December 2019

The Supplemental Curriculum Bazaar: Is What's Online Any Good?

By Morgan Polikoff with Jennifer Dean

Foreword and Executive Summary by
Amber M. Northern and Michael J. Petrilli

THOMAS B. FORDHAM INSTITUTE
ADVANCING EDUCATIONAL EXCELLENCE

ABOUT THE FORDHAM INSTITUTE

The Thomas B. Fordham Institute promotes educational excellence for every child in America via quality research, analysis, and commentary, as well as advocacy and exemplary charter school authorizing in Ohio. It is affiliated with the Thomas B. Fordham Foundation, and this publication is a joint project of the Foundation and the Institute. For further information, please visit our website at www.fordhaminstitute.org. The Institute is neither connected with nor sponsored by Fordham University.

SUGGESTED CITATION FOR THIS REPORT

Polikoff, Morgan with Jennifer Dean. *The Supplemental-Curriculum Bazaar: Is What's Online Any Good?* Washington, DC: Thomas B. Fordham Institute (December 2019). https://fordhaminstitute.org/national/research/supplemental-curriculum-bazaar.

ACKNOWLEDGMENTS

This report was made possible through the support of the Chan Zuckerberg Initiative and our sister organization, the Thomas B. Fordham Foundation.

We are especially grateful to lead author Morgan Polikoff whose scholarly talents—as well as supply of efficiency, responsiveness, and amicability—make him an ideal research partner. Jennifer Dean ably led the review team and also conducted teacher interviews that greatly strengthened the final product. Expert reviewers Jenni Aberli, Sarah Baughman, Bryan Drost, and Joey Hawkins were each thoughtful in their approach to the work and willing to go the extra mile.

Thanks also to advisers Carey Swanson and Katie Keown at Student Achievement Partners and Shakiela Richardson at UnboundEd. Liisa Potts (EdReports) and Michael Goldstein (Match Education) provided initial thoughts on the project's direction. Project staff at ReadWriteThink, Share My Lesson, and Teachers Pay Teachers also kindly addressed our questions.

At Fordham, we extend our gratitude to Chester E. Finn, Jr. for reviewing drafts, Victoria McDougald for overseeing media relations, Olivia Piontek for handling funder communications, and Jonathan Lutton for developing the report's layout and design. Fordham research interns Pedro Enamorado and Tran Le provided assistance at various stages in the process. Former interns Sophie Sussman and Jessica McBirney did the heavy lifting to prepare the materials for review, collected the metadata, and generally helped to keep the project on track. Finally, we thank Pamela Tatz, who copyedited the report, as well as Xavier Arnau Serrat of GettyImages.com for the cover photo.

Contents

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.

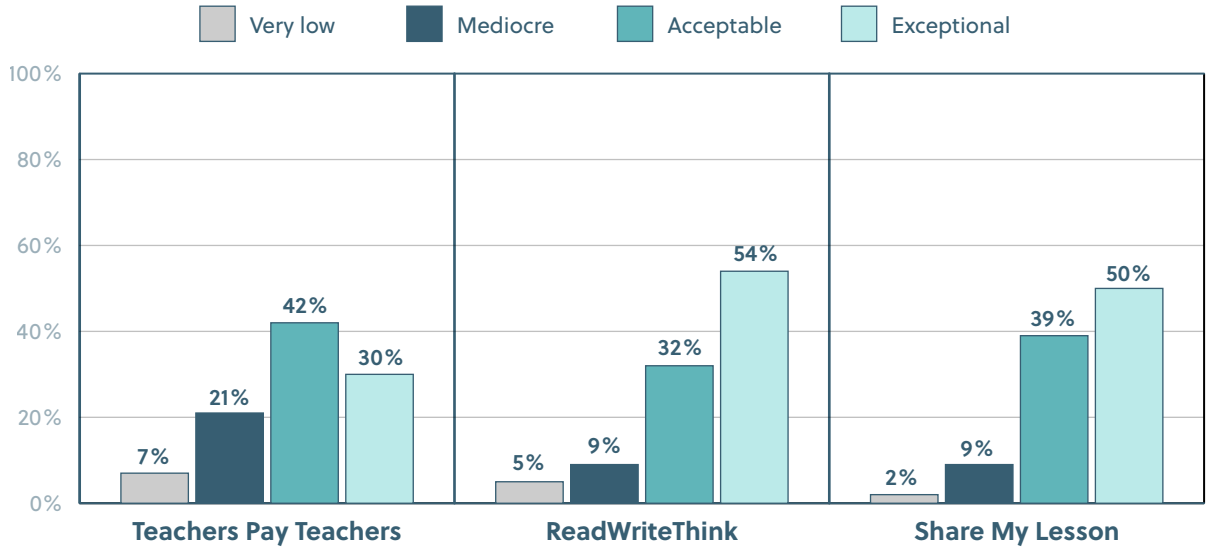
The study yields nine findings, including two strengths and seven weaknesses.

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.

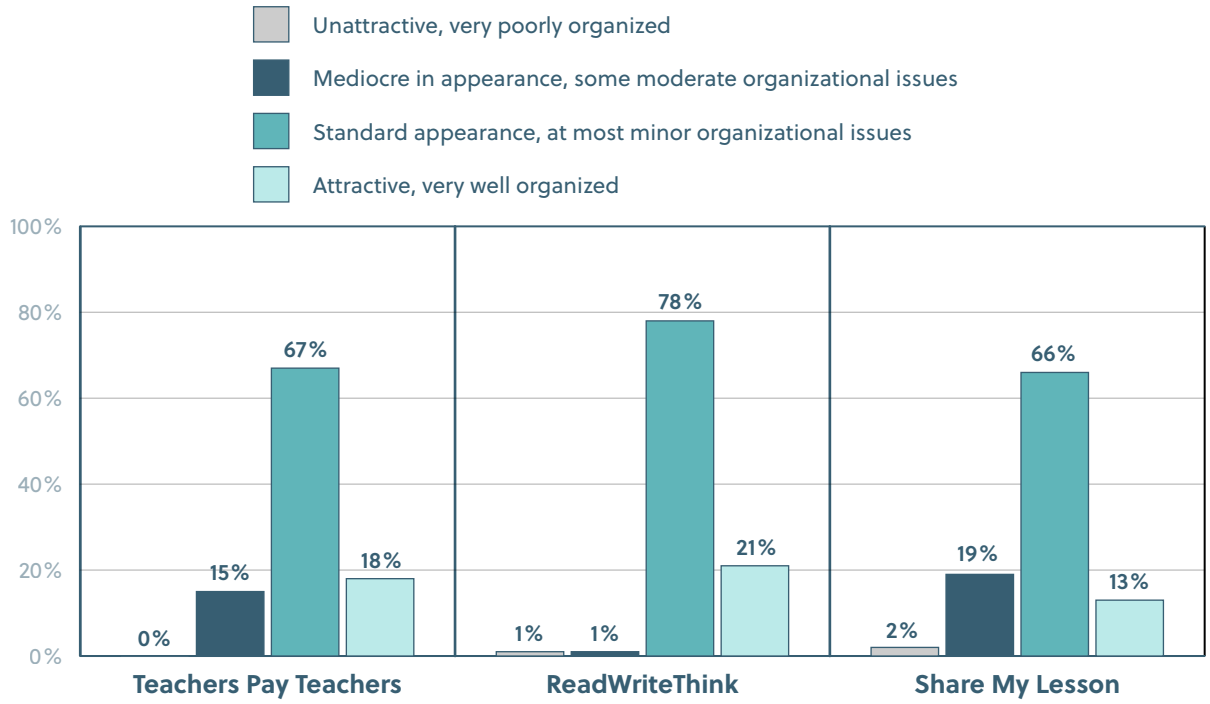


Note: Full scale is as follows. 0 = very low quality—poorly written, little to no grade-level subject-matter content, unimportant; 1 = mediocre quality—average writing, some grade-level subject-matter content, of mediocre importance; 2 = acceptable quality—good writing, appropriate grade-level subject-matter 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.

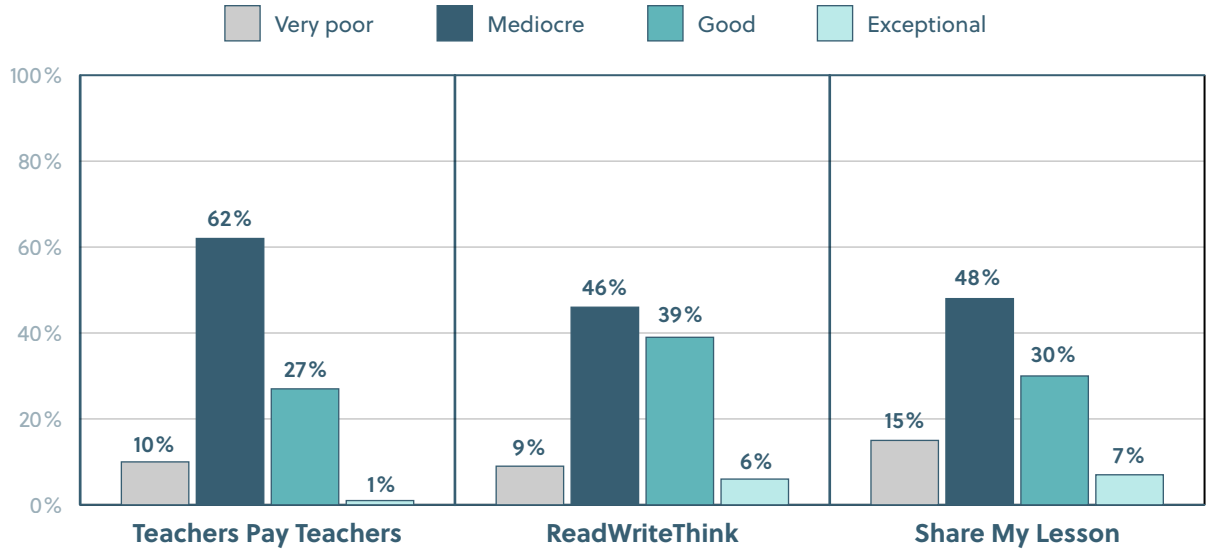
vi. Full scale is as follows: 0 = unattractive, very poorly organized; 1 = mediocre in appearance, some moderate organizational issues; 2 = standard appearance, at most minor organizational issues; and 3 = attractive, very well organized.

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.



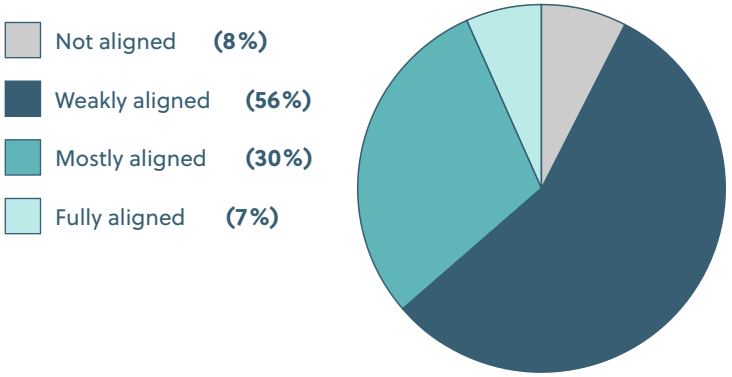
Note: Full scale is as follows. 0 = very poor, teachers should not use this material; 1 = mediocre, has some good and some bad components (for example, well organized but not on important content or covering diverse perspectives but using weak tasks), probably not worth using; 2 = good, overall a high-quality 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).

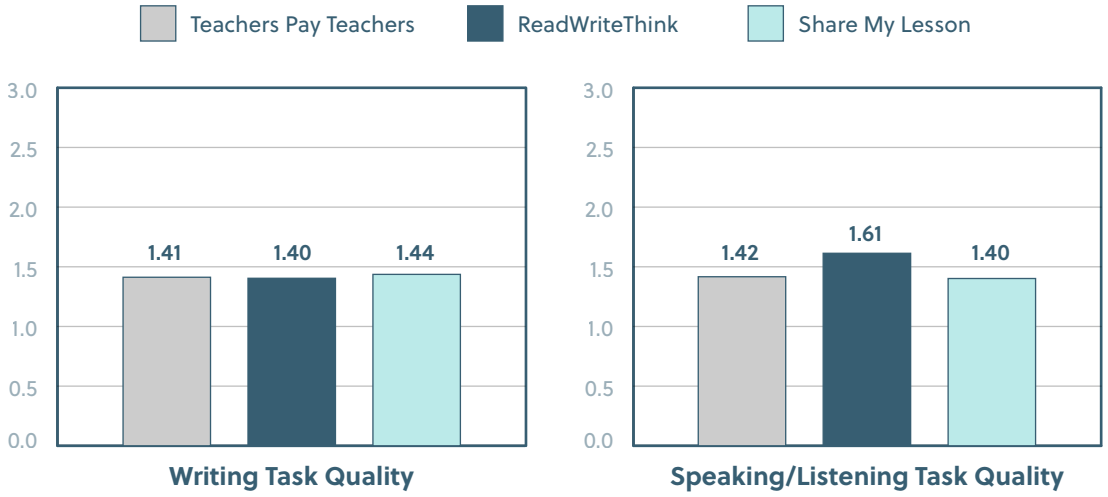
There was a speaking and listening task in 43 percent of materials, and the scale used to judge quality was the same as the writing task.[x] The quality of the speaking and listening tasks is only slightly better than that of the writing tasks, with a mean score of 1.48. As shown in Figure ES-5b, there 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.



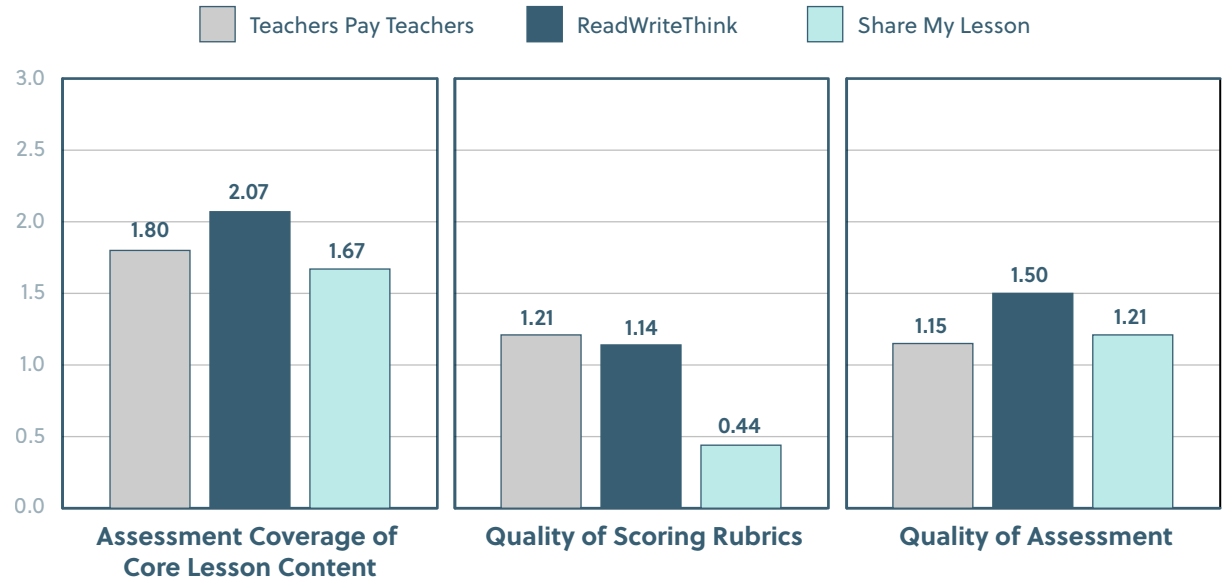
Note: Full scale is as follows. 0 = very low quality—task is unclear to student or task is unimportant (frivolous, silly) or far too easy for the grade level; 1 = mediocre quality—task likely to be clear to student but of limited importance or not very challenging for the grade level; 2 = acceptable 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.

