Project Goals

The purpose of this project is to develop and test a new web-based platform to increase the quality of teacher-administered tests in science classrooms. It draws widely on wellspring of classroom teacher knowledge while employing the rigorous statistical methods used in standardized test creation and validation. The content focus is on the disciplinary core ideas for grades 6-8 physical science in the Next Generation Science Standards (NGSS). Teachers now spend an estimate 20% of their time in assessment yet have relatively few tools to draw upon when creating them. Over time, they learn adapt items from available curriculum materials and textbooks. On the other hand, standardized test developers have the benefit of expert item writers, long development cycles, a large, diverse student population, and sophisticated psychometric tools. This project combines these two approaches, drawing upon teachers to contribute their best items, then immediately piloting them using crowdsourced subjects. Psychometric analysis generates measures of item quality and then "recycles" items to participating teachers for improvement. In this way, a large test item bank will be constructed utilizing teacher input with each item possessing: appropriate reading levels, NGSS alignment, scientific accuracy, appropriate difficulty, high statistical discrimination, and minimal difference by gender, race, or ethnicity. Involvement in this project has potential benefits for teachers lacking formal training in assessment, familiarizing participants with the Next Generation Science Standards, and with the elements of high-quality test development.

The project will gauge the merits of a novel collaborative system for the development and validation of high-quality test items and assessment instruments. It will measure the degree to which teachers can generate effective items and improve existing items exhibiting problematic issues when given the guidance of rigorous psychometric measures that estimate item quality. It will build on earlier research showing that an adult, crowd-sourced sample works well as an initial proxy for grade 6-8 science students, allowing for extremely rapid feedback on item quality (often overnight), with item response theory computation used to establish item difficulty, item discrimination, guessing levels, and differential item functioning (gender and racial/ethnicity bias). In addition, computed measures of misconception strength, scientific correctness, reading level, and match to the NGSS will help to guide revision by teachers. Use of Bayesian futility analysis will "triage" items, minimizing costly testing of items when deemed unlikely to meet item quality criteria, lowering costs. Field testing with a large sample of grade 6-8 students will provide a final check on item quality. The project is a direct outgrowth of the proposer's earlier efforts, including the seminal video A Private Universe and psychometric modeling of misconceptions using multiple-choice assessments (receiving the Journal of Research in Science Teaching Award). Items will be developed much more inexpensively than by methods used for standardized test development. Two pairs (public-release and secure for chemistry and physics) of assessment instruments will be constructed and be freely available to science teachers for classroom use and by education researchers and curriculum developers. A system that provides quick feedback on item quality could potentially revolutionize university instruction and professional development opportunities in assessment. While starting with selected response (multiple-choice) items, the project will be able to implement a larger variety of formats in the future, incorporating automated approaches as they become available.

Major Accomplishments to Date:

1. Recruitment of educators to revise items
During the fall and winter of 2022/23, we recruited additional educators to revise 60 original assessment items. By January 2023, we received 200 revised versions of these items to test. Even with extensive recruitment efforts, only three teachers submitted their own items. These were included in the pilot and field tests.

2. Pilot testing of revised items

Two hundred revised items were compiled into 8 test forms of 25 items each, and two additional forms of 25 items each were created to include 50 original test items. Each of the ten forms also included a qualifying test that was created with ten original items as well as demographic questions. (60 total). All ten forms were then posted on the Amazon Mechanical Turk (AMT) website during March 2023. Data were collected from close to 150 subjects for each test form (see data in supplement).

3. Preparation of field tests

Results for each item tested on AMT were then analyzed using Classical Test Theory to determine psychometric profiles. All revised items with improved characteristics, their matching original item, and one matching item whose characteristics did not improve were included (see example in supplement.) Five test forms of 30 items each (150 items total) were created to be included in the field test.

4. Recruitment of middle school physical science educators for field test
During spring 2023, we recruited educators to participate in the field test. Approximately 10,000 emails were sent to middle school science teachers across the country, using email addresses provided from Market Data Retrieval. Teachers were also recruited at the 2023 NSTA National Conference in Atlanta. As a result of these efforts, 124 educators agreed to participate in the field test.

