Validity Evidence for Forced-Choice and Mixed-Format Knowledge Assessments

Cari F. Herrmann Abell

BSCS Science Learning

Part of the Division D Working Group Roundtable Session "Practical guidance and best practices for gathering validity evidence based on assessment type"

2021 AERA Virtual Annual Meeting

April 8-12, 2021

Abstract:

In the last twenty-five years, the discussion surrounding validity evidence has shifted both in language and scope, from the work of Messick and Kane to the updated Standards for Educational and Psychological Testing. However, these discussions haven't necessarily focused on best practices for different types of instruments or assessments, taking into account specific concerns for each instrument type. The purpose of the working group roundtable was to (1) illustrate validation activities and evidence for 4 distinct types of measures (forced-choice and mixed format knowledge assessments, Likert/Rating Scale instruments, performance assessments, and observation protocols), (2) facilitate a discussion of best practices for each of the measurement contexts, and (3) foster possible collaborations for dissemination of best practices for these contexts. This paper focuses on validation activities for forced-choice and mixed format knowledge assessments.

Introduction:

The Assessing Students' Progress on the Energy Concept (ASPECt) project is the context for the first paper in a roundtable session focused on illustrating validity work for different types of measures and facilitating discussions around best practices. The goal of the ASPECt project is to develop assessments to measure late elementary, middle, and high school students' understanding of energy. It currently has two main components.

The first component is a set of three instruments that were constructed from a bank of 359 distractor-driven, multiple-choice items to assess students' progress on core energy ideas. The items are aligned to a learning progression for energy that is made up of three levels of complexity. The complexity levels progress from a basic level focusing on simple energy relationships and easily observable effects of energy processes to an intermediate level focusing on more complex energy concepts and applications to an advanced level focusing on still more complex energy concepts, often requiring an atomic/molecular model to explain phenomena (Herrmann-Abell & DeBoer, 2018).

The second component is a set of scenario-based tasks (SBTs) that are aligned to the three dimensions of science learning described in the Next Generation Science Standards (NGSS Lead States, 2013). Unlike the content-focused items in the first component of the ASPECt project, the SBTs require students to use science and engineering practices and crosscutting concepts along with disciplinary core ideas to make sense of energy-related phenomena (Herrmann-Abell, Hardcastle, & DeBoer, 2020). The tasks present a phenomenon or scenario followed by a series of constructed-response and multiple-choice items. The items within a SBT move students through a process of phenomenon/problem introduction, sense-making, and final resolution.

Validity framework. Assessment validation is typically viewed as the development of an evidence-based argument (Kane, 2006; 2013) in which claims about the use of assessments are made and supported by evidence. The updated *Standards for Educational and Psychological Testing* list five types of evidence for a validity argument (AERA, APA, & NCME, 1999/2014).

1. *Evidence based on test content* involves an analysis of the relationship between the test content and the targeted construct to be measured and is commonly considered an alignment analysis.
2. *Evidence based on response processes* involves analyzing student responses to investigate the extent to which the processes students use in responding fit the expectations that are part of the construct.
3. *Evidence based on internal structure* focuses on the relationship between items and how those items map on to the construct and may also include an investigation into differential item functioning.
4. *Evidence based on relations to other variables* involves the relationship between test scores and other variables external to the test, including test scores on other tests that measure the same construct.
5. *Evidence based on consequences of testing* focuses on whether the intended consequences of testing are realized, which in our case, is whether instruction using materials aligned to NGSS lead to improved performance on SBTs and content-focused items.

In terms of the types of validity evidence described above, our validity work has focused on gathering evidence based on test content, response process, and internal structure. Table 1 summarizes the types of evidence, the validity argument, and the methods we have used to obtain this evidence. Our work has not yet focused on evidence based on relations to other variables or consequences of testing. The last two rows in Table 1 describe the methods we will use to collect evidence based on relations to other variables and consequences of testing.

