IES Institute of Education Sciences

The Effect of School Report Card Design on Usability, Understanding, and Satisfaction

Regional Educational Laboratory Mid-Atlantic

At Mathematica

REL 2021–101
U.S. DEPARTMENT OF EDUCATION

A Publication of the National Center for Education Evaluation and Regional Assistance at IES

The Effect of School Report Card Design on Usability, Understanding, and Satisfaction

Jesse Chandler, Jacob Hartog, Erin Lipman, and Jonathan Gellar

July 2021

Education policymakers view transparency and accountability as critical to the success of schools. To support these goals, the District of Columbia Office of the State Superintendent of Education (OSSE) has developed an online school report card for communicating information about the characteristics and performance of schools. To support OSSE's interest in making report cards more usable, this study assessed the effect of different designs on how easy the report cards are to use and understand, how easy it is to find information in them, and whether users would recommend the site to others.

The study found that moving the link to details of the district's School Transparency and Reporting (STAR) framework from the top of the page to beneath the STAR score improved the site's usability and that reporting the number of points possible for each metric led to a better understanding of how the score is calculated. The combination of design features that produced the best performance on all measures included these two design changes. Other designs had mixed effects. In particular, making year-over-year change in school performance salient made it easier to identify which schools had improved the most, but participants disliked this feature (demonstrated by lower ratings for usability and satisfaction). In general, participants who accessed the site with mobile devices had more difficulty using it. This study illustrates how policymakers and practitioners in other states can efficiently test school report card design changes at scale.

Why this study?

Under the Every Student Succeeds Act (ESSA), every state education agency must shape its own accountability system in consultation with the U.S. Department of Education. ESSA requires states to establish goals for student performance (including a measure of student performance that is broader than test scores) and hold schools accountable for student achievement. As part of this system, every state must issue annual report cards for every school. Each state is free to develop its own accountability system, provided it meets ESSA standards. The federal government provides detailed requirements about the minimum acceptable contents of a school report card. These include the following:

- A summative determination of school quality. In the District of Columbia this is called the School Transparency and Reporting (STAR) framework.
- Individual performance indicators that feed into the summative determination, such as test results, English learner student proficiency rates, one other academic indicator, and one nonacademic indicator.
- The subgroups for which these indicators must be disaggregated and reported, such as key racial/ethnic groups.
- School-level inputs, such as teacher qualifications and per-student spending, and certain details about the methods used to collect data, such as student participation rates in assessments and the number of students who completed alternative assessments.

States have considerable freedom to collect and report information beyond these minimum requirements. The current District of Columbia school report card has more than 150 data elements, including those required by ESSA and others added in response to community feedback.

Often overlooked in the debate over what to report about school performance is how best to communicate the information to stakeholders. ESSA requires states to present information in a concise and user-friendly way and to consult parents

For more information, including background on the study, technical methods, supporting analyses, and sensitivity analyses, access the report appendixes at https://go.usa.gov/x6tCt.

and other stakeholders while designing report cards. Yet ESSA does not specify how to determine whether report cards are user friendly, and states have wide latitude to decide the final form of their school report cards. In response to extensive consultation with community stakeholders, the District of Columbia Office of the State Superintendent of Education (OSSE) has built a website that provides a comprehensive portrait of all public schools in the district, both charter and district-run.

OSSE partnered with the Regional Educational Laboratory Mid-Atlantic to test the effects of different design choices on users' perceptions of report card usability, understanding of report card content, and satisfaction with the report card. School report cards are the sum of many design choices, each of which influences the user experience. As designers of commercial websites have observed, the aggregate effect of many small design changes can be substantial even when an individual design choice has a small effect (Thomke, 2020). The current study examined design choices individually, and the combination of choices with the most positive impact is identified for each outcome measure.

OSSE can use the results of this study to inform decisions on how to depict information in its school report cards, with the goal of ensuring that school report cards provide clear and accessible information so that parents and other stakeholders can understand the performance of their schools. In addition, the study's findings can inform state education agencies across the country as they design or refine their own school report cards or decide how to approach evaluating potential school report card designs.

