p by i by r"). This fully crossed design, each of the variance components, and their interactions is represented by the Venn diagram in Figure 3.

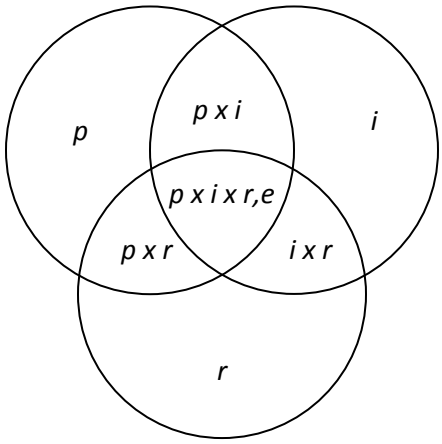


Figure 3. Venn diagram representing variance components in this p x i x r design G Study

Once each of these variance components is estimated in the G Study, two types of error variance can be calculated in the G and D Studies: *absolute error variance* and *relative error variance*. Absolute error variance is denoted by upper case delta (Δ), and is the sum of all variance components which involve the *facets* of measurement ($\hat{\sigma}_I^2, \hat{\sigma}_R^2, \hat{\sigma}_{pI}^2, \hat{\sigma}_{pR}^2, \hat{\sigma}_{iR}^2, \hat{\sigma}_{pIR,e}^2$)[2]. The square root of the absolute error variance is the absolute standard error of measurement (SEM Δ). Absolute error variance is of interest when making criterion-referenced decisions, such as whether or not a person achieves a pre-determined SK score that is representative of some level on the SK construct. Relative error variance is denoted by lower case delta (δ), and is the sum of all variance components which involve the *object* of measurement ($\hat{\sigma}_{pI}^2, \hat{\sigma}_{pR}^2, \hat{\sigma}_{pIR,e}^2$). Again, the square root of this error variance is the SEM (δ). Relative error variance is of interest when comparative decisions are being made, such as distinguishing one respondent from another. Each of these error variances, errors, and SEMs are calculated for a particular

[2] The "hat" (^) over each variance component in the D Study indicates that they are new estimates of the parameters obtained in the G Study. Also, note the capitalized subscripts for items and raters (*I* and *R*) indicating that these are variance components associated with mean scores rather than individual observations (as in the G study).

D Study, and can therefore be compared across different D Studies which specify different numbers of items and/or raters.

These errors are also summarized in two different reliability coefficients: the generalizability coefficient (ρ^2) and the dependability coefficient (Φ) (Brennan, 1992). ρ^2 is analogous to a reliability coefficient in classical test theory, which is a characterization of *relative* error. This coefficient is of interest when making comparative decisions. The generalizability coefficient takes into account the error associated with all components that involve the object of measurement (see equation 2). In this equation, σ_τ^2 represents the universe score variance. In contrast to the conceptualization of score reliability presented in equation 1 which had only a single term in the denominator denoting error variance, the equation for ρ^2 has three separate terms (in the context of this study) in the denominator which represent sources of error. Therefore, for a single facet G Study ρ^2 would be equivalent to Cronbach's alpha. But for a multiple facet design (as in this study), they are not comparable. Classical test theory partitions variance into only two sources (true score and error score), while G Theory partitions variance into multiple sources (universe score and each variance component involving the object of measurement) (Thompson, 2003).

$$\rho^2 = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \hat{\sigma}_{pI}^2 + \hat{\sigma}_{pR}^2 + \hat{\sigma}_{pIR,e}^2} \quad (2)$$

G Theory analyses also provide another estimate of score reliability, the dependability coefficient (Φ). This coefficient is of interest when making absolute decisions. The dependability coefficient Φ is therefore a characterization of *absolute* error, and as such takes into account the error associated with each facet of measurement (equation 3). Because Φ takes into account the same variance components in the equation for ρ^2 plus the variance components that are associated with the

facets and their interactions, the value for Φ is generally lower than that for ρ^2. In this sense, it is a more critical estimate of score reliability.

$$\Phi = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \widehat{\sigma}_I^2 + \widehat{\sigma}_R^2 + \widehat{\sigma}_{pI}^2 + \widehat{\sigma}_{pR}^2 + \widehat{\sigma}_{IR}^2 + \widehat{\sigma}_{pIR,e}^2} \quad (3)$$

G Study Estimates of the Variance Components.

Figure 4 shows the percentage of observed score variance attributable to the object of measurement and to each facet for the FA dimension of SK as estimated in the G study. Table 4 shows the specific variance components estimates and the percentage that each component contributes to the overall variance for both the FA and SCI dimensions. While relatively little variance is attributable to the items themselves, a very large portion of the variance in observed scores (almost 60%) is in the person by item (*p x i*) interaction. This indicates that a respondent's FA scores across the different raters depend heavily on the particular item being sampled from the universe of items (i.e., the teaching scenario to which they are responding). This finding is consistent with related analyses of students' responses to science performance assessments (Ruiz-Primo & Shavelson, 1996; Shavelson, et al., 1993). Very little variance in observed FA scores is attributable to the raters, and almost none to the *p x r* and *i x r* interaction terms. This indicates that the raters were fairly consistent in their ratings across persons and items, an observation that is supported by the relatively strong rater agreement in scoring.

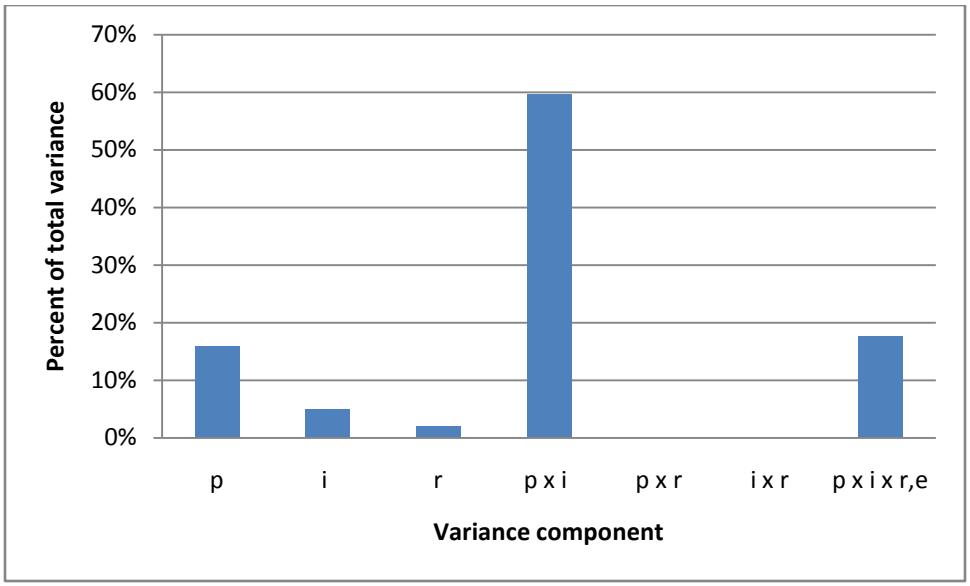


Figure 4. Sources of variability in FA scores

Table 4.

Variance estimates and percentage of total variance, FA and SCI dimensions

	FA		SCI	
Component	Variance	% of total	Variance	% of total
P	0.064	15.9%	0.053	7.1%
i	0.02	5.0%	0.315	42.5%
r	0.008	2.0%	0.014	1.9%
p x i	0.24	59.6%	0.234	31.5%
p x r	0	0.0%	0.006	0.8%
i x r	0	0.0%	0.002	0.3%
p x i x r,e	0.071	17.6%	0.118	15.9%

Figure 5 shows the percentage of observed score variance attributable to the object of measurement and to each facet for the SCI dimension of SK. In contrast to the findings for the observed FA scores, a large part of the variance in observed SCI scores (42%) can be attributed to the items themselves. In other words, the mean score for one randomly selected item (across all persons and raters) is expected to be quite different from the mean score for all items in the universe (across all

persons and raters). Conceived of in this way, this error due to items can be thought of as *sampling error*. A particular item (a teaching scenario) which is sampled from the universe of potential items and included on the survey is likely to be much more difficult or easy for respondents than the average across all of the potential items. Consistent with the variance in observed FA score discussed above, there is also a large amount of variance in observed SCI scores (32%) that can be attributed to the person by item (*p x i*) interaction. This means that a respondent's SCI score across all raters depends on the particular item to which they are responding. Finally, a relatively small amount of the variance in observed SCI scores is attributable to the raters (~2%). Again, this is consistent with the rater agreement on the SCI dimension which was quite good, but not as good as that on the FA dimension.

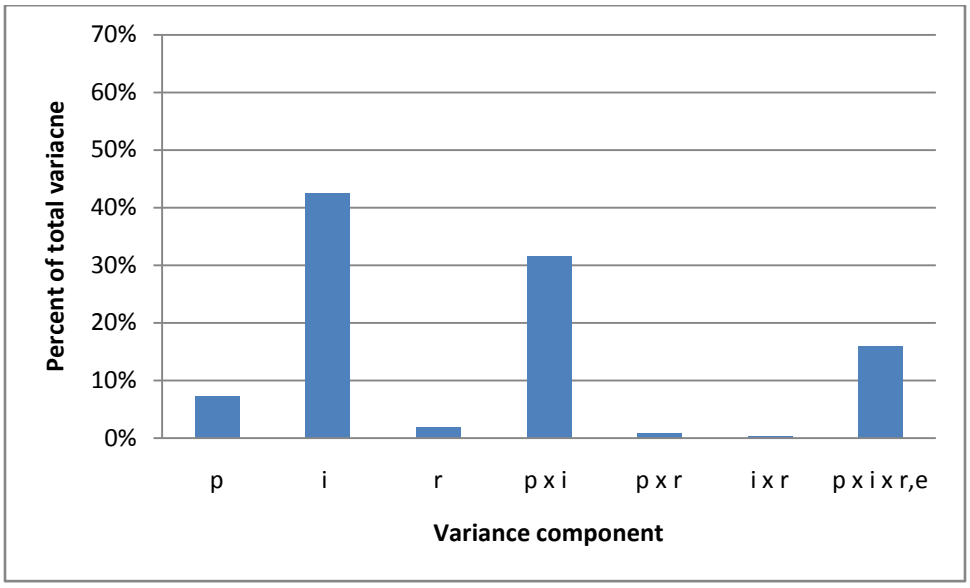


Figure 5. Sources of variability in SCI scores

The central point to highlight from both of these analyses is the importance of the items chosen for use on the instrument. Respondent performance can vary quite a bit depending on the particular item or set of items to which they respond, and the mean score of all persons across any one item is likely to be very different than the mean score of all persons across a larger set of items. It is also important to emphasize that the items in this case are the teaching scenarios, not the constant item

prompts to which individuals respond within each scenario. Again, these findings are consistent with previous work on science performance assessments, where the researchers found that item-sampling variability was a considerable source of measurement error (Ruiz-Primo & Shavelson, 1996; Shavelson, et al., 1993). The implication therefore is that a large number of items are needed on the instrument in order to obtain a reliable measure of respondent performance.

Minimizing the Role of Error Variance in Score Interpretations.

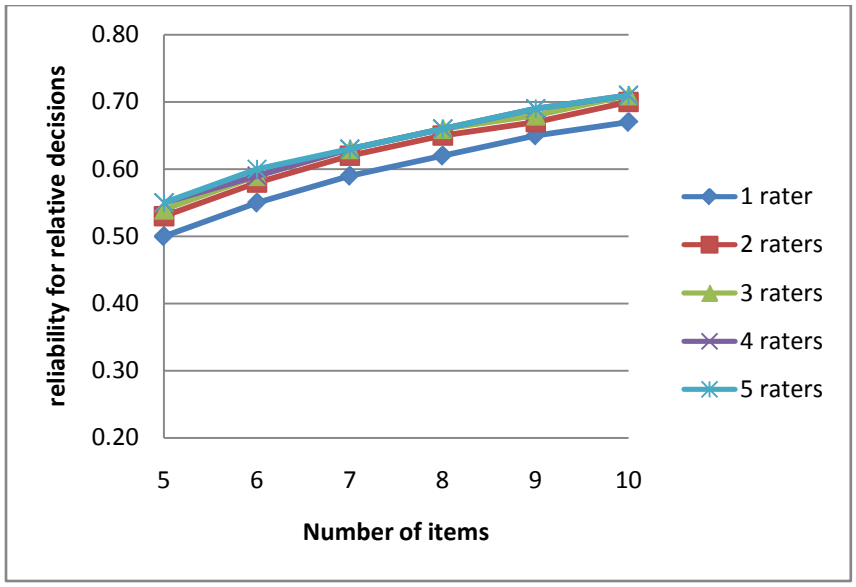
The purpose of the G study was to estimate variance components that were associated with a single observation from the *universe of admissible observations* (e.g., a person's score on one item rated by one rater). These estimates were discussed above in Figures 4 and 5. In the subsequent D studies, the context now shifts to the *universe of generalization*: a specified universe of measurement procedures to which one wants to generalize based on the results of the study. The universe of generalization in this case is specified as different combinations of items and/or raters (beyond the observed five items and three raters). In the D study, variance estimates are generated for *mean* scores of persons, in this case across items and raters. These D study variance components are estimated using the G study variance estimates. For example, the G study variance estimate for items on the SCI dimension was 0.315. This estimate is based on a single person-item-rater observation. The associated D study variance component for a five item specification on the SCI dimension would be 0.315/5 = 0.063. This estimate is for mean scores across items. The expected value of these mean scores for a specified measurement procedure is a person's *universe score*. These D study variance components are the basis for computing measurement error and score reliability that would theoretically result from these different measurement specifications (i.e., item/rater combinations). For both the FA and SCI dimensions of strategic knowledge, I specified D studies for all combinations of five to ten items and one to five raters, and generated both relative and absolute error variances and reliability estimates for each D study.

The left-hand panel of Figure 6 shows how the generalizability coefficient (from here on referred to as "relative reliability") varies as a function of number of items and raters for the FA dimension. For the baseline five item/three rater combination, relative reliability was estimated to be 0.54, which is lower than the value for Cronbach's alpha discussed above (0.62 for this study). This value is smaller than that for Cronbach's alpha because the relative reliability estimate now takes into account facet interactions (*p x i*, *p x* r, and *p x i x r,e*), which are in the denominator of the reliability coefficient calculation, and therefore decreases the value of the estimate. These two facet interactions accounted for an estimated 77% of the variance in observed scores based on the G Study (see Table 4). In this context, by not taking the rater facet interactions (*p x r* and *p x i x r,e*) into account, we would be over-estimating score reliability. This is precisely why the value for Cronbach's alpha is higher.

Not surprisingly, the results from this analysis indicate that the best way to improve FA score reliability comes from adding items to the instrument, rather than from adding raters. Beyond two raters, there is not much increase in reliability for a given number of items. This is consistent with Figure 4 above, which shows the largest amount of variability in observed scores was in the person by item interaction, indicating that a respondent's FA score across raters depends on the particular item being sampled. In order to ameliorate that effect, a larger number of items would be needed on the instrument. Increasing the number of items "sampled" from the universe of items would decrease the standard error of measurement for FA scores. This will be illustrated further below when specifically discussing the standard errors of measurement. Specifically, by doubling the number of items to ten and keeping the number of raters at three, relative reliability could be increased to about 0.71. A similar increase could also be realized by doubling the number of items and having just two raters rather than three. Because relatively little variability in observed scores was attributable to the raters, adding raters (i.e., increasing the size of the rater sample) would not decrease the overall role that error variance

plays. In other words, averaging over more raters would not produce the same effect as averaging over more items because of the low variability attributable to raters compared to that of the items.

The dependability coefficient (from here on referred to as “absolute reliability”) is of interest when making absolute or criterion-referenced decisions, which is consistent with one of the proposed uses of the FASCI instrument. In the D Studies, estimates of absolute reliability are generated for different combinations of items and raters. For five items and three raters, absolute reliability for FA scores is 0.52. The right-hand panel of Figure 6 shows the estimates of absolute reliability as a function of number of items for different rater combinations on the FA dimension. Comparing this to left-hand panel of Figure 6 (relative reliability estimates) shows that these estimates for absolute reliability are similar, though a bit lower. They are slightly lower due to the fact that when making absolute decisions, more facet conditions are taken into account: those involving the facets and their interactions as well as those involving the object of measurement (see equations 2 and 3 above). Because of this, measurement error is increased and the related reliability estimates are lower (i.e., more critical) than those for making relative decisions.



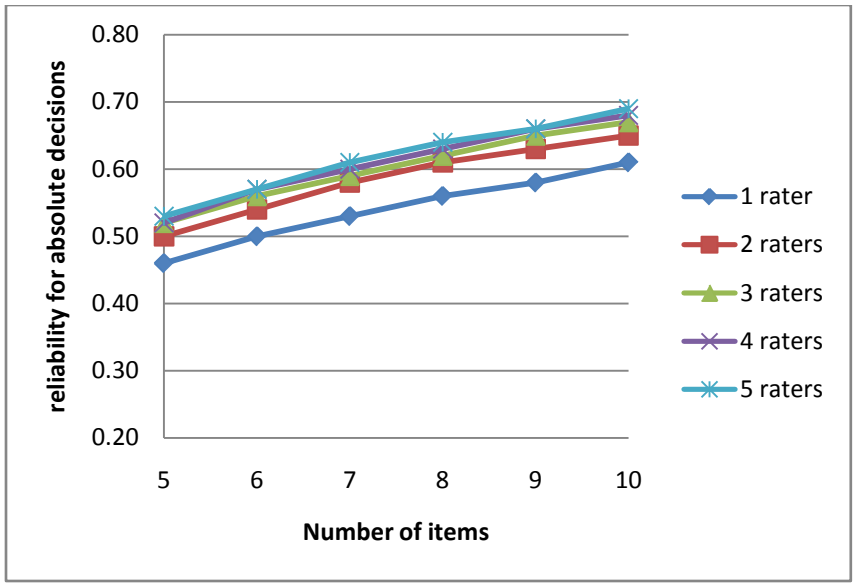
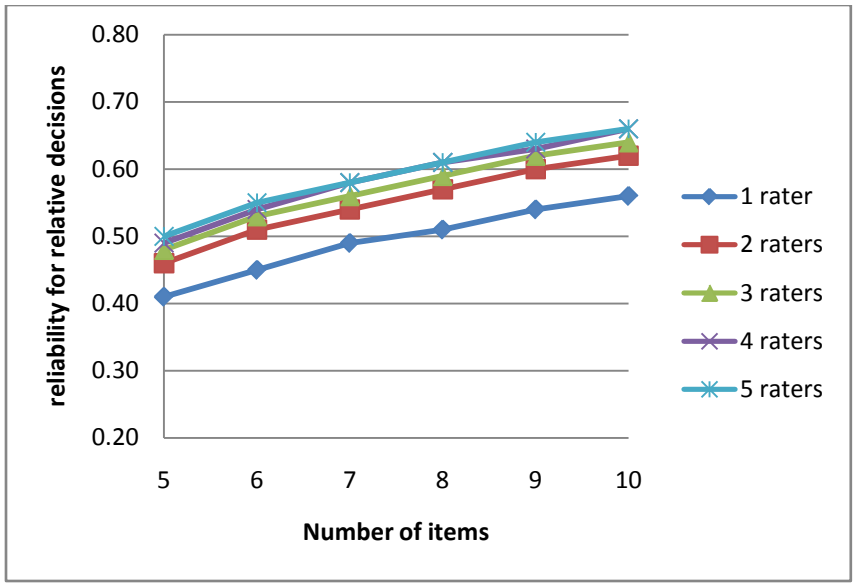


Figure 6. Relative and absolute reliability estimates as a function of number of items and number of raters, FA dimension

Figure 7 shows the same plots for the SCI dimension: relative reliability estimates as a function of number of items and raters in the left-hand panel and absolute reliability estimates in the right-hand panel. Because there was more variance attributable to raters and rater interactions on this dimension (consistent with the slightly lower rater agreement on SCI compared to FA), a larger increase in relative reliability (compared to the FA dimension) is realized when scores are averaged over additional raters, though once again there are clearly diminishing returns. As with the FA dimension, the largest increases in SCI relative reliability can be realized by adding items to the instrument. For five items and three raters, relative reliability is estimated to be 0.48, which is lower than the value for Cronbach's alpha for these same scores (0.51). By increasing the number of items to ten and keeping the number of raters at three, relative reliability could theoretically be increased to about 0.64.

