Foreword

By Amber M. Northern and Michael J. Petrilli

As we were putting the final touches on this report, Amazon unveiled a "new storefront" called Amazon Ignite. The site will allow educators to earn money by publishing—online, of course—their original educational resources (lesson plans, worksheets, games, and more).

The e-commerce titan's entry into the curricular marketplace is obviously motivated by a perceived market opportunity—and that's not wrong. The vast majority of teachers are supplementing their core curriculum or don't have a core curriculum to start with, so it's no surprise that they often frequent the online arena to obtain the materials with which to meet their instructional needs.[i]

In fact, recent studies by RAND found that nearly all teachers report using the Internet to source instructional materials, and many of them do so quite often. For example, 55 percent of English language arts (ELA) teachers said they used Teachers Pay Teachers for curriculum materials at least once a week.[ii,iii] That site reports that one billion resources have been downloaded—a massive number, to be sure.

Yet we know almost nothing about the quality of such supplementary materials. Although several organizations have stepped up to offer impartial reviews of full curriculum products,[iv] to our knowledge, there's no equivalent when it comes to add-on resources. Therefore, we set out to answer a simple question: are popular websites supplying teachers with high-quality supplemental materials?

We recruited University of Southern California associate professor Morgan Polikoff to lead the review. He has conducted numerous studies on academic standards, curriculum, and assessments (including a previous Fordham study on Common Core–era tests), and he co-leads a federal research center on standards implementation. Jennifer Dean, an expert in assessment, standards alignment, and ELA content, served as lead reviewer of materials and assisted with report writing. She was joined by four other expert reviewers with backgrounds in teaching ELA, developing curricula and assessment items, and/or leading instructional teams.

i. Thomas J. Kane, et al., *Teaching higher: Educators' perspectives on Common Core implementation* (Cambridge, MA: Center for Education Policy Research, February 2016), http://cepr.harvard.edu/files/cepr/files/teaching-higher-report.pdf.

ii. Because the response categories on the survey changed across years, direct comparisons from 2015 to 2017 are not possible. But the general point applies. Julia H. Kaufman, V. Darleen Opfer, Michelle Bongard, and Joseph D. Pane, *Changes in what teachers know and do in the Common Core era: American Teacher Panel findings from 2015 to 2017* (Santa Monica, CA: RAND, 2018), https://www.rand.org/pubs/research_reports/RR2658.html.

iii. Julia H. Kaufman, Lindsey E. Thompson, and V. Darleen Opfer, *Creating a coherent system to support instruction aligned with state standards* (Santa Monica, CA: RAND, 2016), https://pdfs.semanticscholar.org/1c0f/998365b9b80edad157d7f8bd1d049ceed101.pdf.

iv. See for instance, the work of EdReports: https://www.edreports.org.

Morgan and Jennifer and their team, with the help of external advisers, developed a rubric that captured both the overall dimensions of quality in curriculum materials—things like rigor and usability—and more discrete dimensions that reflected the key instructional shifts called for by the new generation of states' ELA content standards: things like regular practice with complex texts and reading and writing tasks grounded in evidence from the text. In all, they examined over three hundred of the most downloaded materials found on three of the most popular supplemental websites: Teachers Pay Teachers, ReadWriteThink, and Share My Lesson.

As you will see in the following pages, this crackerjack review team unearthed a wealth of valuable information (encapsulated in *nine* key findings) that has important implications for district, school, and instructional leaders everywhere, as well as for classroom instructors themselves.

Sadly, the reviewers concluded that the majority of these materials are not worth using: more precisely, 64 percent of them should "not be used" or are "probably not worth using." On all three websites, a majority of materials were rated 0 or 1 on an overall 0–3 quality scale.

That's sobering to say the least, particularly given the popularity of these sites and the materials we reviewed. It suggests a major mismatch between what the experts think teachers should (and shouldn't) use in classrooms and what teachers themselves are downloading for such use—and, in some cases, paying for.

That's not necessarily a criticism of the teachers. They may be finding value in these materials in ways that we "experts" need to better understand. In interviews, teachers told us that they use the materials to fill instructional gaps, meet the needs of both low and high achievers, foster student engagement, and save them time. They rarely use the materials as is. Much adapting goes on as they choose and modify items to fill specific needs—needs that likely take precedence day to day over whether particular materials are aligned to state standards or incorporate high cognitive demand (or some other quality valued by experts).

We're not suggesting that teachers' views and judgments should yield to those of experts. Why not weigh both? Consider how this works on Rotten Tomatoes, the popular website that reviews the quality of movies and other entertainment. Their Tomatometer is based on the opinions of hundreds of film and television critics and is a trusted go-to for millions of viewers. When at least 60 percent of the critics' reviews of a movie or TV show are positive, it receives a red tomato, meaning it's "fresh." Less than 60 percent and it gets a green splat, meaning it's "rotten."

Those could reasonably be termed expert judgments. But Rotten Tomatoes also provides Audience Scores, which are just that. When at least 60 percent of viewers give a movie or TV show a star rating of 3.5 or higher, a full popcorn bucket indicates that it's "fresh" from the audience's perspective. When less than 60 percent, a tipped-over popcorn bucket reveals it's "rotten."

So the moviegoer and television watcher can readily access two different ratings—one from professional critics and another from the audience. Often they're similar, but not infrequently, they diverge. It's hard to say who is "right," but potential viewers get more information by seeing both ratings than they would from just one.

Same thing here. By definition, we looked at materials with high "Audience Scores," which is to say these were materials that had been downloaded the most. Yet in a majority of cases, our expert critics gave them a green splat, even though teachers rewarded them with a full popcorn bucket.