The challenges of designing information displays

School report cards are the product of many intentional and unintentional design choices, all of which matter for the user experience (for an overview of the relevant literature on the effects of design on the user experience and understanding and for additional background information about this study, see appendix A). Choices must be made about how to structure the report card, including whether to use one or more pages to show data and where to place data elements on each page. Choices must also be made about whether to display each data element using a table, a graph (and if so, what type of graph), icons, or something else. Design decisions often beget more decisions. For example, if information is to be depicted in bar graphs, the bars could be stacked or positioned side by side, use complementary or contrasting colors, have labels that report natural frequencies or proportions, and so forth. As a further complication, these decisions must consider that users access electronic report cards on devices with different screen sizes and interfaces, such as computers, tablets, and smartphones.

A good design must satisfy the needs of stakeholders with different levels of ability (literacy, numeracy, and digital literacy), subject matter expertise, and reasons for use. Although school report cards can be intended to reduce inequality by giving under-resourced communities valuable information, researchers have expressed concern that parents with higher education levels are better able to take advantage of school report cards (Figlio & Lucas, 2004; Hasan & Kumar, 2019).

Designers must pay attention to the different ways that users respond to a website. A central concern is whether users find a website usable (Hornbæk, 2006)—that is, whether they subjectively find it to be pleasant or unpleasant to use. A closely related concept is whether users like a website. Users must also correctly interpret the information that a website presents (an outcome that is only weakly related to subjective experience; Hornbæk & Law, 2007). Other outcomes might also matter depending on the specific goals for a report card site, such as encouraging users to contact schools to arrange tours or obtain more information. Sometimes, a design can have both positive and negative outcomes, forcing designers to make tradeoffs when deciding whether it is "better" than an alternative design.

One recent study began to address the effect of design on school report cards by examining how five design decisions might affect choice of school, user satisfaction, and user understanding (Glazerman et al., 2020).[1] Users were asked to imagine that they were moving to a new town and wanted to select the best school for their child using a set of school report cards presented in the study. Users understood school information better when it was presented with only numbers than when numbers were supplemented with graphs or icons, but they were less satisfied with numbers alone. Users' satisfaction increased when more information was provided and when schools were sorted by distance from their home. An important finding was that, although most design effects were small, the effects were cumulative, allowing the combination of several small changes to produce larger effects. The current study expands on this research by evaluating untested design factors that OSSE considered including in the 2019/20 update of its school report card. (For a summary of updates made to the report card, see District of Columbia Office of the State Superintendent of Education, 2018.)

Tested design factors and outcome measures

The current study extends existing research by testing novel design modifications that might influence the usability and understandability of displays of school information. It examined the effects of five potential modifications prioritized by OSSE. The goal was to determine which designs participants found easier to use (measured through an assessment of overall usability and ratings of the ease of finding specific information presented in the report card), which designs made the information easier to understand (measured through the number of questions about the report card contents that participants answered correctly), and which designs participants liked (measured indirectly through willingness to recommend the site to others; Reichfeld, 2003). Descriptions of the rationale and potential effect of the modifications are in table 1.

Table 1. Tested design factors

Factor	Business as usual	Alternative design	Rationale
Report card organization (figure 1)	STAR rating appears on the main page, and a link to the explanation of the STAR rating appears on the top ribbon.	STAR rating appears on the main page, and a link to the explanation of the STAR rating system appears under the rating.	Linking from the STAR rating to the STAR framework might make it easier to find information about the purpose of the STAR rating and how it was calculated.
Details about the calculation of the STAR rating (figure 2)	The raw scores for each metric are displayed along with the floor and targeta for that metric.	The number of STAR rating points earned and the points possible for that metric are displayed.	State education agencies have prioritized both the possible number of points and floors and targets. Both are important determinants of how individual metrics contribute to the overall STAR rating.
Proficiency score chart format (figure 3)	Proficiency scores are displayed as a bar, with a line indicating the district average for that metric.	Proficiency scores and the district average are displayed as clustered bars for each metric.	It might be easier to use stacked bars to compare proficiency scores with the district average.
Change over time (figure 4)	Raw scores for the current and previous years are available by clicking a metric.	The difference in scores between the current and previous years for each metric are displayed.	Parents might want to see change over time (Mikulecky & Christie, 2014). Identifying schools that have improved might be easier when users do not have to calculate differences between scores (Vessey, 1991).
School offerings (figure 5)	Only the amenities a school offers are listed.	All potential school offerings are listed, with a checkmark next to the amenities offered by the school and an X next to those not offered.	State education agencies have used both approaches. A list of only the amenities a school offers might be easier to understand (Schkade & Kleinmuntz, 1994) but could make it harder to compare schools (Gentner & Markman, 1997; Zhang & Markman, 2001).