The right-hand panel of Figure 7 shows the absolute reliability estimates for the SCI dimension. Comparing this to left-hand panel of Figure 7 again shows that the estimates of absolute reliability are lower than those for relative reliability. For five items and three raters, absolute reliability for SCI scores is 0.30. And comparing the right-hand panel of Figure 7 to the right-hand panel of Figure 6 shows that the absolute reliability estimates for SCI are *much lower* than those observed for the FA dimension. This is because of the larger item and rater variance component for SCI, and indicates that it would be very difficult to make any absolute decisions on the basis of SCI scores, a finding that is further supported in the discussion of standard error of measurement (below).



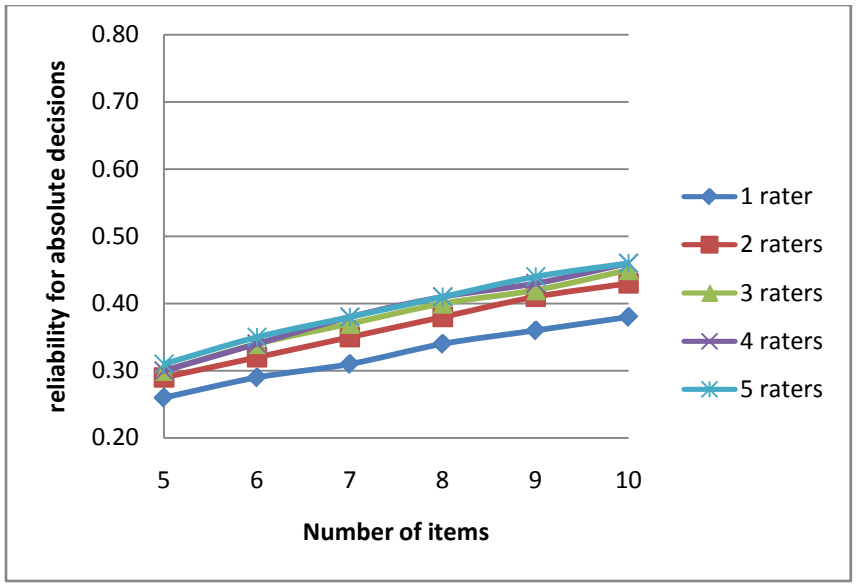


Figure 7. Relative and absolute reliability estimates as a function of number of items and number of raters, SCI dimension

It should be noted that in these analyses, it is assumed that all items in the universe of generalization have been well-specified and as such are of similar quality. Therefore if a poorly discriminating item has been developed and is one of the items sampled from the universe when increasing the number of items, then the expected gain in reliability will not be realized. In other words, adding poor quality items will not help. I estimated item discrimination by calculating the Pearson correlation between the average item score for an individual (across all raters) and their total average score (cf., Crocker & Algina, 1986). SCI item discriminations were 0.63, 0.61, 0.42, 0.66, and 0.65. FA item discriminations were 0.74, 0.59, 0.63, 0.49, and 0.65. Note that for each dimension, there was one relatively poorly discriminating item (item three on SCI and item four on FA).

Increasing the number of items and/or raters on the instrument also comes at a cost of respondent time and/or rater time and money to pay raters. Such changes would need to be evaluated relative to available resources, proposed uses of the instrument and scores, and other factors. For example, the current respondent time burden for a five item set is about 35 minutes. Several respondents noted that they became frustrated with the repetitive item prompts from one scenario to the next (e.g., *"I didn't know in the last question and I don't know now"* ID 1977205, in response to item 5). Based on this observation, it is reasonable to believe that respondent fatigue could play a role when increasing the number of items. If this were the case, then doubling the number of items is not likely to result in these theoretical increases in reliability.

In order to examine the impact of increasing the number of items on measurement error, I plotted the standard errors of measurement for absolute decisions (SEM Δ) and that for relative decisions (SEM δ) as a function of number of items. Figure 8 shows this plot for the FA dimension, and Figure 9 for the SCI dimension. In both plots, the number of raters is held constant at three (though these values are very similar for two raters).

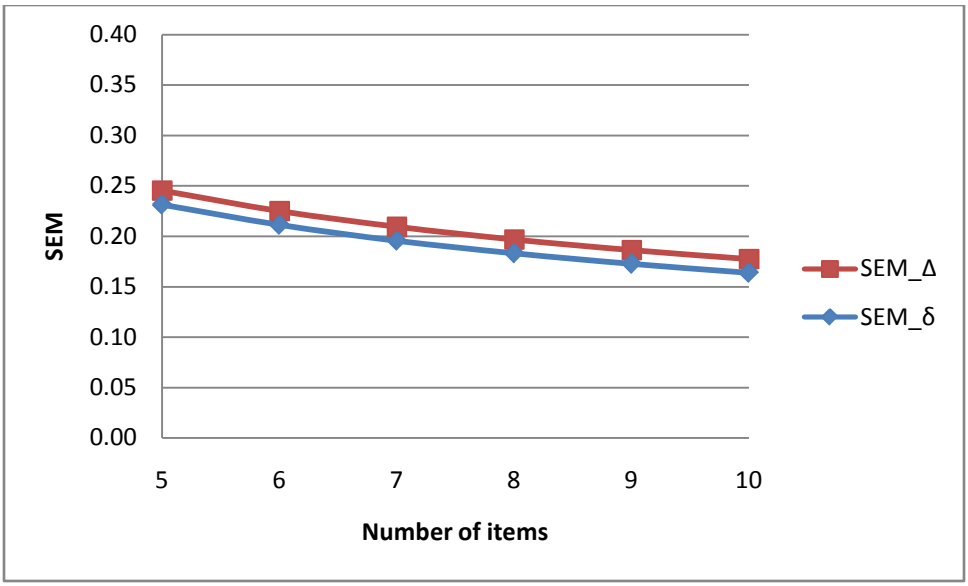


Figure 8. Plot of standard errors of measurement for absolute (Δ) and relative (δ) decisions, FA dimension

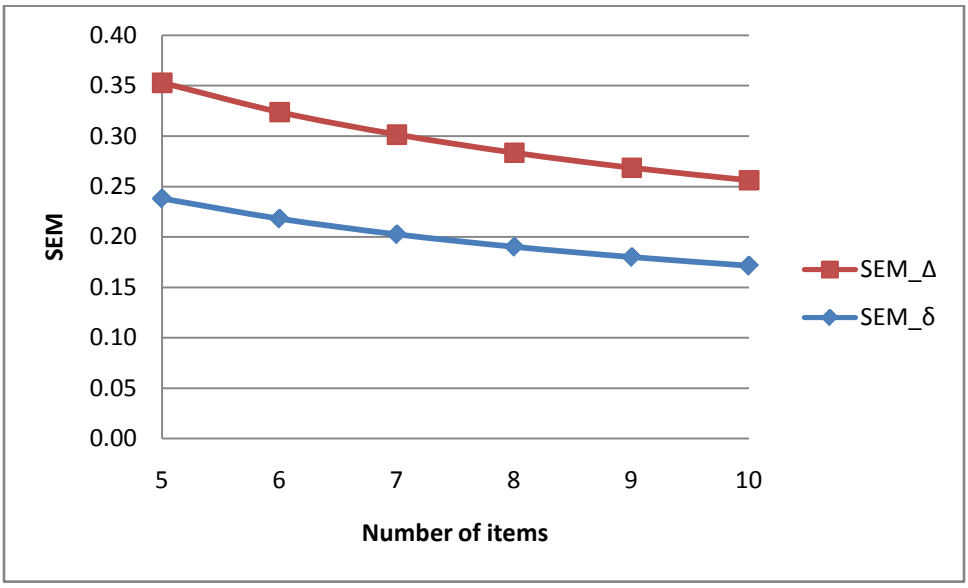


Figure 9. Plot of standard errors of measurement for absolute (Δ) and relative (δ) decisions, SCI dimension

In each of these plots, one can see that increasing the number of items decreases the standard error of measurement for both absolute (criterion-referenced) and relative (norm-referenced) decisions.

This is consistent with the sampling framework conception: as sample size increases, SEM decreases. Because the SEM is inversely related to the square root of the sample size, increasing the sample size by a factor of four will decrease the SEM by a factor of two. Therefore in the present context, I would expect doubling the number of items in the sample of items to decrease the SEM by a factor of $\sqrt{2} = 1.41$.

For relative decisions, the standard error of measurement (SEM δ) as a function of number of items is very similar for FA and SCI. This is because for both dimensions, a large amount of the variance was observed to be in the person by item interaction, and very little was in the person by rater interaction. This statistic is based on the variance components which involve the object of measurement (persons) as discussed above.

The standard error of measurement for making absolute decisions (SEM Δ) is quite different between the FA and SCI dimensions. For the SCI dimension, there is a larger standard error of measurement (by about 30%) for making absolute decisions as compared to that for the FA dimension. This is due to the fact that this statistic is based on the variance components which involve the facets of measurement, and for the SCI dimension a large amount of the variance was observed to be in the items themselves (which was not the case for the FA dimension). Some SCI items were either very easy or very hard for respondents. Therefore on the SCI dimension it would be particularly difficult to make absolute (criterion-referenced) decisions about respondent knowledge with a small number of items, an observation that was made above when examining the absolute reliability for this dimension (Figure 7). For example, the SEM Δ for five items and three raters on the SCI dimension is about 0.35. This means that a 95% confidence interval around a respondent's SCI score would be roughly +/- 0.70. This is a very large range with respect to the observed range of SCI scores, which is from 0 to 1.67 (see Table 5). Even if the number of items was doubled to ten, the SEM Δ is still about 0.25 (a decrease by a factor of $\sqrt{2}$), and the 95% confidence interval would be about +/- 0.50. Such broad confidence intervals would make

it difficult to claim that the SCI score observed for some respondent had not occurred by chance. In other words, a very different SCI score would likely be observed for that same individual had they responded to a different set of SCI items.

Table 5.

Range, mean, and SD of observed FA and SCI scores

	FA	SCI
minimum	0	0
maximum	1.44	1.67
mean	0.64	0.67
SD	0.37	0.36

To further illustrate the magnitude of SEM Δ for SCI scores, consider Figure 10 which shows all 60 mean SCI scores plotted with error bars representing a 95% confidence interval of +/- 0.70 (based on the SEM Δ for five items and three raters of 0.35). If a score of 1 on SCI (circled in Figure 10) represents the middle level of that construct (the teacher views the situation as an opportunity for interactive teaching and learning, but does not articulate a rationale why it is viewed as such), could one say with confidence that someone with a mean SCI score of 1 was emblematic of that level on the construct? With a 95% confidence interval of +/-0.70 around that mean score of 1, such a claim could not be made. The 95% confidence interval around this score ranges from 0.30 to 1.70. However, for lower-stakes decisions, a more relaxed confidence interval (e.g., 68% corresponding to +/- 1 SEM) could be used which would make it easier to distinguish respondents.

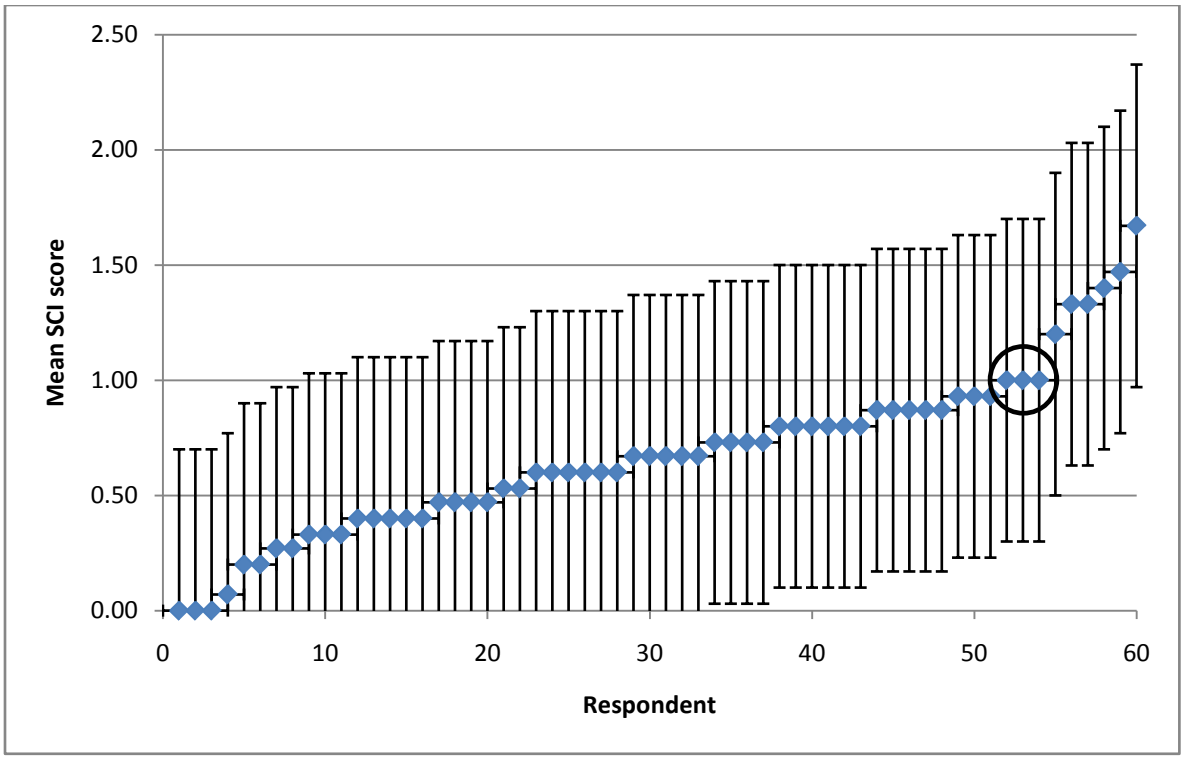


Figure 10. Mean SCI scores from all respondents with error bars representing a 95% confidence interval of +/-0.70 (SEM Δ = 0.35)

The SEM δ (for relative decisions) is very similar for both FA and SCI, as stated above. Each is about 0.23 for five items and three raters, meaning that a 95% confidence interval around a respondent's score would be +/- 0.46. As an example, consider the range of this 95% confidence interval for five items and three raters (0.92) around a particular observed SCI score. The range of observed SCI scores is 0 to 1.67 (Table 5). Therefore I can distinguish about two groups (0.92/1.67 = 0.55) on SCI with this confidence interval. By doubling the number of items to ten, SEM δ decreases to about 0.17 for each dimension, meaning that a 95% confidence interval now spans +/- 0.34. However, even if this decrease in SEM δ were realized, I could still not distinguish three groups on SCI (0.68/1.67 = 0.41). This lower SEM for relative decisions is because score precision does not need to be quite as high as that for making absolute decisions.

As a further example, consider the plot of all 60 mean SCI scores surrounded by a 95% confidence interval based on SEM δ = 0.23 (Figure 11). Many respondents in the top quartile of SCI scores (which ranges from 1.11 to 1.67) overlap with many respondents in the middle and bottom quartiles. In other words, there is not enough precision in measurement to distinguish each of these quartiles and therefore to make relative claims about the SCI scores of most respondents. This plot looks much the same for FA (since SEM δ is the same for that dimension; see Figure 12) but relative comparisons are even more tenuous on that dimension due to the smaller range of observed scores (0 to 1.44). On the FA dimension, the bottom end of the top quartile (which ranges from 0.96 to 1.44) overlaps with many scores in the bottom quartile (which ranges from 0 to 0.48).