Table 1: *Types of validity evidence and methods for obtaining evidence (Methods in italics will be completed in the future.)*

Evidence based on:	Validity argument	Methods for obtaining evidence
Test content	Assessments are appropriate for measuring the targeted construct	Review by external panel of experts using evaluative criteria
Response processes	Assessments tap into the intended cognitive processes	Think-aloud protocols
Internal structure	Relations among tasks and items reflect those expected from theory	Rasch Model Fit and correlations between items and tasks
Relations to other variables	Relations of scores to other variables are consistent with those expected from theory	*Correlations with* • *Reading & writing ability* • *Performance on content-focused items*
Consequences of testing	NGSS-aligned instruction leads to increased performance on SBTs	*Instructional sensitivity using pre- and post-tests*

Validity evidence based on test content:

The primary source of evidence based on test content for both ASPECt components was obtained from an expert review. A panel of experts consisting of educators, scientists, and assessment specialists were consulted to evaluate (1) the appropriateness and scientific accuracy of the assessment contexts, (2) the alignment of the assessments to the learning goals, (3) the fairness and comprehensibility of the assessments, and (4) the fairness, grade-appropriateness, and usability of the scoring rubrics. Overall, reviewers agreed with our alignments to the targeted learning goals and thought that the assessment contexts were appropriate and engaging. Feedback that they provided about the comprehensibility of the assessments and usability of the rubrics was used to inform revisions to the assessments and rubrics.

Validity evidence based on response processes:

For the multi-format scenario-based tasks, we conducted think aloud interviews with students to ensure that they were using the targeted practices, crosscutting concepts, and disciplinary core ideas when responding to the items within the tasks. A total of nine students from two elementary schools, six students from one middle school, and nine students from one high school participated in the interviews. Interviews ranged from 15 to 50 minutes long. During the interview, students were asked to think aloud as they responded to between 1 and 4 assessment tasks. Overall, students found the task scenarios to be familiar and engaging, and the way in which they articulated their thought processes indicated that they were using their understanding

of the targeted practices, crosscutting concepts, and disciplinary core ideas while completing the tasks. During the interviews, the students also described comprehensibility issues with the way some of the questions were worded and pointed out challenges with using a drawing tool implemented in some tasks. We used this feedback to revise the tasks before field testing them with a larger student population.

Validity evidence based on internal structure:

We conduct field tests with students from across the U.S. to obtain data that can be used to provide evidence based on internal structure. The content-focused multiple-choice items from the first component of the ASPECt project were field tested with over 20,000 students in grades four through twelve. We used Rasch analysis to estimate item and person measures. Figure 1 shows the Wright map for the three instruments that are part of this component (Hardcastle, Herrmann-Abell, & DeBoer, 2017). The map shows that each test contains items with a range of difficulties well matched to the students' abilities and the tests become increasingly difficult as one progresses from basic, intermediate, to advanced. Additionally, the results largely supported our hypothesized learning progression that indicates students' growth of understanding progresses from an understanding of forms and transformations of energy to energy transfer to conservation while also progressing along a separate dimension of cognitive complexity (Herrmann-Abell & DeBoer, 2018).

Field tests of the mixed-format scenario-based tasks from the second component of the ASPECt project are currently being conducted but we have started to collect validity evidence about the rubrics used to evaluate students written explanations and arguments (Herrmann-Abell, Hardcastle, & DeBoer, 2020). The rubrics were based on the claim, evidence, reasoning (CER) framework (McNeill & Krajcik, 2011) but included a science ideas category. Therefore, our framework has four categories: a claim that answers the question, evidence that supports the claim, stating or using relevant science ideas, and reasoning that links the evidence and science ideas to the claim. Using Rasch analysis of pilot test data, we evaluated the rubric categories and found that rubric categories fit well to the Rasch model. As shown in the Wright map in Figure 2, categories clustered in a hierarchy of difficulty in which reasoning and applying science idea categories were more difficult than citing evidence, which were more difficult than making a claim (Hardcastle, Herrmann-Abell, & DeBoer, 2021). The observed hierarchy in difficulty of categories is consistent with other studies (e.g., Gotwals & Songer, 2013; Jin, Yan, Mehl, Llort, & Cui, 2020) and adds to the validity argument for the tasks as measures of students' ability to write scientific explanations and arguments about energy-related phenomena.