STAR is School Transparency and Reporting.

a. The floor is the score below which schools receive no points regardless of score; the target is the score above which schools earn full points.

Source: Authors' compilation.

1. The user satisfaction and understanding measures were similar to those used in the current study.

Each participant was shown a site containing the report cards of seven different schools, and each set of seven report cards was formatted in 1 of 32 randomly assigned combinations of design choices. These formats (referred to as treatments) represented all possible combinations of the five design choices being tested. Participants were asked to familiarize themselves with the report cards before completing the outcome measures by reviewing the schools and picking the ones with the strongest academic performance and the most positive school environment. More details about the study design are in box 1.

Research questions

The study team examined the effect of these design decisions on the following research questions:

1. Which design choices influence how users engage with report cards?

   1a. Which design choices influence the usability of the report cards, as measured by assessments of overall usability and the ease of finding the information affected by the design choice?

   1b. Which design choices lead to differences in understanding of the information in the report cards?

   1c. Which design choices influence participants' willingness to recommend the site to others?

2. Do design choices have different effects on different subgroups of users as defined by demographic characteristics, type of device used, prior experience using school report card sites, and method of recruitment into the study?

3. Do different subgroups of users have different average ratings of usability, understanding, ease of finding specific information, and willingness to recommend the site to others?

Box 1. Data sources, sample, and methods

Data sources. This study analyzed responses to a randomized factorial survey experiment conducted online by the District of Columbia Office of the State Superintendent of Education (OSSE). All participants provided biographical and demographic information. The study used school report card data of real District of Columbia high schools, as displayed on the OSSE website, and de-identified them by using new names, pictures, and geographic locations.

Sample. Participants were a convenience sample of 824 U.S. residents older than 13 recruited from three sources: OSSE community outreach (6 percent of the analytic sample), a market research panel consisting mostly of District of Columbia residents (58 percent of the sample), and Amazon Mechanical Turk (an online labor market where people complete tasks such as surveys in exchange for pay; 36 percent of the sample, all of them U.S. residents). Recruitment is described in appendix B. Across the analytic sample 55 percent of participants were district residents, 57 percent were parents, and 7 percent self-identified as educators (see table C1 in appendix C). The sample was not statistically representative of D.C. residents or report card users because it was self-selecting, but it was diverse in terms of respondents' racial/ethnic identity, income, and education level.

Methods. The study was a randomized factorial experiment—one that examines several factors (see table 1 in the main text). Each tested factor consisted of two designs: a business-as-usual design that displayed information the way OSSE displayed it during the 2018/19 academic year and an alternative design that displayed the same information in a different way that OSSE was considering implementing.

Because the experiment simultaneously tested five design factors, each participant saw 1 of 32 ($2 \times 2 \times 2 \times 2 \times 2$) different treatments. The study team then estimated the effect of each factor (and the interactions between factors) simultaneously both overall and within subgroups.[1] Because random assignment ensured that there were no systematic differences in the people

assigned to respond to different treatments, any differences in outcome measures between the treatments are a result of the design factors tested.

Outcome measures. Participants' responses to 26 items and questions about the report cards were used to calculate four outcome measures (see appendix B for the methods used):

- **Usability.** A measure of overall usability based on responses to 13 self-report items that focused on how easy the school report cards were to use ("The school report card site was easy to use") and aesthetics ("I found the website to be attractive"). For each item participants used a six-point scale to indicate whether they disagreed strongly, disagreed, disagreed slightly, agreed slightly, agreed, or agreed strongly. The study team examined average effects on responses to these items.
- **Understanding.** A measure based on responses to six comprehension questions with factually correct answers that could be determined from the school report cards.[2] Each factor was tested using one or two questions about information that the business-as-usual design and the alternative designs displayed in different ways.
- **Ease of finding specific information.** A measure based on responses to six self-report items that focused on how easy or difficult participants felt that it was to find specific information. Each factor was tested using one or two items about the ease of finding information affected by that factor.
- **Willingness to recommend the site to others.** A measure based on responses to a single question: "On a scale from 1 (not likely at all) to 10 (extremely likely), how likely are you to recommend the school report card website to a friend who is interested in learning about public schools in DC?"