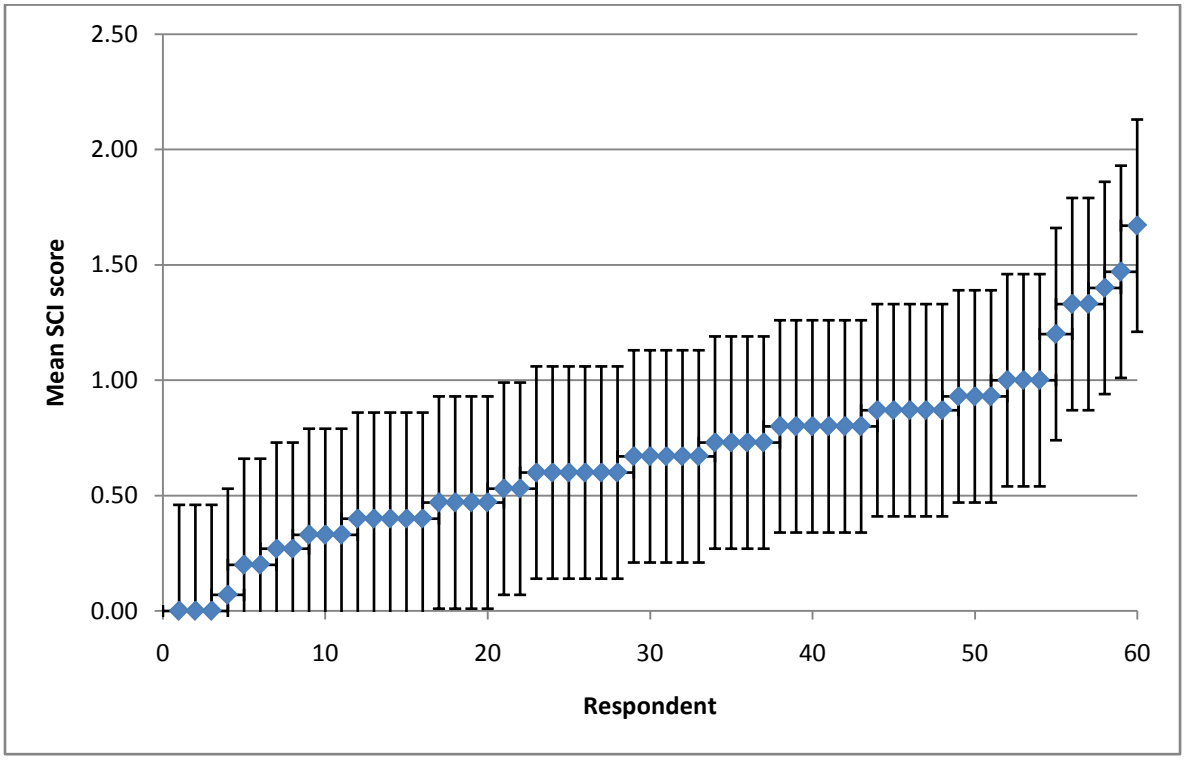


Figure 11. Mean SCI scores from all respondents with error bars representing a 95% confidence interval of +/-0.46 (SEM δ = 0.23)

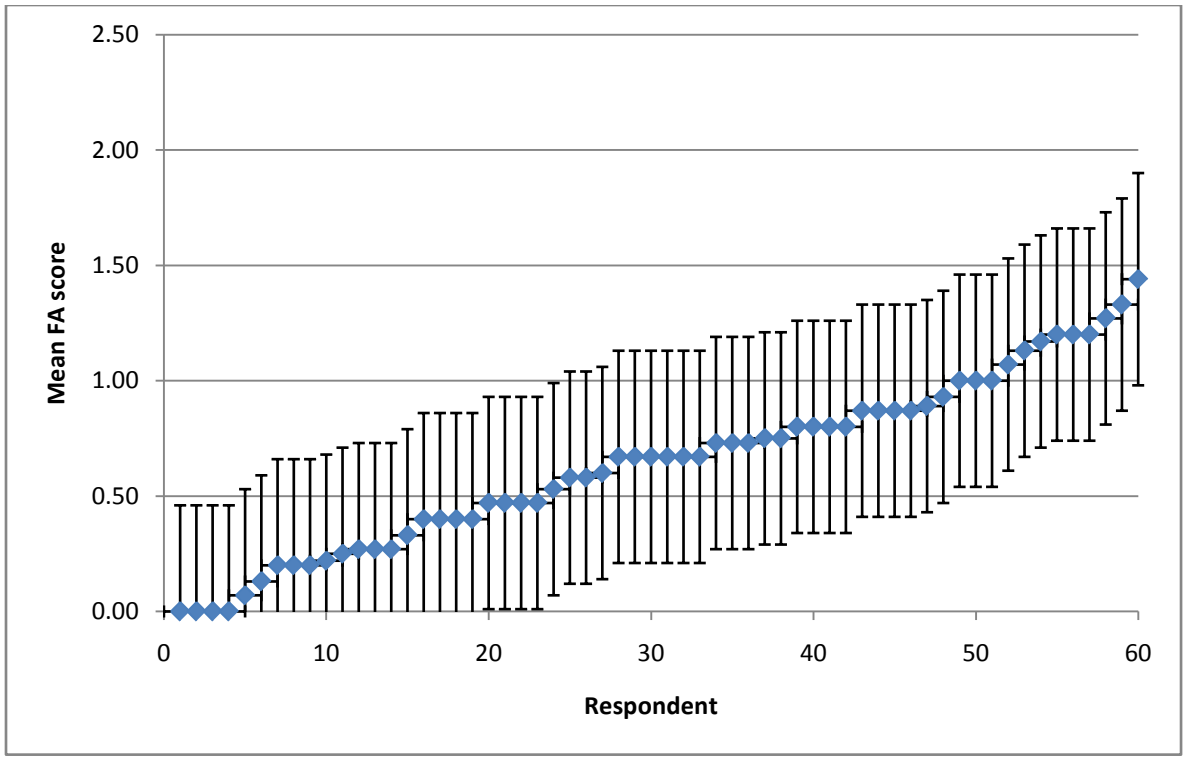


Figure 12. Mean FA scores from all respondents with error bars representing a 95% confidence interval of +/-0.46 (SEM δ = 0.23)

Discussion

Score reliability is often thought of in terms of the agreement between raters on scores of open-ended responses or observable actions. In some instrument development projects, high enough rater agreement might constitute "reliability." The RTOP (Sawada, et al., 2002) is an example of this. Because previous checks of agreement between my scores and those of another rater were relatively low, three new raters were trained and these individuals re-scored the FASCI responses. Scoring agreement between these new raters was observed to be much better than in previous efforts with the FASCI. But this agreement provides only a very coarse picture of score reliability.

As a first step into a more critical examination, I also examined score reliability in terms of the ratio of true score variance to observed score variance. The starting point for this examination was

Cronbach's alpha, which only considers one source of measurement error. Reliability as characterized by Cronbach's alpha was observed to be far below the target value of 0.80. A deeper examination of score reliability is realized by taking a G Theory approach, in which error variance can be decomposed into multiple sources. Further, using G Theory, score reliability for both relative and absolute decisions can be characterized.

In the D studies, I was able to estimate how score reliability could potentially increase if different measurement procedures were specified (i.e., different combinations of items and/or raters). Based on the G study finding that much of the variance in observed scores was found in the items and person by item interaction, it is clear that the number of items included on the instrument is the driving force in reliability. This was found to be the case for both dimensions. Adding more raters will not increase reliability estimates appreciably, but adding items to the instrument will. If the number of items on the instrument was doubled (from five to ten), relative reliability estimates increase to about 0.71 on the FA dimension and 0.64 on the SCI dimension. However, these values are still rather low compared to the target value of 0.80 presented above. Absolute reliability estimates were lower than the relative reliability estimates, because they include both the facets and facet interactions as potential sources of error. These findings raise a serious issue about the reliability (both relative and absolute) of SK scores from the FASCI. Even doubling the number of items would not yield high estimates of score reliability, assuming that these increases could be realized. As mentioned above, this increase in number of items comes at the cost of respondent time burden. Average time for respondents to complete the five-item version of the FASCI was around 35 minutes. Requiring 70 minutes of respondent time may be asking too much of their participation, given that they will likely become fatigued and response quality will decrease.

In order to further illustrate these issues with this lack of precision in measurement, I estimated the standard errors of measurement for both relative and absolute decisions as a function of number of

items. The SEM for relative decisions (SEM δ) is similar for FA and SCI, and decreases as the number of items increases. For both dimensions, it would be difficult to make relative decisions (e.g., comparing the top and bottom quartiles) with the current number of items and raters. The SEM for absolute decisions (SEM Δ) is much higher for the SCI dimension due to the fact that SCI items were either very easy or very hard for respondents (i.e., they were not equally discriminating). It would be very difficult to defend an absolute (criterion-referenced) decision for an individual on the SCI dimension. This finding is also reflected in the estimates of the absolute reliability for the SCI dimension.

In the next chapter, I will further discuss score reliability disaggregated by FASCI version (neutral and physics). The main purpose of that chapter will be to investigate proposition six: when specific science content is added to the FASCI scenarios, SK score interpretations change. In evaluating evidence for this proposition, I will compare mean scores from each version, examine qualitative item responses, and present some results from think-aloud interviews with a subsample of respondents.

Chapter 5: The Content Test

Introduction

Proposition five of the FASCI validity argument states that when specific science content is added to the FASCI scenarios, score interpretations change. I would expect this to be the case for physics experts on the physics version of the FASCI, because this instrument would be better able to access their subject specific, more sophisticated SK which is based on their content area. This hypothesis was discussed in chapter two and is based in part on the construct of pedagogical content knowledge (Shulman, 1986). As discussed in chapter three, my method for gathering evidence to evaluate this proposition is the *content test*. In this study, I compare the scores from two versions of the FASCI instrument, the n-FASCI and the p-FASCI. The n-FASCI is "content neutral" in that the items do not reference any specific science content, while the p-FASCI specifically places those same scenario-based items within the content of physics. The p-FASCI also includes other context meant to further embed physics content into the item context, including physics learning objectives and physics content questions. I expect that the addition of this information on the p-FASCI will lead to the elicitation of a more sophisticated SK for physics experts, and should manifest itself in higher scores for these individuals on the p-FASCI.

In order to put this proposition to the test, I compare mean FA and SCI scores from each version of the instrument. The scores used in these analyses are the average item scores across all three newly trained raters. I compare scores between instrument versions in two different ways in order to examine the potential effect of missing data. As well as comparing the mean scores from each dimension and version with tests of statistical significance, I express these differences in terms of effect size units and discuss the statistical power of these tests of significance.

To specifically investigate the hypothesized differences in scores between versions, I compare scores by the physics expertise of respondents. In these analyses, I compare mean FA or SCI scores

between versions for those who are characterized as physics experts based on one of three classifications: 1) subject they plan on teaching, 2) physics content knowledge scores, and 3) number of physics courses taken. The last part of these quantitative analyses involves an examination of score reliability between each version, and item difficulties by dimension and version. If the addition of specific science content does elicit this more sophisticated SK, then one might expect the reliability of those scores to be different than that of the scores from the neutral version of the FASCI. Also, one would expect item difficulty to vary by version.

I also present a qualitative analysis of item responses to prompt a) of each scenario, upon which the SCI scores are based. And finally, I discuss some of the respondent comments from think-aloud interviews to further examine any difference in FA scores between versions. These two data sources are directly related to the response process.

The main finding that results from these analyses is that SCI responses differ in a qualitatively distinct way between versions, and that the quantitative differences are large and statistically significant. However, this difference in scores is not attributable to the accessing of a more sophisticated knowledge base by the p-FASCI. Rather, the p-FASCI seems to be eliciting construct irrelevant responses on the SCI dimension. Differences in FA scores are slight, and evidence from response processes is mixed, suggesting that these responses are actually quite similar between versions. It seems that one of the determining factors for FA scores for the novice pre-service teachers in this sample may be their lack of teaching experience.

Comparing Scores from the n- and p-FASCI

I compare scores on each version and administration of the FASCI in two ways: 1) mean 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). I conduct the second analysis in order to check the sensitivity of the first score comparisons to the existence of missing data. In comparing results from the two analyses, I found them to be quite similar. However, the first method has the advantage of using all response sets and will therefore yield higher statistical power. As discussed in chapter three, respondents were randomly assigned to take one version or the other.

Table 1 shows the mean scores for each dimension (FA and SCI) by version[3]. 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 1.

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

Dimension	n-FASCI sample size	n-FASCI	p-FASCI sample size	p-FASCI	Difference between p-FASCI and n-FASCI	p-value from t-test
FA	26	0.57 (0.42)	34	0.69 (0.32)	0.12	0.19
SCI	26	0.78 (0.36)	34	0.58 (0.34)	-0.20	*0.03

*significant at p<0.05

The distribution of the mean FA and SCI scores on the n- and p-FASCI are shown in Figure 1. 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 1, 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.

[3] 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, I conducted an analysis where I replaced missing scores with the mode score for that particular item and compared the means calculated 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 I conclude that the averaged mean scores presented in Table 1 are not biased due to differences in the difficulty of items for which responses are missing.

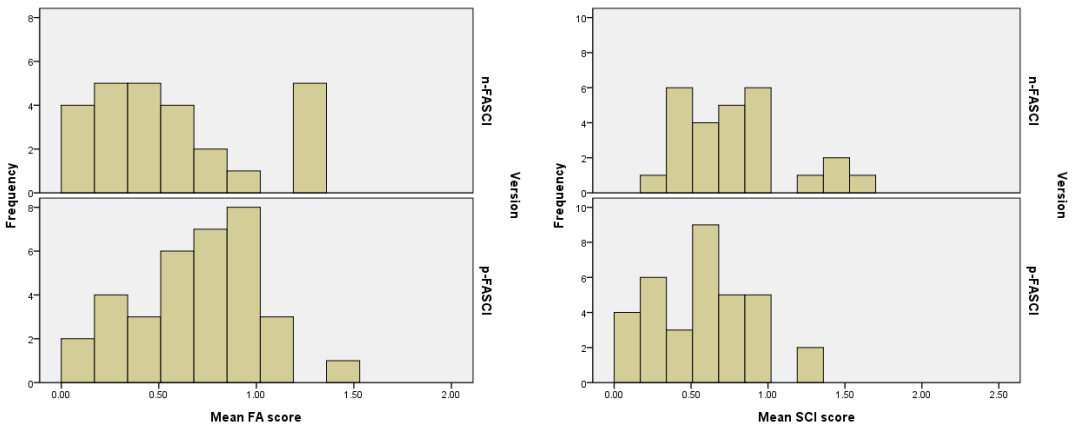


Figure 1. Distribution of mean FA and SCI scores averaged by number of responses for n- and p-FASCI

There were a number of incomplete response sets on each version. It is reasonable to think that this missing data would affect the mean score comparisons presented in Table 1 above, where scores were averaged based on the number of items answered by each respondent. As a check of the sensitivity of the above mean score comparisons to this missing data, I compare the mean scores for each dimension by version for those individuals who had complete response sets across both dimensions (no missing data). This data is presented in Table 2. One shortcoming of this approach is that it ignores partial response sets, thereby reducing sample size. Again, scores on the p-FASCI were higher on FA but lower on SCI and the difference was statistically significant for SCI. The magnitude of these differences is very similar to those in Table 1.

Table 2.

Mean FA and SCI scores (SD) by version for complete response sets only

Dimension	n-FASCI sample size	n-FASCI	p-FASCI sample size	p-FASCI	Difference between p-FASCI and n-FASCI	p-value from t-test
FA	24	0.53 (0.40)	23	0.68 (0.30)	0.15	0.14
SCI	25	0.75 (0.31)	27	0.53 (0.31)	-0.22	*0.02

*significant at p<0.05

In short, it appears that these score comparisons yield the same findings as those presented in Table 1. Therefore I conclude that the mean score comparisons in Table 1 are not significantly biased by the averaging method, and take advantage of a larger number of responses in each case. The remainder of the analysis is conducted with reference to the mean scores presented in Table 1.

Statistical Power and Effect Size.

Statistical power was calculated *post hoc* for the tests of significance between mean score comparisons (Table 1). 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. This is the *inverse* of a Type II error (failing to reject the null hypothesis when it is false). Therefore a higher value for statistical power indicates a higher probability of rejecting the null hypothesis when it is false. For all power calculations, alpha (the probability of falsely accepting the alternative hypothesis when the null hypothesis is true) was set at 0.05. 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 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 differences in mean scores can also be expressed in terms of an effect size (in this case, as Cohen's d). For this analysis, I express change in terms of effect size units (ES), calculated as shown in 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.

Examination of Scores 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, in accordance with the hypothesized difference. I classified respondents 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 (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 classifications are shown in Table 3. 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 3.

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 (t-test)	SCI mean (SD)	p-value (t-test)
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 otherwise) 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 (about 0.20 in favor of the n-FASCI, see Tables 1 and 2) 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, in the next section I compare score reliabilities. After that, I will discuss the differences in the difficulty of items between each version of the FASCI instrument.

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 4. These reliability estimates (reported as both generalizability (ρ^2) and dependability (Φ) coefficients) are based on the average item and total scores across all newly trained raters. Note again that the reliability estimates which characterize absolute error (Φ) are lower/more critical than those which characterize relative error (ρ^2). These are much lower for the SCI dimension, again highlighting the finding from the previous chapter that making absolute decisions on the basis of SCI scores is not warranted. The reliability of SCI scores derived from the n-FASCI is higher (by 0.05 to 0.09) than that found in previous pilot testing, and reliability of FA scores is within the range of that found in previous pilot testing (almost equivalent to the reliability of FA scores from FASCI pilot Test 1). The SD of these observed scores is actually much lower than that from previous pilot testing, but this should be interpreted with caution as scores from those pilot tests were based on one rater (me) while these scores are based on those from three raters. Because these latter scores are averaged over three raters, there is less variability.

Table 4.

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

Version	Sample size	FA reliability		SD of Observed FA score	% FA missing	SCI reliability		SD of Observed SCI score	% SCI missing
		ρ^2	Φ			ρ^2	Φ		
n-FASCI	26	0.67	0.65	0.42	7.7	0.50	0.27	0.35	3.8
p-FASCI	34	0.37	0.34	0.32	32.4	0.37	0.24	0.34	20.6

As seen in Table 4, p-FASCI score reliability is much lower than that for the n-FASCI. Two findings discussed below help to explain this large difference: 1) a larger percentage of missing response data on the p-FASCI (discussed in the next section), and 2) the elicitation of construct irrelevant responses on the p-FASCI (discussed further below).

Missing Data.

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. In order to examine this issue further, I looked at the percentage of incomplete response sets on each version by university, which is shown in Table 5.

Table 5.

Percentage of incomplete response sets by university and version

University	n-FASCI	p-FASCI
Southeast Coastal University (SCU)	21%	45%
Northwest Pacific University (NPU)	0%	50%
Northeast Queen's University (NQU)	N/A	38%
Western State University (WSU)	0%	27%
Central Research University (CRU)	0%	0%

In examining the percentage of incomplete response sets by university, I observed that for universities which had respondents taking both versions of the survey, there were at least twice as many incomplete response sets to the p-FASCI as there were on the n-FASCI. 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 NQU sample, for which there was a very large percentage of incomplete response sets (38%). Respondents from this university mostly describe themselves as "physics people", 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.

Item Difficulties.

Based on the observation that scores differ between versions (significantly for SCI), it is reasonable to think that the items were easier or more difficult for respondents on one version or the other. For example, since SCI scores were significantly lower on the p-FASCI, one would expect the SCI items to be more difficult on that version compared to the n-FASCI. To see if the observations of score differences are supported in this way, I calculated classical item difficulties for each dimension by FASCI version. As with the above analyses, scores for this analysis are based on those from the three newly trained raters. This item difficulty (also known as a p-value, not to be confused with a p-value from a statistical test of significance) is the proportion correct on an item, with higher values indicating an easier item. For polytomous items (such as FA and SCI), the p-value is calculated as the proportion of possible points awarded.

The item difficulties were calculated based on the all completed item responses (not including incomplete responses) and are shown in Table 6. Differences in item difficulties (the right-most column

[4] I replaced missing data on the p-FASCI with mean scores for each rater-item combination and computed reliability estimates (as generalizability coefficients) again. Doing so changes FA score reliability only slightly (from 0.37 to 0.38) and SCI score reliability changes from 0.37 to 0.45.

in Table 6) are expressed as p-FASCI item difficulty minus n-FASCI item difficulty. Therefore, positive values indicate that the item was easier on the p-FASCI, while negative values indicate the opposite.

Table 6.