What then? Should we search for ways to block or deter teachers from using materials that experts don't like? Some on our team would welcome such a heavy-handed approach to monitoring supplemental resources, perhaps by empowering district leaders to enforce stringent policies about which supplemental resources would be allowed in their schools. We understand that impulse. It recalls an argument we often have with libertarians over school choice, wherein we think it's sometimes necessary to close really bad schools even though parents may like them.

In this case, however, we think a better solution is simply to provide teachers with more information, Tomatometer style. In addition to providing user reviews or comments to teachers, or highlighting and promoting the most popular lessons, the platforms should also make expert reviews available.

Two additional points are worth mentioning.

First, as our title indicates, the online marketplace is a bustling bazaar of cacophonous activity with myriad offerings of every sort. We cannot claim that our results apply to the thousands of other online resources out there for educators nor even to everything on the sites that we did evaluate. There's no way to evaluate it all, and undoubtedly, much of what's on offer is worth using. Yet we can state with some confidence that most of the most popular items leave much to be desired.

Second, not everyone will agree with our criteria and methods for assessing these materials. Even within our review team, not everyone was satisfied with every part of the process or with the conclusions about some materials. In some cases, we may have been too easy on the materials. In evaluating alignment, for instance, we simply asked whether the materials aligned to the standards that the teacher developers said that they aligned to. Similarly, a key expectation with assessments was that they cover the key content of the lesson.

In other cases, maybe the bar was too high. For example, we looked for cultural diversity by seeking the inclusion of multiple authors from diverse groups and/or topics of diverse cultural importance. Whether that's a reasonable expectation for any one supplemental item (versus a full-fledged curriculum) is certainly debatable. Ditto in expecting supplementary lessons to offer supports for most or all student subgroups, given how inadequately many full-bore curricula handle differentiation.

Regardless of their quality, one of the things that can get lost when teachers go trawling for supplemental materials is curricular coherence. As such, we agree with Morgan and Jennifer that school leaders and department heads should pay more attention to what's actually taught in classrooms by way of supplemental materials. What they learn could inform an array of subsequent strategies for improvement, from offering teachers training in how to identify high-quality materials to publishing a list of curated supplemental resources and addressing shortcomings and gaps in their core curriculum (the work of the Louisiana Department of Education may be instructive here).

Teachers are understandably hungry for instructional stuff, but the sites they're turning to are often providing subpar versions of it. We hope that they make improvements going forward. And we also hope that Amazon, the "most valuable company on the planet," will learn from its predecessors and strive to beat them at the quality game.

Executive Summary

Where teachers were once limited to traditional textbooks, informational texts, novels, and materials passed along by others, today the online marketplace is wide open, flush with copious materials that teachers might choose, often at little or no cost. But practically nothing is known about what these supplemental instructional materials actually look like and whether they are any good. Do they truly help educators deliver a high-quality curriculum?

In the current study, University of Southern California associate professor Morgan Polikoff and educational consultant Jennifer Dean led an analysis of supplemental materials for high school English language arts (ELA), an area where teachers are highly likely to supplement their core curriculum materials—sometimes because they do not have a core curriculum at all. Polikoff and Dean partner with four expert reviewers with experience in evaluating ELA curricula and assessments to examine over three hundred of the most downloaded materials across three of the most popular supplemental websites: Teachers Pay Teachers, ReadWriteThink, and Share My Lesson. Their analysis addresses two sets of questions:

1. What types of materials are teachers downloading most frequently? What kinds of content do they include?
2. How do experts rate the quality of these materials? What are their strengths and weaknesses, and what is the relationship (if any) between how experts view the quality of the materials and how teachers using them do?

Supplemental materials are evaluated on both overall dimensions of curriculum quality (such as rigor and usability), as well as more discrete criteria that loosely reflect the key instructional shifts of the new generation of ELA content standards.

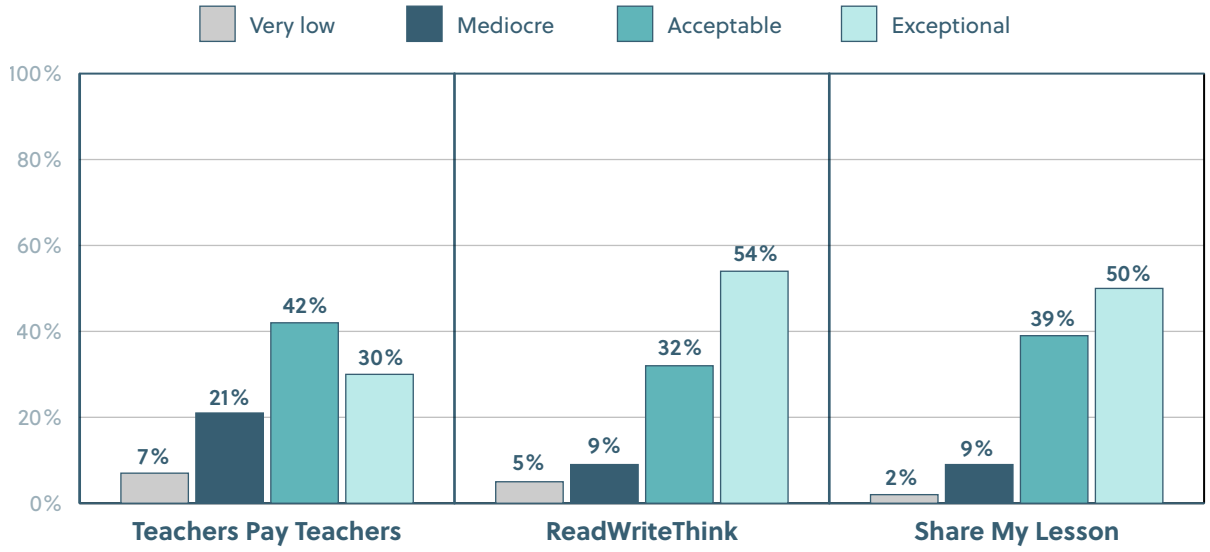
The study yields nine findings, including two strengths and seven weaknesses.