Analysis. The study team used hierarchical Bayesian analyses to analyze the data. Usability and ease of finding specific information were modeled as ordinal variables, but for simplicity differences in the proportion of participants who agreed at least slightly with these items are reported. Understanding was treated as a binary variable (correct or incorrect responses), and willingness to recommend the site to others was treated as a continuous variable (see appendix B for details).

Like traditional methods, Bayesian analysis estimates the effect of design choices and the standard deviations around them. Unlike traditional methods, which represent uncertainty solely though larger standard deviations, Bayesian models also shrink estimates toward each other. Estimates are shrunken closer together when there is more uncertainty (measures completed by fewer participants or measured with fewer items; see appendix B for details). Bayesian models also provide information to make probability statements about the size and direction of the effects (for example, "There is a 70 percent probability that the effect is greater than 0" or "greater than 5 percent"). These probabilities are called posterior probabilities. Differences greater than 1 percent are rounded to the nearest whole percent.

Limitations. Because the sample is not representative of report card users, the treatment effects might not generalize to the desired population of potential report card users, though randomized controlled trials performed on representative and nonrepresentative samples usually produce estimates of the same direction and approximately the same magnitude (Coppock, 2019; Coppock et al., 2018).

Note

1. The subgroup analysis compared District of Columbia residents with nonresidents, mobile device users with computer users, participants who had at least a four-year college degree with those who had some college, participants who had no college degree with those who had some college, participants who spoke a language other than English at home with those who spoke only English at home, participants recruited from Amazon Mechanical Turk with those recruited from the market research panel, participants recruited by OSSE with those recruited from the market research panel, participants who spent less time looking at the site with those who spent more time looking at the site, and participants who had used school report card sites before with those who had not. See appendix B for details.

2. Seven questions were designed to assess understanding, but participants were exposed to only six (see appendix B for details).

Findings

This section reports the effects of the five design changes. Effects were estimated using Bayesian modeling, which does not provide a cutoff that determines whether a finding is significant. Users of the model results must make this decision based on the level of risk they are willing to tolerate for a specific decision: a high-stakes decision such as changing a report card vendor might require a 95 percent or even 99 percent probability of improvement, and a trivial design change might require only a 51 percent probability.

To highlight the findings that are most likely to be actionable, only effects with a 70 percent probability of having any favorable effect or a 70 percent probability of having any unfavorable effect (that is, effects that are likely to be nonzero) are reported (see appendix B for a justification of using 70 percent as a cutoff). The report also presents the posterior probability that an outcome has at least a 5 percent change in odds for categorical measures (or a change of 0.1 standard deviation for willingness to recommend). Effects that exceed this threshold are referred to as substantial. This second threshold is used because some situations in which changing a report card incurs a cost (such as an unplanned change order) might require evidence of a larger effect. For all reported effects the report provides estimated differences in percentages or means after covariates in the model are adjusted for. Detailed results are in appendix C, and a sensitivity analysis is in appendix D.

Changing the placement of the School Transparency and Reporting rating link increased self-reported usability of the report cards