Item Difficulties calculated based on all item responses

Dimension	Item	n-FASCI	p-FASCI	Difference (p-n)
FA	1	.35	.42	.07
	2	.21	.26	.05
	3	.33	.39	.06
	4	.19	.28	.09
	5	.26	.38	.12
SCI	1	.83	.62	-.21
	2	.14	.04	-.10
	3	.10	.06	-.04
	4	.21	.15	-.06
	5	.59	.54	-.05

Scenario 1: "Students are working in groups of four to discuss a conceptual problem"
Scenario 2: "You are working out an example problem up on the board"
Scenario 3: "You having just finished giving a presentation on a complicated topic"
Scenario 4: "You have given your students a quiz to assess their understanding of a difficult topic"
Scenario 5: "You are talking one on one with a student who has a misconception"

As can be seen in Table 6, FA items were consistently less difficult and SCI items were consistently more difficult on the p-FASCI. The magnitude of many of these differences is slight. For example, a value of 0.05 means that there was a 5% difference in the number of respondents on each version who achieved the maximum score on that item. Differences that are this small to indicate comparable item difficulty between versions, especially given the relatively small sample sizes. Further, the magnitude of the average difference in item difficulties is similar for the SCI dimension (-0.09) and the FA dimension (0.08).

There are two notable exceptions to this observation of similarity between versions: item 5 on the FA dimension and item 1 on the SCI dimension. FA item 5 shows a difference in item difficulty between versions of 0.12, still not a large value but greater than the other differences. In previous pilot testing, SCI item 1 has consistently been the easiest item, with p-values in the .80-.95 range. In this

study, SCI item 1 was more difficult for p-FASCI respondents (p-value = 0.62). Taken together with the observations that all SCI items were more difficult on the p-FASCI, and that the magnitude of the differences in mean SCI scores between versions (about 0.20) was greater than that for FA, the SCI item responses deserve some deeper qualitative investigation.

Evidence from the Response Process

In this section, I will present and discuss two sources of data that bear on the response process: 1) the open-ended responses to prompt a) of each scenario upon which SCI scores are based, and 2) respondent comments from think-aloud interviews, which bear mostly on FA scores.

Qualitative Analysis of SCI Responses.

To further investigate the significant difference in SCI scores and the observed difference in SCI item difficulty between the n- and p-FASCI, I examined the prompt a) item responses from each version. While qualitatively examining these responses, I found that many p-FASCI respondents discussed the content provided in the scenario, rather than discussing the students. I interpret this finding 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.

I coded each prompt a) response from both versions as discussing either the content or the students. Recall that prompt a) was designed to elicit from respondents a discussion of the students so that a score for SCI could be given. Table 7 shows a comparison of the prompt a) item responses on the n- and p-FASCI at both administrations. 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%

My 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 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 to 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 seems that this individual sees 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.

Both of these responses are examples of discussing the content rather than the students, and were also given a score of 0 for SCI.

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) seems to be 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 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 item difficulties on each version of the FASCI. All of this evidence (from score, reliability, item difficulty, and response comparisons) supports the proposition at hand for the SCI dimension: when specific science content is added to the scenarios, score interpretations change. But the change is *not* observed only for physics experts, and seems to be due to the elicitation of construct irrelevant responses.

Think-Aloud Interviews.

Three individuals who took the n-FASCI and three who took the p-FASCI were interviewed after both the pre and post-test administrations[5]. In these think-aloud interviews, respondents mostly discussed their responses with respect to the FA dimension, therefore an examination of the specific comments can help to shed light on any differences in FA scores between the n- and p-FASCI. Recall that FA scores were slightly higher on the p-FASCI (but this difference was not statistically significant) and FA items were consistently easier on the p-FASCI (though these differences in difficulty were small). Also, FA score reliability was lower on the p-FASCI (0.37) than it was on the n-FASCI (0.67). Much like the quantitative findings, results from the think-aloud interviews is not clear cut in supporting any differences between versions. There appears to be no clear difference between the articulated response processes of n- and p-FASCI respondents. This is interesting because I would have expected to

[5] As discussed in chapter 3, although the n- and p-FASCI instruments were administered pre and post-semester's instruction, only the pre-test scores are used in quantitative analyses. However, I draw on both sets of interviews (after pre and after post-test) in discussing the responses process.

observe a similar elicitation of construct irrelevant information, much like that observed for the SCI dimension.

For example, in the post-test interview with Drew about his responses to the p-FASCI, he indicates that his lack of experience in teaching physics makes it difficult for him to think of what he would do next (prompt c)) in item five ("you are talking one-on-one with a student who has a misconception"):

> [Drew]: *I don't know. Because, like, I don't really know a lot about the whole physics of it. That was kind of a tough one, I thought.*
>
> [Interviewer]: *Because of the physics, or because of the situation?*
>
> [Drew]: *Because of the situation.*
>
> [Interviewer]: *What about the situation do you think made it hard? The one-on-one aspect?*
>
> [Drew]: *Yeah, because, like, it didn't work, what are you gonna do? And then it didn't work, so what are you gonna do? I don't really have too much real-life experience in teaching physics, so it's hard for me to say what I would really do.*

Drew states that it was hard for him to know what he would do next (if the approach he tried didn't work) because of his lack of experience teaching physics.

Similarly, Liz (and n-FASCI respondent) brings up the idea that her ability to cite relevant contextual factors in making strategic decisions is dependent on her experience as a teacher:

> *[if I had more experience] I feel like I would have a lot more examples of actual students and things that I've actually tried. So then I might be more likely for, "if that didn't work, what would*

> *you do", for some of my students I would actually be able to draw on that personal experience to say, "I tried this with some students and it worked well, but it didn't work with these students, so I did this with these students."*

Liz saw the structure of the FASCI prompts as being able to "distinguish who has the experience" from those who do not. Like Drew, it was her lack of experience teaching that limited her ability to articulate what she would do next. Perhaps in both cases (n- and p-FASCI) lack of teaching experience is the main factor in determining how novice pre-service teachers respond to prompts b) and c), and therefore determines their FA score.

This lack of classroom teaching experience can be seen in other interviewee comments. For example, Julie (a p-FASCI respondent) indicated a certain amount of anxiety created by an unfamiliar teaching context:

> *A lot of the ones where it says, "If the approach you described doesn't work, what would you do in the next class session?" for me, I think that was really throwing, because right now, with the [university-level] recitations I have a lot of flexibility in what I do, but I know when I'm a high school teacher I'm gonna be worried about getting through all the curriculum, and unfortunately I have to teach to the CSAP[6] tests and whatnot, so depending on what I'm teaching, I was thinking, gosh*

Julie scored highly on the physics content knowledge measure, had high mean FA score, but she was still "thrown" by prompt c): "If the approach you described doesn't work, what would you do in the next class session?" Further, she was worried that depending on the situation, she might "run out of

[6] Colorado Student Assessment Program (CSAP)

ideas." Again, for her it seems that lack of experience (not necessarily the physics content) was a determining factor in her responses, or at least it was a part of her thinking and response process.

In summary, results from think-aloud interviews are mixed and do not seem to indicate that the physics content had as much of an effect on the FA response process as did lack of teaching experience, at least for these novice pre-service teachers. As mentioned above, I would have expected to observe a difference in responses between versions based on elicitation of construct irrelevant information (in this case, physics knowledge).

Discussion

In evaluating proposition five (when specific science content is added to the FASCI scenarios, score interpretations change), results from the *content test* are mixed. The main finding is that SCI responses differ both quantitatively and qualitatively between versions, and that the quantitative differences are statistically significant. But these differences are for the aggregate sample, not just for physics experts as I had hypothesized. And 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 I had originally hypothesized. Differences in FA scores are slight, and qualitative evidence from response processes is mixed, suggesting that the responses between versions are actually quite similar. This qualitative evidence also suggests that one of the determining factors for FA scores for the novice pre-service teachers in this sample may be their lack of teaching experience.

One particular interesting observation was the amount of missing response data on the p-FASCI relative to the n-FASCI. As discussed above, the physics content may have put off some respondents. This is based on a few comments from think-aloud respondents. A second explanation from this study relates to the NQU sample, which consisted of "physics people" who were frustrated with the structure of the instrument in general. In their case, they may have also stopped responding to the n-FASCI had

they taken that version. So it is unclear whether or not the physics content was the main factor driving the missingness, or if it was the structure of the instrument itself as appeared to be the case for the NQU sub-sample.

Based on these findings, it appears that the effect of embedding specific science content into the scenarios is to detract from a novice teacher's ability to see a teaching situation as an opportunity for interactive teaching and learning. SCI scores were significantly lower on the p-FASCI than on the n-FASCI, and the SCI items were more difficult for p-FASCI respondents. Also, the magnitude of differences in these scores is rather large: effect size = 0.58. Qualitative analyses of item responses show that a larger percentage of p-FASCI respondents are discussing the content rather than the students in response to item prompt a): "How might this activity facilitate student learning?" which I interpret as a finding of construct irrelevant variance. Based on these findings, the proposition at hand is well-supported for the SCI dimension, but not for the reason that was hypothesized. That is to say, when specific science content is added to the scenarios, SCI score interpretations *do* change due to the elicitation of construct irrelevant responses rather than due to the accessing of a more sophisticated knowledge base. However, the proposition *does not* appear to be supported for the FA dimension. This finding is unexpected. I would have expected strategic choice to be heavily influenced by adding specific science content to the items, but this was not observed to be the case.

Although I interpret the difference in SCI scores as being attributable to the elicitation of construct irrelevant responses, this finding cannot be generalized beyond this sample of novice pre-service science and mathematics teachers. It could be that more experienced teachers or individuals who are more expert in physics would respond differently to the p-FASCI. Or it could be that this "more sophisticated" (perhaps subject specific) SK does not exist. Despite these possibilities, the observed difference was not due to the reason hypothesized, and therefore does not support the suggested empirical resolution of PCK as discussed in chapter two.

In the next chapter, I will compare scores from the FASCI instrument to scores from observations of teaching practice for a separate sample of beginning science and mathematics teachers. This evidence will be used to evaluate proposition three: SK can be observed in teaching practice. The evidence will also be evaluated to see if the observations scores can be used as a source of convergent validity evidence for SK scores from the FASCI instrument.

Chapter 6: Observing Strategic Knowledge in Practice

Introduction

In order to support the proposed score interpretation for the FASCI instrument (SK scores of novice science and mathematics teachers can be compared and distinguished), it is important that SK can be observed in teaching practice. This was presented as proposition three in the validity argument outline, and provides an important source of *convergent validity* evidence for SK scores. In this chapter, I evaluate that proposition by comparing SK scores from the neutral version of the FASCI instrument to scores from the Reformed Teaching Observation Protocol (RTOP; Sawada, et al., 2002) for a sample of novice science and mathematics teachers.

The Reformed Teaching Observation Protocol (RTOP)

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 (American Association for the Advancement of Science, 1990) 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).

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 which were conducted as part of a study comparing traditional and reformed teaching.

Reliability.

Rater reliability was established based on ratings from two observers on each of 16 math and physics lessons (32 observations total). Regressing total scores for one observer onto those of another, r-squared was found to be 0.954. R-squared describes the proportion of variability in the data that is

accounted for in regressing one set of scores on the other. It characterizes how well one rater's scores can be predicted by the other rater's scores. For each of the subscales, reliability between the two raters (again expressed as r-squared) was found to range from 0.670 to 0.946, with the mean being 0.862. Data from a second set of paired ratings (conducted in biology classrooms) also demonstrated similar rater agreement, with r-squared being 0.803 for total scores.

One limitation of characterizing scoring agreement with the r-squared statistic is that it only summarizes *relative* score position. It does not characterize *absolute* agreement in scoring, as do the percent agreement and Cohen's kappa statistics that I reported in chapter four for my study. For example, when a set of scores 1, 2, and 3 are regressed on a set of scores 2, 3, and 4, r-squared is 1. However, in this case the percent agreement is 0%. Also, having good rater agreement does not mean that score reliability will be high. For example, in chapter four I also discussed the fact that the rater agreement for FASCI responses was good, but that score reliability was still somewhat low. No other rater or score reliability information (e.g. percent agreement, Cronbach's alpha, etc) is provided in the RTOP reference manual (Piburn, et al., 2000).

Validity.

As presented in chapter two, evidence for the construct validity of the RTOP is presented in three forms: 1) correlations between RTOP total score and subscale scores, 2) correlation between student learning gains (as measured by a concept inventory, such as the Force Concept Inventory) and their instructor's RTOP scores, and 3) factor analyses based on the observations from 153 classrooms. For the first two analyses, correlations were found to be high. More relevant to the current analyses in my study are the results from the factor analyses. An initial principle component analysis 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 evaluators 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 3 items load at 0.50 or greater, 2 of which also cross-load on factor 1). Further, when using a cut-off value of 0.30 (rather than 0.50) for significance in factor loadings, the authors identify *five* factors rather than three. These five factors seem to represent the most meaningful statistical groupings of the items and each of these factors was operationalized and described by the authors. Therefore it is these factors (and their item groupings) that I use as a basis 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 principle 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 principle 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 PCK (and cite 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 principle 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 principle component analysis. This factor describes a classroom which triggers divergent thinking where the teacher encourages student exploration.

The items associated with each of these factors can be seen in the RTOP in Appendix E.

The RTOP is strongly based in the literature of reform 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 discussed above. So even though these two instruments (the RTOP and the FASCI) are designed to accomplish a similar task (characterize science teacher's knowledge of practice) they have undergone different development processes and go about the task in different ways. But because of the similarity between the RTOP and SK constructs, we should expect to see a positive correlation between scores on each instrument.

Sample

As discussed in chapter three, the sample for this analysis consists 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. 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 SK scores with RTOP scores for these individuals.

Each of these individuals responded to the n-FASCI and was observed at least two times during the spring semester of 2009. These observations were conducted by me and other members of the WSU-LA research team. The version of the FASCI to which they responded consisted of six scenario-based items, rather than five (as in the version of the n-FASCI used in the *content test* discussed in the previous chapter). This is the same version of the FASCI used in Pilot Test 2, which was referenced in the discussion about score reliability in chapter four.

Data Sources

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 (Appendix E) consists of 25 five point Likert scale items[7] 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 which use this instrument. As a sort of rule of thumb, 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. Below, I discuss the RTOP design and structure in more detail, and present the way in which I compare RTOP data to FASCI scores.

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 use scores on each of the five factors discussed above in this comparison. My goal in these comparisons is to identify cases where teaching characterizations based on each instrument were consistent (i.e., rated similarly on both instruments), or where they were inconsistent (i.e., rated dissimilarly). Once these cases are identified through the descriptive statistical analysis, I am 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

[7] Total possible score on the RTOP is 100, as the lowest category for rating on each item is zero.

factor, across all observations available for that individual[8]). Note that the FASCI scores used are based on only my ratings, as these responses were not scored by the newly trained raters. This is because the version taken by this sample of respondents was that used in pilot test two, which included six scenarios, only two of which were common with the version used in the other analyses. Correlations between these scores are shown in Table 1. In general, there is only one notable correlation between mean FA score and any of the RTOP factor scores, that being RTOP Factor 1 (Inquiry Orientation). However, in all cases there are stronger correlations between mean SCI score and RTOP factor scores than there are between mean FA and RTOP factor scores. Note that for all correlations, the sample size is small (n = 18).

Table 1.

Correlations between 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

*significant at $p < 0.05$
**n = 18 for all correlations

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 which do not fit these trends, I examine scatterplots from each of these relationships (Figures 1 through 6).

[8] I chose to average RTOP scores across all observations (either 1, 2, or 3 depending on the individual) because they were purposefully observed more than one time in order to account for the possible effects of observing an atypical lesson.

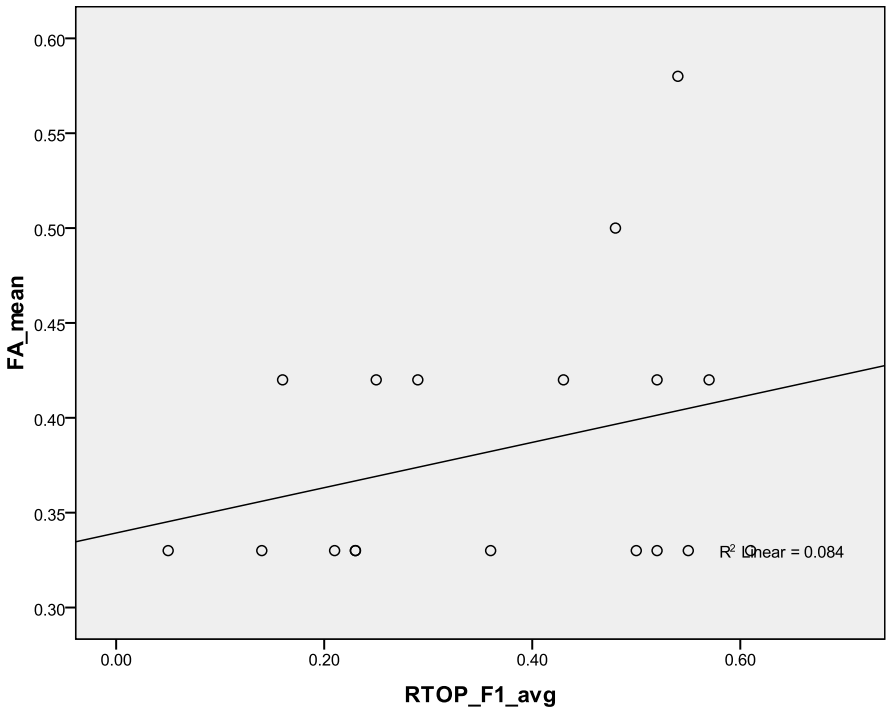


Figure 1. Mean FA score vs. RTOP Factor 1 (Inquiry Orientation) score

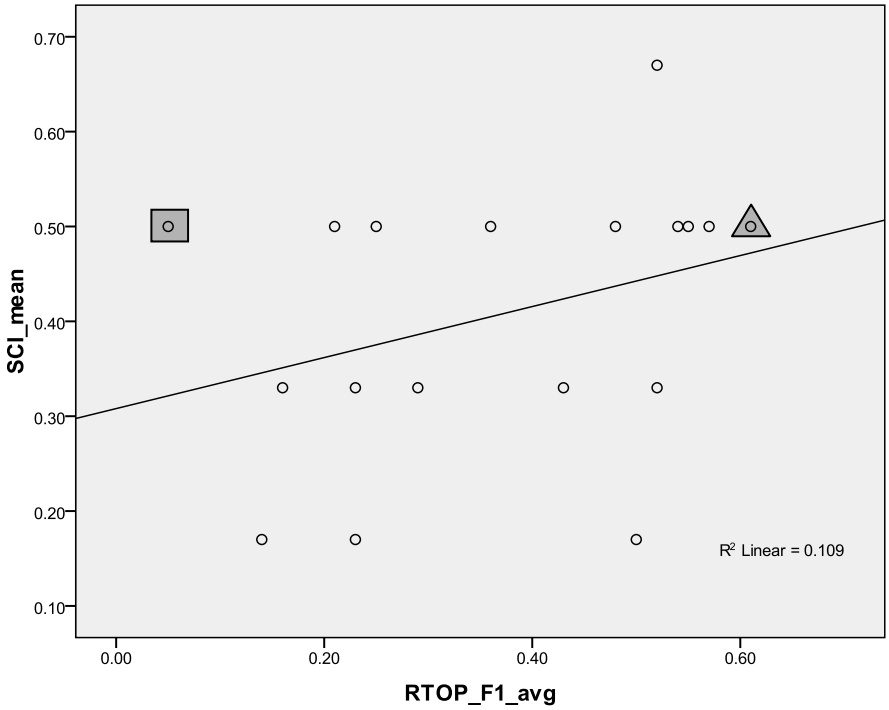


Figure 2. Mean SCI score vs. RTOP Factor 1 (Inquiry Orientation) score

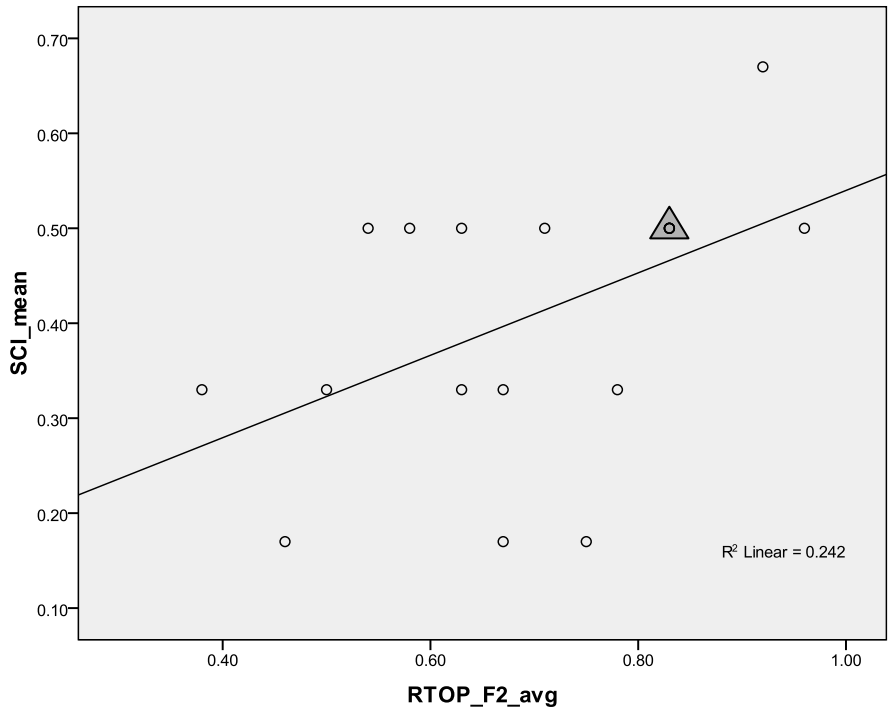


Figure 3. Mean SCI score vs. RTOP Factor 2 (Content Propositional Knowledge) score

