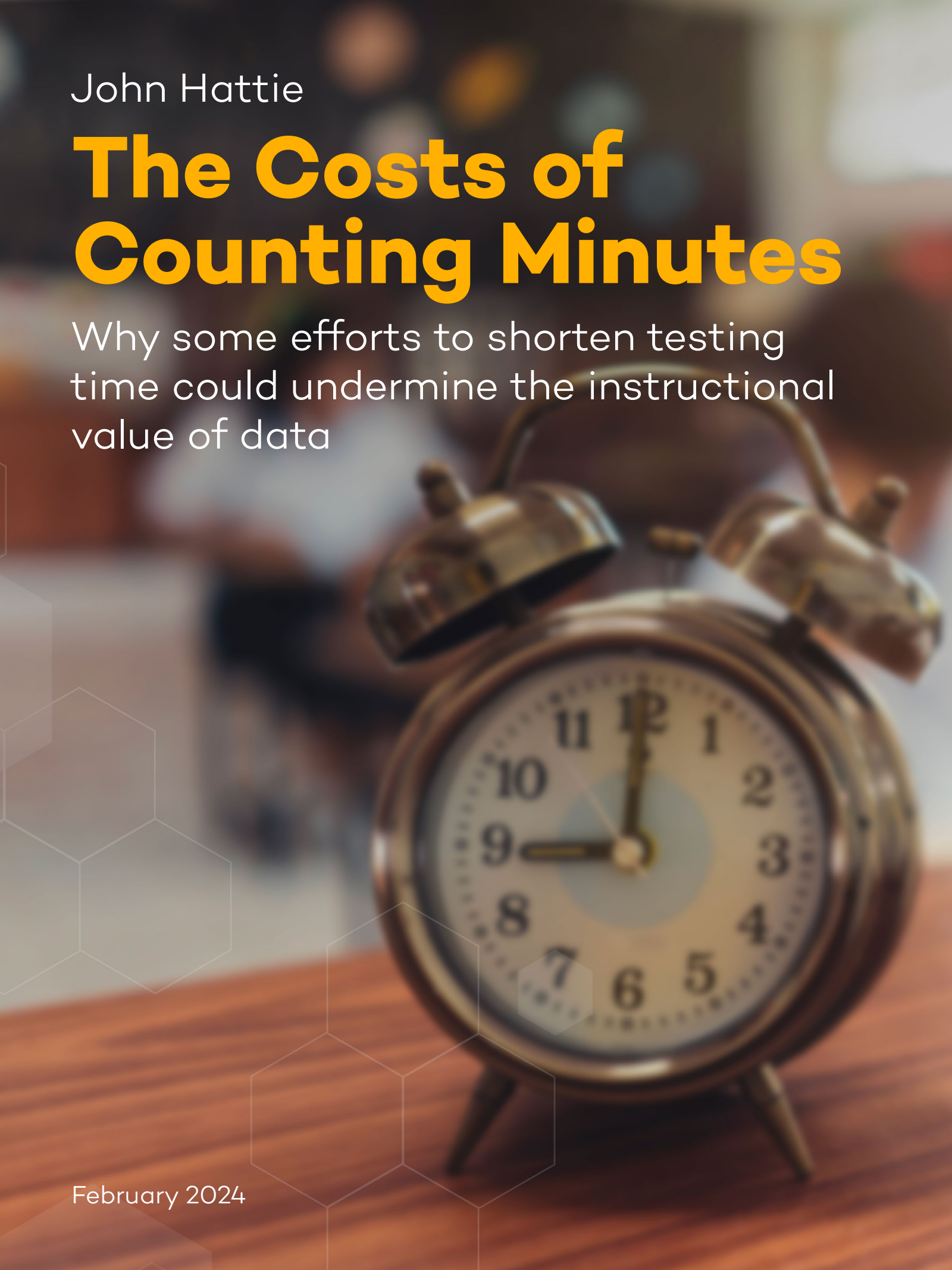
John Hattie

The Costs of Counting Minutes

Why some efforts to shorten testing time could undermine the instructional value of data



February 2024

The Costs of Counting Minutes

By: John Hattie

Following unprecedented disruptions to education due to COVID-19, parents, teachers, and education leaders want to know how students are doing. National assessments, like the Nation's Report Card (NAEP), do not reflect what is happening locally, and results from state summative assessments are often slow to arrive for both parents and teachers. As a result, some of the most widely used measures of student achievement and student growth are interim assessments, administered up to three times a year, that are meant to help gauge how students are doing and to provide timely data to enable teachers to drive instruction. These tests vary in length and coverage, impacting what can be done with the results.

Not all tests, however, are created equal.

Decisions about local assessment systems and interim assessments must center on the purpose of the assessment and relate to instruction and content area domains.

Those decisions should be focused on three things:

1. **Precision of the score.** How well does the test tell teachers how students are doing?
2. **The focus of the test.** Is there sufficient coverage of the key ideas?
3. **Actionable insights.** Can 1 and 2 provide the teacher sufficient information to drive instruction?

There is always a trade-off though – we do not want schools to over-assess and take critical time away from the teaching and learning. We must use fit-for-purpose tests that will reliably and dependably lead to optimal information and decisions. The tests should not be too long and over-precise nor too short and under-precise.

To consider the relationship between a test's length and the precision or reliability of the results, consider the three most commonly used interim assessments. Interim assessments such as i-Ready Diagnostic, NWEA MAP, and Renaissance Star have been used in K-8 classrooms for years, and each of these assessments are designed to have high reliability, meaning they do well in telling teachers how kids are doing.