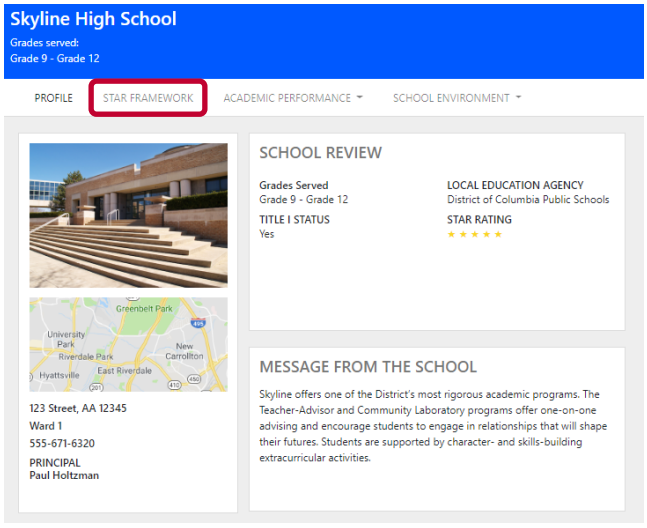
The business-as-usual OSSE report card had a tab in the top ribbon that displayed STAR ratings. In the alternative design users can access STAR ratings through a hyperlink under the STAR rating on the front page (figure 1).

The proportion of participants who thought that the report card was usable (that is, who rated it above the midpoint of the scale) was higher among those who saw report cards with the explanation link under the STAR rating (78.5 percent) than among those who saw report cards with the link in the top ribbon (77 percent). There is an 86 percent chance that placing the link under the STAR rating increases usability and a 73 percent chance that it increases the odds of rating the report card as more usable by more than 5 percent (see table C2 in appendix C).

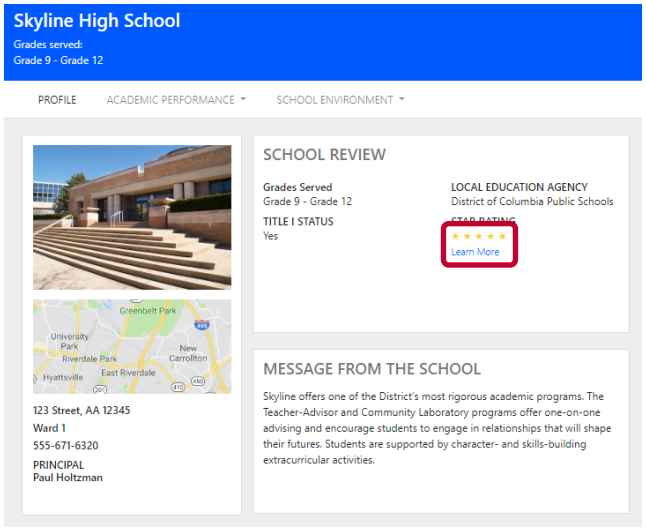
Placing the link under the STAR rating did not change participants' ability to correctly identify the school with the higher STAR rating (see table C5 in appendix C), reported ease of identifying that school (see table C7), or willingness to recommend the site to someone else (see table C9).

Figure 1. Comparison of the business-as-usual and alternative location of the STAR rating link

Business-as-usual design



Alternative design



STAR is School Transparency and Reporting.

Note: The business-as-usual design reflects the layout used by the District of Columbia Office of the State Superintendent of Education in the 2018/19 academic year. The red boxes have been added for emphasis.

Source: Business-as-usual and alternative designs created by Tembo, Inc.

Showing the points earned and points possible for the School Transparency and Reporting rating metrics increased users' understanding of the STAR rating and their willingness to recommend the site to others

The original OSSE report card reported STAR rating metrics along with the scoring range, which indicates the floor (below which schools receive no points regardless of score) and target (above which schools earn full points) for each metric. The alternative design showed how many points the school earned for each metric and how many points were possible for that metric (figure 2).

The proportion of participants who understood the influence of different metrics on the STAR rating (that is, answered the comprehension question correctly) was higher among participants who saw points possible for each metric (14.5 percent) than among those who saw information about floors and targets (13.5 percent). There is a 74 percent chance that displaying points possible increases understanding of the influence of different metrics on the STAR rating (see table C5 in appendix C). Notably, displaying points possible did not reduce understanding of the function of metric floors and targets.

Participants who saw report cards that displayed points possible were more willing to recommend the site to others (by 0.1 scale point; see table C9 in appendix C). There is a 79 percent chance that displaying points possible increases willingness to recommend the site to others.