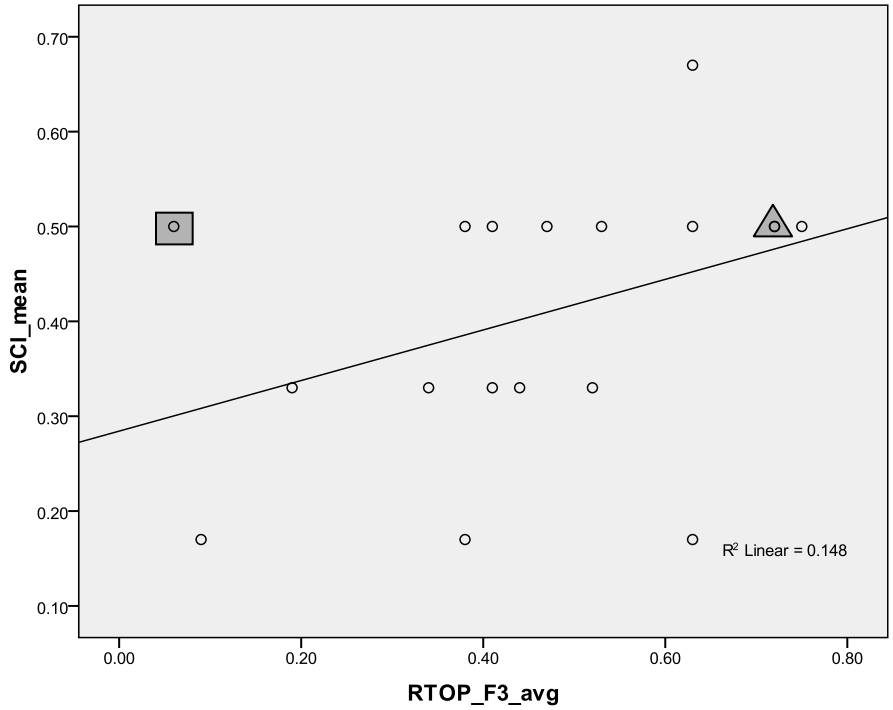


Figure 4. Mean SCI score vs. RTOP Factor 3 (Content Pedagogical Knowledge) score

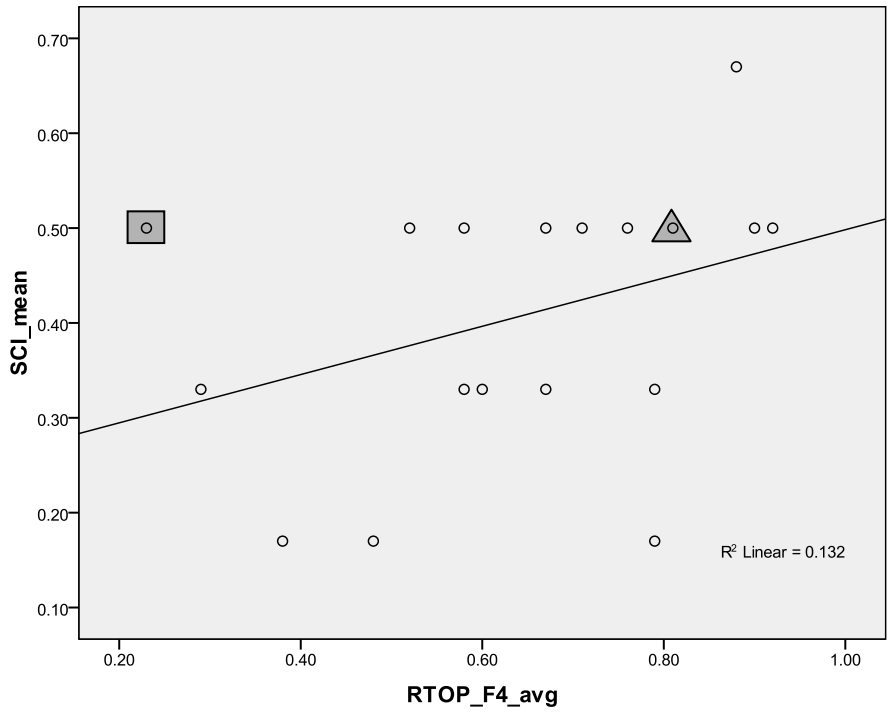


Figure 5. Mean SCI score vs. RTOP Factor 4 (Community of Learners) score

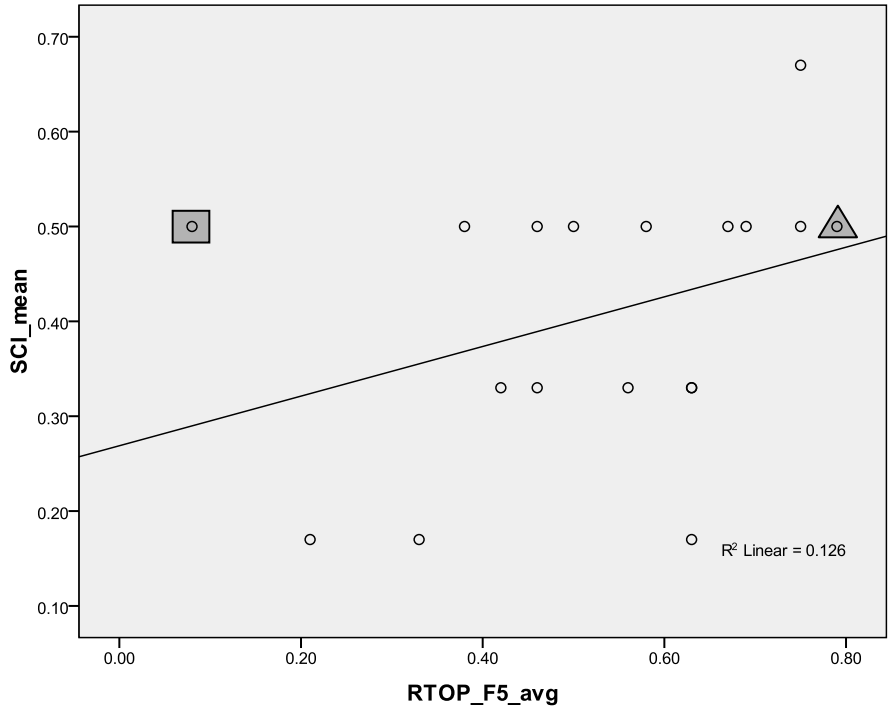


Figure 6. Mean SCI score vs. RTOP Factor 5 (Reformed Teaching) score

Although each of these scatterplots shows the general trend in relationship between each of the variables compared, they are more useful in identifying both outliers (cases which 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 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. Also, 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 than 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 frequently characterized

consistently (e.g., low on both FASCI and RTOP, or high on both). Because there were no "high" FA categorizations, for this comparison I consider the middle level for FA (1) to be "high." The pattern of these consistent/inconsistent characterizations is shown in Table 2. 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 2.

Table 2.

Comparison of FA and SCI Category Rating to RTOP Factor Category Rating, all individuals in 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	.	.

Cases Identified for Further Analysis

Individuals of particular interest for qualitative analysis are listed by their pseudonym in Table 2. Jason and Ellie are chosen because they always scored high on the FASCI and on each of the RTOP factors. Laura is chosen because she was consistently low on both measures. James is 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 scatterplots) is 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). FA, SCI, and RTOP factor scores (expressed in standard units) for each of the cases identified are shown in Table 3. In analyzing each of these cases, I present commonalities and differences in the rating comparisons identified above.

Table 3.

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

Individual Case Analyses

Consistent Cases.

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, RTOP Factor 3, etc). 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 in 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 five-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 prompt a) ("How might this activity facilitate student learning?") responses 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 the 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 are 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 show that her low FA score is due to the fact that she repeats the same strategy over when faced with a potential obstacle. For example, she often writes that she would "go over another similar problem." In summary then, Laura's FASCI and RTOP characterizations were consistently low based on quantitative and qualitative comparisons. I infer that she does not appear to be very student-centered, nor does 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 tenth, eleventh and twelfth 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 takes time to talk to a group that does not want to work together, and presents an alternative for them in which they work independently but discuss 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 uses 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 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 writes 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 writes 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 infer that Ellie is student-centered in her thinking, and that she has 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 out to class. In one classroom observation, this 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 is evidenced by the fact that 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 engage them in the lesson, 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 Cases.

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, 2010) notebook structure. He began class with a warm-up activity projected onto the front board which 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 couple of 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 2). 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 which 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 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, off-task conversations, etc). 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, 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, it appears that the biggest inconsistency in George's characterization was that 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 seems as though 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 which correlated most highly with 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.

In comparing the consistent and inconsistent cases with respect to FASCI and RTOP ratings, one potential 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 which bore on 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.

Discussion

In evaluating proposition three (SK can be observed in teaching practice) I compared scores from the RTOP to scores from the FASCI. Though the SK and RTOP constructs were conceived of and developed differently, the constructs are similar and this comparison is warranted in order to identify cases which rate either consistently or inconsistently on each measure. Results of these comparisons are mixed, but the examination of both consistently rated and inconsistently rated cases can provide some further insight with respect to this proposition. Again, it should be noted that all of these observations are for novice science and mathematics teachers so the same findings cannot be extrapolated to more experienced teachers.

Three consistently rated cases were identified, one that rated very low on both constructs (Laura) and two that rated very high on both (Jason and Ellie). Each of these cases provides convergent validity evidence for the FASCI instrument, but each represents an extreme case in terms of the two scales. None of them could be considered “average” on the RTOP or the FASCI. Laura was very low on all RTOP factors and on SCI and FA, and Jason was very high on all RTOP factors and SCI and FA. Ellie was

very high on four of the five RTOP factors and high on both FA and SCI (refer to Table 3 for specific standardized scores).

Although only two inconsistent cases were noticed and discussed, both shared a common aspect which is potentially driving these score comparisons: contextual difference between their teaching practice and that of the FASCI scenarios. If this difference is a factor in causing the observed inconsistency in RTOP and FASCI ratings, then the FASCI scenarios may be contextually limited in that they do not always capture a respondents' actual practice.

Based on these findings, I conclude that SK as defined by the FASCI construct can be observed in novice science and mathematics teacher practice *to a limited degree*. Limitations exist for some teachers, perhaps based on their teaching context or based on teaching experience. This limited observation could be due to a lack of specificity in the middle region of the SK construct. This finding should be interpreted cautiously, as it is based on a relatively small amount of observational data (18 cases). I will discuss the implications of this potential limitation and of the small sample in the next chapter.

Chapter 7: Discussion and Implications for the Development of Measures of Science and Mathematics Teacher Knowledge

Introduction

In this research I outlined a structure for a validity argument for the Flexible Application of Student-Centered Instruction (FASCI) instrument which was designed to measure science and mathematics teachers' Strategic Knowledge (SK). I described the types of evidence needed to evaluate each proposition in that argument, and collected and analyzed that evidence. Foundational to this argument is the proposed score interpretation: the Strategic Knowledge (SK) of novice science and mathematics teachers can be compared and distinguished both relatively and absolutely. This interpretation supports the intended use of the instrument: to evaluate the effects of a teacher education program on novice science and mathematics teachers' Strategic Knowledge (SK). I outlined a set of five propositions that support the proposed score interpretation, and therefore the instrument use. These propositions were the basis for the specific studies that I carried out in this research.

This score interpretation, the related propositions, and associated validity evidence all contribute to addressing my specific research questions:

1. To what degree is the FASCI instrument valid for comparing and distinguishing the strategic knowledge of novice science and mathematics teachers?
2. What are some potential obstacles to developing valid and reliable measures of science and mathematics teachers' strategic knowledge?

In this chapter I will discuss each of the propositions, the evidence gathered to investigate each, and the findings. I will also evaluate each proposition with respect to these findings. Finally, I will discuss the implications of this work, and present and discuss a set of recommendations for others who are pursuing similar work.

The FASCI Validity Argument: Evidence, Findings, and Evaluations

A summary of the propositions underlying the FASCI validity argument, evidence related to those propositions, findings related to each, and an evaluation of these propositions is shown in Table 1. Below I will discuss each of these elements in detail.

Table 1.

FASCI Instrument Validity Argument Evidence, Findings, and Evaluation

Score Interpretation: the SK of novice science and mathematics teachers can be compared and distinguished both relatively and absolutely
Instrument Use: to evaluate the effects of a teacher education program on novice science and mathematics teachers' SK

Proposition	Evidence	Findings	Evaluating the Proposition
1. SK is one type of knowledge required to be a quality science or mathematics teacher	• Test Content	• The SK construct is related to previous research and the literature on expertise, elicitation of students' ideas, and formative assessment (chapter two, p. 38).	• The evidence from the literature and previous research *supports* the proposition.
2. SK exists across all domains of science and mathematics teaching (e.g., biology, chemistry, physics, math, etc.)	• Test Content	• Science and mathematics teachers use strategic approaches which can be classified similarly, and therefore what constitutes a change in strategic approach is common for these teachers (chapter two, p. 38-9).	• The evidence from the literature and from the scenario-based item content *supports* the proposition.
3. SK can be observed in teaching practice	• Relations to other Variables	• Consistency in SK and RTOP scores were observed at high and low levels of SK, but not in the middle of the SK construct (chapter six, Table 3 and p. 131). • Inconsistencies in SK and RTOP scores were observed for two teachers with teaching contexts that differed from those on the FASCI instrument (chapter 6, Table 3 and p. 135).	• The evidence and findings are based on a small number of observations (18) and is therefore weak. The proposition *is not well supported*.

4. SK can be measured reliably with a scenario-based survey	• Response Processes • Internal Structure • Test Content	• Well-trained raters of similar background can score FASCI response reliably (80-91% agreement on FA, 76-88% agreement on SCI) (chapter 3, p. 73). • Score reliability estimates for relative decisions were low (0.54 for FA, and 0.48 for SCI). Those for absolute decisions were lower (0.52 for FA, and 0.30 for SCI) (chapter 4, p. 85-6, 88). • Relative reliability could be increased appreciably (to 0.71 for FA and 0.64 for SCI) by doubling the number of items on the instrument (chapter 4, p. 85-6, 88), but it is not likely that these increases could be fully realized. • SK scores are very sensitive to the particular items included on the instrument (chapter 4, p. 90).	• The evidence and findings *do not support* the proposition. Though rater agreement was good, there are serious issues with the reliability of SK scores, and it is not clear that these issues can be readily resolved.
5. SK score interpretations change when specific science content is added to the survey items	• Test Content • Response Processes • Internal Structure	• SCI response differences are statistically and practically significant (effect size = 0.58) (chapter 5, p. 101, 104, 111). • FA responses are similar (chapter 5, p. 101, 116).	• The evidence and findings *support* the proposition for the SCI dimension, but *do not support* the proposition for the FA dimension.

SK is required to be a quality Science or Mathematics teacher.

The first proposition put forth states that SK is one type of knowledge required to be a quality science or mathematics teacher. This proposition is based on the idea that the construct being measured needs to be one that matters with respect to the proposed instrument use. A clear link between the construct and previous research and theory is necessary in order to support the idea that the measure is of something that is meaningful. In chapter two, I discussed the SK construct and FASCI instrument content and linked the FA and SCI dimensions to the literature on expertise, elicitation of students' ideas, and formative assessment. The FA dimension is sported by the concept of expertise (e.g., Chi, 2006), and that of *adaptive expertise* specifically (Hatano & Inagaki, 1986). The SCI dimension is supported by the literature on the elicitation of students' ideas (e.g., Bransford, et al., 1999; Fosnot, 1996; Greeno, et al., 1996). I also discussed the similarities between SK and formative assessment, stating that a teacher who is able to elicit students' ideas and then conditionally select a strategic approach is engaging in part of the formative assessment process. The literature and previous research *support* this proposition.

SK exists across all domains of Science and Mathematics teaching.

The second proposition states that SK exists across all domains of science and mathematics teaching (e.g., biology, physics, mathematics, etc). Evidence needed to evaluate this proposition again comes from the FASCI instrument content and links to the literature. The strategic approaches used by all science and mathematics teachers can be classified similarly. In classifying these strategic approaches (and therefore in scoring the FA dimension), a framework by Treagust (2007) was used. More generally, being student-centered, possessing a repertoire of strategies, and being able to enact formative assessment practices are important aspects of teaching within any science or mathematics

domain. Evidence from the literature and from the scenario-based item content *support* this proposition.