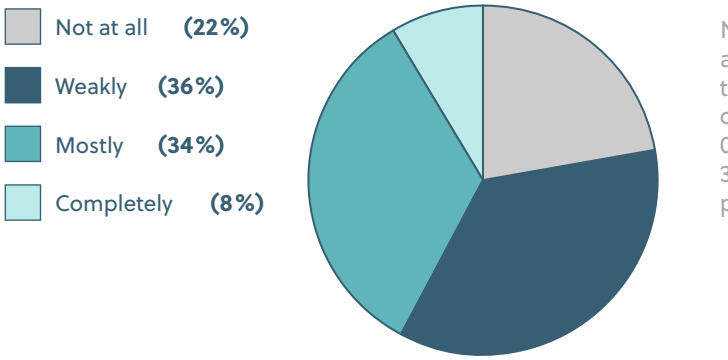
Note: Full scale is as follows. 0 = no rubric available; 1 = rubric available but of poor quality; 2 = rubric available and of adequate quality; and 3 = rubric available and of high quality.

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.

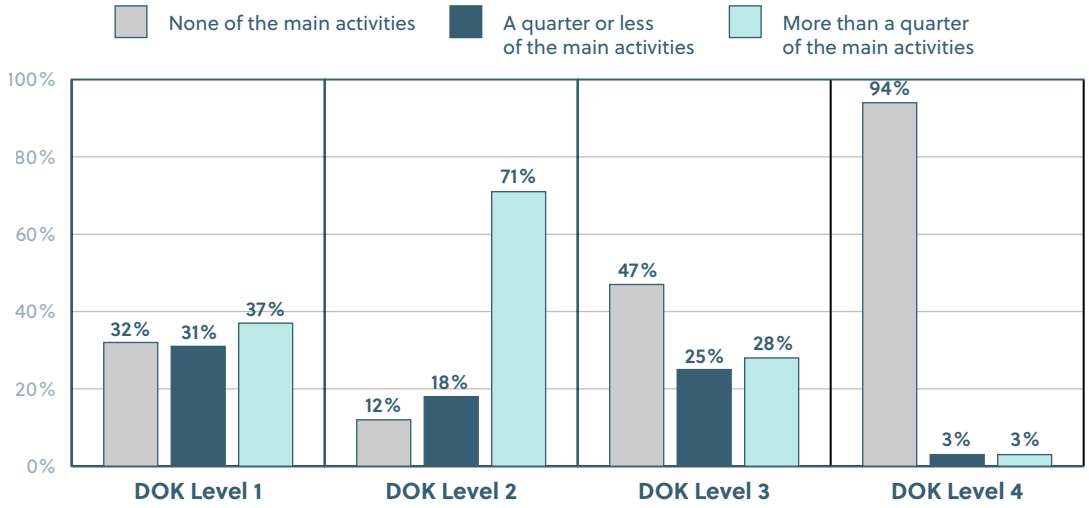
Figure ES-7. Of all units, 58 percent "not at all" or only "weakly" build student knowledge.



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.

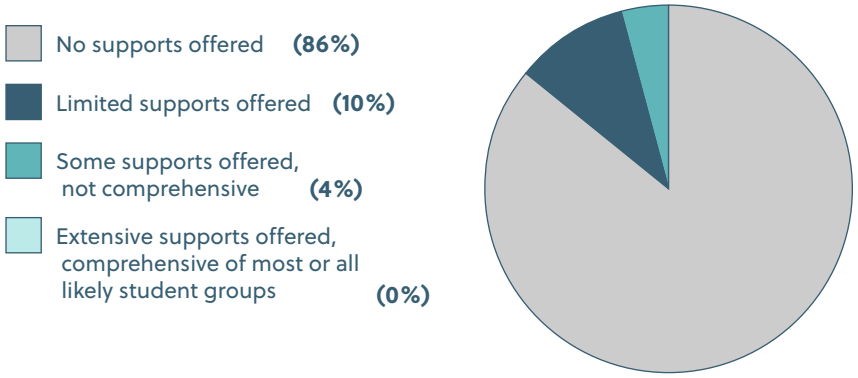


Note: Numbers may not sum to 100 percent due to rounding.

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.

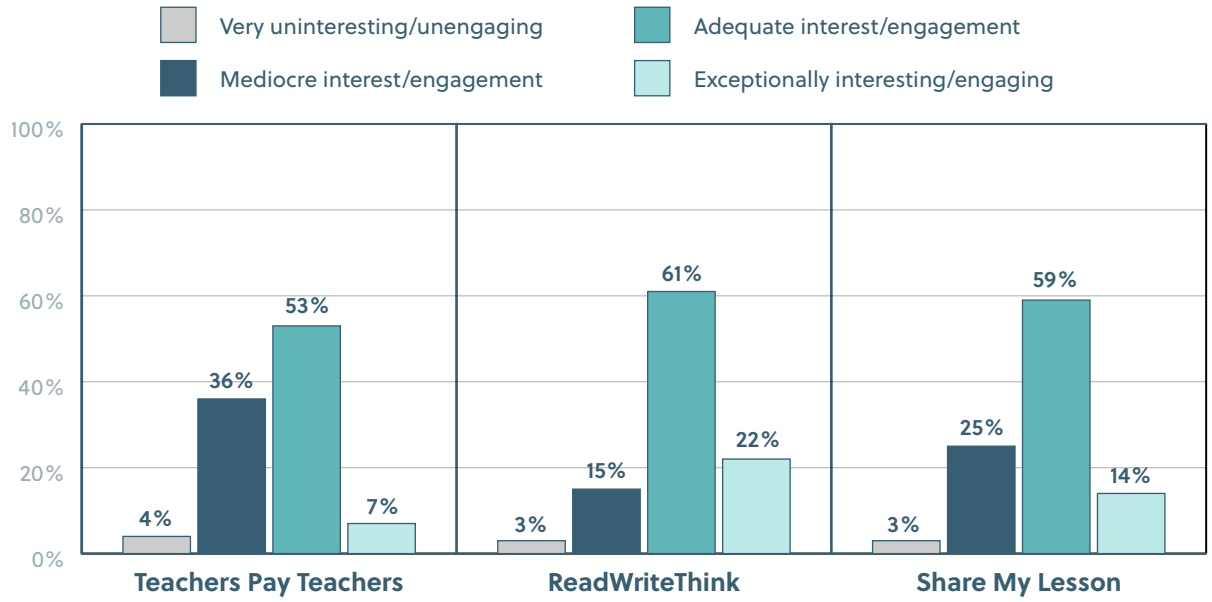


Note: Numbers may not sum to 100 percent due to rounding.

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).

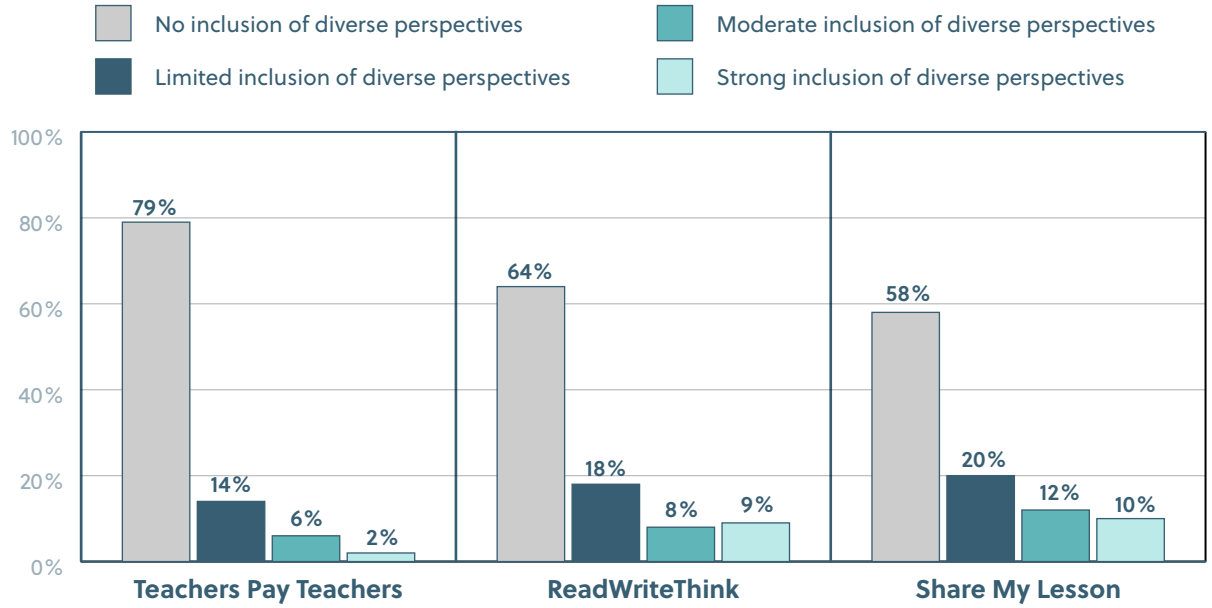
Figure ES-10. Most materials are rated as having adequate interest/engagement, but 18–40 percent of materials (depending on the site) are rated as mediocre interest or very uninteresting.



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.

I. Introduction

Recent years have seen growing attention to curriculum and its role in instruction and student achievement. How much difference does curriculum actually make? Several studies[1] have provided convincing evidence that textbooks can affect student achievement, but the most recent and largest study on mathematics found no such effects.[2]

Responding to these confusing findings and hoping to help districts make better decisions about which materials to adopt, several worthy outside organizations such as EdReports.org and the Learning List offer impartial reviews of popular curricular products. Several states also now conduct their own reviews in order to provide districts with much-needed information on curricular quality, content, and standards alignment.[3] Louisiana, in particular, has been posting reviews of instructional materials on its state website for several years, focusing mostly on core curricula and ranking them by overall quality and alignment to state standards. Other groups have developed rubrics and evaluation tools intended to help education leaders vet the quality and alignment of textbooks, units, and lesson plans, including Educators Evaluating the Quality of Instructional Products (EQuIP), Instructional Materials Evaluation Tool (IMET), and Student Achievement Partners' Publishers' Criteria.[4] Even Amazon has entered the curricular marketplace, launching a platform for educators that features free resources and teacher ratings and reviews.

1. Roberto Agodini, et al. *Achievement effects of four early elementary school math curricula: Findings for first and second graders*, NCEE 2011–4001 (Washington, D.C.: National Center for Education Evaluation and Regional Assistance, U.S. Department of Education, Institute of Education Sciences, 2010), https://files.eric.ed.gov/fulltext/ED512551.pdf; Rachana Bhatt and Cory Koedel, "Large-scale evaluations of curricular effectiveness: The case of elementary mathematics in Indiana," *Educational Evaluation and Policy Analysis* 34, no. 4 (2012): 391–412, https://www.jstor.org/stable/23357020; Rachana Bhatt, Cory Koedel, and Douglas Lehmann, "Is curriculum quality uniform? Evidence from Florida," *Economics of Education Review* 34, no. 1 (2013): 107–21, doi:10.1016/j.econedurev.2013.01.014; and Cory Koedel, et al., "Mathematics curriculum effects on student achievement in California," *AERA Open* 3, no. 1 (2017): 1–22, doi:10.1177/2332858417690511.

2. David Blazar, et al., *Learning by the Book: Comparing math achievement growth by textbook in six Common Core states* (Cambridge, MA: Center for Education Policy Research, Harvard University, 2019), https://cepr.harvard.edu/files/cepr/files/cepr-curriculum-report_learning-by-the-book.pdf.

3. Lindsey Tepe and Teresa Mooney, *Navigating the New Curriculum Landscape: How states are using and sharing open educational resources* (Washington, D.C.: New America, 2018), https://s3.amazonaws.com/newamericadotorg/documents/FINAL_Navigating_the_New_Curriculum_Landscape_v3.pdf.

4. Achieve, "ELA," accessed November 8, 2013, https://www.achieve.org/our-initiatives/equip/tools-subject/ela; Achieve the Core, "Instructional Materials Evaluation Tool," August 21, 2013, https://achievethecore.org/page/1946/instructional-materials-evaluation-tool; and Achieve the Core, "Revised Publishers' Criteria for ELA/Literacy," August 23, 2013, https://achievethecore.org/page/227/revised-publishers-criteria-for-ela-literacy.

Despite the focus on core curriculum materials in most of this research and work, many teachers must still improvise their own, in part because they do not have a core curriculum at all.[5] Moreover, virtually all of them supplement whatever core materials they do have, often quite extensively. A recent multistate survey, in fact, found that 95 percent of all teachers report using materials sourced from the Internet—and about half use such materials in at least one-quarter of their lessons.[6]

The online marketplace is wide open, flush with copious materials that teachers might choose.[7] But practically nothing is known about what these materials actually look like and whether they are any good. Will they truly help educators deliver a high-quality English curriculum? This report offers a first cut at analyzing supplemental materials for high school English language arts (ELA), where teachers are highly likely to supplement core curriculum materials (again, perhaps because they do not have a core curriculum at all).[8] We examine 328 materials across three of the most popular websites—Teachers Pay Teachers (TPT), ReadWriteThink (RWT), and Share My Lesson (SML)—to address two sets of questions:

1. What types of materials are teachers accessing? 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?

We address these questions for the materials as a whole and also for each website. In the rest of the report, we share more about our motivation for the work and describe both our evaluation rubric and the process for selecting and reviewing the materials. Then, we present the findings and conclude with recommendations for policymakers at the state and district levels.

5. In an ongoing state-level study with researchers at RAND that was codirected by Morgan Polikoff (not yet published), just 34 percent of Massachusetts teachers and 42 percent of Massachusetts school district leaders reported that their district required or even recommended core curriculum materials for high school ELA teachers. In a 2016 study, about one-third of high school ELA teachers reported that they did not use any materials that were required or recommended by their district. V. Darleen Opfer, Julia H. Kaufman, and Lindsey E. Thompson, *Implementation of K–12 state standards for mathematics and English language arts and literacy: Findings from the American Teacher Panel* (Santa Monica, CA: RAND, 2016), https://www.rand.org/content/dam/rand/pubs/research_reports/RR1500/RR1529-1/RAND_RR1529-1.pdf.
6. Blazar, et al., *Learning by the Book*.
7. In fact, EdWeek's survey of district leaders around the country found that no product—in either ELA or math—has more than about 15 percent of the market.
8. Opfer, Kaufman, and Thompson, *Implementation of K–12 state standards*.

II. Background

Curriculum materials are essential for instruction, offering a bridge between state academic standards and the enacted curriculum that students experience in the classroom. They affect the content of instruction—e.g., what topics are included and excluded, which are emphasized, and how they are sequenced. Several studies of elementary mathematics suggest that the choice of materials can also directly affect student achievement, while a larger and newer study found no such effects.[9] Owing in part to the evidence of potential effects and to their clear role as a key policy lever for implementing standards, curriculum materials have seen heightened interest in the policy and philanthropic communities.