Strengths

FINDING 1: The quality of the texts is good to excellent, and students are often asked to provide textual evidence when analyzing a text.

Reviewers generally thought that the main text referenced in the materials was of good quality, with a mean of 2.21 on a 0–3 scale. In fact, exceptional quality is the most common rating (Figure ES-1). Just 5 percent of main texts receive the lowest rating of very low quality. Important differences arise across sites, however: ReadWriteThink and Share My Lesson have higher-quality texts (means of 2.34 and 2.36, respectively) than does Teachers Pay Teachers (mean of 1.96). The grade-level appropriateness of a text was one factor consistently associated with lower ratings.

Figure ES-1. All three websites have high-quality texts, but the texts on ReadWriteThink and Share My Lesson demonstrate "exceptional quality" more often than the texts on Teachers Pay Teachers.

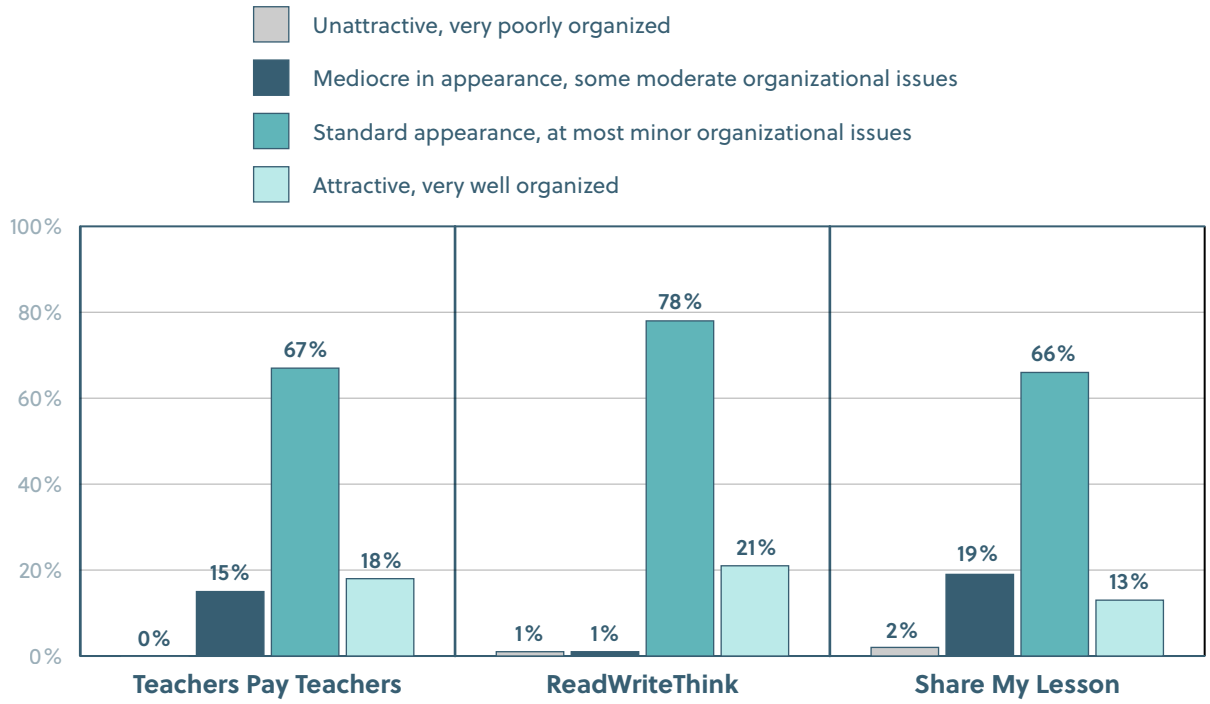


Note: Full scale is as follows. 0 = very low quality—poorly written, little to no grade-level subject-matter content, unimportant; 1 = mediocre quality—average writing, some grade-level subject-matter content, of mediocre importance; 2 = acceptable quality—good writing, appropriate grade-level subject-matter content, an important text; and 3 = exceptional quality—exceptional writing, rich in grade-level subject-matter content, an exceptionally important text. Numbers may not sum to 100 percent due to rounding.

FINDING 2: The materials are generally free from errors and well designed.

Reviewers found that the materials were generally free from errors that might affect student understanding. On a 0–3 scale,[v] the mean score is 2.75. Across all sites, just 2 percent of materials are rated as having major or moderate errors, while 77 percent are rated as having no or very few errors. ReadWriteThink has the fewest errors (mean = 2.92), while Share My Lesson has the most (mean = 2.53) and Teachers Pay Teachers is in the middle (mean = 2.79). Materials also rated well in terms of their visual appearance and organization (Figure ES-2). On a 0–3 scale,[vi] the mean across sites is 2.04, with 87 percent of all materials earning 2 or 3 on this dimension. Across sites, Share My Lesson materials were rates as least attractive and least organized (mean = 1.89), and ReadWriteThink was rated the most attractive and most organized (mean = 2.19).

Figure ES-2. Most materials across all three sites are reasonably attractive and well organized.