Related to this proposition is the importance of operationalizing specific domain boundaries when measuring teacher knowledge. Some questions that should be considered are: a) for what type of teacher is this knowledge construct important?, b) what makes this knowledge construct uniquely important to this group of teachers and not some other group?, c) with respect to the proposed instrument use, what grain size or level of domain boundary is appropriate? In regards to the last of these, for the FASCI instrument it is important that it be useful for *all* science and mathematics teachers, not just those of one specific discipline or a set of disciplines (e.g., the physical sciences), given the structure of the University of Colorado at Boulder Learning Assistant Program (LA Program; Otero, et al., 2006).

SK can be observed in teaching practice.

The third proposition on this validity argument states that SK can be observed in teaching practice. Evidence for evaluating this proposition comes from comparing SK scores to scores from the Reformed Teaching Observation Protocol (RTOP; Sawada, et al., 2002). Though the RTOP was not designed specifically to measure SK, the "reformed teaching" construct is related to SK in a way that warrants this comparison (as discussed in chapter six). Consistency between scores on each for novice science and mathematics teachers would provide a source of convergent validity evidence for the FASCI. I found that these consistencies in ratings were observed for teacher who were high or low on SK, but not for teachers who were in the middle. Further, two inconsistently rated cases shared one characteristic: a difference between their teaching context and that presented in the FASCI scenarios. However, this evidence and the findings are somewhat weak since they are based on a small number of observations (n=18). Therefore the proposition is *not well supported*. In order to better evaluate this proposition, a larger sample size or other convergent validity evidence is needed.

SK can be measured reliably.

The fourth proposition states that SK can be measured reliably with a scenario-based instrument. Though rater agreement was good, score reliability was rather low. The newly trained raters were able to reach higher levels of scoring agreement than those observed in previous pilot testing of the FASCI. In working to reach this agreement, these new raters further clarified some of the scoring language and added a new highest level to the SCI construct map. Their pair-wise agreement was between 80-91% on the FA dimension and 76-88% on the SCI dimension.

The implication of the good rater agreement is that rater recruitment and training are of paramount importance. In the present study, the raters that were recruited were all thought to be high on the SK construct, and were experienced secondary science or mathematics teachers. In previous pilot testing and scoring moderation, it is not clear that all raters shared these characteristics. Further, the training of these new raters was successful in large part due to their initiative in wanting to reach agreement and in working to clarify scoring language.

The score reliability resulting from the set of scenario-based items on the FASCI (as characterized by Cronbach's alpha) was not very high. For the FA dimension, score reliability was 0.62 and for the SCI dimension it was 0.51. By using the G Theory framework, score reliability was able to be examined more critically. It was found that much of the variance in observed scores was due to the items and person by item interactions. For example, on the FA dimension, about 60% of the variance in observed scores is in the person by item interaction. This indicates that a respondent's FA scores across the different raters depend heavily on the particular item being sampled (i.e., the item to which they are responding). And on the SCI dimension, a large part of the variance in observed SCI scores can be attributed to the items themselves (about 42%). In other words, the mean score for one randomly selected item (across all persons and raters) is expected to be quite different from the mean score for all items in the universe (across all persons and raters). The person by item interaction was also a large

score of observed score variance (over 30%) for SCI scores. Further estimation of variance components based on different measurement specifications (in the D studies) shows that reliability could be increased appreciably by doubling the number of items on the instrument (from five to ten). This would increase relative FA score reliability to about 0.71, and relative SCI score reliability to about 0.64. But neither of these values is likely high enough, given that a good target value is 0.80. And further, this potential increase is not likely to be fully realized if the number of items is doubled. Individuals exhibit fatigue in responding to just five items. Therefore it is reasonable to believe that respondents would be unlikely to complete a set of ten items, or that their response quality would degrade substantially. Based on this evidence and the findings, the reliability proposition is *not supported*.

These findings indicate the importance of item choice on the instrument. Because so much of the variance in observed scores is due to the items, the inclusion or exclusion of a particular item could have a large effect on the observed score which is particularly problematic for making absolute decisions. Accordingly, in its current form the FASCI instrument should not be used in potentially high-stakes applications. However, its use may be warranted in more formative applications. The implication here is that item design should be well-specified, and there should be a pool of *quality* items from which item sets can be drawn. Also these sets should be piloted in order to empirically determine the best combination of items which reduces the item-level variance in observed scores and increases the variance attributable to persons.

SK score interpretations change when specific content is added.

The final proposition in the FASCI validity argument (proposition five) states that SK score interpretations change when specific science content is added to the scenario-based items. In chapter two, I discussed my reason for hypothesizing this difference. Based on the construct of PCK, one might think that each teacher has a more sophisticated way of thinking about teaching in their own discipline.

In turn, I would expect that an instrument which bases the items in this particular discipline would do a better job accessing this sophisticated knowledge in an individual who has expertise in that discipline.

Evidence for evaluating this proposition comes from the comparison of item responses and scores from two versions of the FASCI: the content-neutral and physics-specific versions. Statistically and practically significant differences in SCI responses were found. The qualitative analysis of SCI responses showed that when physics content was added to the scenario-based items, respondents were more likely to discuss the content rather than the students. This difference indicates that the physics-FASCI is eliciting from respondents other information in addition to SK in formulating their responses to prompt a) (upon which SCI scores are based). This is a source of *construct-irrelevant variance* since this other information is not related to the construct of interest. This difference is observed for the aggregate sample of respondents, not for physics experts. Therefore the proposition is partially *supported* for the SCI dimension, but not for the reason that was hypothesized. The addition of content to the items did not serve to access a more sophisticated strategic knowledge in respondents.

FA scores and item responses were similar between versions, and did not differ when disaggregated by physics expertise. Therefore the proposition is *not supported* for the FA dimension.

Validity of the FASCI Instrument and Future Directions

My first research question stated: To what degree is the FASCI instrument valid for comparing and distinguishing the strategic knowledge of novice science and mathematics teachers? In addressing this research question, it might not be appropriate for me to say that the FASCI is "valid" or is "not valid" for anyone to use in the evaluation of their teacher education program. And it is difficult to make a summary judgment of "yes" or "no" on the FASCI validity with respect to the proposed score interpretation and use. The user of an instrument is the one who is ultimately responsible for evaluating any validity evidence in relation to the particular setting in which they will use that instrument (AERA, et al., 1999). Clearly the FASCI has some issues, since some of the propositions were not supported. Based

on this validity argument and on the evidence gathered and evaluated, I think it is appropriate to make claims about the strengths and weaknesses of the FASCI with respect to its validity, and about where further observations need to be made and how development needs to proceed in order to support a stronger claim of instrument validity.

The SK construct is based on the literature and on the experiences of pre-service teachers in the CU Boulder Learning Assistant (LA) program. As such, from a construct standpoint the FASCI instrument is appropriate for use in the evaluation of that program. But further work needs to be done in seeking convergent validity evidence for the SK construct. The limited number of observations in this study does not provide enough information to fully evaluate the proposition related to observing SK in practice. Other sources of convergent validity evidence should be considered as well in order to further evaluate this aspect of the validity argument.

The scenario-based item format of the FASCI instrument seems to offer both affordances and limitations. The main affordance is that the items are capable of eliciting responses which are based on an individual's SK. Responses can be scored with acceptable levels of agreement (by well-trained raters) on the SK construct. However, the choice of particular scenario-based items to include on the instrument is a critical one. Because so much of the variance in observed scores was in the items and the person by item interactions, item specification and choice are central to achieving reliable SK scores. This item sensitivity has been observed in other research as well, namely that on science performance assessments (e.g., Shavelson, et al., 1993). Thinking of this in a slightly different way that is more consistent with the G Theory framework used in chapter four, item *design* (rather than just item choice) is the important factor. Because the scenario-based items can be conceived of as being randomly sampled from a universe of admissible items, then the scenario-based structure itself is the critical component. Some items will always be more difficult than others, but a better specification of what

those scenarios entail and how they are structured is needed. This is the critical piece, in addition to a larger number of items from which to create (sample) item sets.

Finally, the FASCI items are sensitive to the elicitation of construct irrelevant responses when specific science content is embedded into the items. Based on comparing scores between versions of the FASCI, future uses of this instrument should use item content that is consistent with the proposed instrument use. In this case, that intended use is for teachers from all science and mathematics disciplines. The SK construct has been operationalized with this purpose in mind: content-neutrality across science and mathematics disciplines. When content is embedded into the items something different is elicited. This is not the construct of interest for the FASCI instrument. If the aim is to measure SK then this other elicitation is confounding.

In summary then there remain some serious challenges with respect to the validity of claims based on SK scores from the FASCI, especially related to the reliability of SK scores. Further item design specification may help to ameliorate these issues. Before claiming a high degree of validity of the FASCI for the proposed score interpretation, these challenges need to be addressed. These findings and the development of this validity argument suggest some concrete recommendations for others who are engaged in similar measurement efforts.

Recommendations for Related Measure Development Efforts

My second research question states: What are some potential obstacles to developing valid and reliable measures of science and mathematics teachers' strategic knowledge? Based on having been involved in the FASCI development process, the experience of having conducted this validity argument for the FASCI instrument, and on the findings from these investigations, I present a set of four recommendations for others who are undertaking related work. These recommendations come from encountering some of these obstacles and focus on the areas of construct definition, item design and development, rater recruitment and training, and the validation process itself.

One of the big challenges to any measurement effort lies in specifying and defining the construct of interest. The SK construct was initially developed based on theoretical work, the previous experiences of researchers involved in the development, and those of pre-service teachers in the target teacher education program. It has been refined and further specified at various stages during the instrument development process. Establishing the qualitatively distinct levels of this construct has been challenging. For example, the SCI construct only had two levels for a long time. It was not until a new group of raters were trained that a third level emerged as being distinct from the others. This was based on an examination of item responses from the fresh perspective of these new raters. Also during the development of this validity argument, I found that it was difficult to observe convergent validity evidence (in the form of consistent RTOP-FASCI score comparisons) for persons in the middle of the SK continuum. This is likely due in part to the small sample in that study, but may also be indicative of the difficulty of specifying the middle levels of the construct of interest in detail. Based on this, my recommendation for others is to gather evidence early in the development process which will help to define the qualitatively distinct levels of the construct, especially those levels in the middle of the continuum. This evidence could come not only from previous work and participant or developer experience, but also from item responses and think-aloud responses gathered during early development. It may be tempting to see this latter type of evidence as contributing chiefly to item design, but it is equally important for construct definition. In this study, the item responses proved essential in defining the third level of the SCI construct.

Once the construct of interest has been at least preliminarily defined, items are designed to elicit responses that can be scored on that construct. The format that these items take should initially be defined by the construct one is trying to measure. For example, the SK construct can be thought of as complex and related to actual practice, in that it seeks to distinguish how individuals *conceive of* and *respond to* various teaching situations. This construct lends itself to the scenario-based items which

represent actual classroom situations. This type of item (which can be considered a sort of performance task) carries with it a set of challenges. First, if not carefully aligned with the construct of interest, these items can elicit information other than that related to the construct of interest (i.e., they can introduce construct irrelevant variance). Second, they can elicit information from the respondents inconsistently, and therefore yield relatively low score reliability. In similar situations, my recommendations would be to specify item design carefully, to develop a pool of items from which the set included on the instrument could be drawn, and to pilot these items early in the process. The set of items that yields optimal score reliability should be determined as early as possible so that pilot tests can be based on these items. Developers should allocate sufficient resources for this item design and development.

Once item responses from pilot testing have been gathered, rater recruitment and training are central to obtaining reliable scores. In this study, a strong group of raters was recruited and trained. These individuals were all thought to rate highly on the SK construct, and they took the initiative to work on agreement among them and to modify the scoring guides as necessary. Based on this experience, my recommendation would be for others to be deliberative about how raters are recruited and trained. Does it make sense to have raters who would rate highly on the construct of interest themselves? In this case, it did make sense, but perhaps not in all cases. The point is that some defensible criteria for rater recruitment should be established and used. Raters should not come from some sort of convenience sample. Rater training should be well-planned, but should also be flexible enough so that it can be tailored to the specific group of raters.

My final recommendation has to do with the validation process itself. It would be useful to conceive of the instrument development process, from the outset, within a validity argument structure. The proposed instrument use and score interpretation should be articulated very early on in the development process, and supporting propositions should be identified and evidence gathered to evaluate those propositions all throughout development. Those propositions are likely to change and

new ones will be formed during development, but by beginning instrument development with these things in mind evidence from all steps within the process can be used in evaluating each proposition. By making this validation effort explicit from the outset, it will be apparent how the evidence gathered can inform the development itself. For example, had the decomposition of observed score variance been analyzed earlier in the FASCI development, the scenario-based item design may have been further specified and new items may have been piloted in order to achieve more reliable scores.

Conclusion

As stated in the opening chapter of this dissertation, it is imperative that we are able to measure science and mathematics teachers' strategic knowledge in order to evaluate the effectiveness of our teacher education programs. When engaging in such measurement efforts, the stakes are potentially high. Teacher education programs could be deemed ineffective and lose enrollment, funding, or accreditation. Based on the relatively low score reliabilities observed in this study, it would be difficult to justify using the FASCI instrument for these high-stakes applications. However, the FASCI instrument should not be abandoned as a lost cause. In this study, respondent sample sizes were small. If the sample size was increased dramatically (perhaps into the hundreds), score reliability would increase. The increase would lead to more confident norm-referenced SK score comparisons.

A second situation is also promising. As discussed in the findings chapters, responses to the scenario-based items provide a rich picture of respondent knowledge. This information was the basis for the deeper, qualitative examination of SCI responses which served to further define the observed differences in SCI scores between versions of the FASCI. Using this same item structure and further specifying the item design would help to lessen the role that item-level variance plays in measurement error. Doing so could increase score reliability enough to justify high-stakes uses, while also providing the qualitative response data which would be useful to teacher educators. This latter use would be formative in nature, and could help teacher educators to diagnose their students' SK and inform

instructional decisions. In this way, the FASCI instrument could be used for both high and low-stakes applications. Even without this further item specification, the FASCI instrument could currently be used in this formative way. However, if further item specification is undertaken, new studies would need to be conducted in order to examine the changes in score reliability and instrument validity.

The *process* of instrument validation involves the evaluation of the plausibility and appropriateness of the proposed instrument use and score interpretations. I emphasize the word *process* because validation is not a one-time event. As the construct and instrument are refined or further developed, and as more evidence and observations are made available, validation efforts should proceed. This gathering of further evidence is particularly important to the process. As Kane states, "the challenge is to make the connection between limited samples of observations and the proposed interpretations and uses" (Kane, 2006, p. 17). As further observations become available from future testing and administration of the FASCI, this validation process should continue and the above propositions (and perhaps new ones) should be re-visited. As well, if the FASCI instrument is to be used for a different purpose then a new validation effort must be undertaken.

In this study I have provided an example of an instrument validation effort. I have also suggested some future directions for the further development of that instrument. And finally, I have presented a set of recommendations for others who are engaged in similar work, based on the obstacles encountered in this study. If measuring an individual's knowledge was as simple as using a meter stick, none of these elements would constitute a contribution to the field. But any dimension of knowledge is far more complex that 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.

References

Advancement Via Individual Determination. (2010). Intro to the AVID Program Retrieved October 5, 2010, from http://www.avid.org/intro.html

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

American Association for the Advancement of Science (Ed.). (1990). *Science for all Americans* (New ed.). New York: Oxford University Press.

Angoff, W. H. (1988). Validity: An Evolving Concept. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 19-32). Mahwah, N.J.: Lawrence Erlbaum Associates.

Atkin, J. M., Black, P., & Coffey, J. (2001). Classroom Assessment and the National Science Education Standards. In D. C. C. f. E. National Academy of Sciences - National Research Council Washington (Ed.), (pp. 126).

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

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74.

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

Bransford, J., Brown, A. L., & Cocking, R. R. (1999). *How people learn: brain, mind, experience, and school*. Washington, D.C.: National Academy Press.

Brennan, R. L. (1992). An NCME Instructional Module on Generalizability Theory. *Educational Measurement: Issues and Practice, 11*(4), 27-34.

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.

Burry-Stock, J. A., & Oxford, R. L. (1993). Expert Science Teaching Educational Evaluation Model (ESTEEM): Measuring Excellence in Science Teaching for Professional Development. *Journal of Personnel Evaluation in Education, 8*, 267-297.

Chi, M. T. H. (2006). Two Approaches to the Study of Experts' Characteristics. In N. Charness, P. Feltovich & R. Hoffman (Eds.), *Cambridge Handbook of Expertise and Expert Performance* (pp. 21-30). Cambridge: Cambridge University Press.

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

Clermont, C. P., Borko, H., & Krajcik, J. S. (1994). Comparative study of the pedagogical content knowledge of experienced and novice chemical demonstrators. *Journal of Research in Science Teaching, 31*(4), 419-441.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt Rinehart and Winston.

Cronbach, L. J. (1988). Five Perspectives on Validity Argument. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 3-15). Mahwah, N.J.: Lawrence Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin, 52*(4), 281-302.

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

Dwyer, C. A. (1998). Psychometrics of Praxis III: Classroom Performance Assessments. *Journal of Personnel Evaluation in Education, 12*(2), 163-187.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 1149-1160. doi: DOI 10.3758/BRM.41.4.1149

Fosnot, C. T. (1996). *Constructivism : theory, perspectives, and practice*. New York: Teachers College Press.

Greeno, J. G., Collins, A., & Resnick, L. B. (1996). Cognition and Learning. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of Educational Psychology* (pp. 15-46). New York: Macmillan.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational Measurement* (Fourth ed., pp. 65-110). Westport, CT: Praeger.

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

Hammer, D. (1996). More than misconceptions: Multiple perspectives on student knowledge and reasoning, and an appropriate role in education research. *American Journal of Physics, 64*(10), 1316-1325.

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.

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

Hatano, G., & Inagaki, K. (1986). Two Courses of Expertise. In H. Stevenson, H. Azuma & K. Hakuta (Eds.), *Child Development and Education in Japan* (pp. 262-272). New York: Freeman.

Heller, J. I., Daehler, K. R., Shinohara, M., & Kaskowitz, S. R. (2004, April). *Fostering Pedagogical Content Knowledge about Electric Circuits through Case-Based Professional Development.* Paper presented at the National Association for Research in Science Teaching, Vancouver.

Hestenes, D., Wells, M., & Swackhammer, G. (1992). Force Concept Inventory. *Physics Teacher, 30*(3), 141-158.

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

Horizon Research, I. (1999a). *Local Systemic Change through Teacher Enhancement Classroom Observation Protocol*.

Horizon Research, I. (1999b). *Local Systemic Change through Teacher Enhancement Science 6-12 Teacher Questionnaire*.

Horizon Research, I. (2000). Validity and Reliability Information for the LSC Classroom Observation Protocol

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*.

Kane, M. T. (2001). Current Concerns in Validity Theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kane, M. T. (2004). Certification Testing as an Illustration of Argument-Based Validation. *Measurement, 2*(3), 135.