Though most curriculum research to date has understandably focused on core textbooks—their adoption, implementation, and effects on instruction and achievement—these materials represent just a slice of what teachers typically employ in their classrooms. By all accounts, teachers make extensive use of supplemental instructional materials, by which we mean any instructional materials outside the teacher's core curriculum, especially those that the teacher or her colleagues have a hand in selecting (where teachers have no core curriculum, what we term "supplemental" may in fact serve as the bulk of instructional materials).

Where teachers were once limited to traditional textbooks, informational texts, novels, and other materials passed along by others, such as lecture notes and lesson plans, today they can access many websites that offer additional materials, often at little to no cost. The RAND Corporation's American Teacher Panel survey found that nearly all teachers report using the Internet to source instructional materials.[10] Of the teachers surveyed, 95 percent said they ever used Google, 77 percent ever used Pinterest, and 73 percent ever used TPT. The 2017 survey found these materials were also used with great frequency by most teachers. For example, 55 percent of ELA teachers said they used TPT for curriculum materials at least once a week.[11] More generally, teachers report supplementing their formal instructional materials with additional resources created by themselves or their colleagues.[12] (For more information, see "What do we know about the quality of supplemental materials?")

9. Agodini, et al., *Achievement effects*; Bhatt and Koedel, "Large-scale evaluations of curricular effectiveness"; Bhatt, Koedel, and Lehmann, "Is curriculum quality uniform?"; Blazar, et al., *Learning by the Book*; and Koedel, et al., "Mathematics curriculum effects."

10. 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.

11. 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.

12. 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.

What do we know about the quality of supplemental materials?

The research base on supplemental materials is thin, despite their widespread use. Two large quantitative studies examined which website metadata (e.g., number of comments and overall ratings) predicted teachers' online resource selections. One used data from TPT and found that sales were more strongly predicted by the number of ratings and comments on a resource than by the average user rating itself.[13] Another study accessed and rated curricular resources from TFANet, an online network for Teach for America corps members. It found that several variables predicted resource downloads, including the number of ratings and comments, the mean rating, the number of characters in the description of the resource, expert-generated ratings, whether the file format was easily edited, and whether the author was a current TFA corps member.[14] As in the previous study, the number of ratings better predicted a resource's popularity than did its mean rating. One possible explanation is that the number of ratings is a common way to sort choices from the database (or could be the default), so that users select from a smaller pool of possible materials.

Three other studies investigated how teachers select web-based resources. One found that they used several criteria: alignment to Common Core standards, students' learning needs, features of the resources themselves, and teachers' own needs.[15] Another investigated the websites that preservice teachers used during their field experiences and found that teachers preferred sites that tagged materials with the academic standards to which they were aligned; engaged students in real-world problems; provided support for struggling students; and included easy navigation tools.[16] At times, teachers valued accessibility of the website (such as easy navigation) over the quality of the tasks.

The third study examined the use of Internet resources in 158 lesson plans from two teacher-education programs. It found that preservice teachers did not evaluate resources based on content but rather on their ease of access or popularity.[17] For example, one teacher used the number of "pins" on Pinterest as a means of identifying valid lesson plans (a similar approach to relying on the number of comments or ratings). Together, these studies demonstrate that teachers often value the signals that they can easily glean about a material's popularity or the fact that a website is easy to use more than the quality of the site's materials—perhaps because they cannot see the material in full until they have purchased it.

13. Samuel Abramovich and Christian Schunn, "Studying teacher selection of resources in an ultra-large scale interactive system: Does metadata guide the way?" *Computers & Education* 58, no. 1 (2012): 551–59, doi:10.1016/j.compedu.2011.09.001.

14. Samuel Abramovich, Christian D. Schunn, and Richard J. Correnti, "The role of evaluative metadata in an online teacher resource exchange," *Educational Technology Research and Development* 61, no. 6 (2013), 863–83, doi:10.1007/s11423-013-9317-2.

15. Corey Webel, Erin E. Krupa, and Jason McManus, "Teachers' evaluations and use of web-based curriculum resources in relation to the Common Core State Standards for Mathematics," *Middle Grades Research Journal* 10, no. 2 (2015): 49–64.

16. Joanne Caniglia and Michelle Meadows, "Pre-service mathematics teachers' use of web resources," *International Journal for Technology in Mathematics Education* 25, no. 3 (2018), doi:10.1564/tme_v25.3.02.

17. Amanda G. Sawyer and Joy Myers, "Seeking comfort: How and why preservice teachers use Internet resources for lesson planning," *Journal of Early Childhood Teacher Education* 39, no. 1 (2018): 16–31, doi:10.1080/10901027.2017.1387625.

III. Description of Sites

In this section, we briefly describe similarities and differences between the three focal websites.

Teachers Pay Teachers

Teachers Pay Teachers is a privately owned for-profit website where former and current teachers can create and upload instructional materials, which they offer for free or for sale, often for just a few dollars. According to national surveys, 55 percent of teachers used materials from TPT once or more per week in 2017.[18] The website reports that it contains over 3 million resources and that more than 5 million teachers (including more than two-thirds of all U.S. teachers) have used it to access materials.

Teachers Pay Teachers allows users to filter by grade level, content area, subdomain (e.g., close reading, literature, or poetry), resource type (e.g., worksheets, lesson plans, and so on), price point, and—more generally—whether the resource is free or not. Users can also provide ratings and leave comments. The average material that we reviewed on TPT had more than 300 comments and an average user rating of 3.98 on a 1–4 scale. There is no standard organizational format for materials posted on TPT because each teacher vendor creates their own. Thus, some materials are labeled with standards they purportedly align to (at the discretion of the individual author) and some are not.

ReadWriteThink

ReadWriteThink is a joint project of the International Literacy Association and the National Council of Teachers of English. According to Kaufman et al., 30 percent of ELA teachers used the website to access curriculum materials at least once a week in 2017.[19] The resources on RWT are 100 percent free and open source—no account is needed to access them. They are created by a limited number of teacher developers (fewer than 100),[20] who are compensated by the site developers. All materials are peer reviewed before being published.

Unlike TPT, resources on RWT all have the same general format. A landing page for each resource includes an overview of the material and "tabs" that list the standards covered, what resources and preparation are needed, the instructional plan, and related resources, as well as a place for users to enter comments (the average material has just two comments; user ratings are not included). Each lesson or unit typically comes with several downloadable resources that are retrieved through the various tabs. All of the RWT resources for this study were labeled with standards, and users can access the alphanumeric references for the Common Core and for other states' standards as well.

18. Kaufman, et al., *Changes in what teachers know and do.*
19. Ibid.
20. ReadWriteThink, "About Us," accessed November 8, 2019, http://readwritethink.org/about/authors/index.html.

Share My Lesson

Share My Lesson is a project of the American Federation of Teachers. It is free (like RWT) and covers all subject areas (like TPT). It is the least used of the three websites under study. According to Kaufman et al, 2 percent of teachers use it to access curriculum materials at least once a week.[21]

As with RWT, some SML resources are posted by an approved list of providers; however, anyone with an account can post materials too. Share My Lesson is somewhere between RWT and TPT in terms of standardization. Each resource's landing page is similar, with a short description, links to all the downloadable resources, and comments and ratings. Materials developed by a given content creator are often similar in look and feel. Share My Lesson resources are generally multiday and contain numerous materials—PowerPoint slides, worksheets, and so on. As with TPT, only some resources are labeled with standards. Our analysis revealed that the average material has just three comments and a mean rating of 4.6 on a 1–5 scale.

Key differences among sites are summarized in Table 1.

Table 1. Key Differences Among Focal Websites

	Teachers Pay Teachers	ReadWriteThink	Share My Lesson
Owner	Privately owned	International Literacy Association/National Council of Teachers of English	American Federation of Teachers
Free or paid	Some free, some paid	Free	Free
Percent of teachers indicating that they used the site once a week or more[22]	55%	30%	2%
Subject areas covered	All	ELA only	All
Account needed to access materials	Yes	No	Yes
Who can post materials	Anyone with an account	Approved providers	Anyone with an account
Standard format	No	Yes	Landing page, some materials
Labeled with standards?	Some materials	All materials	Some materials
Allows user ratings?	Yes	No	Yes

21. Kaufman, et al., *Changes in what teachers know and do.*

22. Ibid.

IV. Methods

This section summarizes the criteria used to evaluate materials, how websites and materials were selected for review, how expert reviewers were trained, and how the analysis was conducted and data analyzed.

Evaluation Criteria

We sought to develop a rubric that would capture both overall dimensions of curriculum material quality (like rigor and usability) and more discrete criteria that loosely reflected the key instructional shifts of the new generation of ELA content standards.

These instructional shifts are as follows:

1. regular practice with complex texts and academic language;
2. reading, writing, and speaking grounded in evidence from text, both literary and informational; and
3. building knowledge through content-rich nonfiction.[23]

Additionally, our measure needed to be flexible and concise enough that reviewers could apply it expeditiously to hundreds of materials that varied widely in scope and purpose, from single lessons to multiweek units that spanned multiple topics.

Most of the criteria were gleaned or adapted from existing instruments. We started with criteria used to evaluate the quality of state assessments in a 2016 study, for which one of us served as coauthor.[24] They required that students demonstrate a range of higher-order thinking skills, use evidence from the text to defend their responses, demonstrate research and inquiry skills, and synthesize information from multiple sources, among other areas. We also consulted the EQuIP rubric, EdReports.org rubrics, and the rubrics used in Louisiana to evaluate curriculum materials.[25] We found broad agreement across these resources in terms of their focus, and we constructed the rubric to capture the essential dimensions emphasized across these different rubrics and tools.

23. For more information, see the following: *Common Core State Standards Initiative*, "Key Shifts in English Language Arts," accessed November 8, 2019, http://www.corestandards.org/other-resources/key-shifts-in-english-language-arts.

24. Nancy Doorey and Morgan Polikoff, *Evaluating the content and quality of next generation assessments* (Washington, D.C.: Thomas B. Fordham Institute, February 2016), https://files.eric.ed.gov/fulltext/ED565742.pdf.

25. Achieve, "EQuIP," accessed November 8, 2019, https://www.achieve.org/our-initiatives/equip/equip; Ed Reports, "Rubrics & Evidence Guides," accessed November 8, 2019, https://www.edreports.org/reports/rubrics-evidence; and Louisiana Department of Education, "Online Instructional Materials Reviews," accessed November 8, 2019, https://www.louisianabelieves.com/academics/ONLINE-INSTRUCTIONAL-MATERIALS-REVIEWS.

We revised it several times in response to feedback from the project leads, Fordham staff, members of the review team, and content experts from several organizations.[26] We made further revisions after applying the rubric to a handful of sample lessons that served as a pilot.

The rubric's final dimensions were grouped into the following ten areas (see *Appendix A* for the final rubric, which offers more details about each dimension and rating scale):

1. Descriptive data
2. Alignment to standards
3. Depth of knowledge
4. Text complexity and quality
5. Close reading and evidence from the text
6. Writing task quality
7. Speaking and listening task quality
8. Usability
9. Assessment quality
10. Knowledge building and cultural responsiveness
11. Overall rating

Several subratings are nested under each dimension (for total of twenty-eight ratings) and described in the findings that follow. Descriptive elements were obtained from website metadata (e.g., the number of comments and ratings and the category of the material) and entered into spreadsheets by project staff rather than reviewers.

In general, we asked reviewers to rate materials on one of two scales. Some items were yes/no (e.g., "Is there a writing task?"). Others were rated on a four-point Likert scale (e.g., 0 = very low quality; 1 = mediocre quality; 2 = acceptable quality; and 3 = exceptional quality). We include the specifics for each scale in *Appendix A*.

26. External reviewers included staff from Student Achievement Partners, UnboundED, and EdReports.

Choosing Websites

We started with a list of most commonly used ELA supplemental materials websites as reported on the 2015 RAND survey.[27] The top sites included Google (95–98 percent reported ever using, depending on school characteristics), Pinterest (76–80 percent), Teachers Pay Teachers (73–77 percent), Readworks (45–56 percent), Newsela (21–28 percent), Share My Lesson (18 percent), and ReadWriteThink (45–49 percent). We eliminated several of these sites from the study for various reasons:

- Google and Pinterest were eliminated because there was no obvious way to identify the set of available materials for high school ELA nor to sort them by downloads or usage.
- Readworks and Newsela were eliminated because they serve a narrower purpose of strengthening students' reading comprehension, rather than providing ELA instructional materials writ large.[28]

As indicated, this left us with three final selections: Teachers Pay Teachers, ReadWriteThink, and Share My Lesson.

Choosing Materials

Our goal was to analyze the most downloaded high school lesson or unit plans that included content in reading, writing, or speaking and listening, which comprise the core of the high school ELA curriculum. Thus, we excluded lesson or unit plans with an exclusive focus on grammar, spelling, or other aspects of language. We also excluded individual worksheets or games unless they were a part of a lesson or unit plan—as the latter (one would hope) are more reflective of a teacher's daily instruction.

How we identified the most downloaded lesson or unit plans varied slightly by website. Because all materials on Share My Lesson and ReadWriteThink are free, and because they either used download data in their sorting algorithms (SML) or provided us with lists of the most downloaded materials (RWT), it was easy to identify them. For Teachers Pay Teachers, sorting strictly on downloads would have resulted in selecting mostly individual lessons, because free materials are the most downloaded and most multiday units are not free. Thus, we first stratified the available materials by units and lessons and also by free and paid. We identified the top fifteen free units, the top fifteen free lessons, the top forty-seven paid units, and the top forty-seven paid lessons, which resulted in an approximately 75/25 split between paid and free resources.

27. Opfer, Kaufman, and Thompson, *Implementation of K–12 state standards*.

28. For instance, they supply teachers with similar content at varying grade levels, as well as content that is or can be organized as a series of text sets. See the following for an example: Newsela, "Support Article: Text Sets and Collections," accessed November 8, 2019, https://support.newsela.com/item/supportArticle/text-sets-and-collections.

Training the Raters

We identified five expert raters with previous experience in test-item development, alignment studies, and item review. They were selected in part because they were generally familiar with the dimensions included in the rubric, having served in ELA leadership roles themselves and/or participated in other evaluations using the same or similar dimensions (see reviewers' bios in *Appendix B*.)

The initial training was a two-hour webinar that included an overview of all rating dimensions, as well as a detailed explanation of the criteria used for scoring. One or more example lessons were scored for each criterion to illustrate the types of evidence used to substantiate particular ratings (gathered mostly from off-grade-level materials on Share My Lesson). Reviewers discussed the examples and ratings, including offering dissenting opinions for further discussion. In addition to having the training slides as a reference, we provided reviewers with a manual that provided elaboration about exemplar materials and various "rules of thumb" to use when evaluating the different dimensions.[29]

The webinar was followed by a meeting to discuss preliminary ratings once the reviewers had gotten underway but before completing their final reviews. Reviewers also met on their own as a group and in pairs as questions arose, with the project leads addressing questions throughout the seven-week review process.

Conducting the Reviews

We assigned two reviewers to code each material. We divided the materials on each website into fifths alphabetically by title and then assigned each fifth to two raters, spiraling the raters across fifths (so that each rater had to review 40 percent of the materials from each website, with those reviews evenly split between two fellow reviewers). In total, we reviewed 328 materials—100 from RWT, 104 from SML, and 124 from TPT. Each reviewer reviewed approximately 130 materials.