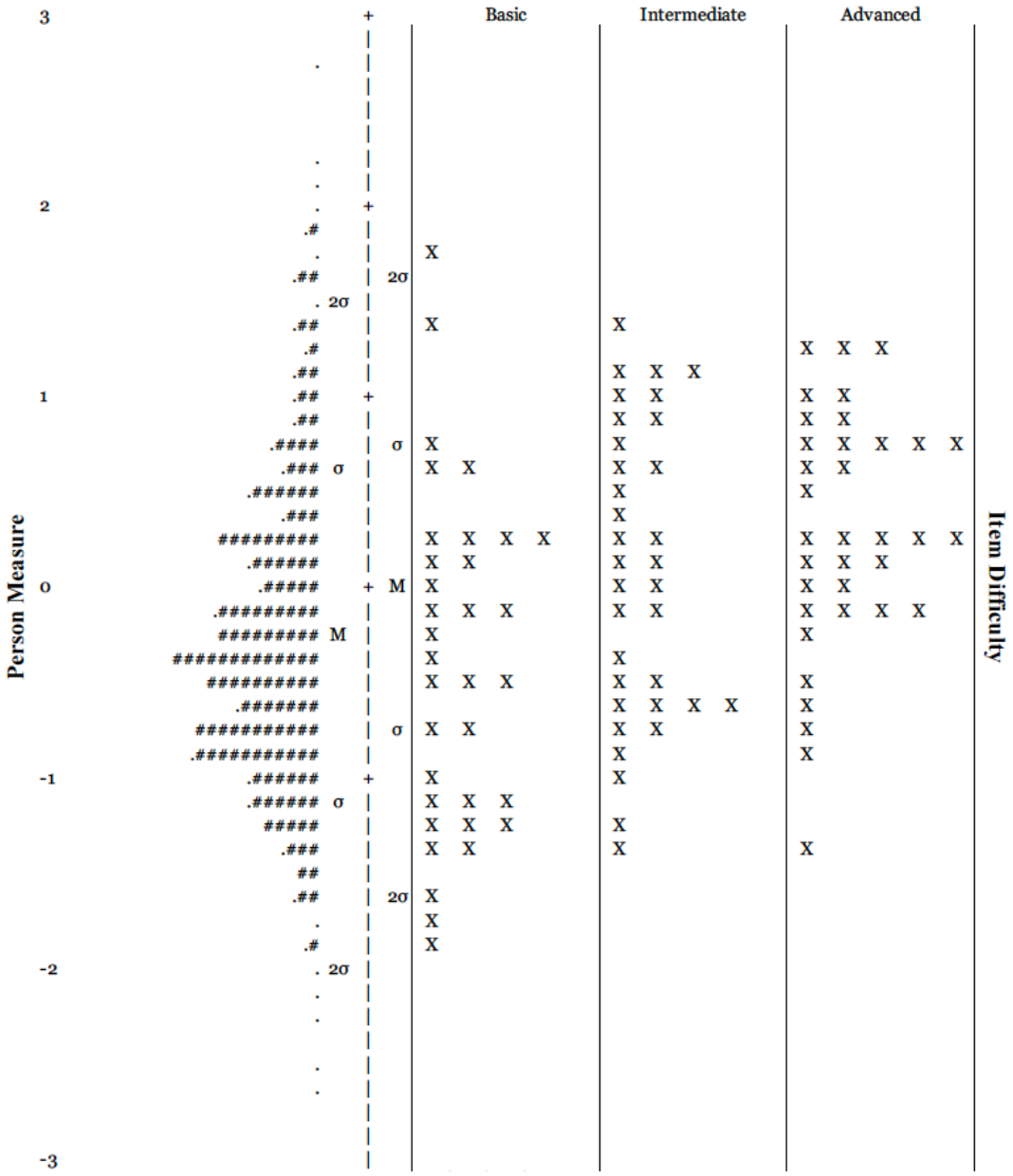

Figure 1: Wright Map of the basic, intermediate, and advanced instruments from the first ASPECt component. X = item, "#" = 8 students, "." = 1-7 students, M = mean, σ = standard deviation (Hardcastle, et al., 2017).