Similar proportions of participants thought that both report card designs were usable (see table C2). However, among participants who had not attended college, the proportion who thought the report card was usable was higher among those who saw report cards with points possible (79 percent) than among those who saw report cards with floors and targets (78 percent). For this subgroup there is a 76 percent chance that displaying points possible makes the report card more usable. Among participants with some college or at least a bachelor's degree, the proportion of participants who thought the report card was usable was similar.

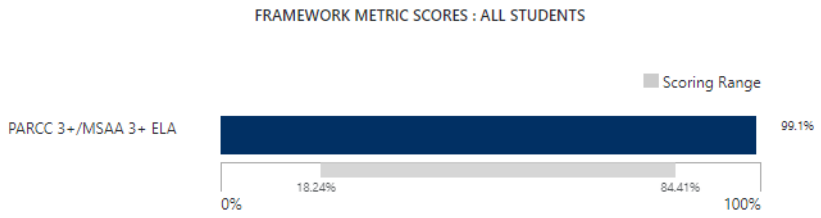
The self-reported ease of finding information about how the STAR rating is calculated was similar for both designs (see table C7 in appendix C).

Changing the display of school average proficiency reduced self-reported ease of use

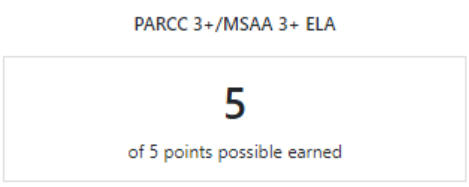
The original OSSE report card reported the school's percentage proficient on state tests as a bar and the District of Columbia average for percentage proficient as a line superimposed over the bar. In the alternative design the school score and district average appeared as clustered bars for each metric (figure 3).

Figure 2. Comparison of the business-as-usual and alternative depiction of STAR rating metrics

Business-as-usual design



Alternative design



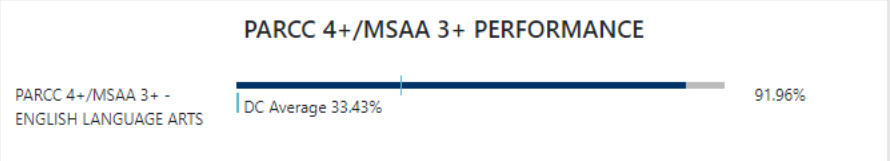
STAR is School Transparency and Reporting.

Note: The business-as-usual design reflects the layout used by the District of Columbia Office of the State Superintendent of Education in the 2018/19 academic year.

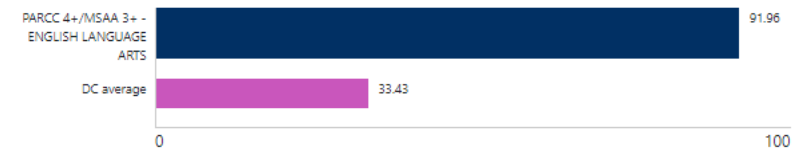
Source: Business-as-usual and alternative designs created by Tembo, Inc.

Figure 3. Comparison of the business-as-usual and alternative depiction of school average proficiency

Business-as-usual design



Alternative design



Note: The business-as-usual design reflects the layout used by the District of Columbia Office of the State Superintendent of Education in the 2018/19 academic year.

Source: Business-as-usual and alternative designs created by Tembo, Inc.

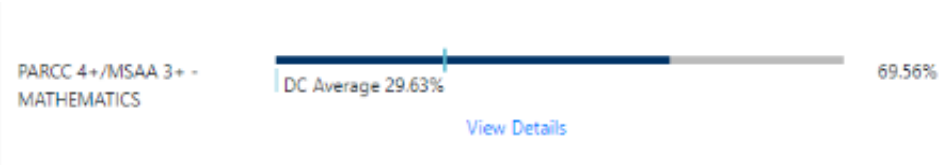
Contrary to expectations, the proportion of participants who thought that school proficiency information was easy to find was lower among those who saw report cards with clustered bars (77.7 percent) than among those who saw report cards with a single bar and a line (78 percent). There is a 72 percent chance that this design change is worse than the business-as-usual design (see table C7 in appendix C). Both designs led similar proportions of participants to think the school report card was usable (see table C2) and to understand the report card (see table C5), and participants were equally willing to recommend both designs to others (see table C9).