We asked reviewers to conduct their evaluations in three waves. In the first wave, they coded all of the SML and RWT lessons independently. Then the project leader analyzed their data and identified areas of discrepancy. Specifically, ratings were flagged if (a) the pair of reviewers disagreed on any yes/no ratings and (b) the pair disagreed by two or more points on any four-point scale. In the second wave, we sent the SML and RWT ratings back to reviewers and compelled them to come to consensus in pairs on all of the yes/no ratings.[30] We also encouraged them to discuss and make adjustments to the four-point scale ratings if possible but did not require consensus. Finally, in the third wave, reviewers analyzed the TPT materials.

29. All of these materials are available upon request.

30. Project leadership initially assumed that it would be fairly straightforward to ask reviewers to agree on yes/no questions, such as, "Is there a writing task included in the material?" As it turned out, such questions were not so simple, as reviewers had to define the minimum threshold for what constituted a reading, writing, and speaking/listening task. For instance, in the speaking/listening area, did group work, with a student moderator providing a summary of the group's discussion to the class, constitute a speaking task?

Again, reviewers had to come to consensus on the yes/no items and were encouraged to discuss any two-point discrepancies on the four-point scales, just as they had done for the other sites' materials.

Analyzing the Data

We analyzed the data using straightforward descriptive methods. Below we present distributions of rating scores, overall and by site. We also conducted simple tests of mean differences where we make claims about differences among the sites. Any time we describe a difference between two sites, the difference was statistically significant in a two-sample t-test at $p < 0.05$. (For an overview of all results, see the *Discussion* section, Tables 2–4.)

To gain additional insights about our study, we conducted interviews with seven educators who are experienced in accessing online curriculum materials. Five were teachers, one was a literacy coach, and one was a technology coach. All were either currently teaching ELA or recently taught the subject. The teachers hailed from four different states: Kentucky, Texas, Utah, and Vermont. We interviewed each teacher individually by telephone, asking why, how, and how often they access supplemental materials and what kinds of ELA materials they search for. Interview data are presented in several sidebars throughout the report.

V. Findings

Let's first review the types of materials and content that teachers are accessing (for more on this topic from teacher interviewees, see "Why Teachers Supplement").

The most common materials on the three websites vary considerably in their content, form, and structure. However, there are several clear patterns that emerge across the 328 materials that we reviewed.

First, teachers go to these websites primarily to retrieve lessons and units focused on reading and writing, as evidenced by both how the materials were catalogued and the nature of what students were asked to do.[31] As we gathered materials for the study, we tagged them as focused on reading, writing, and/or speaking and listening (not mutually exclusive). These identifiers were based either on content descriptions offered on the respective websites or on a quick perusal of the material (to rule out that it was focused solely on language). In all, we tagged 95 percent of the materials as writing focused, 79 percent as reading-comprehension focused, and 52 percent as speaking and listening focused. These results do not appreciably differ across sites; writing was the dominant focus on all three (though SML was more likely than the other sites to have materials tagged as covering multiple areas).

This pattern is corroborated based on reviewers' ratings of whether the lessons included reading, writing, or speaking and listening tasks.[32] Based on those ratings, 82 percent of materials have a writing task, 73 percent a reading-comprehension task, and 43 percent a speaking and listening task. These results differ somewhat by website. About 80 percent of SML materials have a reading-comprehension task, as compared to just 60 percent of RWT materials, and 47–49 percent of RWT and SML materials have a speaking and listening task, as compared to just 34 percent of TPT materials.

31. Recall that we constrained materials to units or lessons focused on reading, writing, or speaking and listening, so we cannot say anything about whether other content areas or other types of materials are more common than these. But of the materials we observed, reading and writing are most common.

32. Tasks were defined as follows: For reading comprehension, there must be a reading passage or a visualization for which the student must engage. There must also be one or more questions to which the student must respond that demonstrate whether the student has understood the passage/visualization. For writing, students must be required to write at least one paragraph in response to a prompt of some kind. Fill-in-the-blank (even if a paragraph in length) or individual-sentence answers do not count. For speaking/listening, students must listen to an audio recording and respond to it in some way; watch a video recording and respond to it in some way; or give a speech or other oral presentation. Responding to static images does not count, nor does reading aloud, though presenting a scene (e.g., from a novel) in a dramatic fashion would count.

The most common format of the materials is a multiday unit focused on one or more core texts, with both reading comprehension and writing as a major focus. The specific texts vary considerably, and very few texts are used in more than a handful of materials. The only real exception is *Romeo and Juliet*, which appears as the focal text in eleven sets of materials. The texts referenced range from classics (e.g., Shakespeare, *Canterbury Tales*, and *The Color Purple*) to modern texts (e.g., *St. Lucy's Home for Girls Raised by Wolves*, *Down These Mean Streets*, and *Brown Girl Dreaming*). They include fiction, nonfiction, and poetry—though fiction is far more common than nonfiction (and the most common nonfiction materials being speeches and memoirs).

Teachers clearly have a preference for longer units than individual lessons. On average, the RWT materials indicate they cover five days' worth of instruction, while the SML materials indicate they cover seven. On both those websites, just 10 percent of the top-downloaded materials are intended for one day's instruction. Teachers Pay Teachers materials cover longer periods of time, on average—although, recall that our selection criteria for TPT ensured a large proportion of the materials were multiday.[33] On average, across all of the sites and all of the materials, the mean length of time the materials are intended to cover is 8.8 days.[34]

33. As a reminder, our selection approach for TPT ensured that half of the materials would be units.

34. Materials come with many downloadable items for teachers to use (e.g., lesson plans, slide, assessments, and worksheets). The mean number of downloadable items is 9.4. There are some interesting between-site differences, however—TPT materials report covering longer periods of time (13.5 days) but have fewer items (7.1). The other two sites report covering shorter periods of time (RWT 5.0 days, SML 7.2) but include many more items on average (11.3 and 10.5, respectively).

Why Teachers Supplement

According to our interviews, teachers turn to online supplemental materials for four main reasons: to increase student engagement, to meet their students' diverse needs, to fill instructional gaps, and as a way to save time. We touch on each below.

Most often, teachers report searching online sites for new classroom activities or innovative materials that will spark student engagement. Several teachers stressed that today's students need stimulation and varied instructional approaches in the classroom—and online materials help fill that need.

- *"I look for things that are more engaging so I can get my kids to buy in a little more. Textbooks are boring a lot of the time. I like bringing in more engaging activities. I'll download fun activities for the first couple weeks of school, and end-of-year stuff like review materials for the end-of-year assessment. I look for hands-on, task-y stuff like scavenger hunts, which are really big right now."*
- *"I'm looking for activities more and more because that's where teaching is going—there's a push to have kids create and a push to have kids work more in small groups. Therefore, teachers need the ability to do different things. Also, without variety, kids can decide they are bored."*

Teachers also say they go online to find ways to meet their students' diverse needs. They search for materials for students who need more enrichment or more practice; frequently, they are looking for ways to unpack important concepts for struggling learners. Consequently, they see the opportunity to download off-grade materials—both below and above grade level—as a significant advantage to online sites. Interestingly, they don't expect a specific lesson or unit to include multiple ways to reach diverse learners; rather, they shoulder the task of locating appropriately challenging lessons by searching widely among multiple online offerings.

- *"I have a set of three units, and they are amazing. Even so, I still need activities to go with these units because I might need multiple ways to reach certain students . . . or some students may need more practice. When I have those kinds of needs, then I go to TPT."*
- *"I'm always looking at college-level and middle-school-level materials to reach my higher and lower students."*
- *"I'll download a whole unit, but I'm often just looking for activities to cover something that kids aren't getting."*

Why Teachers Supplement, cont.

Teachers also report looking for content that fills an instructional gap. Several teachers mentioned searching for online resources before beginning a literature unit, including searching for deeper analyses and interpretations of a text.

- *"If I use something from online, it has to fit where I need to go. I'm asking, 'Does it fit the skill I want to teach?' For example, if my students are having problems summarizing, will this activity reinforce that skill?"*
- *"Generally, I look for literature and writing. I may research a particular novel or look for sources for literary analysis. I don't look online every week, maybe once a month or once a unit. Or I might say, 'What I'm doing right now isn't working; I've exhausted all my ways of trying, so let me look for something new.'"*
- *"I found some materials to help me teach academic discourse. They show how to have an academic debate, which is definitely lined up well with speaking and listening standards and also aligns with opinion/argument writing."*

Finally, teachers report accessing materials to save time. They find a lot of lessons that they too could have developed, had they enough time, but downloading ready-made resources saves hours. At first, they reported feeling awkward—even guilty—when they started using online lessons and activities, but it was short-lived after they learned how much time they were saving. Even if they have to modify the material, it reduces preparation time, and the cost is seen as well worth the expense.

- *"I used to think I had to do everything myself and it was wrong to use others' materials. But that becomes exhausting. It's not plagiarism; it's part of the profession."*
- *"A lot of teachers say, 'I can't believe you bought it; I can't believe you didn't create your own.' My response would be, 'It saves me time, saves me stress, and it's about the implementation of the materials I buy.' I don't use them as is; I modify them to meet my needs."*
- *"Two dollars is a lot cheaper than two days of my time."*

Strengths

Our reviewers identified two main strengths in the materials.

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

We asked reviewers to provide an overall rating of the quality of the reading texts in each material, essentially asking the question, "Is this something students should read?" More specifically, readers considered whether the text was important, well written, and contained grade-level content. Reviewers generally thought the texts were of good quality, with a mean of 2.21 on a 0–3 scale.[35] In fact, as shown in Figure 1, "exceptional quality" is the most common rating. 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 Teachers Pay Teachers (mean of 1.96).

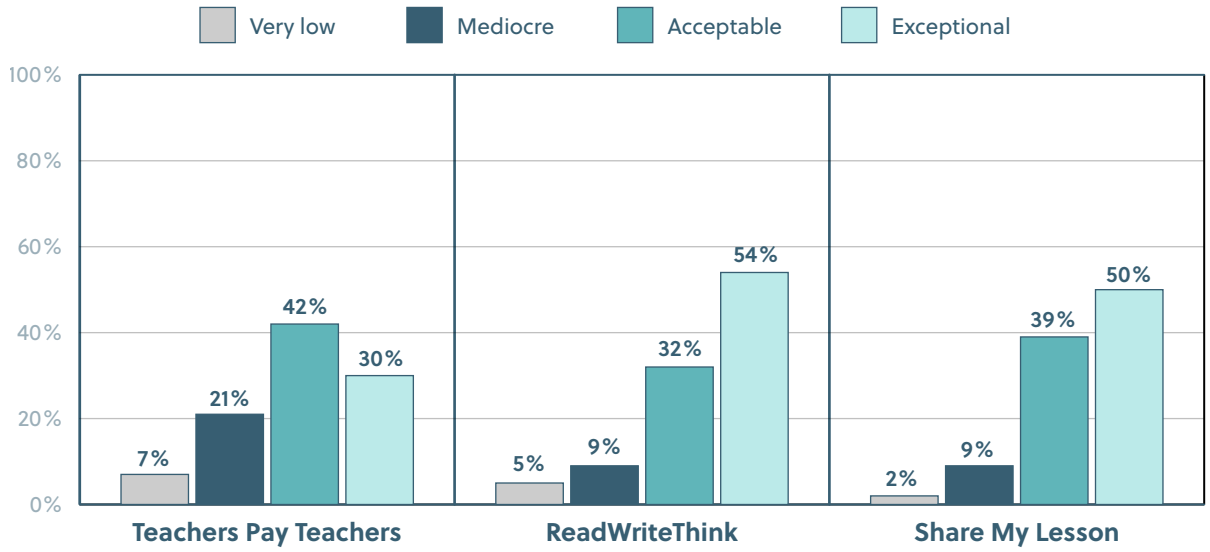
While the overall ratings for text quality are high, one factor was often associated with lower ratings: the grade-level appropriateness of a text. For example, one lower-rated text was a Roald Dahl book that reviewers thought was more appropriate for middle schoolers; another was the young adult book *Freak the Mighty*, which is aimed at grades 4–7. In fact, for those texts for which we could identify a Lexile level, the average was 998, which is about the 50th percentile for finishing fifth graders. Just 25 percent of the texts were at a Lexile level of 1,155 or above, which is the 50th percentile for finishing eighth graders. Recall that the sole criterion we used for text quality was holistic and included three dimensions (quality of the writing, whether the text included grade level content, and the importance of the text). We anticipate had we separated those measures, the texts would have performed very poorly on inclusion of grade-level content.[36]

35. 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.

36. That said, low Lexile levels tend to be less of a concern for literary text than for informational text. In fact, it is not unusual for literary texts to measure low on the Lexile scale but still carry an appropriate level of high-school complexity in terms of themes, organizational structures, characterization and other literary elements. For instance, the work of John Steinbeck, Jack London, and Ernest Hemingway are often cited as examples. On the other hand, if informational texts used in high school are not in the 1000 to 1300 range, that's a serious concern since comprehension of such materials is crucial for college readiness.

Our reviewers also examined the reading-comprehension tasks in terms of how they ask students to engage with the texts, focusing on several key shifts called for in ELA state standards: requiring students to use evidence from the text, focusing on central ideas or important particulars, and requiring close reading and analysis.[37] They found that the materials demonstrate some but not overwhelming coverage of these key shifts. For instance, as to requiring students to use evidence from the text, on a 0–3 scale,[38] the average score across the three websites is 1.77, with SML (mean = 1.96) and TPT (mean = 1.81) outscoring RWT (mean = 1.49). Scores are slightly lower for whether the task has a focus on central ideas/important particulars (mean = 1.73) and whether the task requires close reading and analysis (mean = 1.58).

Figure 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.

37. The guidance for teachers is as follows. For close reading and analysis, the key point is that students analyze the text, not merely that they use it as a springboard to answer questions. For focus on central ideas or important particulars, the key point is that tasks help students understand the gist of the reading or that if the task asks about details, they are important details. For requiring evidence from the text, the key point is whether the task requires textual justification for students' responses. For more information, also see the appendix in Doorey and Polikoff, *Evaluating the Content and Quality of Next Generation Assessments*.

38. Full scale is as follows: 0 = not at all; 1 = yes, to a small extent; 2 = yes, to a moderate extent; and 3 = yes, to a major extent.

