Crowdsourcing can help create better science tests cheaper

August 31 2016

When it comes to developing test questions, there's the ordinary way and the fancy way. The ordinary way is to just make up questions and put them on the test. However, this can lead to questions that are misleading, confusing, or simply don't test for the knowledge you're trying to measure. The fancy way takes a lot of possible questions, tries them out on students, and whittles them down to the most useful. But this process is both time-consuming and expensive.

A group of researchers at the Harvard-Smithsonian Center for Astrophysics (CfA) has found a way for schools, professors, textbook publishers, and educational researchers to check the quality of their test questions that turns out to be both fast and cheap. It invokes the power of crowdsourcing.

"Crowdsourcing opens up a whole new possibility for people creating tests," says lead author Philip Sadler. "And instead of taking a semester or a year, you can do it in a weekend."

The CfA group has had a long-standing program of developing methodologically rigorous tests for various sciences and grade bands. The researchers evaluate new multiple-choice questions in a two-step process. First, they conduct pilot testing of lots of questions, developed by content experts, on a large number of students. Then they conduct field testing on 1,000-2,000 students. Using statistical analyses, they select the best questions for the final tests.

Sadler and his team investigated whether it was possible to replace the first step, pilot testing, with crowdsourcing. Crowdsourcing websites, like Amazon's Mechanical Turk, assign thinking tasks to a global community who receive small payments in return. For this study, the task of each participant was to answer a set of 25 multiple-choice life science questions developed for middle-school students.

The team evaluated a total of 110 multiple-choice questions both ways - traditional pilot testing and crowdsourcing - and compared the results. Since crowdsourcing participants are adults, while pilot testing was conducted with a sample of the target population (middle-school students), they wondered if the results would be similar. Perhaps surprisingly, the best test questions identified by crowdsourcing turned out to be high-quality questions for students too. Low-quality questions were poor for both adults and kids.

Sadler emphasizes that crowdsourcing can't entirely substitute for studying the target student population when producing high-quality tests. However, by using it as an early step, questions can be quickly evaluated for deletion, revision, or acceptance. Surviving questions can then undergo more rigorous testing.

"The key to creating good standardized tests isn't the expert crafting of every test question at the outset, but uncovering the gems hidden in a much larger pile of ordinary rocks," says co-investigator Gerhard Sonnert. "Crowdsourcing, coupled with using commercially available test-analysis software, can now easily identify promising candidates for those needle-in-a-haystack items."

A number of test developers could benefit from this new approach. For example, some schools are moving to standardize their exams and share them across the school system. Testing questions on their own students would let students know exactly what questions to expect on future

exams. Crowdsourcing offers a low-budget alternative.

In addition, curriculum developers and textbook authors can rapidly test and refine the questions they include in their materials. Educational researchers will be able to produce questions that more effectively measure changes in student knowledge. And professional development programs that now have teachers produce assessment questions for their students can, overnight, measure the performance of those questions.

The full results of this study are newly published in the journal *Educational Assessment* in a paper by Sadler, Sonnert, Hal Coyle (CfA) and Kelly Miller (Harvard John A. Paulson School of Engineering and Applied Sciences).

Provided by Harvard-Smithsonian Center for Astrophysics