‘Interim’ Tests Are Used Everywhere. Are They Good Measures of Student Progress?

By Sarah Schwartz — July 21, 2023 | Corrected: July 24, 2023

8 min read

E+/Getty

Corrected: A previous version of this story incorrectly described part of EdReports’ assessment review process. Independent psychometric experts score the tests on technical quality.

MAP, iReady, Star—these every-few-months tests are some of the most commonly used tools among the nation’s school districts, for a wide array of different purposes.

When school leaders say they champion “data-driven” instruction, they’re often talking about the data from these assessments: Some use the tests to get a sense of how well kids are likely to score on year-end exams; some, to figure out a subset of skills students struggle with that teachers should prioritize for reteaching; still others use them to try to gauge whether a new tutoring initiative is working. (Or all three.)

But for all of their heavy use, there’s little external evaluation of their technical properties, or how well they align to the most commonly used curriculum. Nearly all of them are proprietary, owned by private companies.

Now, a recent attempt to look under the hood of these tools, officially known as “interim assessments,” is on hold after several providers of the tests wouldn’t commit to participation.

The curriculum reviewer EdReports, known for its ratings of curriculum series, set its sights on evaluating interim assessments, a project that the group started working on in 2016.

The process that EdReports undertook, and the response from publishers, demonstrate the ways in which a key part of teaching and learning can be opaque for schools and the teachers who rely on the results.

The stakes are even higher now: Many districts are using the data from these tests to drive decisions about pandemic recovery.

“Districts are relying on this interim assessment data even more so to really gauge what happened during the pandemic,” said Eric Hirsch, the executive director of EdReports.

What interim assessments measure

Because most of these tests don’t have external evaluations to demonstrate their effectiveness, EdReports wanted to take a closer look. It planned to include educator evaluators—similar to its curriculum reviews—and feedback from testing experts through collaboration with the Center for Assessment, a New Hampshire-based consulting group.

But unlike curriculum reviews, the assessment reviews would have required publisher participation to gain access to internal information about assessment design—participation that EdReports didn’t get, for the most part.

Despite the clear value for district leaders, there's little incentive for publishers to participate in such a review, said Gregory Cizek, a professor of educational measurement and statistics at the University of North Carolina Chapel Hill.

"You have to be a pretty large, well-funded publisher who's pretty confident that you're going to look good," he said. "Because you don't want these independent published reviews to say that your test is terrible."

One wrinkle concerns the wide range of tools that fall under the moniker "interim assessment." Different interim assessments serve different purposes.

Some claim that they gauge how students are doing in a particular subject area, like math or science, in such a way that would predict student performance on a state summative assessment.

Others are more geared toward providing information that can directly influence teaching. These assessments might evaluate students' proficiency with specific skills—number sense, for example—so that teachers or interventionists can focus on those skills with students, said Erika Landl, a senior associate with the Center for Assessment who worked with EdReports.

One of the goals of the EdReports analysis was transparency—to make it clear to district leaders what these assessment providers were claiming they could do, and what they weren't.

"I think there's this black box about, 'What does this mean, and what should I do with it?'" said Landl, about interim assessment data.

She gave an example: As students progress throughout the school year, they will likely show improvement on an interim assessment that predicts performance on an end-of-year, summative test. But even though students are growing, they still might not have grown enough to reach the "proficient" level on that summative assessment—a distinction that can be confusing.

EdReports also wanted to determine whether these tests actually measured the things that they claimed to measure. If an assessment provider says that its test predicts performance on an end-of-year state test, does it? If the company says it can pinpoint the skills that were causing students to struggle, can it?

A few organizations do this kind of analysis. The Buros Center for Testing at the University of Nebraska-Lincoln has reviewed more than 2,800 commercially available tests, and their reviews are available for purchase. But many tests aren't included, said Cizek.

"They often struggle to get some of the most high profile testing programs to participate. For example, the SAT is not in there. The NAEP is not in there," he said, referring respectively to the college-admissions tests and the federally administered "nation's report card."