Kane, M. T. (2006). Validation. In E. National Council on Measurement in & R. L. Brennan (Eds.), *Educational Measurement* (4th ed., pp. 17-64): Praeger.

Kind, V. (2009). Pedagogical content knowledge in science education: perspectives and potential for progress. *Studies in Science Education*, 169-204. doi: DOI 10.1080/03057260903142285

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.

Lee, E. (2005). *Conceptualizing pedagogical content knowledge from the perspective of experienced secondary science teachers.* PhD, University of Texas, Austin, TX. Retrieved from http://hdl.handle.net/2152/391

Lee, E., Brown, M. N., Luft, J. A., & Roehrig, G. H. (2007). Assessing Beginning Secondary Science Teachers' PCK: Pilot Year Results. *School Science and Mathematics, 107*(2), 52-61.

Lee, E., & Luft, J. A. (2005). *Capturing the Pedagogical Content Knowledge of Experienced Science Teachers.* Paper presented at the Annual conference of the Association for Science Teacher Education, Colorado Springs, CO.

Loughran, J., Milroy, P., Berry, A., Gunstone, R., & Mulhall, P. (2001). Documenting Science Teachers' Pedagogical Content Knowledge through PaP-eRs. *Research in Science Education, 31*(2), 289-307.

McDermott, L. C. (1991). What we teach and what is learned: Closing the gap. *American Journal of Physics, 59*(4), 301-315.

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

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

Minstrell, J. (1991). *Facets of Students' Knowledge and Relevant Instruction.* Paper presented at the Research in Physics Learning Workshop at the University of Bremen: Theoretical Issues and Empirical Studies, Kiel, Germany: Institute fur die Padagogik der Naturwissenshaften.

National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics*. Reston, VA: NCTM.

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

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

Otero, V., & Nathan, M. J. (2008). Preservice Elementary Teachers' Views of Their Students' Prior Knowledge of Science. *Journal of Research in Science Teaching, 45*(4), 497-523.

Peterson, R., & Treagust, D. (1995). Developing Preservice Teachers' Pedagogical Reasoning Ability. *Research in Science Education, 25*(3), 291-305.

Piburn, M. D., Sawada, D., Turley, J., Falconer, K., Benford, R., Bloom, I., & Judson, E. (2000). Reformed Teaching Observation Protocol (RTOP) Reference Manual: Arizona State University.

Rowan, B., Schilling, S. G., Ball, D., & Miller, R. (2001). Measuring Teachers' Pedagogical Content Knowledge in Surveys: An Exploratory Study: Consortium for Policy Research in Education.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: An update. *Journal of Research in Science Teaching*, 1045-1063.

Sadler, D. R. (1989). Formative Assessment and the Design of Instructional Systems. *Instructional Science, 18*(2), 119-144.

Sawada, D. (2003). Reformed Teacher Education in Science and Mathematics: An Evaluation of the Arizona Collaborative for Excellence in the Preparation of Teachers (pp. 344). Tempe, AZ: Arizona State University.

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.

Schilling, S. G., & Hill, H. C. (2007). Focus Article: Assessing Measures of Mathematical Knowledge for Teaching: A Validity Argument Approach. *Measurement: Interdisciplinary Research and Perspectives, 5*(2-3), 70-80.

Sfard, A. (1998). On Two Metaphors for Learning and the Dangers of Choosing Just One. *Educational Researcher, 27*(2), 4-13.

Shavelson, R. J., Baxter, G. P., & Gao, X. H. (1993). Sampling Variability of Performance Assessments. *Journal of Educational Measurement*, 215-232.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory : a primer*. Newbury Park, Calif.: Sage Publications.

Shin, M.-K., Yager, R. E., Oh, P. S., & Lee, M.-K. (2005). Changes in Science Classrooms After Experiencing an International Professional Staff Development Program. *International Journal of Science and Mathematics Education, 1*(4), 505-522.

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.

Taylor, J., & Gess-Newsome, J. (2007, January). *Exploring Tools and Methods for Measuring Pedagogical Content Knowledge.* Paper presented at the Annual conference of the Association for Science Teacher Education, Clearwater, FL.

Thompson, B. (2003). *Score reliability : contemporary thinking on reliability issues*. Thousand Oaks, Calif.: SAGE Publications.

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.

Traub, R. E. (1994). *Reliability for the Social Sciences: Theory and Applications*. Thousand Oaks, CA: Sage.

Treagust, D. (2007). General Instructional Methods and Strategies. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of Research on Science Education* (pp. 373-391). Mahwah, NJ: Lawrence Erlbaum Associates.

United States Department of Education. (2007). Improving Teacher Quality Retrieved April 23, 2007, from http://www.ed.gov/teachers/nclbguide/improve-quality.html

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.

van Zee, E., & Minstrell, J. (1997). Using Questioning To Guide Student Thinking. *Journal of the Learning Sciences, 6*(2), 227-269.

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.

Wilson, M. (2005). *Constructing measures : an item response modeling approach*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Appendix A: Versions of the FASCI

Content-Neutral (n-) FASCI

The first set of questions pertains to your personal information

1) Which statement best describes you?
   a) I am a Learning Assistant
   b) I am a student in the teacher education program, but not in the LA program
   c) I am a Practicing K-12 teacher
   d) Other (Please specify)

2) What university do you attend?
   a) Southeast Coastal University
   b) Northeast Queen's University
   c) Northwest Pacific University
   d) Western State University
   e) Central Research University
   f) Other (Please specify)

3) What is your age?

4) What is your gender?
   a) Female
   b) Male

5) What is your race? Choose all that apply.
   a) American Indian or Alaska Native
   b) Asian
   c) Black or African American
   d) Hispanic or Latino
   e) Native Hawaiian or Other Pacific Islander
   f) White
   g) Other (Please specify)

6) What is your subject area specialty that you plan on teaching in the near future (when you complete teacher education, or in your next year of teaching practice)?
   a) Astronomy
   b) Biology
   c) Chemistry
   d) Geology/Earth Sciences
   e) Math
   f) Physics
   g) Other (Please specify)

The next set of questions pertains to your *most recent* teaching experience

7) What was the setting of your most recent teaching experience
   a) Elementary School (K-5)
   b) Middle School (6-8)
   c) High School (9-12)
   d) Instructor for an undergraduate course
   e) Learning Assistant or Teaching Assistant for an undergraduate course
   f) Instructor for a graduate course
   g) Learning Assistant or Teaching Assistant for a graduate course
   h) I have never taught before
   i) Other (Please specify)

8) In what content area were you teaching? Choose all that apply.
   a) Astronomy
   b) Biology or Life Science
   c) Chemistry
   d) Engineering
   e) Geology or Earth Science
   f) Physics
   g) Mathematics
   h) Statistics
   i) I have not taught before
   j) Other (Please specify)

9) What was the approximate number of students in your class(es)?
   a) Less than 10
   b) 11-20
   c) 21-30
   d) 31-50
   e) 51+
   f) I have not taught before

10) What was the gender distribution of your class(es)?
    a) More females than males
    b) About equal distribution of females and males
    c) More males than females
    d) I have not taught before

11) What was the racial/ethnic composition of your class(es)?
    a) Primarily African-American
    b) Primarily Asian
    c) Primarily Hispanic
    d) Primarily white/Caucasian
    e) I have not taught before
    f) Other (Please specify)

12) How many years of experience have you had in the role of a teacher, in any setting?

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.

13) 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?

14) You are working out an example problem up on the board.

a) How might this activity facilitate student learning?

You accidentally make a mistake in solving the problem but don't realize this until you get to the end of your solution and realize that the answer doesn't make sense. No one in the class has said anything, so you're not sure if they caught the mistake or 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?

15) You have just finished giving a presentation on a complicated topic.

a) How might this activity facilitate student learning?

You notice that many of the students in the class have very confused expressions on their faces.

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?

16) You have given your students a quiz to assess their understanding of a difficult topic.

a) How might this activity facilitate student learning?

Many of your students are discouraged after performing poorly on the quiz.

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?

17) In talking with one of your students you discover that they have a misconception about a central topic presented in that week's class. You attempt to address the misconception by having a one-on-one conversation with the student.

a) How might this activity facilitate student learning?

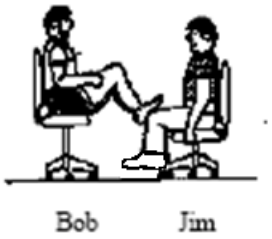
Despite your conversation, the student maintains the same misconception.

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?

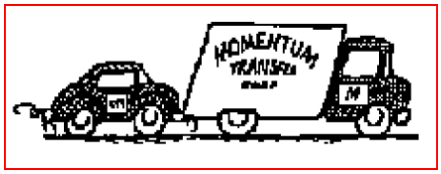
To assess your students' understanding of Newton's 3rd Law, you have given them the following conceptual questions:

18) Two students sit in identical office chairs facing each other. Bob has a mass of 95 kg, while Jim has a mass of 77 kg. Bob places his bare feet on Jim's knees, as shown to the right. Bob then suddenly pushes outward with his feet, causing both chairs to move. In this situation, while Bob's feet are in contact with Jim's knees,



a) Neither exerts a force on the other
b) Bob exerts a force on Jim, but Jim doesn't exert any force on Bob
c) Each student exerts a force on the other, but Jim exerts the larger force
d) Each student exerts a force on the other, but Bob exerts the larger force
e) Each student exerts the same amount of force on the other
f) None of these answers is correct

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, as in the picture below.



Pick one of the choices a) through f) which correctly describes the forces between the car and the truck for each of the descriptions in the questions below.

19) The car is pushing on the truck, but not hard enough to make the truck move.
   a) The force of the car pushing against the truck is equal to that of the truck pushing back against the car.
   b) The force of the car pushing against the truck is less than that of the truck pushing back against the car.
   c) The force of the car pushing against the truck is greater than that of the truck pushing back against the car.
   d) The car's engine is running so it applies a force as it pushes against the truck, but the truck's engine isn't running so it can't push back with a force against the car.
   e) Neither the car nor the truck exert any force on each other. The truck is pushed forward simply because it is in the way of the car.
   f) None of these descriptions is correct.

20) The car, still pushing the truck, is **speeding up** to get to cruising speed.
   a) The force of the car pushing against the truck is equal to that of the truck pushing back against the car.
   b) The force of the car pushing against the truck is less than that of the truck pushing back against the car.
   c) The force of the car pushing against the truck is greater than that of the truck pushing back against the car.
   d) The car's engine is running so it applies a force as it pushes against the truck, but the truck's engine isn't running so it can't push back with a force against the car.
   e) Neither the car nor the truck exert any force on each other. The truck is pushed forward simply because it is in the way of the car.
   f) None of these descriptions is correct.

21) The car, still pushing the truck, is at cruising speed and continues to travel at the **same speed**.
   a) The force of the car pushing against the truck is equal to that of the truck pushing back against the car.
   b) The force of the car pushing against the truck is less than that of the truck pushing back against the car.
   c) The force of the car pushing against the truck is greater than that of the truck pushing back against the car.
   d) The car's engine is running so it applies a force as it pushes against the truck, but the truck's engine isn't running so it can't push back with a force against the car.
   e) Neither the car nor the truck exert any force on each other. The truck is pushed forward simply because it is in the way of the car.
   f) None of these descriptions is correct.

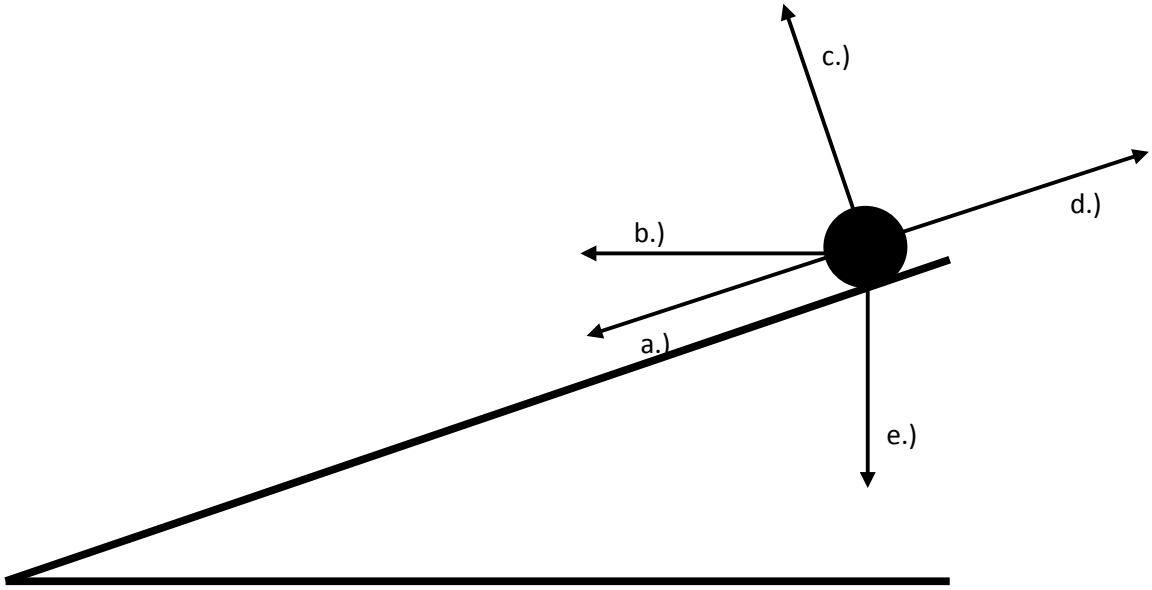
22) The car, still pushing the truck, is at cruising speed when the truck puts on its brakes and causes the car to **slow down**.
   a) The force of the car pushing against the truck is equal to that of the truck pushing back against the car.
   b) The force of the car pushing against the truck is less than that of the truck pushing back against the car.
   c) The force of the car pushing against the truck is greater than that of the truck pushing back against the car.
   d) The car's engine is running so it applies a force as it pushes against the truck, but the truck's engine isn't running so it can't push back with a force against the car.
   e) Neither the car nor the truck exert any force on each other. The truck is pushed forward simply because it is in the way of the car.
   f) None of these descriptions is correct.

To assess your students' understanding of Newton's 2nd Law, you have given them the following conceptual questions:

You throw a coin vertically downward. Please indicate whether each of the following statements is true or false. Ignore air resistance and note that more than one of the statements could be true.

23) After the coin leaves your hand, the net force on the coin is increasing
24) After the coin leaves your hand, the acceleration of the coin is increasing
25) After the coin leaves your hand, the speed of the coin is increasing
26) After the coin leaves your hand, the net force on the coin is zero
27) After the coin leaves your hand, the speed of the coin is constant
28) After the coin leaves your hand, the acceleration of the coin is constant

29) A ball is rolled up a ramp and has reached its highest point as indicated in the figure below. Which vector represents the net force on the ball when it is at its highest point and is just beginning to roll back down the ramp? Consider friction and air resistance to be negligible.



30) Approximately how many Physics course have you taken at the post-secondary level?
   a) 0
   b) 1
   c) 2
   d) 3
   e) 4
   f) 5
   g) more than 5

Physics-Specific (p-) FASCI

The first set of questions pertains to your personal information

1) Which statement best describes you?
   a) I am a Learning Assistant
   b) I am a student in the teacher education program, but not in the LA program
   c) I am a Practicing K-12 teacher
   d) Other (Please specify)

2) What university do you attend?
   a) Southeast Coastal University
   b) Northeast Queen's University
   c) Northwest Pacific University
   d) Western State University
   e) Central Research University
   f) Other (Please specify)

3) What is your age?

4) What is your gender?
   a) Female
   b) Male

5) What is your race? Choose all that apply.
   a) American Indian or Alaska Native
   b) Asian
   c) Black or African American
   d) Hispanic or Latino
   e) Native Hawaiian or Other Pacific Islander
   f) White
   g) Other (Please specify)

6) What is your subject area specialty that you plan on teaching in the near future (when you complete teacher education, or in your next teaching practice)?
   a) Astronomy
   b) Biology
   c) Chemistry
   d) Geology/Earth Sciences
   e) Math
   f) Physics
   g) Other (please specify)

The next set of questions pertains to your *most recent* teaching experience

7) What was the setting of your most recent teaching experience
   a) Elementary School (K-5)
   b) Middle School (6-8)
   c) High School (9-12)
   d) Instructor for an undergraduate course
   e) Learning Assistant or Teaching Assistant for an undergraduate course
   f) Instructor for a graduate course
   g) Learning Assistant or Teaching Assistant for a graduate course
   h) I have never taught before
   i) Other (Please specify)

8) In what content area were you teaching? Choose all that apply.
   a) Astronomy
   b) Biology or Life Science
   c) Chemistry
   d) Engineering
   e) Geology or Earth Science
   f) Physics
   g) Mathematics
   h) Statistics
   i) I have not taught before
   j) Other (Please specify)

9) What was the approximate number of students in your class(es)?
   a) Less than 10
   b) 11-20
   c) 21-30
   d) 31-50
   e) 51+
   f) I have not taught before

10) What was the gender distribution of your class(es)?
   a) More females than males
   b) About equal distribution of females and males
   c) More males than females
   d) I have not taught before

11) What was the racial/ethnic composition of your class(es)?
    a) Primarily African-American
    b) Primarily Asian
    c) Primarily Hispanic
    d) Primarily white/Caucasian
    e) I have not taught before
    f) Other (Please specify)

12) How many years of experience have you had in the role of a teacher, in any setting?

Topic 1: Newton's 3rd Law

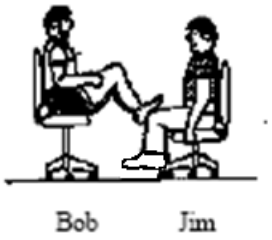
For the questions and scenarios that follow, please assume that you are teaching a high school course in physics to a class of 25-30 students. You have defined the following learning objectives for this class:

- 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.

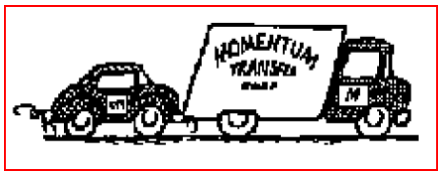
To assess your students' understanding of this content, you have given them the following conceptual questions:

13) Two students sit in identical office chairs facing each other. Bob has a mass of 95 kg, while Jim has a mass of 77 kg. Bob places his bare feet on Jim's knees, as shown to the right. Bob then suddenly pushes outward with his feet, causing both chairs to move. In this situation, while Bob's feet are in contact with Jim's knees,



    a) Neither exerts a force on the other
    b) Bob exerts a force on Jim, but Jim doesn't exert any force on Bob
    c) Each student exerts a force on the other, but Jim exerts the larger force
    d) Each student exerts a force on the other, but Bob exerts the larger force
    e) Each student exerts the same amount of force on the other
    f) None of these answers is correct

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, as in the picture below.



Pick one of the choices a) through f) which correctly describes the forces between the car and the truck for each of the descriptions in the questions below.