Note: Full sacale as shown. Numbers may not sum to 100 percent due to rounding.

v. Full scale is as follows: 0 = major errors that are likely to affect student understanding; 1 = moderate errors that may or may not affect student understanding; 2 = minor errors that are unlikely to affect student understanding; and 3 = no or very few errors.

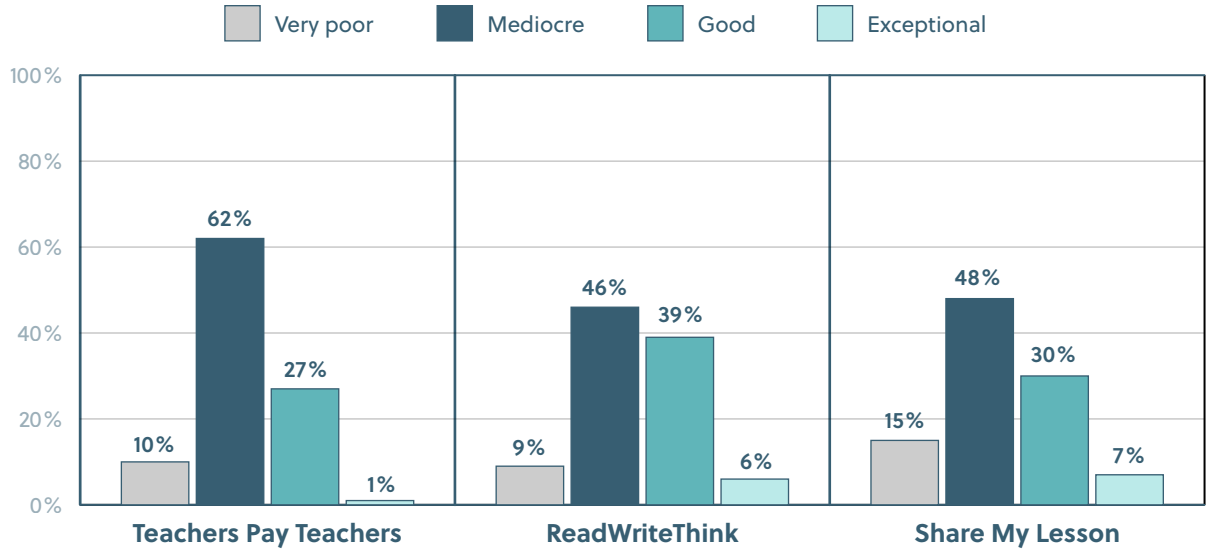
vi. Full scale is as follows: 0 = unattractive, very poorly organized; 1 = mediocre in appearance, some moderate organizational issues; 2 = standard appearance, at most minor organizational issues; and 3 = attractive, very well organized.

Weaknesses

FINDING 3: Overall, reviewers rate most of the materials as "mediocre" or "probably not worth using." Clarity and instructional guidance are weak. At best, there's modest evidence that the quality of the material predicts teachers' use of it.

On a 0–3 scale, with 2 or higher corresponding to materials that reviewers thought teachers should use, the mean score for materials is 1.28, with reviewers recommending that 64 percent *not be used* or are *probably not worth using*. No website has a majority of materials earning an *exceptional* rating (Figure ES-3), but ReadWriteThink receives a slightly higher overall rating on average (mean = 1.41) than Share My Lesson (mean = 1.29) or Teachers Pay Teachers (mean 1.18). A major contributing factor to the poor overall ratings is the lack of clarity of the guidance offered to teachers. On a 0–3 scale,[vii] with 2 intended to represent standard guidance, the mean across the three sites is 1.61.

Figure ES-3. On all three websites, most materials receive an overall rating of very poor or mediocre. Less than 10 percent of materials on each site are rated exceptional.



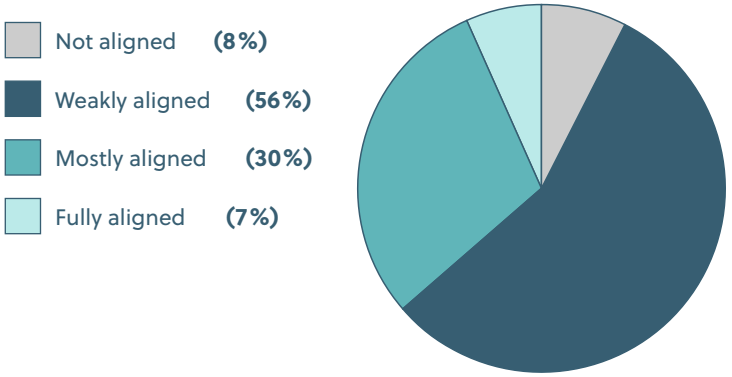
Note: Full scale is as follows. 0 = very poor, teachers should not use this material; 1 = mediocre, has some good and some bad components (for example, well organized but not on important content or covering diverse perspectives but using weak tasks), probably not worth using; 2 = good, overall a high-quality material, well organized and usable, covering important content, likely to contribute to a quality curriculum; and 3 = exceptional, unusually well crafted, rich with content, highly likely to contribute to a quality curriculum. Numbers may not sum to 100 percent due to rounding.

vii. Full scale is as follows: 0 = very unclear or no guidance offered; 1 = some lack of clarity or limited guidance offered; 2 = adequate clarity and guidance offered; and 3 = exceptionally clear, complete guidance offered.

FINDING 4: The materials are weakly to moderately aligned with the standards to which they claim alignment.