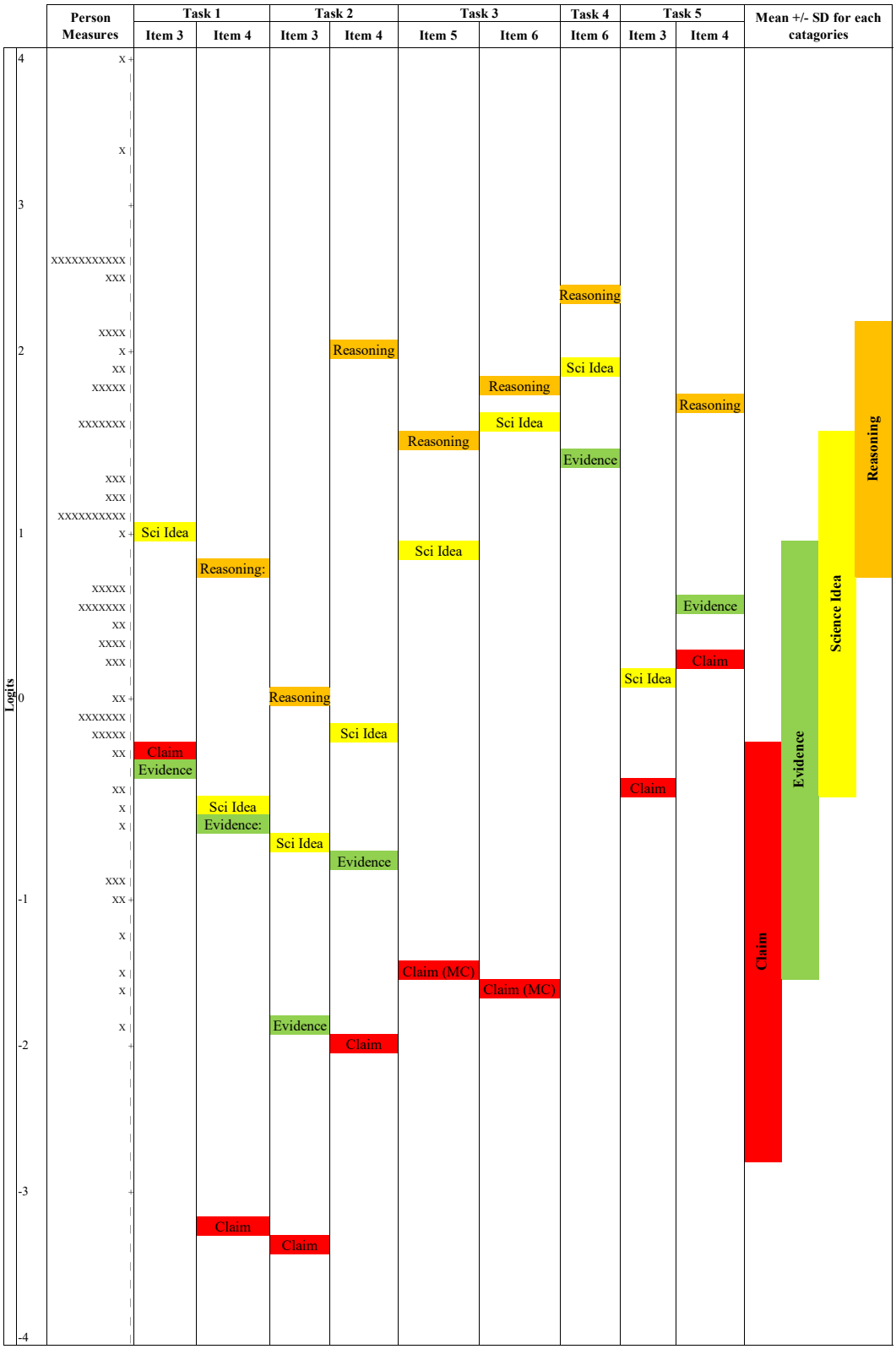

Figure 2: Wright Map showing the difficulties of claim, evidence, science idea, and reasoning categories from SBTs about chemical reactions and energy (Hardcastle, et al., 2021).

Validity evidence based on relations to other variables:

As mentioned above, we have not yet collected evidence base on relations to other variables. We plan to address this with the data collected from the current field test which includes the mixed-format scenario-based tasks and a subset of the content-focused multiple-choice items that focus on the same science ideas as the SBTs. We will look at the correlation coefficients of students' scores on the SBTs and the content-focused items, and we will use Rasch modeling to investigate the overall dimensionality of the field tests.

We also hope to explore the relationship between performance on SBTs and English reading and writing ability in a future project. This is especially important in the case of SBTs because past research has shown that these types of assessments, which have increased reading and writing demands, are more difficult than traditional science assessments (e.g. Penuel, Turner, Jacobs, Van Horne, & Sumner, 2019; Gane, McElhaney, Zaidi, & Pellegrino, 2018). According to the *Standards for Educational and Psychological Testing*, ‘‘for all test takers, any test that employs language is, in part, a measure of language skills,’’ (AERA, APA, & NCME, 1999, p 91). However, we do not want English language fluency to get in the way of students being able to demonstrate their three-dimensional understanding of science when responding to the SBTs.

Validity evidence based on consequences of testing:

A major challenge NGSS assessment developers have faced in studying the validity of these mixed-format scenario-based tasks is finding student populations that have experienced three-dimensional instruction and have had the opportunity to learn the knowledge and skills required to be successful on NGSS-aligned assessments. A review of the literature has shown that very few projects involving the development of NGSS-aligned assessments have concentrated on providing empirical evidence based