



**THE MEASUREMENT, EVALUATION, STATISTICS, AND ASSESSMENT
DEPARTMENT**

Invite you to Attend the Final, Public Dissertation Defense for:

Courtney Castle

Titled:

**MEASURING MULTIDIMENSIONAL SCIENCE LEARNING: ITEM DESIGN, SCORING,
AND PSYCHOMETRIC CONSIDERATIONS**

The defense will be held on

Thursday, March 29th at 2pm

in

Campion 124

Dissertation Committee:

Dr. Henry Braun (Chair), Dr. Nathaniel Brown (Reader),
Dr. Katherine McNeill (Reader), Dr. Sue Doubler (Reader)

Abstract:

The Next Generation Science Standards propose a multidimensional model of science learning, comprised of Core Disciplinary Ideas, Science and Engineering Practices, and Crosscutting Concepts (NGSS Lead States, 2013). Accordingly, there is a need for student assessment aligned with the new standards. Creating assessments that validly and reliably measure multidimensional science ability is a challenge for the measurement community (Pellegrino, et al., 2014). Multidimensional assessment tasks may need to go beyond typical item designs of standalone multiple-choice and short-answer items. Furthermore, scoring and modeling of student performance should account for the multidimensionality of the construct.

This research contributes to knowledge about best practices for multidimensional science assessment by exploring three areas of interest: 1) item design, 2) scoring rubrics, and 3) measurement model. This study investigated multidimensional scaffolding and response format

by comparing alternative item designs on an elementary assessment of matter. Item variations had a different number of item prompts and/or response format. Observations about student cognition and performance were collected during cognitive interviews and a pilot test. Items were scored using a holistic rubric and a multidimensional rubric, and interrater agreement was examined. Assessment data was scaled with multidimensional scores and holistic scores, using unidimensional and multidimensional Rasch models, and model-data fit was compared.

Results showed that scaffolding is associated with more thorough responses, especially among low ability students. Students tended to utilize different cognitive processes to respond to selected-response items and constructed-response items, and were more likely to respond to selected-response arguments.

Interrater agreement was highest when the structure of the item aligned with the structure of the scoring rubric. Holistic scores provided similar reliability and precision as multidimensional scores, but item and person fit was poorer. Multidimensional subscales had lower reliability, less precise student estimates than the unidimensional model, and interdimensional correlations were high. However, the multidimensional rubric and model provide nuanced information about student performance and better fit to the response data.

Recommendations about optimal combinations of scaffolding, rubric, and measurement models are made for teachers, policymakers, and researchers.