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

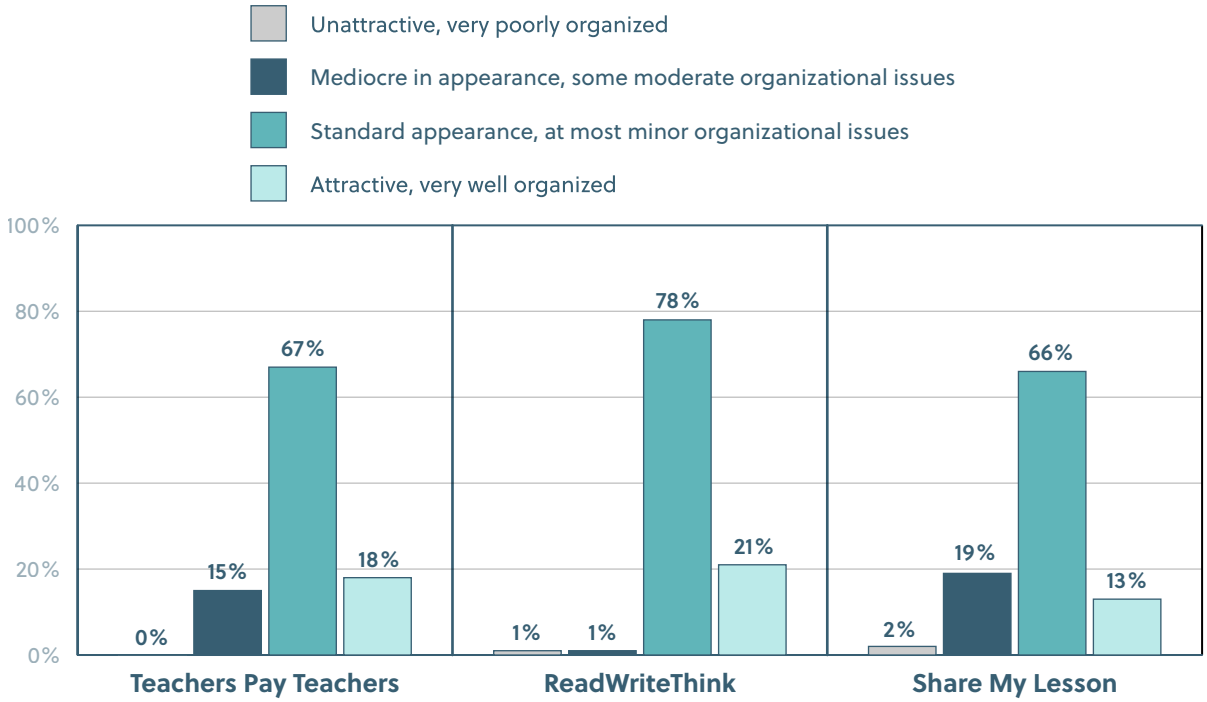
Reviewers examined five dimensions of usability (i.e., interesting and engaging; free from errors; visual design and organization; clarity of guidance; and supports for diverse learners). On two of the five, the materials scored favorably. At the most basic level, reviewers found that the materials were generally free from errors that might affect student understanding. On a 0–3 scale,[39] 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 SML has the most (mean = 2.53), with TPT in the middle (mean = 2.79). When there are errors, they tend to be typographical in nature, and although that carelessness frustrated reviewers, they conceded that typos are not likely to affect student understanding.

Materials also rated well in terms of their visual appearance and organization. On a 0–3 scale,[40] the mean across sites is 2.04, and 79 percent of all materials earn a 2 or a 3 on this dimension. There are modest differences across sites, with SML materials scored as the least attractive and least organized (mean = 1.89) and RWT the most (mean = 2.19). As shown in Figure 2, RWT has very close to zero materials earning a lower score on this dimension (likely because they have the most standardized formats and the smallest number of authors), while about 20 percent of SML materials earn a 0 or 1 (for more, see "On Visual Appeal and Ease of Use").

39. 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.

40. 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.

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



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

On Visual Appeal and Ease of Use

According to our interviews, teachers prize formats that are eye-catching and engaging. Many believe that attractive formats will increase student interest and involvement and that the formatting of their own materials falls short. Yet formatting that is too juvenile alienates students. Most all teachers remarked that "cutesy" or "showy" formatting is no substitute for rigor.

Teachers also want materials that are clearly written and easy to use; they will bypass anything that is not well explained or seems too complicated. For instance, one teacher looked at an activity that required her to print several pages and then cut out a hundred small cards; she said she would either forego that particular lesson or find a way to adapt it to make it more practical for her.

- *"If it's not friendly or interesting, students will shut down. We will always do some activities that are not visually interesting, as when we are annotating texts or doing peer reviews [of student writing]. But these need to be mixed up with things that are a lot more appealing to the eye. I don't think that fonts and design are the only reason to buy, but it matters; it does show that the teacher cares. Otherwise, I can do it myself. Goes back to making my life easier."*
- *"Format is a double-edged sword. It's good to use materials that are standards aligned and appropriate to grade level, but visual appeal can be the siren call of something that's just cutesy and shouldn't be used."*
- *"You learn that the work of some teachers online is just cutesy, without substance. And I don't really need all the clip art and fonts. What I look for is, 'I love this and it would take me forever to make it, so I'm going to purchase it.'"*
- *"I follow teachers on Instagram, and I know who has which niche. One teacher is all about grammar. Another teacher is about writing and also novels like The Great Gatsby. Another does workbooks; that's her thing. She takes a novel and does questions for each chapter. The workbooks are visually engaging and created with electronic tools I would have had to purchase myself. Instead of doing it all themselves, teachers can buy a set of workbooks for $11 or $12."*

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.

After providing separate ratings for each dimension, reviewers provided an overall or capstone rating for each material at the end. It's helpful to provide this big-picture snapshot first.

On a 0–3 scale,[41] with 2 or higher corresponding to materials our reviewers thought teachers should use, the mean score for materials is 1.28, with reviewers recommending that 64 percent of them *not be used* or, similarly, are *probably not worth using*. No website has a majority of materials earning a positive (2 or 3) rating, but RWT receives a slightly higher overall rating on average (mean = 1.41) than SML (mean = 1.29) or TPT (mean = 1.18). Figure 3 shows that on all three websites, a majority of materials is rated *mediocre* on this scale.

If busy teachers are going to take the time to look for supplemental materials, they certainly want to know (quickly) how to use them (for more, see "The Process of Accessing Materials"). Hence, 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,[42] with a 2 intended to represent standard guidance—for example, a reasonably detailed lesson plan that would easily convey how teachers were to use the materials—the mean across the three sites is 1.61. ReadWriteThink materials by far provide the clearest guidance, with a mean of 2.00; in contrast, TPT materials (mean = 1.50) and SML materials (mean = 1.37) are rated much lower. As an example of poor guidance, consider a CliffsNotes-style lesson on a Kurt Vonnegut short story, which includes the story itself and a guide summarizing the plot, characters, and so on, yet the material has no instructions at all for teachers about how to use it.

We also examined whether reviewers' overall ratings corresponded with the available evidence about teachers' perceptions of the quality of the materials. We correlated the overall ratings with the available metadata on the number of downloads, number of comments, and average ratings—each of which could be a proxy for quality (though comments could, of course, be both positive and negative). We ran correlations separately by website, since the averages were so different from site to site.

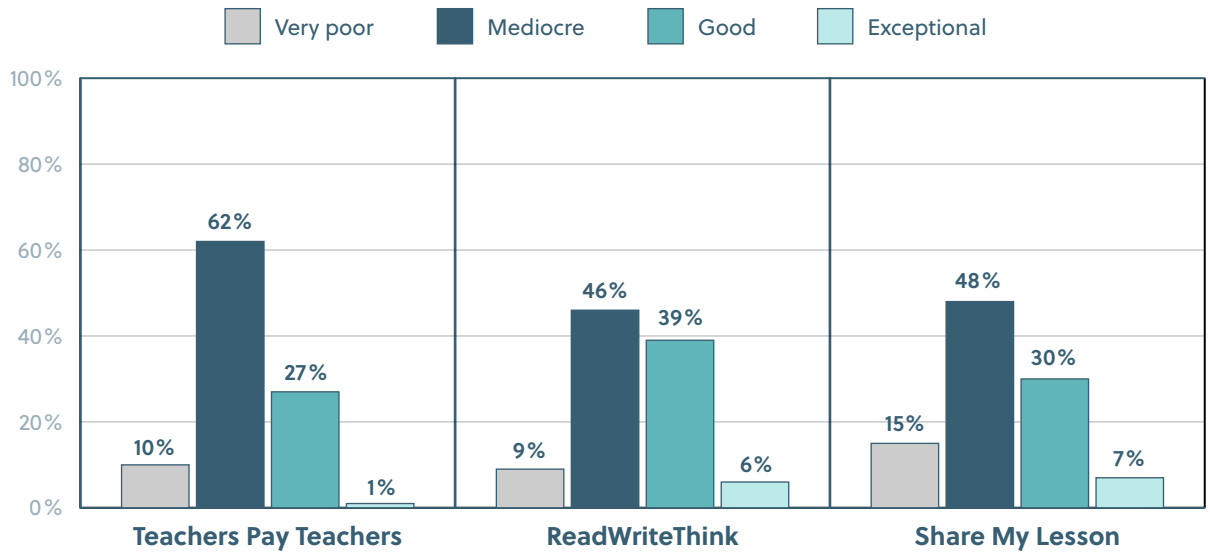
41. Full scale is as follows: 0 = very poor, teachers should not use this material; 1 = mediocre, has some good and some bad components (e.g., well organized but not on important content or covering diverse perspectives but using weak tasks), probably not worth using; 2 = good, overall a high-quality 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.

42. 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.

In short, there is little relationship between our ratings and these metadata. Two of the eight correlations are greater than 0.20 in absolute value. For TPT, our quality ratings are moderately positively correlated with the number of downloads (recall that TPT download data were only available for the free materials), with a correlation of 0.38. In other words, we did see modest evidence that the more downloaded materials on TPT are higher quality as judged by our ratings.

For SML, our quality ratings are weakly negatively correlated with the number of comments on the materials, with a correlation of −0.21. This could indicate that the weaker materials see more comments that are constructive or critical. None of the other correlations are as large as these two. In short, there is at best modest evidence that material quality (as judged by our overall ratings of quality) predicts teacher use of materials—and then only on some websites.

Figure 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.

The Process of Accessing Materials

When it comes to looking for materials online, teachers take different approaches. Some search regularly a few hours a week; others search when starting a unit or when they feel a need for something new and different. Some start with a favorite website; others just begin with Google and see what appears. Still others use Instagram as a filter to streamline and focus their online searches. Teachers follow content developers on Instagram whose work they've already used and receive tips about related materials ("If you liked this, you may also like. . ."). Alternately, some teachers regularly read the blogs of their favorite teacher developers.

- *"I tend to focus on literature. I look online for activities related to the literature I'm going to be teaching. I either Google it or look in my favorite sites."*
- *"I often search for what's new. I have a few teachers who think the same way I do, and I follow them on Instagram. I can always find some teachers who are on the same wavelength as I am."*
- *"Instagram also has videos of how the materials are used: if I buy a product, Instagram shows me how the materials have been used. I can hear the author explain her materials, see examples of student work, or even see live recordings of her class. I follow teachers who have 'TPT teacher' at the top of their page. And they provide links to their TPT sites. TPT is overwhelming: 'How do I search? What do I look for?' Instead, Instagram allows you to follow teachers you're familiar with and teachers whose work you respect."*
- *"It takes me a long time to search, usually three hours. If I just go to one site, like TPT, it probably takes an hour or hour and a half. If I go to other sites, it takes longer."*
- *"I'd describe myself as a moderate to heavy user of the sites. I spend about five or six hours a week searching (I do it as I'm watching TV)."*

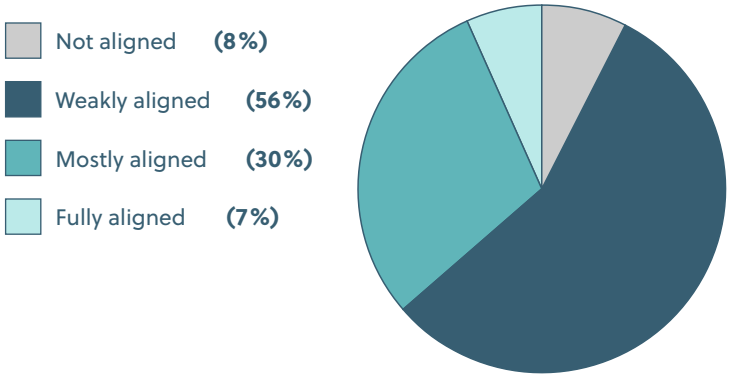
FINDING 4: The materials are weakly to moderately aligned with the standards to which they claim alignment.

We asked reviewers to evaluate the alignment of the materials against the standards to which they claim alignment. Thus, materials that did not include standards alignment were excluded (about 40 percent of TPT and SML materials and about 10 percent of RWT materials did not indicate alignment). Even when judged against this standard, the materials on the three sites do not fare well.

Respondents used a 0–3 scale that ranged from *not* to *fully aligned*.[43] The average alignment rating is 1.35 or about one-third of the way between weakly and mostly aligned. As shown in Figure 4, 56 percent of all materials score a rating of 1, 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."[44] These low alignment ratings occur because most materials claim alignment to a very large number of standards (sometimes even standards at multiple grade levels!), presumably to make the materials more likely to show up in a search.

Although high-quality literacy instruction naturally connects multiple standards across the reading and writing domains, the concern is that the materials are including any and all standards that might loosely apply, rather than addressing a few key standards. To wit, reviewers report that the materials are aligned to some of the listed standards but very rarely aligned to most or all. Average alignment ratings are identical for TPT and RWT (mean = 1.28), but they are higher for SML (mean = 1.56). (For more information, see "How well do online materials align to standards?")

Figure 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.

43. 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.

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

How well do online materials align to standards?

The teachers we interviewed recognized that a lack of standards alignment can be a significant flaw in online materials. Several noted that a lesson may be interesting and appealing but not at all aligned to college- and career-readiness standards or even to the right grade level. Some teachers cited examples of colleagues who were led astray from standards-based instruction by the engaging nature of an online activity or lesson.

Others expressed frustration with the fact that online materials often claim alignment to a dozen or more individual ELA standards. Although it's conceivable that a set of lessons could touch on that many standards given the interdependence of reading and writing, it's impossible to align deeply to all of them. In these cases, teachers say the alignment work falls on them.

- *"Any time I looked for lessons or content on [X website], I would find things that didn't seem fully aligned. Here's the standard I'm looking for, but the lessons say they cover all the standards. The materials were not specifically geared toward the particular standard that I'm looking for. I can see that the eighteen standards listed are included in the lessons, but there's no particular focus on specific standards. And it's the same thing with most websites."*
- *"As a literacy coach, I tell my teachers that just because it's cute doesn't mean it's good . . . . I find a lot on [X site] that doesn't meet the standards, but it is cute."*
- *"Typically, I will download a bunch of materials at the beginning of a unit and then see if I need them. So I purchase two or three things, and then after I review them, I look at the standards they are supposed to align to, and then I modify the materials to fit the framework I need."*

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

WRITING TASKS

Though the quality of texts is good, the quality of the writing and speaking and listening tasks is much weaker. Recall that 82 percent of materials 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,[45] the writing tasks average 1.42.[46] Just 6 percent of writing tasks earn a score of 3, while 51 percent earn a score of 0 or 1. As shown in Figure 5a, there are no differences across sites in writing-task quality—all three sites are between 1.40 and 1.44. We also asked reviewers to rate on a 0–3 scale the extent to which tasks required writing to a text.[47] Here TPT (mean = 1.60) and SML (mean = 1.55) fare the best, while RWT scores lower (mean = 0.98).

Strong writing tasks often score highly on both the overall rating and writing to a text (the correlation between the two ratings is 0.56). For example, one well-rated writing task asks students to respond to texts and videos about nonviolence using the following prompt: "In a well-written essay, choose one Principle of Nonviolence and defend how [character] demonstrates it (through his words and actions), citing at least three pieces of evidence to support your idea." Weaker writing tasks often fall short in several ways: they are vaguely worded, largely focused on personal feelings, or—most importantly—not text dependent. For example, one poorly rated writing task asks students to select five important events from their life and write an autobiography but offers little guidance beyond that. Another in a unit about Veteran's Day asks students to choose a war and write a narrative including one day in the life of a veteran before, during, and after the war, using details and event sequences (but no assigned text).