Respondents used a 0–3 scale that ranged from *not* to *fully aligned*. The average alignment rating is 1.35. Of all the materials, 56 percent score a rating of 1 (see Figure ES-4), which technically means "lesson partly aligns to some of the listed standards or fully aligns to a few (but not the majority) of the listed standards."[viii] These low alignment ratings occur primarily because most materials claim alignment to a very large number of standards.

Figure ES-4. The majority of materials are rated as weakly aligned with the standards to which they claim alignment.



Note: Full scale is as follows. 0 = not aligned to the target standards; 1 = weakly aligned to the target standards; 2 = mostly aligned to the target standards; and 3 = fully aligned to the target standards. Numbers may not sum to 100 percent due to rounding.

FINDING 5: The overall quality of writing and speaking and listening tasks is weak.

Of all the materials, 82 percent have a writing task that requires students to write a paragraph or more. On a 0–3 scale, ranging from *very low* to *exceptional* quality, the tasks average 1.42.[ix] Just 6 percent of them earn a score of 3, while 51 percent earn a score of 0 or 1. There are scarcely any differences across the three sites, with all scoring between 1.40 and 1.44 (Figure ES-5a).

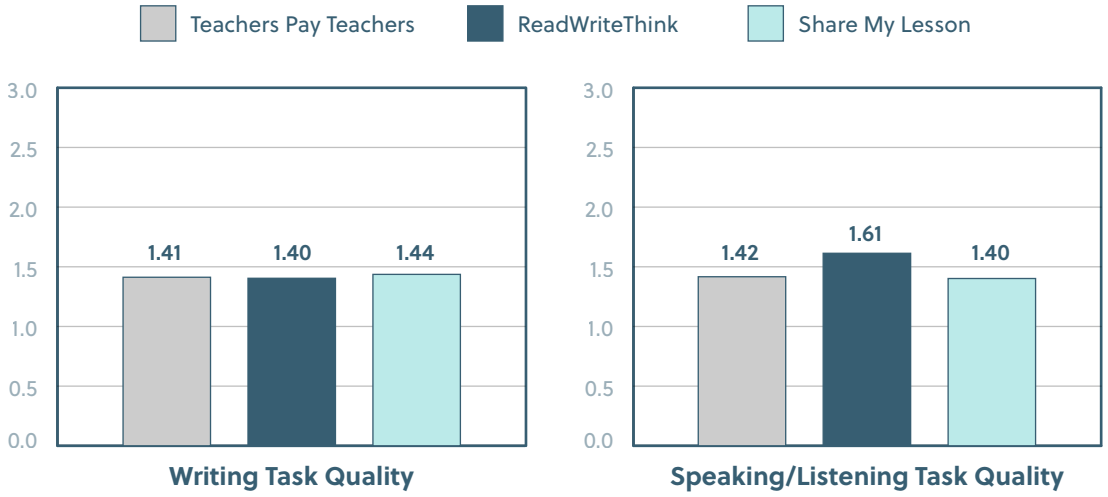
There was a speaking and listening task in 43 percent of materials, and the scale used to judge quality was the same as the writing task.[x] The quality of the speaking and listening tasks is only slightly better than that of the writing tasks, with a mean score of 1.48. As shown in Figure ES-5b, there is a small difference favoring ReadWriteThink, with a mean of 1.61 (versus 1.42 and 1.40 for Teachers Pay Teachers and Share My Lesson, respectively).

viii. Reviewers received additional guidance in a scoring manual that explained in more detail what each score point represented for each indicator.

ix. The rubric mandated that in order to score 3, the task had to require writing to a text.

x. The rubric mandated that in order to score 3, the task had to require speaking or listening to a text.

Figure ES-5a-b. Writing and speaking and listening tasks demonstrate moderate quality across all three sites.



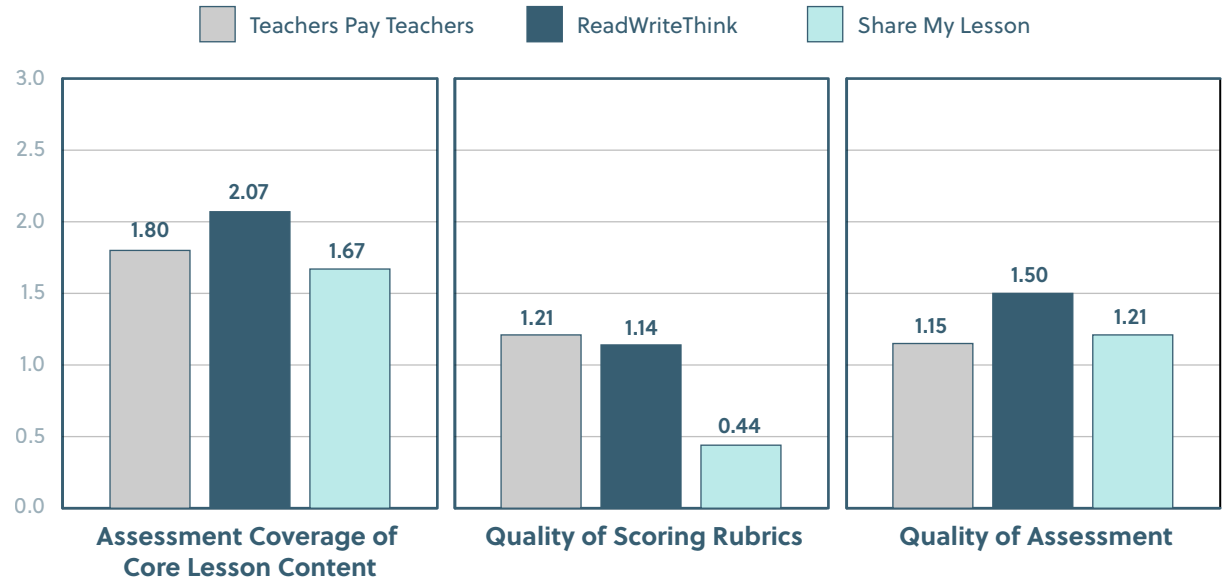
Note: Full scale is as follows. 0 = very low quality—task is unclear to student or task is unimportant (frivolous, silly) or far too easy for the grade level; 1 = mediocre quality—task likely to be clear to student but of limited importance or not very challenging for the grade level; 2 = acceptable quality—clear, important, and adequate challenge for the grade level; and 3 = exceptional quality—clear, highly important, and challenging for the grade level (note that 3 can only be awarded if the task requires writing to a text).