The National Center on Intensive Intervention also evaluates some interim assessments for validity—that they measure what they say they measure—and reliability: whether the tool produces consistent results. But similarly to the Buros Center reviews, not every interim assessment provider participates.

To that landscape, EdReports review would have included educator feedback in addition to evaluation from psychometricians. Just as with curriculum, Hirsch said, it's important for assessments to also be reviewed by "those who actually utilize them"—and to package it in a user-friendly way.

"Most districts lack the capacity and the expertise to put these different component pieces together, interpret the technical manuals, and put these in a synthesized report," he said.

EdReports' proposed assessment 'gateways'

Similar to how it designs curriculum reviews, EdReports planned to organize the assessment review around different categories, or "gateways."

The first assessed alignment to college- and career–ready standards, as well as accessibility.

A second gateway examined technical quality, scored by independent, national psychometric experts. These questions probe whether the tests actually measure what they claim to measure.

The third gateway would have evaluated the clarity and usability of score reports, a task undertaken by both educators and technical experts.

The process asked publishers to submit a wealth of information, some proprietary: item examples, the algorithms that power adaptive tests, technical guides, and evidence guides.

The ask of companies was “show us your work,” Hirsch said.

The process was “resource intensive and well worth it,” said Kristen Huff, the vice president of assessment and research at Curriculum Associates, which had signed on to be part of the review before it was tabled in May.“It forced us to take all of these rich rationales that we had in our minds and in our words and put them on paper.”

Other companies, though, were more hesitant. NWEA, which makes the suite of MAP tests, was invited to participate and declined. The company declined to comment. Renaissance, another assessment provider, hadn’t yet decided whether to participate in the review when EdReports paused the process.

“There’s value in this particular type of project because districts, states, others are very interested in how they evaluate one assessment provider over another. So I think it’s helpful to look at what those claims would be and whether those claims are met,” said Darice Keating, the senior vice president of government affairs at Renaissance.

But other existing processes already gather much of this information—like state requests for proposals, or the National Center on Intensive Intervention, noted Andy Frost, vice president of assessment product management at Renaissance. And it’s a “significant” effort to assemble all of the information that EdReports is requesting, he said.

“There would certainly be some utility. EdReports and the Center [are] tremendous organizations,” he said. “There would absolutely be some incremental value added. I think the question is how much and at what cost.”

Hirsch acknowledged that the process could be “onerous.”

“I certainly don’t want to state in any way that this is an easy process on a publisher,” he said.

Passing the baton to districts

Cizek, the UNC Chapel Hill professor, ultimately said getting companies to sign on is an uphill battle.

“It’s really a lose-lose for interim assessment providers,” he said. Determining alignment and validity takes time and money. After seeing the review criteria, publishers might know that they don’t meet some of the requirements—around accessibility, for example—and avoid the process as a result, he said.

“If their customers are happy with their product, why would they bother participating in this hassle?” he asked.

Several companies that chose not to participate in EdReports’ reviews said that the criteria have still affected their decisionmaking.

“We are looking to those criteria as we build our new assessments. That’s big,” said Laura Slover, the chief executive officer of CenterPoint Education Solutions, a nonprofit that works with districts to develop coherent instructional systems and creates assessments. She cited timing as the reason why the organization didn’t participate.

“To that extent, I think EdReports is already starting to have the impact, perhaps indirectly that they were hoping to have,” Slover said.

The time and effort EdReports and the Center for Assessment put into the project isn’t wasted, either.

Landl and Susan Lyons, a principal consultant for Lyons Assessment Consulting, recently released guidance for district leaders based on the criteria the group had put together for the EdReports reviews. Its aim is to be a version of the review process that districts can use to inform their decisionmaking—even if they don’t have access to testing experts, Landl said.

The tool is a way to pass the baton to districts Hirsch said, “and say, ‘You may not be able to answer all of the questions we were trying to answer. But you may be able to answer some of them.”