14) The car is pushing on the truck, but not hard enough to make the truck move.
   a) The force of the car pushing against the truck is equal to that of the truck pushing back against the car.
   b) The force of the car pushing against the truck is less than that of the truck pushing back against the car.
   c) The force of the car pushing against the truck is greater than that of the truck pushing back against the car.
   d) The car's engine is running so it applies a force as it pushes against the truck, but the truck's engine isn't running so it can't push back with a force against the car.
   e) Neither the car nor the truck exert any force on each other. The truck is pushed forward simply because it is in the way of the car.
   f) None of these descriptions is correct.

15) The car, still pushing the truck, is **speeding up** to get to cruising speed.
   a) The force of the car pushing against the truck is equal to that of the truck pushing back against the car.
   b) The force of the car pushing against the truck is less than that of the truck pushing back against the car.
   c) The force of the car pushing against the truck is greater than that of the truck pushing back against the car.
   d) The car's engine is running so it applies a force as it pushes against the truck, but the truck's engine isn't running so it can't push back with a force against the car.
   e) Neither the car nor the truck exert any force on each other. The truck is pushed forward simply because it is in the way of the car.
   f) None of these descriptions is correct.

16) The car, still pushing the truck, is at cruising speed and continues to travel at the **same speed**.
    a) The force of the car pushing against the truck is equal to that of the truck pushing back against the car.
    b) The force of the car pushing against the truck is less than that of the truck pushing back against the car.
    c) The force of the car pushing against the truck is greater than that of the truck pushing back against the car.
    d) The car's engine is running so it applies a force as it pushes against the truck, but the truck's engine isn't running so it can't push back with a force against the car.
    e) Neither the car nor the truck exert any force on each other. The truck is pushed forward simply because it is in the way of the car.
    f) None of these descriptions is correct.

17) The car, still pushing the truck, is at cruising speed when the truck puts on its brakes and causes the car to **slow down**.
    a) The force of the car pushing against the truck is equal to that of the truck pushing back against the car.
    b) The force of the car pushing against the truck is less than that of the truck pushing back against the car.
    c) The force of the car pushing against the truck is greater than that of the truck pushing back against the car.
    d) The car's engine is running so it applies a force as it pushes against the truck, but the truck's engine isn't running so it can't push back with a force against the car.
    e) Neither the car nor the truck exert any force on each other. The truck is pushed forward simply because it is in the way of the car.
    f) None of these descriptions is correct.

Please respond to the following questions about your teaching:

18) 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?

19) On the board, you are drawing free-body diagrams of the car and the truck.

a) How might this activity facilitate student learning?

You accidentally make a mistake in drawing these diagrams but don't realize this until after you complete the diagrams and realize that they don't make sense. No one in the class has said anything, so you're not sure if they caught the mistake or 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?

20) You have just finished giving a presentation on Newton's Third Law.

a) How might this activity facilitate student learning?

You notice that many of the students in the class have very confused expressions on their faces.

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?

Topic 2: Newton's 2nd Law

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:

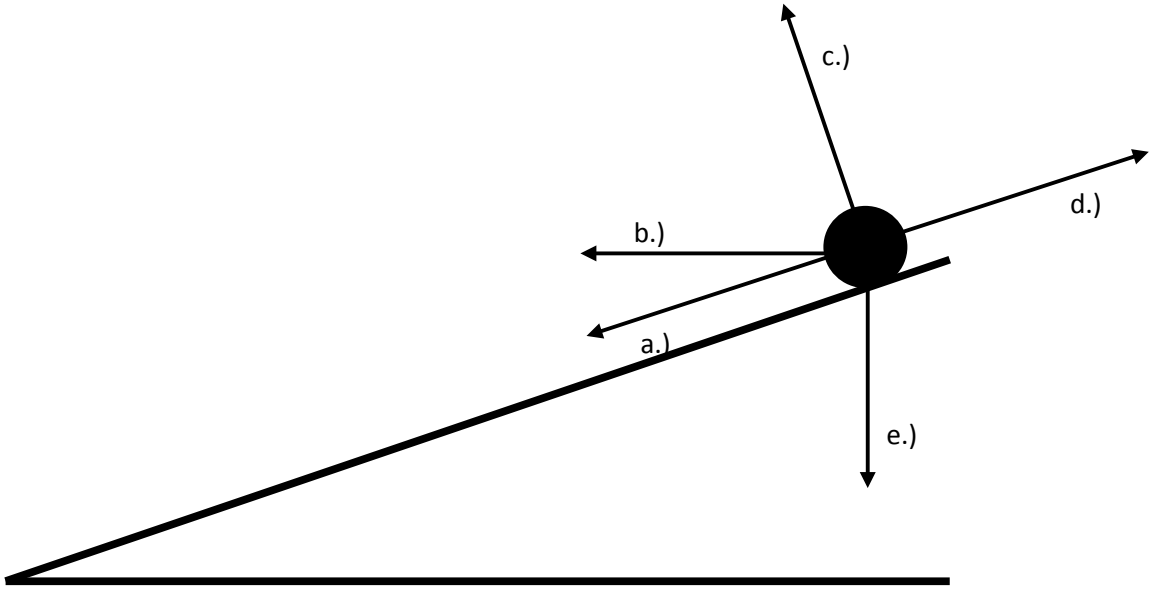
- Students should understand the relation between the net force that is exerted on an object and the resulting change in the object's velocity, so they can:
  - Calculate, for an object moving in one dimension, the velocity change that results when one constant force is exerted on the object over a specified time interval.
  - Determine, for a constantly accelerating object, the average net force that was exerted on the object.
  - Understand how Newton's Second Law applies to an object that interacts with the Earth, and be able to draw a well-labeled free-body diagram of that object.
  - Analyze situations in which an object moves with specified acceleration when one or more forces are exerted on it, and determine the magnitude and direction of the net force.

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

You throw a coin vertically downward. Please indicate whether each of the following statements is true or false. Ignore air resistance and note that more than one of the statements could be true.

21) After the coin leaves your hand, the net force on the coin is increasing
22) After the coin leaves your hand, the acceleration of the coin is increasing
23) After the coin leaves your hand, the speed of the coin is increasing
24) After the coin leaves your hand, the net force on the coin is zero
25) After the coin leaves your hand, the speed of the coin is constant
26) After the coin leaves your hand, the acceleration of the coin is constant

27) A ball is rolled up a ramp and has reached its highest point as indicated in the figure below. Which vector represents the net force on the ball when it is at its highest point and is just beginning to roll back down the ramp? Consider friction and air resistance to be negligible.



Please respond to the following questions about your teaching:

28) You have given your students a quiz about the above questions to assess their understanding of Newton's 2nd Law.

a) How might this activity facilitate student learning?

Many of your students are discouraged after performing poorly on the quiz.

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?

29) In talking with one of your students you discover that they think that there is still a force exerted by your hand on the coin after it leaves your hand. You attempt to address this difficulty by having a one-on-one conversation with the student.

a) How might this activity facilitate student learning?

Despite your conversation, the student stills holds his/her prior idea.

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?

30) Approximately how many Physics course have you taken at the post-secondary level?
a) 0
b) 1
c) 2
d) 3
e) 4
f) 5
g) more than 5

Appendix B: Scoring Guides, Construct Maps, and Rater Agreement

FA Scoring

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 B1. FA scoring guide

Criteria for FA scoring:

- Modification of teaching approach (drawing from their repertoire)
    - Note: the modification needs to take the activity beyond the original activity presented in the scenario. They need to "break out" of the loop. **Remember: we are comparing the strategy that they present in c to that they present in b, not those compared to the original situation.**
- Discussion of contextual factors *that bear on the modification* of the teaching approach, not just context in general.
- Different content/context from b to c does not constitute a new strategy.
- **They cite a specific classroom context for which the new strategy would be well-suited. Context cited and strategy need to be linked.**

*Note: for FA, it doesn't matter if their teaching strategies are student-centered or teacher-centered,

just that they modify or change their approach somehow.

Level	Respondent Characteristics	Example Responses
2	• The teacher has repertoire of strategies that can be used to facilitate student learning within a given class session. • If the teaching strategy comprised of these acts is not producing the desired result, sometimes it can be modified. • The teacher recognizes that the choice of a class activity and associated teaching strategy will depend upon variables specific to the classroom context. *Note: the context cannot be a re-statement of the question (i.e., the obstacle. "It didn't work…")	i1, prompt c.): "Depends on the question and its role in the class. If the question was fundamental and something we needed to build on right away, it's conceivable that I might *discuss* the two solutions, have OTHER groups help argue out which was right (and why) and which was wrong (and why). If time was short, I might just present such a 'discussion' myself. If there was time and other groups were also struggling on the same point, I might build a secondary activity to approach the same concept from a new angle" ID:2191029
1	• The teacher has a repertoire of strategies that can be used to facilitate student learning within a given class session. • If an activity based on a particular teaching strategy is not producing the desired result, the activity can be modified by selecting a different strategy.	I3, prompt b.): "I would start to ask more questions starting with open questions to get students engaged and then relate those open questions to some new closed questions." Prompt c.): "I would find an experiment or activity that demonstrates the same thing as the topic." ID:1977241
0	• The teacher has a limited repertoire of strategies. • Once a particular activity has been selected for a class session, it is not easily modified with a different strategy.	I2, prompt b.): "I would ask the students if they see an error in the problem and if they do fix it from there. If they do not then show them where it is and have them help to fix it." Prompt c.): "Bring up the problem again and show the students how to do it correctly and how I went wrong and show them to be careful of it." ID:1976780

Figure B2. Flexible Application (FA) Construct Map: Respondents and Responses

SCI Scoring

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 B3. SCI scoring guide (new additions italicized)

Criteria/notes for SCI scoring:

- Discussion of *interactive teaching* in the learning activity. The teacher views the students' role in the activity from a constructivist standpoint. They see the learning activity as an opportunity for interacting with the student so that they can identify the students' current conceptions, understanding of the material, or level of engagement with the material. This may include students interacting with the teacher, articulating and defending their ideas with others, asking questions, working on a problem, manipulating apparatus, etc. The teacher takes action to involve the student in the learning process, such as questioning them in order to identify their previous/current conceptions.

Level	Respondent Characteristics	Example Responses
2	• Discussion of interactive teaching which would be observable to the teacher or to an outside "other." • *Discussion of a rationale for why they see this as an opportunity for interactive teaching and learning* Teacher ⟷ Students and/or Students ⟷ Students	i2, prompt a: "it would be better if the students put the diagrams on the board instead of me, but i would ask them what should be drawn, and only draw what they told me. this way, the students are responsible for the direction the diagrams go." ID 3652772
1	• Discussion of interactive teaching which would be observable to the teacher or to an outside "other." *Note: the two-directional interaction needs to be easily inferred or is clearly stated Teacher ⟷ Students and/or Students ⟷ Students	i4, prompt a.): The quiz allows both the students and I to see where they have succeeded in understanding Newton's 2nd Law and where they need to improve. ID:3648350
0	• No discussion of interactive teaching • Teacher primarily views classroom activities as ways to help students make sense of new ideas. Information goes from teacher to student. Teacher → Students	i3, prompt a.): "Allows the students to think about concepts. Allows the students to absorb information from the knowledge of the teacher." ID:1976780

Figure B4. Student-Centered Instruction (SCI) Construct Map: Respondents and Responses (new additions italicized)

FA Rater Agreement

Table B1.

Overall FA Rater Agreement

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

Table B2.

Item 1 FA Rater Agreement

Rater Combination	Percent Agreement	Kappa
r1-r2	82%	.69
r1-r3	77%	.61
r2-r3	92%	.85

Table B3.

Item 2 FA Rater Agreement

Rater Combination	Percent Agreement	Kappa
r1-r2	88%	.78
r1-r3	82%	.66
r2-r3	87%	.73

Table B4.

Item 3 FA Rater Agreement

Rater Combination	Percent Agreement	Kappa
r1-r2	80%	.64
r1-r3	80%	.62
r2-r3	92%	.83

Table B5.

Item 4 FA Rater Agreement

Rater Combination	Percent Agreement	Kappa
r1-r2	82%	.62
r1-r3	78%	.56
r2-r3	93%	.85

Table B6.

Item 5 FA Rater Agreement

Rater Combination	Percent Agreement	Kappa
r1-r2	83%	.65
r1-r3	85%	.68
r2-r3	93%	.84

SCI Rater Agreement

Table B7.

Overall SCI Rater Agreement

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

Table B8.

Item 1 SCI Rater Agreement

Rater Combination	Percent Agreement	Kappa
r1-r2	85%	.72
r1-r3	78%	.62
r2-r3	90%	.83

Table B9.

Item 2 SCI Rater Agreement

Rater Combination	Percent Agreement	Kappa
r1-r2	82%	.34
r1-r3	82%	.29
r2-r3	92%	.54

Table B10.

Item 3 SCI Rater Agreement

Rater Combination	Percent Agreement	Kappa
r1-r2	87%	.39
r1-r3	87%	.42
r2-r3	92%	.40

Table B11.

Item 4 SCI Rater Agreement

Rater Combination	Percent Agreement	Kappa
r1-r2	73%	.40
r1-r3	63%	.21
r2-r3	83%	.41

Table B12.

Item 5 SCI Rater Agreement

Rater Combination	Percent Agreement	Kappa
r1-r2	87%	.76
r1-r3	68%	.47
r2-r3	82%	.69

Appendix C: Think-Aloud Interview protocols

Think-Aloud Interview Protocol 1: After First administration

- What did you think of the survey overall?
- What do you think this survey is trying to measure or get at?
- Do you think it is designed in a way that does this well?
- Can you talk me through you thought process as you were formulating a response to this question? (Prompts and probes from cognitive interview guide)
- What do you think was the most difficult question to respond to?
- What do you think the easiest question was?
- Do you have any suggestions?

Think-Aloud Interview Protocol 2: After Second administration

- What do you remember about taking the survey the first time? Has your perspective changed after having taken it a second time?
- Can you talk me through what you were thinking as you were responding to this question? (Prompts and probes from cognitive interview guide)
- "We're trying to assess science teachers' strategic knowledge with this survey. However, we're not saying that this is a judgment of how good a teacher you are, though we do think these things are related. If we give you too much leading info in the questions it could be problematic for how we assess your responses." (Show them the construct maps and scoring guides, as well as the descriptive stats of how many scored at each level.) Given this information, do you think you were given sufficient opportunity to convey your thoughts and ideas?
    - Do you think the survey is designed in a way that does this well?
- What frustrated you, if anything?
    - Did the frustration have an impact on how you responded to anything?
- What suggestions do you have?

Appendix D: Think-Aloud Interview Coding Framework

An asterisk (*) denotes a code that was generative (identified during coding) rather than *a priori*.

Passages could be assigned more than one coded (i.e., double coded) , and codes with sub-levels (e.g., General Reactions) could have passages coded at the highest and/or lower levels.

- Difficult
- Easy
- FASCI construct
- Response thought processes
    - content dependence
    - context dependence
- Frustrations
- General Reactions
    - negative
    - positive
- *LA Program
- Respondent Background
    - *motivation to teach
    - *practicum experiences
    - *School of Education courses
    - teaching experience
    - want to teach
- Respondent change from pre to post
- Suggestions

Appendix E: The Reformed Teaching Observation Protocol

Reformed Teaching Observation Protocol (RTOP)

Daiyo Sawada External Evaluator
Michael Piburn Internal Evaluator

and

Kathleen Falconer, Jeff Turley, Russell Benford and Irene Bloom
Evaluation Facilitation Group (EFG)

Technical Report No. IN00-1
Arizona Collaborative for Excellence in the Preparation of Teachers
Arizona State University

I. BACKGROUND INFORMATION

Name of teacher _____ Announced Observation? _____ (yes, no, or explain)

Location of class _____ (district, school, room)

Years of Teaching _____ Teaching Certification _____ (K-8 or 7-12)

Subject observed _____ Grade level _____

Observer _____ Date of observation _____

Start time _____ End time _____

II. CONTEXTUAL BACKGROUND AND ACTIVITIES

In the space provided below please give a brief description of the lesson observed, the classroom setting in which the lesson took place (space, seating arrangements, etc.), and any relevant details about the students (number, gender, ethnicity) and teacher that you think are important. Use diagrams if they seem appropriate.

Record here events which may help in documenting the ratings.

Time	Description of Events

III. LESSON DESIGN AND IMPLEMENTATION

		Never Occurred				Very Descriptive
1)	The instructional strategies and activities respected students' prior knowledge and the preconceptions inherent therein.	0	1	2	3	4
2)	The lesson was designed to engage students as members of a learning community.	0	1	2	3	4
3)	In this lesson, student exploration preceded formal presentation.	0	1	2	3	4
4)	This lesson encouraged students to seek and value alternative modes of investigation or of problem solving.	0	1	2	3	4
5)	The focus and direction of the lesson was often determined by ideas originating with students.	0	1	2	3	4

IV. CONTENT

Propositional knowledge

6)	The lesson involved fundamental concepts of the subject.	0	1	2	3	4
7)	The lesson promoted strongly coherent conceptual understanding.	0	1	2	3	4
8)	The teacher had a solid grasp of the subject matter content inherent in the lesson.	0	1	2	3	4
9)	Elements of abstraction (i.e., symbolic representations, theory building) were encouraged when it was important to do so.	0	1	2	3	4
10)	Connections with other content disciplines and/or real world phenomena were explored and valued.	0	1	2	3	4

Procedural Knowledge

11)	Students used a variety of means (models, drawings, graphs, concrete materials, manipulatives, etc.) to represent phenomena.	0	1	2	3	4
12)	Students made predictions, estimations and/or hypotheses and devised means for testing them.	0	1	2	3	4
13)	Students were actively engaged in thought-provoking activity that often involved the critical assessment of procedures.	0	1	2	3	4
14)	Students were reflective about their learning.	0	1	2	3	4
15)	Intellectual rigor, constructive criticism, and the challenging of ideas were valued.	0	1	2	3	4

Continue recording salient events here.

Time	Description of Events

V. CLASSROOM CULTURE

	Communicative Interactions	Never Occurred				Very Descriptive
16)	Students were involved in the communication of their ideas to others using a variety of means and media.	0	1	2	3	4
17)	The teacher's questions triggered divergent modes of thinking.	0	1	2	3	4
18)	There was a high proportion of student talk and a significant amount of it occurred between and among students.	0	1	2	3	4
19)	Student questions and comments often determined the focus and direction of classroom discourse.	0	1	2	3	4
20)	There was a climate of respect for what others had to say.	0	1	2	3	4
	Student/Teacher Relationships					
21)	Active participation of students was encouraged and valued.	0	1	2	3	4
22)	Students were encouraged to generate conjectures, alternative solution strategies, and ways of interpreting evidence.	0	1	2	3	4
23)	In general the teacher was patient with students.	0	1	2	3	4
24)	The teacher acted as a resource person, working to support and enhance student investigations.	0	1	2	3	4
25)	The metaphor "teacher as listener" was very characteristic of this classroom.	0	1	2	3	4

Additional comments you may wish to make about this lesson.