5. Implementation of field test with middle school physical science students. 10,323 forms were mailed to 124 educators in early April 2023. These teachers administered the tests to their students during normal classroom sessions. Because the field tests are designed to be taken by students as near to the end of their classes as possible, administration will conclude mid-June 2023. All field tests should be returned by the end of June. To date, 3,033 tests have been returned (see data in supplement.) Once their students' answer sheets are returned, educators are sent a document with an answer key and information about the NGSS Standard for each item that is being assessed (see sample in supplement.) Educators can use this document to review concepts covered in the assessment with their students.

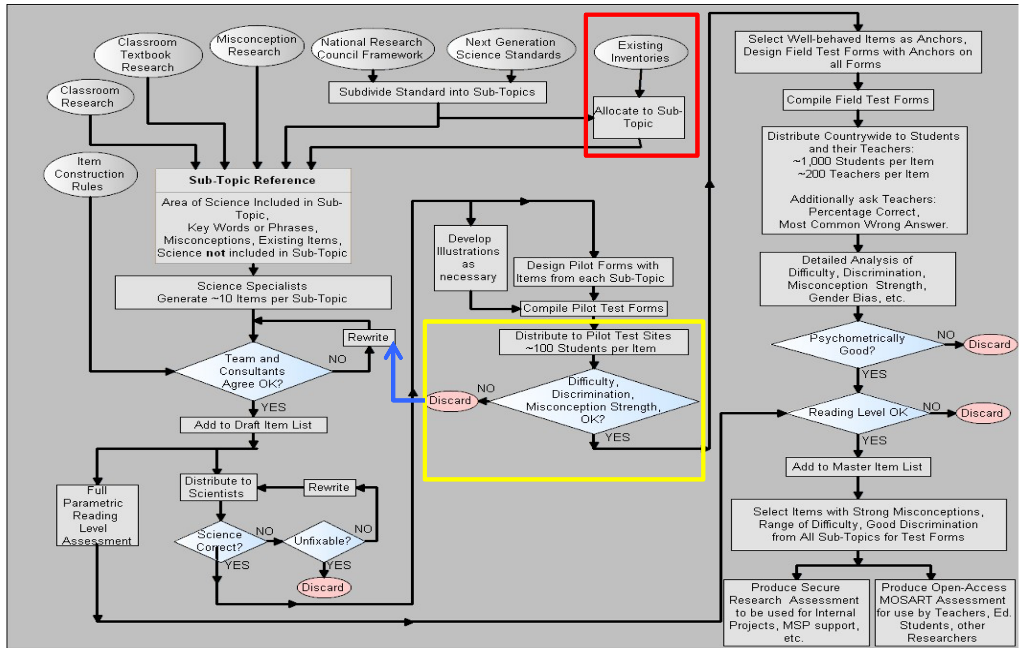
CONSTRUCT

Crowd-sourced Online Nexus for Science Teachers Researching and Upgrading Classroom Tests