on consequences of testing. Most projects have provided evidence based on test content and internal structure by making explicit the details of the development procedure and on having the assessments reviewed by a panel of experts (e.g., Harris, Krajcik, Pellegrino, & DeBarger, 2019; Underwood, Posey, Herrington, Carmel, & Cooper, 2018). A contributing factor to the lack of evidence based on consequences of testing is the fact that NGSS is still relatively new and there are few NGSS-aligned curriculum materials available for use. As more high-quality NGSS-aligned curriculum materials are developed and implemented, more students will have the opportunity to learn the knowledge and skills in NGSS and evidence based on consequences of testing will be easier to obtain.

In a future project, we hope to collect pre- and post-test data from students experiencing high-quality NGSS-aligned instruction and look at the differences in performance to investigate the instructional sensitivity of the SBTs (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002; Polikoff, 2010; Ruiz-Primo, Li, Wills, Giamellaro, Lan, Mason, & Sands, 2012). We are also interested in comparing the relative instructional sensitivity of the SBTs and the content-focused items.

Significance:

A review conducted by Della-Piana, Gardner, & Mayne (2020) identified shortfalls in validity evidence for science achievement assessments over an 11-year period. They found that some types of evidence are more frequently reported than others. In particular, they found that consequences of testing evidence and evidence related to the relationship of the assessment to

external variables to be underreported. We have identified challenges in studying these types of evidence that stem from students' lack of opportunity to learn the knowledge and skills outlined by the newly-established *Next Generation Science Standards* and from increased reading and writing demands of SBTs. We hope that sharing our methods of collecting validity evidence for mixed-format science assessments and the challenges that we face in studying the validity of these types of assessments contribute to a discussion around best practices.

Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A180512 to the BSCS Science Learning and through Grant R305A120138 to the American Association for the Advancement of Science. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

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

Della-Piana, G.M., Gardner, M.K., & Mayne, Z.R. (2020). Validity Evidence in Science Achievement Assessments Found in a Sample of Published Research Articles on Science Teaching. *Journal for STEM Education Research*. *3*, 279–294.

Gane, B. D., McElhaney, K.W., Zaidi, S. Z., Pellegrino, J. W. (2018, March). *Analysis of student and item performance on three-dimensional constructed response assessment tasks*. [Paper Session]. NARST Annual International Conference, Atlanta, GA.

Gotwals, A. W., & Songer, N. B. (2013). Validity evidence for learning progression-based assessment items that fuse core disciplinary ideas and science practices. *Journal of Research in Science Teaching*. *50*(5), 597-626.

Hardcastle, J., Herrmann Abell, C. F., & DeBoer, G. E. (2017, Apr 25). *Validating an Assessment for Tracking Students' Growth in Understanding of Energy from Elementary School to High School* [Paper Session]. NARST Annual International Conference, San Antonio, TX.

Hardcastle, J., Herrmann Abell, C. F., & DeBoer, G. E. (2021, Apr 7 - 10). *Validating a Claim-Evidence-Science Idea-Reasoning (CESR) Framework for use in NGSS assessment Tasks* [Paper Session]. NARST Annual International Conference, Virtual.

Harris, C.J., Krajcik, J.S., Pellegrino, J.W., DeBarger, A.H. (2019). Designing Knowledge-In-Use Assessments to Promote Deeper Learning. *Educational Measurement Issues and Practice*. 38 (2), 53-67.

Herrmann Abell, C.F. & DeBoer, G.E. (2018). Investigating a Learning Progression for Energy Ideas from Upper Elementary Through High School. *Journal of Research in Science Teaching, 55*(1), 68-93.

Herrmann Abell, C. F., Hardcastle, J., & DeBoer, G. E. (2020, Apr 17 - 21). *Developing Next Generation Science Standards–Aligned Tasks to Assess Elementary School Students' Ability to Explain Energy-Related Phenomena* [Paper Session]. AERA Annual Meeting, San Francisco, CA. http://tinyurl.com/v3pp4d9 (Conference Canceled).

Jin, H., Yan, D., Mehl, C. E., Llort, K., & Cui, W. (2020). An Empirically Grounded Framework That Evaluates Argument Quality in Scientific and Social Context. *International Journal of Science and Mathematics Education*. *19*, 681–700.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.) *Educational Measurement* (4th Ed., pp.17-64). Westport, CT: Praeger Publishers.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1–73.

McNeill, K., & Krajcik, J. (2011). *Supporting grade 5-8 students in constructing explanations in science: The claim, evidence and reasoning framework for talk and writing*. Boston, MA: Pearson Education.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. Washington DC: The National Academies Press.

Penuel, W.R., Turner, M.L., Jacobs, J.K., Van Horne, K., & Sumner, T. (2019b). Developing tasks to assess phenomenon-based science learning: Challenges and lessons learned from building proximal transfer tasks. *Science Education*, 103(6), 1367-1395.

Polikoff, M. S. (2010) Instructional Sensitivity as a Psychometric Property of Assessments. *Educational Measurement: Issues and Practice, 29*(4), 3–14

Ruiz-Primo, M.A., Shavelson,R.J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching, 39*(5), 369-393.

Ruiz-Primo, M.A., Li, M., Wills, K., Giamellaro, M., Lan, M-C., Mason, H., Sands, D. (2012). Developing and evaluating instructionally sensitive assessments in science. *Journal of Research in Science Teaching, 49*(6), 691-712.

Underwood, S.M., Posey, L.A., Herrington, D. G., Carmel, J. H., & Cooper, M.M. (2018). Adapting Assessment Tasks to Support Three-Dimensional Learning. *Journal of Chemical Education, 95* (2), 207-217.