Note: Full scale is as follows. 0 = very low quality—task is unclear to student or task is unimportant (frivolous, silly) or far too easy for the grade level; 1 = mediocre quality—task likely to be clear to student but of limited importance or not very challenging for the grade level; 2 = acceptable quality—clear, important, and adequate challenge for the grade level; and 3 = exceptional quality—clear, highly important, and challenging for the grade level (note that 3 can only be awarded if the task requires speaking or listening to a text).

FINDING 6: Assessments included in the materials rank poorly because they sometimes fail to cover key content and rarely provide teachers the supports needed to score student work.

Regarding whether the assessments covered the core content of the lesson or unit, the materials average a 1.84 on a 0–3 scale, where 2 represents assessment of more than half of the core content of the lesson/unit (Figures ES-6a-c). A bare majority of materials (51 percent) include scoring rubrics to help teachers evaluate student performance; the mean score across the three websites is 0.94 on a 0–3 scale, ranging from *no* to a *high-quality* rubric. The assessments rated poorly on an overall evaluation of quality, scoring 1.27 on a 0–3 scale.

Figures ES-6a-c. Assessments are rated highest on covering the core content of the lesson and lowest on the availability of a scoring rubric.



Note: Full scale is as follows. 0 = very poor coverage—fails to assess the core content of the lesson; 1 = mediocre coverage—assesses some core content in the lesson but has some large gaps; 2 = good coverage—assessments most of the content in the lesson, at most small gaps; and 3 = full coverage—assesses the core content in the lesson completely.

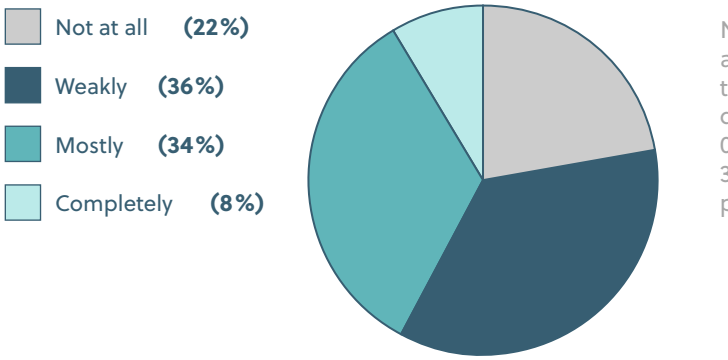
Note: Full scale is as follows. 0 = no rubric available; 1 = rubric available but of poor quality; 2 = rubric available and of adequate quality; and 3 = rubric available and of high quality.

Note: Full scale is as follows. 0 = very low quality—poorly written, containing significant errors, assesses unimportant content; 1 = mediocre quality—minor lack of clarity, containing minor errors, assesses content of mediocre importance; 2 = acceptable quality—well written, no errors, assesses most of the important content; and 3 = exceptional quality—exceptionally well written and challenging, no errors, assesses all of the most important content.

FINDING 7: Lesson units do a poor job of building students' content knowledge, and they are generally not cognitively demanding.

Reviewers evaluated the extent to which multiday units introduced and sequenced knowledge in a way that allowed students to build their understanding of a topic. Of the units scored, 58 percent earned a 1 or 2 on this dimension, indicating that they support students' ability to demonstrate such knowledge *not at all* or *weakly* (Figure ES-7). The mean score on the 0–3 scale is 1.28.

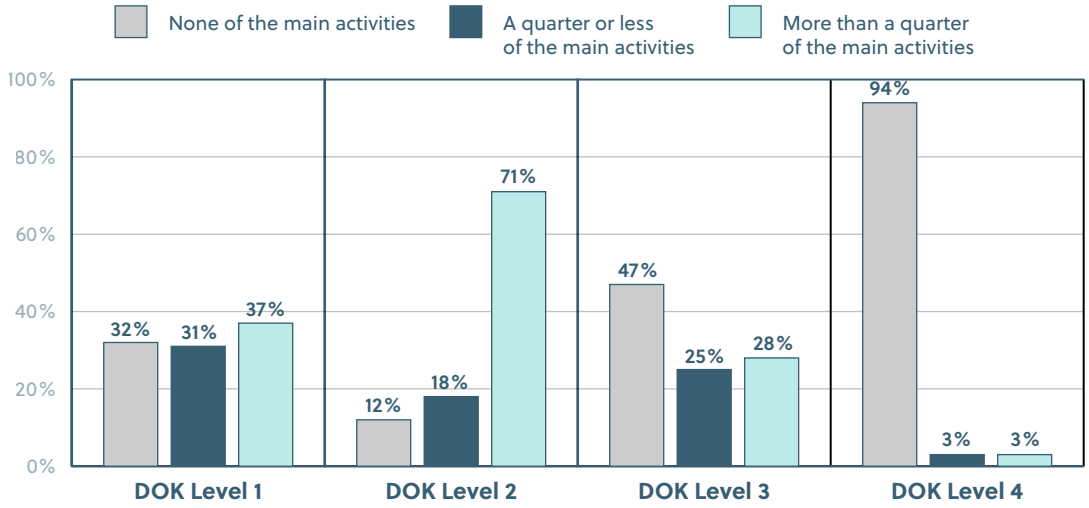
Figure ES-7. Of all units, 58 percent "not at all" or only "weakly" build student knowledge.