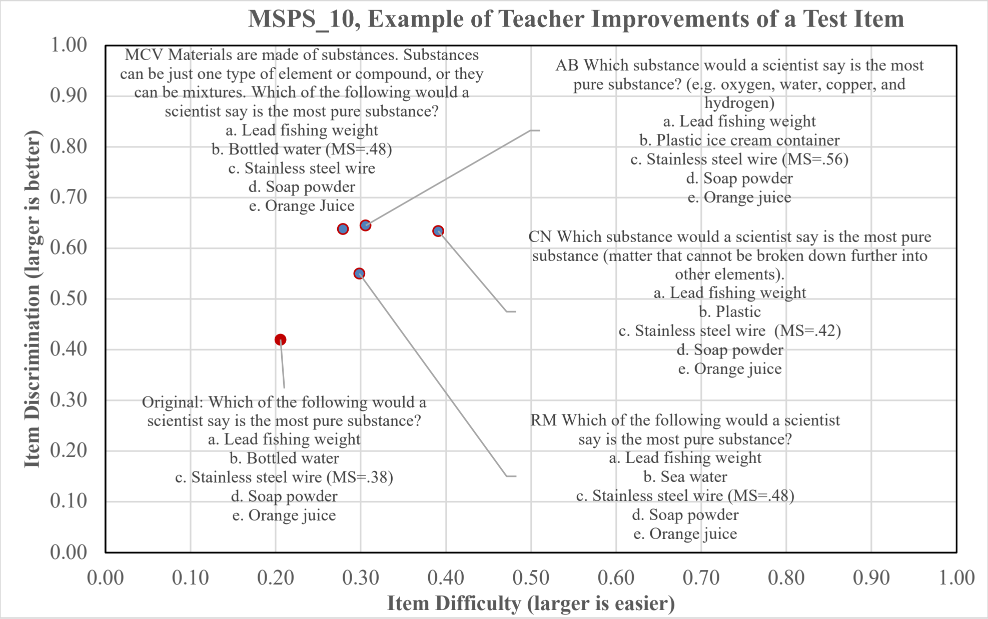
DRK-12, 1621210

Harvard-Smithsonian Center for Astrophysics, Cambridge, MA

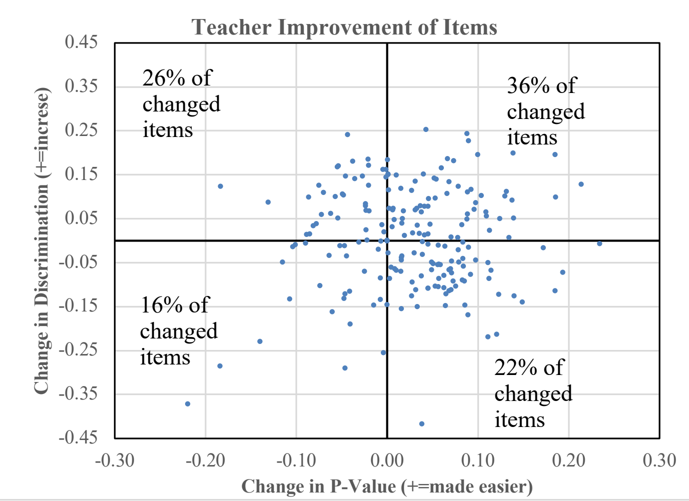
Philip M. Sadler, PI; Gerhard Sonnert, Co-I; Sue Sunbury, Project Manager
Cynthia Crockett, Science Education Specialist; Annette Trenga, research assistant; Rongxiu Wu, Post-Doctoral Fellow