SPEAKING AND LISTENING TASKS

As indicated, 43 percent of materials had a speaking and listening task, and the scale used to judge quality was the same as the writing task.[48] 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. However, even fewer speaking and listening tasks—just 4 percent—earn the top score. For example, one of the top-rated tasks asks students to create a podcast, drawing on contemporary news stories about an issue of their interest; another asks students to create and deliver a spoken-word poem based on the poems they've read in class.

45. 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.

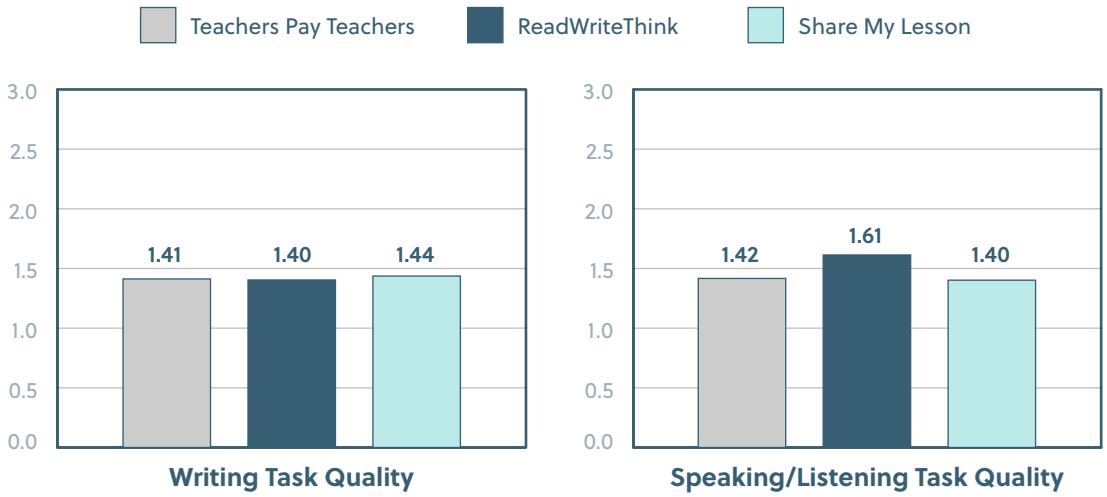
46. The rubric mandated that in order to score 3, the task had to require writing to a text.

47. Full scale is as follows: 0 = not at all; 1 = yes, to a small extent; 2 = yes, to a moderate extent; and 3 = yes, to a major extent.

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

As shown in Figure 5b, there is a small difference favoring RWT on speaking- and listening-task quality, with a mean of 1.61 (versus 1.42 and 1.40 for TPT and SML, respectively). Strong listening tasks generally require analysis of what students are hearing or seeing; weaker listening tasks focus more on personal reactions without analysis. Strong speaking tasks require students to present information with supporting evidence; weaker ones are more appropriate for lower grades or have unfocused, nonspecific directions.

Figure 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.

To ensure that reviewers were evaluating the same assessment within a unit or lesson, the project team located a culminating assessment in each material to be analyzed.[49] This resulted in some of the materials being rated on assessments that were not indicated as such by the content creators (e.g., culminating activities where it was not clear that students were being graded). That said, the assessments, as identified, were not rated strongly by reviewers on any of the three dimensions evaluated. First, we asked whether the assessments covered the core content of the lesson or unit. On a 0–3 scale,[50] where 2 represents assessment of more than half of the core content of the lesson/unit, the materials average a 1.84. Just over two-thirds of materials (69 percent) earn a 2 or a 3 on this scale. As shown in Figure 6a, assessments from RWT (mean = 2.07) are rated as covering more core content than those on TPT (mean = 1.80) or SML (mean = 1.67).

Rubrics that provide clear guidance on how to evaluate student performance on assessments are valuable resources for teachers. However, a bare majority (51 percent) of materials fails to include such rubrics. On a 0–3 scale, ranging from "no" to a "high-quality" rubric,[51] the mean score across the three websites is 0.94. Share My Lesson fares by far the worst on this metric, with a mean of just 0.44, while the mean score is 1.21 for TPT and 1.14 for RWT (see Figure 6b).

Finally, reviewers provided an overall rating of the quality of the assessment, using a similar quality rating scale as for the quality of the text, writing, and speaking and listening tasks.[52] Overall, the assessments are rated poorly, scoring 1.27 on the 0–3 scale. As shown in Figure 6c, the quality of assessments is slightly better on RWT (mean = 1.50) than TPT (mean = 1.15) or SML (mean = 1.21). But very few assessments (just 4 percent) are rated as having *exceptional* quality, and 57 percent are rated as having *very poor* or *mediocre* quality. Assessments tend to score poorly because they are not well aligned to the standard or the lesson, because they are not text dependent, or because they focus on recall or simple forms of comprehension (as opposed to in-depth comprehension and analysis).

49. Note that reviewers could identify a different assessment if they disagreed with the project team's choice.

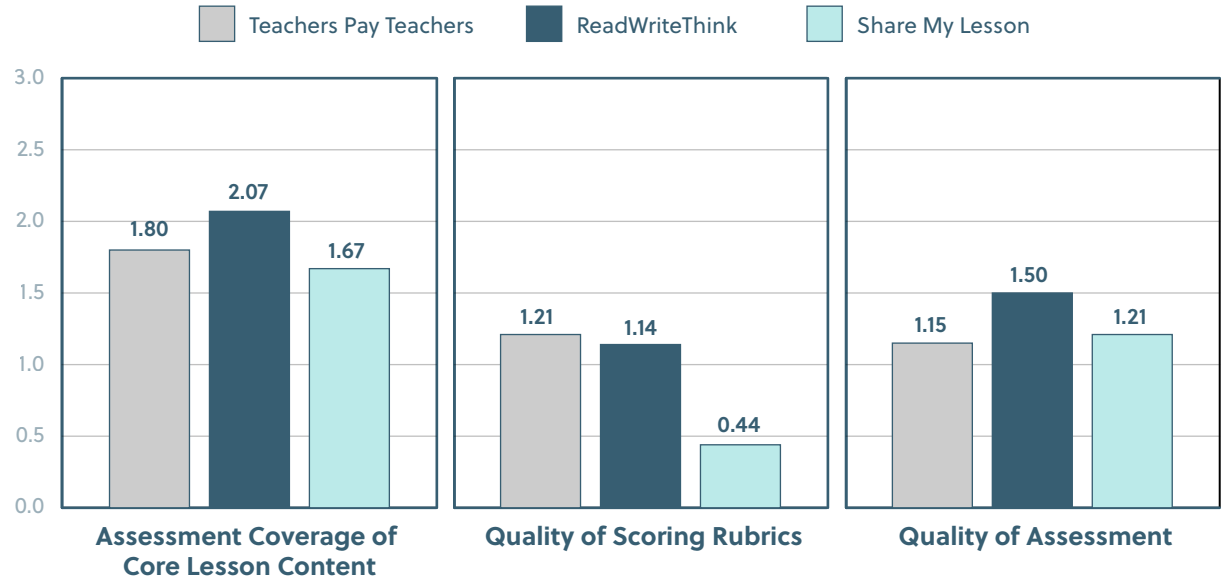
50. Full scale is as follows: 0 = very poor coverage—fails to assess the core content of the lesson; 1 = mediocre coverage—assesses some core content in the lesson but has some large gaps; 2 = good coverage—assesses most of the content in the lesson, at most small gaps; and 3 = full coverage—assesses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

51. 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.

52. 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.

To wit, a weeks-long unit on *The Great Gatsby* includes assessments that are mostly multiple-choice recall and simple comprehension questions (e.g., "Where does Gatsby's reunion with Daisy take place?"). The same can be said for a unit on *Lord of the Flies*, in which students are asked to respond true or false to this statement: "Jack considered fire more important than anything." When short-answer questions are included, they also tend to focus on recall or simple comprehension.

Figures 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.

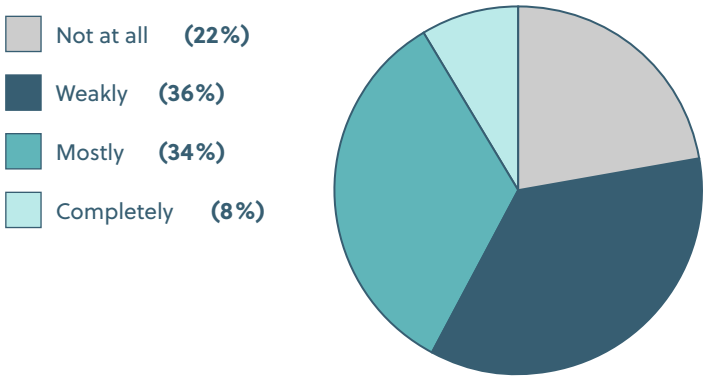
Simply knowing more about a variety of topics has been shown to speed and strengthen reading comprehension.[53] Because it presumably takes more than one lesson to build student knowledge on a topic, we asked reviewers to evaluate only the extent that multiday units did so—as opposed to single day lessons (75 percent of evaluated materials were units). By "building knowledge," we mean the extent to which the unit introduced and sequenced knowledge in a way that allowed students to demonstrate their understanding of a topic in a domain like social studies, science, technology, the arts, and so on. Although some tend to equate such a sequential approach with the elementary grades, research shows that benefits accruing to knowledge and vocabulary from topically connected texts is not limited to the primary years.[54]

We find that some units do in fact build knowledge. Specifically, 42 percent of units score a 2 or a 3 on this dimension, indicating that they support students' ability to *mostly* or *completely* demonstrate their knowledge of a topic (Figure 7). The mean score on the 0–3 scale is 1.28. The units on RWT (mean = 1.50) and SML (mean = 1.54) are rated more highly than the units on TPT (mean = 0.91). However, more than half of all units build knowledge *weakly* or *not at all*. These units tend to be devoid of historical or literary content, focusing instead on skill building, simple recall, or personal interpretation without emphasizing any particular content (e.g., a unit on *Romeo and Juliet* without any reference to Elizabethan England or a unit on *The Great Gatsby* without any reference to the Roaring Twenties).

53. As cognitive scientist Dan Willingham explains, "Background knowledge makes one a better reader in two ways. First, it means that there is a greater probability that you will have the knowledge to successfully make the necessary inferences to understand a text. Second, rich background knowledge means that you will rarely need to reread a text in an effort to consciously search for connections in the text." See, for example, the following: Daniel T. Willingham, "How Knowledge Helps," *American Educator* (Spring 2006), https://www.aft.org/periodical/american-educator/spring-2006/how-knowledge-helps.

54. Donna R. Recht and Lauren Leslie, "Effect of prior knowledge on good and poor readers' memory of text," *Journal of Educational Psychology* 80, no. 1 (1988): 16–20, doi:10.1037/0022-0663.80.1.16; Hugh W. Catts, "The narrow view of reading promotes a broad view of comprehension," *Language, Speech, and Hearing Services in Schools* 40, no. 2 (2009): 178–83, doi:10.1044/0161-1461(2008/08-0035); David A. Kilpatrick, *Essentials of assessing, preventing, and overcoming reading difficulties* (Hoboken, NJ: John Wiley & Sons, 2015); and Walter Kintsch, *Comprehension: A paradigm for cognition* (Cambridge, UK: Cambridge University Press, 1998).

Figure 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. Because the materials are often quite extensive, covering multiple lessons, we bounded the task by focusing only on the DOK of the main activity (which we identified for reviewers, such that all materials had a main activity). We used a four-level DOK scale built off Norman Webb's taxonomy:[55] level 1 means recall and reproduction; level 2 means skills and concepts; level 3 means strategic thinking and reasoning; and level 4 means extended thinking.[56] We asked reviewers to indicate what percentage of the content in the main activity is at each DOK level on a 0–4 scale.[57]

55. Normal L. Webb, "Issues related to judging the alignment of curriculum standards and assessments," Applied Measurement in Education 20, no. 1 (2007): 7–25.

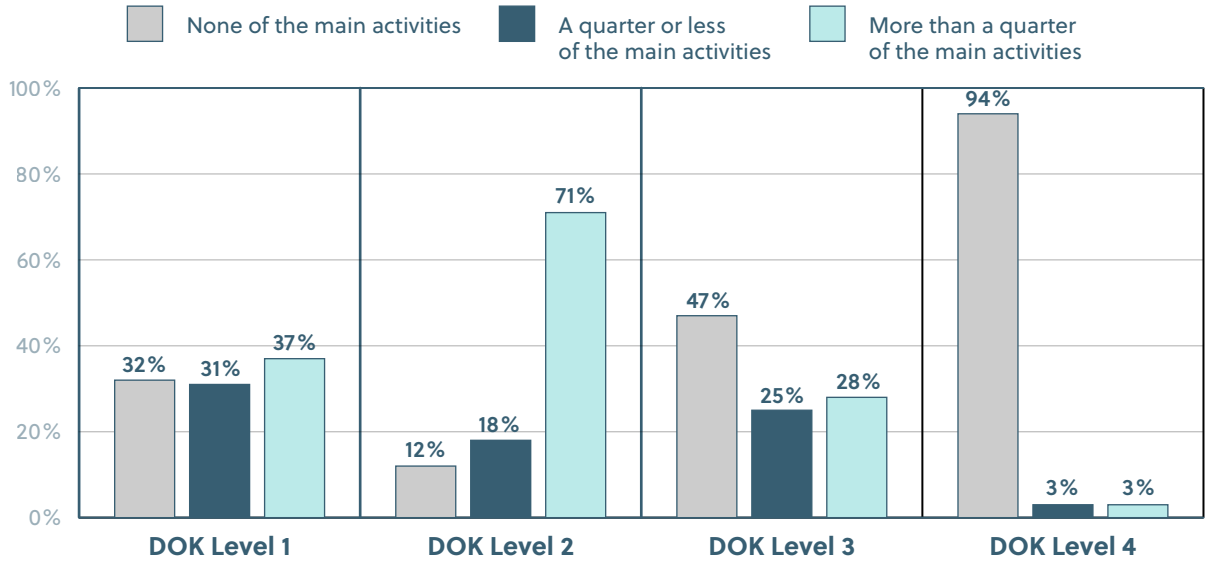
56. For full definitions of the taxonomy, see the following: Karin Hess, *A guide for using Webb's Depth of Knowledge with Common Core State Standards* (Common Core Institute, 2013), https://education.ohio.gov/getattachment/Topics/Teaching/Educator-Evaluation-System/How-to-Design-and-Select-Quality-Assessments/Webbs-DOK-Flip-Chart.pdf.aspx.

57. Full scale is as follows: 0 = 0% of the content of the main activity at that level; 1 = 1–25%; 2 = 26–50%; 3 = 51–75%; and 4 = 76–100%.