Note: Full scale is as follows. The questions and tasks in the unit support students' ability to complete a culminating task in which they demonstrate their knowledge of a topic: 0 = not at all; 1 = weakly; 2 = mostly; and 3 = completely. Numbers may not sum to 100 percent due to rounding.

Reviewers also evaluated depth of knowledge (DOK)—the cognitive demand required for students to successfully engage with the materials. Most of the content included in the main activity of each material is DOK level 1 or 2 (Figure ES-8). Nearly half of the main activities have no DOK level 3 content at all (the grey bar in the third set), and just 6 percent score higher than a 0 for DOK level 4 (the navy and teal bars in the fourth set).

Figure ES-8. About half of all main activities in the materials have no depth of knowledge level 3 content, and less than 6 percent have any DOK level 4 content.

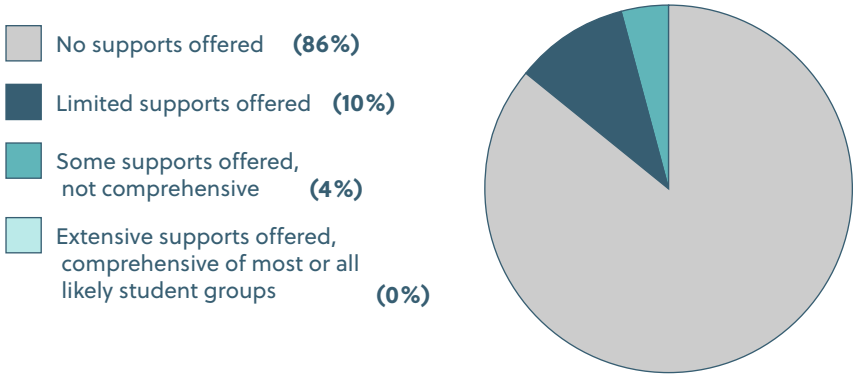


Note: Numbers may not sum to 100 percent due to rounding.

FINDING 8: The materials do a very poor job of offering teachers support for teaching diverse learners.

The level of support provided for teaching diverse learners garners the lowest ratings among all of the evaluated dimensions. We asked how comprehensive were the supports for differentiation with regard to meeting the needs of high- or low-performing students, students with disabilities, and English-language learners. A full 86 percent of the materials score 0 on this dimension, indicating that they offer no support (Figure ES-9). Less than 1 percent of materials score 3, indicating extensive supports for most or all student subgroups. The mean score across the three sites is 0.19, with slightly more differentiation supports on Share My Lesson (mean = 0.34) than the other two sites (means of 0.10 and 0.15).

Figure ES-9. The majority of materials offer no supports for teaching diverse learners.

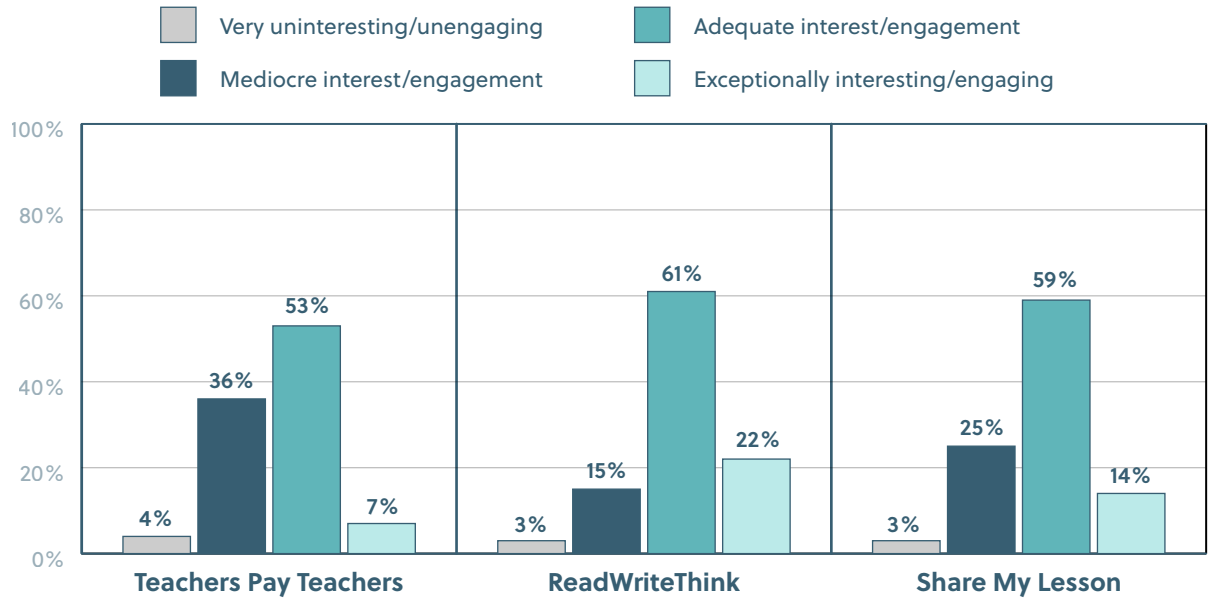


Note: Numbers may not sum to 100 percent due to rounding.