One way to gauge reliability is the standard error of measurement (SEM). The SEM denotes upper and lower bounds around a student score, and the aim is to make this reasonably small. As a result, it will vary based on the number of items on a test and the scale of that test. The i-Ready scale has a larger SEM than other interim assessments, primarily because the i-Ready scale is much greater (100-800) compared to NWEA MAP (100-260).[1] That does not make i-Ready less reliable. For every test, the size of the scale, the length of the test, and the reliability of the test are tightly interrelated.

For tests with similar length and similar construction, like i-Ready Diagnostic and NWEA MAP, the reliability relative to the size of the scale is very similar. Renaissance Star, with a much shorter test, has a greater SEM relative to its scale score range. This is typical of tests of shorter length.

If the tests under consideration are all sufficiently reliable, the content coverage and the data teachers receive become much more critical to making decisions on what is the best interim assessment for the purpose. Tests have different content coverage and different lengths based on the intended purpose and use of the test. As a result, not all tests can provide the same detailed view of how students are doing, and not all tests support instruction in the same way.

"Not all tests can provide the same detailed view of how students are doing, and not all tests support instruction in the same way."

Assessments that tell you generally if a student is on grade level or in the 60th percentile are helpful if you want to know how the student is doing in general. To use data for instruction, teachers need more specificity. For instance, teachers should know not only that a student should work on math but more specifically that a student needs to work on identifying types of triangles.

The i-Ready Diagnostic and NWEA MAP are longer assessments than other commonly used interims or screeners, like Renaissance Star or Fastbridge, and thus should be used when more precision and detail are needed.

The i-Ready Diagnostic has approximately 60 items, depending on content and grade. NWEA MAP, a variable length computer adaptive test, has approximately 40-43 items depending on content and grade.[2] Given its design and its ability to report domain level results, i-Ready Diagnostic is best suited for instruction when actionable details about student performance are needed. With just 34 items, Renaissance Star is a shorter assessment often used for screening students.[3]

Shorter tests, like Renaissance Star, may appropriately measure the broader constructs of reading and math proficiency to provide information on how students are doing overall in reading and math, but they provide less insight into domain level skills in these areas. Reducing a test from 60 items to 34 items in order to reduce testing time means coverage of domain content must be sacrificed. Thus, choosing the right assessment isn't as easy as choosing the shortest assessment. You may gain time but you lose coverage and accuracy.

"Reducing a test from 60 items to 34 items in order to reduce testing time means coverage of domain content must be sacrificed. Thus, choosing the right assessment isn't as easy as choosing the shortest assessment. You may gain time but you lose coverage and accuracy."

When to choose different assessments

The critical trade-off between assessment length (testing time) and precision (the type and accuracy of the results) is an important consideration to keep in mind when choosing the right assessment for teachers and students. Though the shortest assessments might seem to be the most desirable to teachers and instructional leaders to save time, they may not address educators' needs.

Educators often use domain scores from interim assessments for instructional decision-making, for decisions on academic intervention, or to route students into personalized instruction provided on digital platforms. To do this, they need enough detail to guide instruction with enough reliability to ensure no misinterpretation or misidentification. Being behind in Geometry is very different than being behind in Numbers and Operations. The knowledge and skills needed in those mathematics domains are different, and understanding student needs allows educators to decide instruction and intervention for each student.

To make these decisions, educators expect that each domain score provides accurate information about students' needs beyond the information provided by the overall score. The coverage of the domain should be detailed enough so that the score provides unique information about what the student knows and can do and must be reliable to avoid misinterpretation of the student needs.

"The coverage of the domain should be detailed enough so that the score provides unique information about what the student knows and can do and must be reliable to avoid misinterpretation of the student needs."

A longer test, like the i-Ready Diagnostic, meant to provide more precise instructional information, might take 45 minutes and could have upwards of 60 items, allowing for approximately 12-20 domain-specific questions in each content area domain. A shorter 20-30 minute test, like Renaissance Star, may only have a handful of questions in each domain.

The detail and quality of information available from five questions versus 15 questions varies significantly. Five questions might allow a teacher to determine that a student needs more support in math, but 15 questions would provide far more information on the specific math skills where a student needs support. This will hold true for any assessment of comparable length, coverage, or content area, and is not specific to any one short assessment such as Renaissance Star.

"If tests lead to misinterpretations and misclassifications, then a test that is too short may comprise instructional validity. Shorter tests may win back minutes, but those wins may come at a measurable cost."

The more precise the results teachers receive, the more precise a teacher may be in targeting instruction. For instructional decision-making, the interpretation of what a student needs determines the appropriate instructional support for each student.

When the test is too short, there is a higher likelihood of making false or misleading interpretations about what the student needs and the optimal instructional next steps to meet those needs. If tests lead to misinterpretations and misclassifications, then a test that is too short may comprise instructional validity. **Shorter tests may win back minutes, but those wins may come at a measurable cost.**

References

[1] Renaissance Star has recently changed from a Star Enterprise Scale to a Star Unified Scale, making some comparisons challenging.

[2] This information comes from Curriculum Associates and NWEA technical documentation available from each organization.

[3] Star Assessments™ for Math Technical Manual (2023); Star Assessments™ for Reading Technical Manual (2023)

About John Hattie

John Hattie is an emeritus laureate professor at the Graduate School of Education, University of Melbourne, Australia. He is an internationally recognized author, educator, and researcher whose Visible Learning research, synthesizing 15 years of studies involving millions of students, is believed to be the world's largest evidence-based study into the factors that improve student learning.