Results in Figure 8 show that virtually none of the main activities in the materials have any DOK level 4 content. In fact, just 6 percent of the main activities score higher than 0 for DOK level 4 (the navy and teal bars in the fourth set), and just 1 percent of main activities are mostly (greater than 50 percent) DOK level 4. Nearly half (47 percent) of the main activities have no DOK level 3 content at all (the grey bar in the third set), and just 28 percent have more than one-quarter of the activity's content tagged as DOK level 3 (the teal bar in the third set). Indeed, most of the content in the main activities is DOK level 1 or 2. Of course, that is not to say that there is no place for low-level DOK questions but rather that more literal/recall questions should often serve as scaffolding, enhancing students' ability to do more in-depth analysis and deeper cognitive work. Of the main activities, 71 percent have more than one-quarter of their content deemed DOK level 2 (the teal bar in the second set), and 37 percent have more than one-quarter deemed DOK level 1 (the teal bar in the first set).[58]

One activity rated as 100 percent DOK level 1 asks students to read a handout about racial profiling and answer simple factual questions that either draw on their own knowledge or rely on information in the handout. In contrast, a material that is rated as majority DOK level 3+ asks students to read a text and "write an explanatory essay analyzing how the themes of guilt and responsibility interact and develop over the course of the text. Consider the author's organizational choices of story details in both chapters about the man that was murdered."

Figure 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.



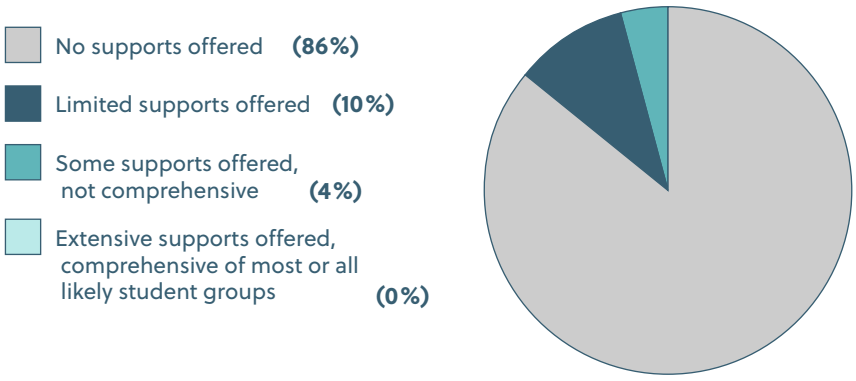
Note: Full scale as shown. Numbers may not sum to 100 percent due to rounding.

58. Evaluating the DOK of the main activity proved to be one of the most challenging assignments for reviewers, mostly because they felt that the cognitive challenge of a lesson stems partly from the materials themselves and partly from what the teacher does with those materials.

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. Specifically, we asked how comprehensive the supports for differentiation were 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 they offer no support for differentiation (Figure 9). Less than 1 percent of materials score a 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 SML (mean = 0.34) than the other two sites (means of 0.10 and 0.15).

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



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

An example of a well-rated material (score = 2) is a unit on the book *The Things They Carried*, where each day's materials offer some guidance about differentiation for students with disabilities and English-language learners. This is a rare exception, however; teachers using the vast majority of the materials on any of these sites will be on their own when it comes to planning supports for diverse learners (that said, our interviewees insisted that they don't use the materials "as is," often customizing them for ability level; see "Some Assembly Required").

Some Assembly Required

In the course of the interviews, all teachers mentioned making changes to activities and lessons as opposed to using them as is. The most frequent refrain was, "After I download it, I still have to make it fit my needs."

Changes can be large or small. One teacher mentioned changing materials to reflect the language of the standards—e.g., referring to "writing opinion and argument" rather than "writing persuasive essays." Another said she advises colleagues to do the assignment first "to make sure it's not broken" or doesn't contain a content mistake that's embarrassing. Another described creating mini-lessons to teach underlying concepts. For example, an online worksheet might query students about the tone of a chapter, so she'll add a mini-lesson about tone for deeper instruction. Modifying lessons to increase rigor is common as well. One teacher asks students to provide textual annotation to reading passages to make the lesson more challenging. At the other end of the spectrum, teachers report frequently fixing typos.

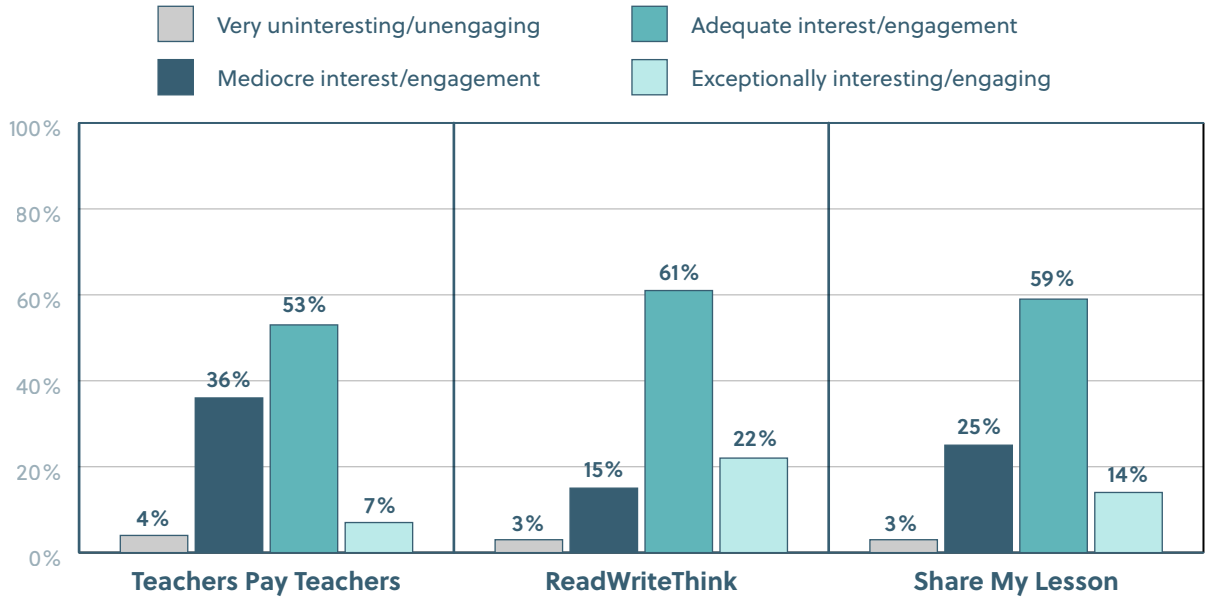
- *"I can easily manipulate the materials to make things at a 'reteach level' versus an 'I've got it level' versus an 'enrichment level.'"*
- *"Now that I've used these sites, it's almost like this is a great starting-off point, but I still need to quality check. Even though I feel like, 'I've given you my money; I should be able to trust that the quality has been taken care of.' I don't just mean things like facts and dates, but sometimes there needs to be a common-sense check. This lesson is presumably for high school, but it looks like something for fourth graders. I just don't use those materials."*
- *"I can find a cool idea online that sparks an idea for myself. For example, I found an escape-room activity that I could use at the end of the year. The content of the activity itself wasn't all that good, but I could use the framework and supplement it to make it better—more rigorous and meaningful . . . .I rarely use [materials] as is."*
- *"On occasion on ReadWriteThink, I like the format or the ideas behind a lesson, so I'll print it out and adapt it, usually to make it more challenging."*

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

Student engagement is, of course, a subjective concept, but the teachers that we interviewed reported seeking supplemental materials in part for that very reason. Thus, we asked reviewers (current and former educators) whether they thought that students would likely care about and be interested in the materials presented to them.

In general, the engagement ratings are not as high as one might hope. Across all the materials, on a 0–3 scale, ranging from *very uninteresting* to *exceptionally interesting*,[59] materials average 1.81 for engagement. Across websites, most are rated as *adequately* interesting (53–61 percent), although a moderate amount (18-40 percent) are rated as *very uninteresting* or of *mediocre* interest (Figure 10). ReadWriteThink materials are deemed most interesting (mean = 2.02), with TPT the least (mean = 1.63) and SML in the middle (mean = 1.83).

Figure 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 as shown. Numbers may not sum to 100 percent due to rounding.

One reason that teachers might visit these websites is to augment their core curriculum with materials that represent more diverse authors or texts about culturally diverse topics. Reviewers 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. Although

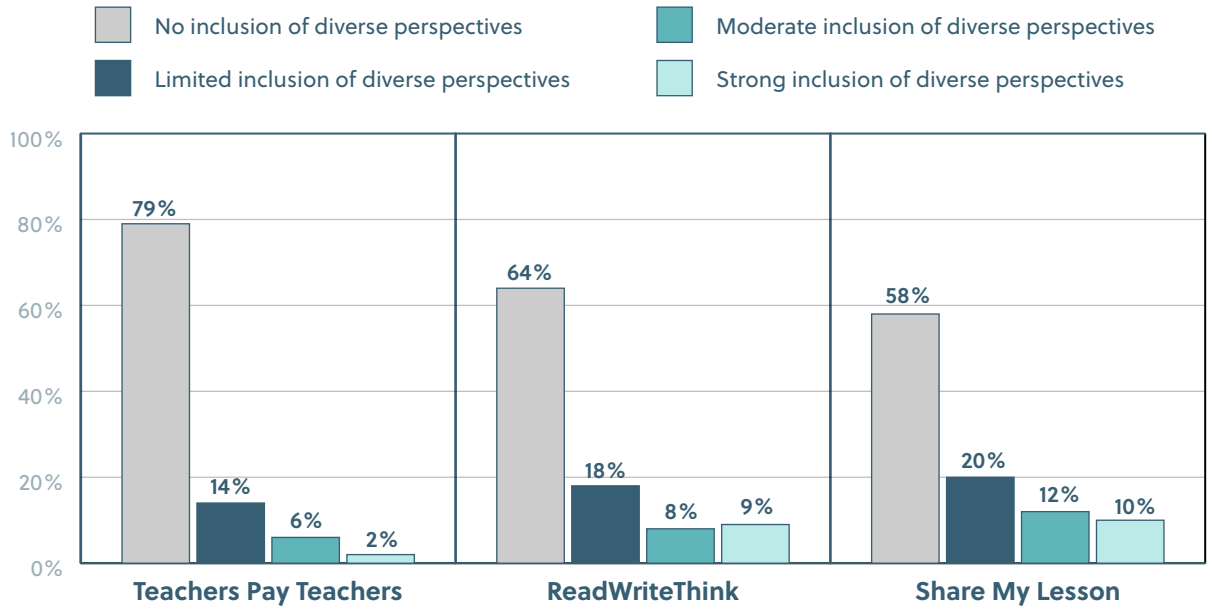
59. Full scale is as follows: 0 = very uninteresting/unengaging—highly boring, very likely to be of limited interest to most students; 1 = mediocre interest/engagement—somewhat boring, may be of interest to some students but likely not most; 2 = adequate interest/engagement—not boring, likely to be of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

some may argue that only full-blown curriculum can be expected to reflect such diversity, our reviewers felt strongly that even individual supplemental materials could and should be expected to meet this standard given the racial and ethnic diversity in American classrooms—if not all, at least some reasonable proportion.[60]

On a scale of 0–3, over two-thirds of materials (67 percent) score 0, meaning they do not include diverse authors or cover culturally diverse topics. Just 16 percent of materials score a 2 or a 3, respectively, 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 a 0.53, but RWT (mean = 0.62) and SML (mean = 0.75) score much higher than TPT (mean = 0.30). Figure 11 shows the distribution of ratings on each website, with a majority on all sites scoring 0. The materials that score highly, for example, include units on literary analysis using books like *Things Fall Apar*t by Chinua Achebe and *Roll of Thunder, Hear My Cry* by Mildred Taylor; explore the topic of gentrification using nonfiction texts; and cover argumentation using Martin Luther King's "I Have a Dream" speech.

Figure 11. A majority of materials on all 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.

60. One external adviser also cautioned: "While the effort to identify diverse materials is laudable, one thing to consider is the integration of diverse cultures and topics into the literary canon. I think we perpetuate stereotypes by looking for isolated units of study primarily focused on people of color, rather than the marginalized perspectives on well-known topics. Moreover, the use of culturally relevant teaching and diversity of authorship are two recent phenomena that do not take into account the role of equity within grade-level work."

VI. Discussion

Online, teacher-sourced, supplemental materials are incredibly popular, but popularity does not necessarily equate with quality. In fact, until now there has been very little evidence about what content these materials cover, what they look like, and how good they are—including how likely they are to contribute to a coherent, high-quality, standards-aligned curriculum. Our reviews produced disappointing results.

Yet they also offer a couple bright spots. The materials use texts that our reviewers rate as mostly high quality, visually appealing, and free from significant errors. However, they are not well aligned with the standards to which they claim alignment, and the quality of the writing and speaking and listening tasks is quite modest. Because writing is a dominant focus in the most frequently downloaded materials, that means the most sought-after materials are also among the weakest.

Moreover, the assessments in the materials are generally of poor quality, rarely covering all of the key content and often lacking helpful guidance for teachers, such as scoring rubrics. Materials fall especially short in providing supports for diverse learners and in representing cultural diversity. There is room also to improve deliberate building of students' content knowledge, as we know that students are often greatly in need of this to strengthen literacy—and what they too often get is fragmentation. On the whole, reviewers find the majority of the materials not or probably not worth using.

In comparing our three supplemental websites, we see a few interesting patterns. First, SML materials are most likely to cover a wider array of content and to emphasize higher levels of cognitive demand than RWT and (especially) TPT (see Table 2).[61] For instance, SML is the highest or tied for the highest in the percentage of the material at DOK levels 3 and 4 and in inclusion of reading-comprehension, writing, and speaking and listening tasks. In short, SML materials appear the most comprehensive in coverage of the three sites.

As for task quality and coverage of key instructional shifts of new standards, the patterns are not as clear (see Table 3). ReadWriteThink has the highest or tied-for-the-highest scores on text quality, speaking and listening task quality, and overall quality (with TPT rating the lowest). But RWT rates the weakest or tied-for-the-weakest on four of the five key instructional shifts we measured (close reading and analysis; focusing on central ideas and important particulars; using evidence from the text; and writing to a text).

61. One reviewer who evaluated SML materials noted that some of them were directly or indirectly gleaned from lessons also used as part of the EngageNY curriculum, which has been favorably reviewed by EdReports. See the following: EdReports, "Engage NY (2016)," accessed November 8, 2019, https://www.edreports.org/reports/overview/engage-ny-2016.

Table 2. Share My Lesson materials tend to be the most aligned and have the most coverage of multiple types of content.

	Overall	Teachers Pay Teachers	ReadWriteThink	Share My Lesson
Rating of alignment	1.35	1.28	1.28	1.56
Percent of materials with > ¼ DOK level 1	37%	48%	30%	35%
Percent of materials with > ¼ DOK level 2	71%	73%	73%	65%
Percent of materials with > ¼ DOK level 3	28%	14%	36%	37%
Percent of materials with > ¼ DOK level 4	3%	0%	6%	4%
Is there a main text (Y)?	68%	65%	58%	81%
Is there a reading-comprehension task (Y)?	73%	78%	60%	80%
Is there a writing task (Y)?	82%	78%	80%	88%
Is there a speaking/listening task (Y)?	43%	34%	49%	47%
Is this a unit (Y)?	75%	72%	83%	69%

Table 3. ReadWriteThink materials rate the highest overall and in terms of text quality but the lowest in terms of key instructional shifts.