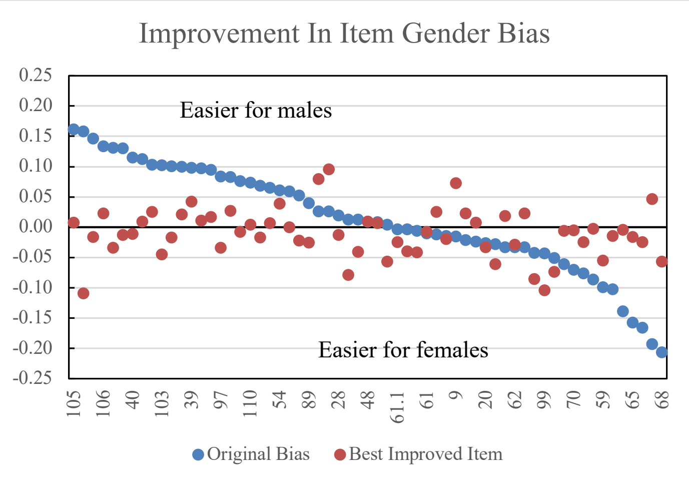

Classroom Textbook Research
Misconception Research
National Research Council Framework
Next Generation Science Standards
Existing Inventories
Classroom Research
Subdivide Standard into Sub-Topics
Allocate to Sub-Topic
Item Construction Rules
Sub-Topic Reference
Area of Science Included in Sub-Topic, Key Words or Phrases, Misconceptions, Existing Items, Science **not** included in Sub-Topic
Science Specialists Generate ~10 Items per Sub-Topic
Rewrite
Team and Consultants Agree OK?
NO
YES
Add to Draft Item List
Full Parametric Reading Level Assessment
Distribute to Scientists
Rewrite
Science Correct?
NO
YES
Unfixable?
NO
YES
Discard
Develop Illustrations as necessary
Design Pilot Forms with Items from each Sub-Topic
Compile Pilot Test Forms
Distribute to Pilot Test Sites ~100 Students per Item
Difficulty, Discrimination, Misconception Strength, OK?
NO
Discard
YES
Select Well-behaved Items as Anchors, Design Field Test Forms with Anchors on all Forms
Compile Field Test Forms
Distribute Countrywide to Students and their Teachers: ~1,000 Students per Item ~200 Teachers per Item
Additionally ask Teachers: Percentage Correct, Most Common Wrong Answer.
Detailed Analysis of Difficulty, Discrimination, Misconception Strength, Gender Bias, etc.
Psychometrically Good?
NO
Discard
YES
Reading Level OK
NO
Discard
YES
Add to Master Item List
Select Items with Strong Misconceptions, Range of Difficulty, Good Discrimination from All Sub-Topics for Test Forms
Produce Secure Research Assessment to be used for Internal Projects, MSP support, etc.
Produce Open-Access MOSART Assessment for use by Teachers, Ed. Students, other Researchers

Item Inventory Development Process. Red box highlights harvesting of items from existing inventories. Yellow box highlights a key "bottleneck" of pilot-testing solved by using Amazon Mechanical Turk Crowdsourcing Platform. Blue arrow shows path for teachers to rewrite items with poor psychometric item parameters. Red box shows entry point for item contributed from teachers' own item inventories.



Middle School Physical Science Test Item improved by 4 Teachers. The original item was deemed too difficult (P-value of 0.21) with low discrimination (D=0.44). Four teachers rewrote the item trying to make it easier and have a higher discrimination. Teacher CN did the best job.





Work Remaining:

1. Preparation and analysis of field test data
2. During the summer of 2023, field test data will be scanned and cleaned in preparation for analysis. During fall and winter of 2023/24, data will then be analyzed using Item Response Theory to determine psychometric properties for each item to answer research questions #2 and #3.
3. Reporting back to participating teachers of classroom results
4. Educators who submitted written or revised questions will be provided with data on the quality of each of their items determined by psychometric analysis (difficulty, discrimination, and differential item functioning - gender and racial/ethnicity bias), based on the results of the field test. Educators whose students participated in the field test will be provided with reports of aggregated field test data. Each report will include the correct answer and accompanying educational standard for each item as well as the percent of students who selected each of five choices for each item.
5. Development of public versions of assessments
6. Two instruments (25 questions each) will be constructed from relevant candidate items, both original and revised. The first test will include items representing the middle school NGSS physical science standard PS1: Matter and its Interactions, and the second test will include items for physical science standards PS2: Motion and Stability: Forces and Interactions, PS3: Energy, and PS4: Waves and Their Applications in Technologies for Information Transfer. Each test will include items with a range of difficulty, discrimination and misconception strength determined using differential item functioning (DIF) analysis, minimizing differences for groups underrepresented in science (e.g., females, ethnic and racial minorities, speakers of English as a second language). These tests, each with an accompanying document that includes an answer key, NGSS DCIs covered, and misconceptions probed, will be posted to the MOSART Self-Service website as a free resource for any interested educator and researcher.
7. Documentation, storage, and sharing of data
8. Dissemination of findings and item writing workshops for educators
9. We will continue to disseminate findings from this study by presenting workshops for educators in item writing and revising. A presentation has been submitted to NSTA for the October 25-28, 2023, conference in Kansas City.