FINDING 9: Materials score fairly low on their potential to engage students and do not reflect the cultural diversity of classrooms.

Reviewers evaluated whether they thought that students would likely care about and be interested in the material presented to them. On a 0–3 scale, ranging from *very uninteresting* to *exceptionally interesting*, materials average 1.81 for engagement (Figure ES-10). Across websites, most are rated as *adequately* interesting (51–60 percent), although 29 percent are rated as *very uninteresting* or of *mediocre* interest. ReadWriteThink materials are deemed most interesting (mean = 2.02) and Teachers Pay Teachers the least (mean = 1.63), while Share My Lesson lands in the middle (mean = 1.83).

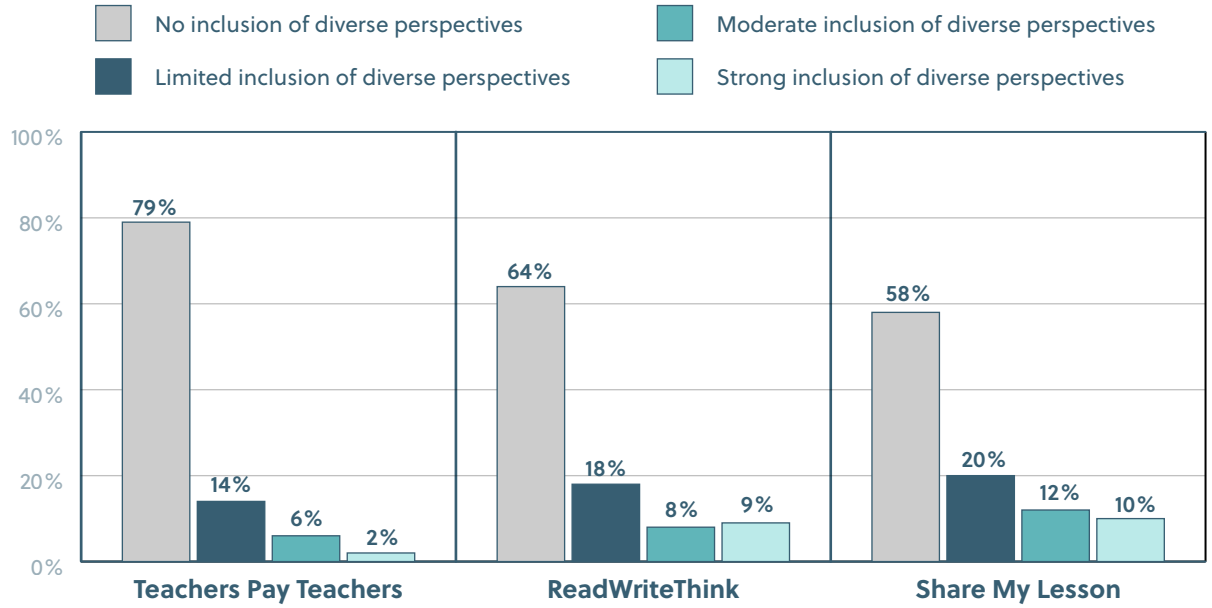
Figure ES-10. Most materials are rated as having adequate interest/engagement, but 18–40 percent of materials (depending on the site) are rated as mediocre interest or very uninteresting.



Note: Full scale is as follows. 0 = very uninteresting/unengaging—highly boring, very likely to be of limited interest to most students; 1 = mediocre interest/engagement—somewhat boring, may be of interest to some students but likely not most; 2 = adequate interest/engagement—not boring, likely to be of interest to most students; and 3 = exceptionally interesting/engaging—very likely to be of high interest to nearly all students. Numbers may not sum to 100 percent due to rounding.

Reviewers also examined both the choice of authors and the texts themselves relative to their representation of cultural diversity, with a focus on race/ethnicity, gender, and culture/national origin. On a scale of 0–3, 68 percent of materials score 0, meaning they do not include diverse authors or cover culturally diverse topics (Figure ES-11). Just 15 percent of materials score 2 or 3, meaning *moderate* or *strong* inclusion of diverse perspectives, including several authors from diverse groups and/or topics of great diverse cultural importance. The overall mean on this item is 0.53, but ReadWriteThink (mean = 0.62) and Share My Lesson (mean = 0.75) score much higher than Teachers Pay Teachers (mean = 0.30).

Figure ES-11. A majority of materials on all three sites do not include diverse authors or cover culturally diverse topics.



Note: Full scale is as follows. 0 = no inclusion of diverse perspectives; 1 = limited inclusion of diverse perspectives—includes one or two authors from diverse groups or topics of some diverse cultural importance; 2 = moderate inclusion of diverse perspectives—includes several authors from diverse groups or topics of great diverse cultural importance; and 3 = strong inclusion of diverse perspectives—includes several authors from diverse groups and topics of great diverse cultural importance. Numbers may not sum to 100 percent due to rounding.

*****

Polikoff and Dean draw five implications from these findings:

1. Supplemental ELA materials on the most popular sites have a long way to go before they can be used to strengthen gaps that exist in high school curricula.
2. The market for supplemental materials is bewildering and begs curation.
3. More supplemental materials need to provide teachers with soup-to-nuts supports, including stronger assessments and supports for diverse learners.
4. We need better sourcing of supplemental materials that focus on diverse authors and cultural pluralism.
5. School and district leaders need to decide whether and how to monitor the enacted curriculum.