	Overall	Teachers Pay Teachers	ReadWriteThink	Share My Lesson
Text Quality	2.21	1.96	2.34	2.36
Close reading and analysis	1.58	1.44	1.52	1.80
Focus on central ideas / important particulars	1.73	1.78	1.49	1.85
Evidence from the text	1.77	1.81	1.49	1.96
Require writing to a text	1.40	1.60	0.98	1.55
Writing task quality	1.42	1.41	1.40	1.44
Require speaking/ listening to a text	1.69	1.69	1.69	1.69
Speaking/listening task quality	1.48	1.42	1.61	1.40
Overall rating	1.28	1.18	1.41	1.29

Note: Comparisons are among the three websites, not of each site to the mean. Light blue values are significantly higher than dark navy values (none shown), which are significantly higher than orange values. Values of the same color are not significantly different from one another.

Finally, in terms of usability, assessment quality, knowledge building, and cultural responsiveness, RWT is the clear winner (see Table 4). It scores significantly higher than one or both of the other sites on nine of the ten ratings under this area. ReadWriteThink materials are rated as the most interesting/engaging, freest from errors, clearest in visual design and teacher guidance, and being the highest quality of assessments (all three dimensions). They also rate higher than TPT (and tied with SML) in terms of cultural responsiveness and knowledge building. Teachers Pay Teachers and Share My Lesson score lower on average across these dimensions, though each site has a bright spot or two. Given these high-level results, we turn next to the implications of this work.

Table 4. ReadWriteThink materials typically fare the best in terms of usability, assessment, knowledge, building, and cultural responsiveness.

	Overall	Teachers Pay Teachers	ReadWriteThink	Share My Lesson
Interesting and engaging	1.81	1.63	2.02	1.83
Free from errors	2.75	2.79	2.92	2.53
Visual design	2.04	2.04	2.19	1.89
Clarity of guidance	1.61	1.50	2.00	1.37
Supports for diverse learners	0.19	0.10	0.15	0.34
Assessment coverage of core lesson content	1.84	1.80	2.07	1.67
Scoring rubrics available/high-quality	0.94	1.21	1.14	0.44
Assessment quality	1.28	1.15	1.50	1.21
Knowledge building	1.28	0.91	1.50	1.54
Cultural responsiveness	0.53	0.30	0.62	0.75

Note: Comparisons are among the three websites, not of each site to the mean. Light blue values are significantly higher than dark navy values, which are significantly higher than orange values. Values of the same color are not significantly different from one another.

VII. Policy Implications

Based on our findings, we offer five takeaways with implications for policy and practice.

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.

Many teachers find that their core curriculum falls short in various ways (that is, if they even have a core curriculum), and they routinely turn to supplemental materials. Although these websites could provide a valuable service to teachers, overall ratings show they fall far short, with reviewers ultimately deeming most materials unsuitable for classroom use. Even on the highest scoring of the three websites—ReadWriteThink—fewer than half of the materials score 2 or greater on overall quality. Ratings are especially low for writing and speaking and listening tasks. Furthermore, the materials are not well aligned to the standards to which they claim alignment, or they claim alignment to such a large number of standards that the alignment has little meaning. More than half of all units do a poor job of building content knowledge. To state the obvious, if there is a need for supplemental materials—and apparently a large majority of teachers think there is—then there is also a need to provide consistently high-quality supplemental materials that are aligned to standards and merit use[62] (ironically enough, interviewed teachers don't see it as their job to improve these materials; see "Communicating with Teacher Developers" for more).

Communicating with Teacher Developers

Despite using online materials regularly, few of our interviewees contacted developers or wrote comments on the websites in an effort to improve lesson quality. One teacher said that she uses Instagram to ask developers whether their materials include a particular focus or element and that they almost always promptly respond. Another teacher contacted a developer about typos. But others were reluctant to make contact, mostly because they feared sounding critical of their colleagues.

- *"I wrote her to say that I loved, really loved, everything she produced, but there were too many typos, and I wasn't going to buy anything else until corrections were made. I told her that my students didn't turn in things that were so sloppy. I even offered, 'I understand you're busy, so if you need a copyeditor, I'll be glad to do it.' I think I insulted her. I never heard back."*
- *"They went way out on a limb to put it out there, and I don't want to seem like I'm criticizing. People don't have a lot of self-confidence in our profession. We get bashed constantly, by administrators, the press, parents, etc., so I can't see making it any worse for anybody."*

62. It is certainly possible that if all teachers were provided with high-quality core curricula, they would feel less need to supplement. However, the prevalence of supplementation suggests that teachers are likely to supplement at least some extent, no matter what materials have been provided.

2. The market for supplemental materials is bewildering and begs curation.

It is incredibly difficult to navigate the plethora of supplemental materials, astutely evaluate what is out there, and ultimately make informed decisions about what to use. There is clearly a role to be played for individuals or organizations to sift through what's on these sites and separate the wheat from the chaff. This service could be provided by the sites themselves, by school district curriculum experts, or by some newly created organization or entity. Such a service could serve two functions—helping to reduce the burden of searching and evaluating available materials and improving the quality of materials ultimately selected for use.

3. More supplemental materials need to provide teachers with soup-to-nuts supports, including stronger assessments and supports for diverse learners.

Our reviewers examined hundreds of materials and found very few that offered substantial supports for differentiation, particularly pertaining to high and low achievers, students with disabilities, and English-language learners. Even modest enrichment supports or scaffolding would be a useful addition to most of the materials we evaluated. The assessments were also lacking essential content and guidance about what constitutes progress and/or mastery. If a material is going to serve as a stand-alone unit that teachers can take off the metaphorical shelf, it must provide teachers with greater instructional support. Otherwise, they will be using the inferior supports that come with the units or spending their time creating new supports from scratch. Again, this feels like a role for the sites themselves—more careful curation of what's posted to ensure that it supports good teaching.

4. We need better sourcing of supplemental materials that focus on diverse authors and cultural pluralism.

The large majority of reviewed materials made limited to no effort to represent the cultural diversity of America's students. When fewer than 45 percent of K–12 students are white,[63] more curriculum materials—core and supplemental alike—should include materials that are written by nonwhite authors and represent the diverse cultures inhabiting our classrooms or address culturally diverse topics. Although supplemental materials could fill this role nicely, this is not the case. Simply put, there is a need for better sourcing of supplemental materials that focus on diverse authors and cultural pluralism. Websites like TPT, RWT, and SML could encourage or develop more resources along these lines and label them as such so that they are easier to find—or another entity entirely may fill this need.

63. National Center for Education Statistics, "Indicator 6: Elementary and Secondary Enrollment," last updated February 2019, accessed November 8, 2019, https://nces.ed.gov/programs/raceindicators/indicator_rbb.asp.

5. School and district leaders need to decide whether and how to monitor the enacted curriculum.

No doubt there are lots of mediocre materials being used in U.S. classrooms, given the poor quality overall of what we evaluated and the fact that it comprises the most popular materials from some of the most popular websites. School and district leaders should think seriously about how they want to handle this issue. They could certainly take a hands-off approach, like nearly all of them now do, and allow teachers to arbitrate on their own the good from the bad. Yet providing a quality curriculum is a primary function of a school. Moreover, the popularity of sites like these and the frequency that they're accessed indicates that these supplementary materials may in fact soon become part of the core curriculum—whether school leaders and department heads realize it or not. The latter are advised then, at the very least, to pay more attention to what's actually taught in classrooms when it comes to supplementing. What they learn as a result could inform an array of subsequent approaches, from offering teachers training in how to evaluate and select high-quality materials to publishing a list of curated supplemental resources to explicitly discouraging the use of unacceptable materials.

*****

The supplementation train has left the station—augmenting core curricula with materials found on the web has become an established part of the way that American teachers teach. In an era of new and more ambitious standards and tests, our findings raise the question of whether existing popular supplementary materials are up to the task. Especially in terms of low-hanging fruit like providing materials aligned to standards and that are relevant to our diverse student bodies, these websites can do much better.

Clearly teachers are on the hunt for particular materials to supplement their core curriculum. Unfortunately, some of the websites they are using are providing numerous subpar offerings. Given the prevalence of supplementation and the likelihood that standards-based reform will continue to guide state and national policy in the coming years, we hope that content creators at these and other websites will see these findings as a clarion call to improve their quality for the betterment of teaching and learning in America's classrooms.

Appendix A: Final Evaluation Rubric

Note: Metadata populated by study team appears in **bold**. Gateway items are *italicized*, meaning the remaining questions in the section do not apply if the response is "no."

1) Descriptive Data

a. **File name**
b. **Title of the material**
c. **Number of materials**
d. **Expected number of days**
e. **Category**
    i. **Reading**
    ii. **Writing**
    iii. **Speaking and listening**
f. **What topics is it focused on?**
g. Lesson/unit purpose (open-ended and at most three sentences)
h. **What texts does it use (list of titles)?**
i. **Author of the material on the Excel sheet**
j. **What is the main text?**
k. **Rank on the website**
l. **Metadata (where available)**
    i. **Number of downloads**
    ii. **Number of comments**
    iii. **Average rating**

2) Alignment to Standards

a. Does it include standards it aligns to? Yes/no
b. To what standards does the main activity say it aligns?
c. Rating of topic alignment: 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.

3) Depth of Knowledge

(4-level Webb DOK taxonomy)

a. DOK level of the main activity: break down the main activity in its coverage of each DOK level. In other words, what percent of the main activity would you say is at each of the following DOK levels, 1–4: 0 percent, 1–25%, 26–50%, 51–75%, or 76–100%?

Lesson/Unit Text/Stimuli

4) Text Complexity and Quality

a. Is there a main text? Yes/no

b. **What is the main text?**

c. **Quantitative measure—Lexile**

d. Overall measure of text quality: 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.

Lesson/Unit Tasks

5) Close Reading and Evidence from the Text

a. *Is there a reading-comprehension task? Yes/no*

b. Does the task require close reading and analysis? 0 = not at all; 1 = yes, to a small extent; 2 = yes, to a moderate extent; and 3 = yes, to a major extent.

c. Does the task focus on central ideas and/or important details in the text? 0 = not at all; 1 = yes, to a small extent; 2 = yes, to a moderate extent; and 3 = yes, to a major extent.

d. Does the task require students to use evidence from the text? 0 = not at all; 1 = yes, to a small extent; 2 = yes, to a moderate extent; and 3 = yes, to a major extent.

6) Writing Task Quality

(applicable for lessons that require students to write a paragraph or more)

a. *Is there a writing task? Yes/no*

b. Does the lesson require writing to a text? 0 = not at all; 1 = yes, to a small extent; 2 = yes, to a moderate extent; and 3 = yes, to a major extent.

c. Overall measure of writing task quality: 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).

7) Speaking and Listening Task Quality

a. *Is there a speaking/listening task? Yes/no*

b. Does the lesson require speaking about/listening to a text/recording? 0 = not at all; 1 = yes, to a small extent; 2 = yes, to a moderate extent; and 3 = yes, to a major extent.

c. Overall measure of speaking/listening task quality: 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).

8) Usability

a. Interesting and engaging: 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.

b. Free from errors: 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.

c. Visual design and organization: 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.

d. Clarity of guidance for teachers: 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.

e. Supports for diverse learners (high or low performers, students with disabilities, or English-language learners). How comprehensive are the supports for differentiation? 0 = none offered; 1 = limited supports offered; 2 = some supports offered, not comprehensive; and 3 = extensive supports offered, comprehensive of most or all likely student groups.

9) Assessment Quality

a. ***Includes some form of assessment to help teachers gauge whether students have mastered the content of the lesson? Yes/no***

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

c. Scoring rubrics available/good: 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.

d. Quality of assessment: 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.

10) Knowledge Building and Cultural Responsiveness

a. *Is this a unit (versus single lesson)? Yes/no*

b. 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.

c. Includes diverse perspectives (racial, gender, culture, national origin): 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.

11) Overall Rating

a. 0 = very poor, teachers should not use this material; 1 = mediocre, has some good and some bad components (e.g., well organized but not on important content and 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.

Appendix B: Contributor Bios

Morgan Polikoff (lead author) is an associate professor of education at the USC Rossier School of Education. He has spent the last decade studying the design, implementation, and effects of standards, curriculum, assessment, and accountability policies; for this work, he was awarded the American Educational Research Association Early Career Award in 2017. He has published over forty articles in peer-reviewed journals and been PI or co-PI on more than $13 million in federal- and foundation-supported research grants. He codirects the USC Center on Education Policy, Equity and Governance and coedits the journal *Educational Evaluation and Policy Analysis*. He received his bachelor's degree in mathematics from the University of Illinois at Urbana Champaign in 2006 and his PhD in education policy from the University of Pennsylvania in 2010.

Jennifer Dean (lead reviewer and co-author) is a freelance consultant in educational assessment with more than thirty years of experience in teaching and educational publishing. She began her career teaching secondary ELA and adult basic education, followed by positions as an ELA content specialist and a project director in the K–12 assessment field. She served as executive director for a K–12 assessment program in a large assessment company, overseeing the development of state and district assessments in all content areas. Before consulting, she worked for Student Achievement Partners, developing standards alignment guidelines, sample assessments, and evaluation tools. In the last several years, she has reviewed numerous text passages, items, and other materials for their overall quality and alignment to the Common Core State Standards.

Jenni Aberli (reviewer) is the high school ELA instructional lead for Jefferson County Public Schools in Louisville, Kentucky. In this role, she develops curriculum frameworks and resources, professional learning opportunities, and in-school supports for teachers that are standards aligned. Prior to working as an instructional lead, Jenni served as an instructional coach and high school English teacher. She is an EdReports Klawe Fellow and an ELA reviewer/writer, a Kentucky Core advocate with Achieve the Core, a Louisville Writing Project Fellow, a Literacy Design Collaborative (LDC) Lead and Learn fellow, and a twenty-year National Board certified teacher. She received a Learning Forward Best Evaluation of Professional Development Award. Jenni holds a bachelor's degree in English, MAT in English education, EdS in administration and supervision, and reading program consultant and administrative certifications from the University of Louisville.

Sarah Baughman (reviewer) is an educational consultant with over a decade of middle and high school English teaching experience in public, charter, online, and international schools. She has worked with Advanced Placement and International Baccalaureate curricula and developed numerous elective and core classes. Sarah is a core advocate who evaluates curriculum for Student Achievement Partners and a humanities content lead with Great Minds.

Dr. Bryan R. Drost (reviewer) is a central office administrator for a school district in northeast Ohio, where he works with curriculum and technology integration in all content areas. He is currently chairperson of the National Council of Measurement Standards for Test Use Committee, an Ohio Department of Education regional data lead, a member of ODE's Fairness and Test Use Committee and Content Advisory Committees, a national supervisor for edTPA, and the math lead for Ohio's core advocates. He has presented throughout the state and country on various topics related to instructional shifts, assessment, world pedagogy, and technology integration.

Joey Hawkins (reviewer) is a national literacy consultant. As a middle-level teacher in a rural Vermont public school, she developed and taught integrated content/literacy curriculum for many years. In Vermont, she was a leader in the use of writing portfolios in assessing student writing and served as an advisor in the development of the New Standards Reference Exam. Joey is a founder of the Vermont Writing Collaborative and the lead author of *Writing for Understanding*, a content- and standards-based approach to writing instruction that is foundational to several national literacy curricula in widespread use. She works with both local school districts and various state- and national-level organizations to support teachers in developing content- and standards-based literacy instruction at all grade levels. Joey is a graduate of Mount Holyoke College with a master's degree from Dartmouth.