2012

Scrutinizing a Measure of Science and Mathematics Teacher Knowledge: Implications for Claims of Validity

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

Scrutinizing a Measure of Science and Mathematics Teacher Knowledge: Implications for Claims of Validity

University of Colorado Denver
School of Education & Human Development

Robert M. Talbot III

Math and Science Learning and Education
COLLEGE OF LIBERAL ARTS AND SCIENCES
UNIVERSITY OF COLORADO **DENVER**

Abstract

Defining and measuring teacher knowledge is a challenging endeavor. The instrumentation from which these measures are derived must be valid. This study provides a critical investigation of the validity of an instrument that was developed to determine the effect of a teacher education program on novice science and mathematics teachers' *Strategic Knowledge* (SK). It was found that there are issues with respect to the validity of this instrument. These issues center on the reliability of scores, the effect of embedding specific content into the instrument, and the lack of convergent validity evidence from observations of practice. These findings and the development of the validity argument suggest some concrete recommendations for others who are engaged in similar measurement efforts.

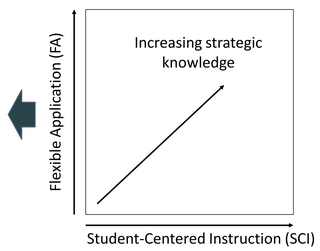
Background

•The Flexible Application of Student-Centered Instruction (FASCI) instrument was designed to measure the Strategic Knowledge (SK) of participants in the CU-Boulder Learning Assistant (LA) Program (Otero et al., 2006)

•The FASCI was designed to be "content neutral" in order to be useful for measuring levels of SK in individuals who come from a variety of STEM disciplines

The Strategic Knowledge Construct

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.



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

References

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

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

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

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

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.

A Framework for Examining Validity

***Validity*: the degree to which evidence and theory support the interpretation of test scores entailed by the proposed test uses (AERA, APA, & NCME, 1999; Kane 2006)**

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 these 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*]

The FASCI Items

The scenario-based items on the FASCI all have a common form

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.

A classroom scenario is presented (e.g. "Students are working in groups of four to discuss a conceptual question you provided them at the beginning of class.")

•How might this activity facilitate student learning?

A potential obstacle to learning is introduced (e.g. "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.")

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

•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?

Adding Content to the Items

Hypothesis: placing the item scenarios in specific science content (e.g., physics) would access a "more sophisticated" SK for those with expertise in that content area (e.g., physics teachers).

To test this hypothesis, a content-neutral version and a content-specific version (based in physics) were administered at random to a sample of pre-service science teachers.

•Adding content introduced a source of construct-irrelevant variance into the items.
•Score reliability was lower for content-specific items relative to the content-neutral items
•Average scores between versions were significantly different on the SCI dimension, but not on the FA dimension.

Observing SK in Practice

A sample (n=18) of practicing novice science and math teachers responded to the neutral-FASCI and were observed three times using the Reformed Teaching Observation Protocol (RTOP; Sawada et al., 2002).

•Consistency in RTOP/FASCI ratings were observed *to a limited degree*.
•Differences in characterizations could be due to differences in the teaching contexts observed.

Score Reliability

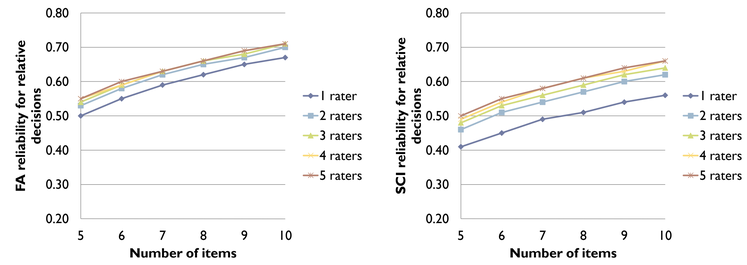
Score reliability was examined in three ways:

1. Inter-rater reliability in scoring (80-91% agreement, 0.63-0.82 kappa between 3 raters on FA dimension; 76-88% agreement, 0.40-0.57 kappa on SCI dimension)
2. Measure of internal consistency (Cronbach's alpha 0.45-0.62 on FA dimension, 0.41-0.51 on SCI dimension)
3. Analysis of error variance in scores based on Generalizability Theory (G-Theory; Brennan, 2001)

Generalizability Theory Analyses

For the FA dimension, the bulk if score variance was in *person x item* interactions. Doubling the number of items from 5 to 10 would theoretically increase score reliability to 0.71

For the SCI dimension, the bulk if score variance was in *items* and *person x item* interactions. Doubling the number of items from 5 to 10 would theoretically increase score reliability to 0.64



Items need to be better specified in order to reduce error variance and increase score reliability

Conclusions

The scenario-based items on the FASCI are sensitive to specification

Recommendations:
•Begin instrument development with the validity argument in mind
•Collect validity evidence early and often throughout development
•For instruments using open-ended items:
•Specify item design carefully and analyze that specification often
•Use defensible criteria for recruiting and training raters

This work is funded by the LA-TEST project, NSF grant ESI-TPC 0554616 (research conducted at CU Boulder)

Special thanks to Derek Briggs and Valerie Otero at CU Boulder