Showing change over time in school performance reduced usability and willingness to recommend the site to others

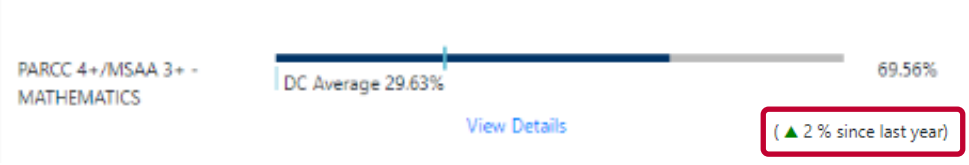
The original OSSE report card allowed users to click on a metric to view time trends for that metric on a line graph with the year on the X-axis. The alternative design maintained this functionality but also showed the year-over-year change in score under each metric (figure 4).

Figure 4. Comparison of the business-as-usual and alternative depiction of change over time

Business-as-usual design



Alternative design



Note: The business-as-usual design reflects the layout used by the District of Columbia Office of the State Superintendent of Education in the 2018/19 academic year. The red box has been added for emphasis.

Source: Business-as-usual and alternative designs created by Tembo, Inc.

The proportion of participants who thought the report card was usable was lower among those who saw report cards displaying change over time (76 percent) than among those who saw report cards without change over time (79.5 percent). There is a 99 percent chance that displaying the change over time decreases usability and a 96 percent chance that it decreases the odds of rating the report card as more usable by at least 5 percent (see table C2 in appendix C). Participants who saw report cards displaying change over time were also less willing to recommend the site to others (–0.19 scale point). There is a 90 percent chance that displaying the change over time reduces willingness to recommend the site to others (see table C9 in appendix C).

The proportion of participants who understood which school improved the most was higher among those who saw change over time (27 percent) than among those who did not (26 percent). There is a 79 percent chance that displaying the change over time increases the proportion of users who identify the most improved school (see table C5). The design choices led similar proportions of participants to think that it was easy to find information about change in performance over time (see table C7).

Showing a checklist of all possible school offerings reduced user willingness to recommend the site to others

Schools in the District of Columbia offer a wide range of programs, such as Advanced Placement, dual enrollment, and interscholastic sports. The original OSSE report card listed only the programs that a school offered. The alternative design listed all possible offerings and included a checkmark next to those that were offered at the school and an X for those that were not offered (figure 5).

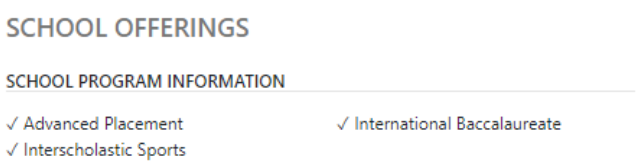
The proportion of participants who correctly identified a school with a specific offering was higher among those who saw report cards that listed all possible offerings (49 percent) than among those who saw report cards that listed only each school's individual offering (47 percent). There is a 74 percent chance that listing all possible offerings increases the proportion of users who correctly identify schools with specific offerings (see table C5 in appendix C). Participants rated this information similarly easy to find for both designs.

The proportion of participants who correctly identified the school with the most offerings was the same for participants who saw different designs, but the proportion of participants who thought that it was easy to find the school with the most offerings was lower among those who saw report cards with all possible offerings (78 percent) than among those who saw only each school's individual offerings (79 percent). There is a 75 percent

Figure 5. Comparison of the business-as-usual and alternative depiction of school offerings

Business-as-usual design

Alternative design



Note: The business-as-usual design reflects the layout used by the District of Columbia Office of the State Superintendent of Education in the 2018/19 academic year.

Source: Business-as-usual and alternative designs created by Tembo, Inc.

chance that listing all possible offerings reduces the proportion of users who say it is easy to find the school with the most extracurricular offerings (see table C7).

Participants who saw all offerings were less likely to recommend the site to others (–0.11 scale point; see table C9). There is a 79 percent chance that listing all offerings reduces willingness to recommend the site to others. Listing all possible school offerings did not influence usability (see table C2).

Some design changes had clearly positive effects on the user experience; others led to tradeoffs

The effects of the proposed design changes on measures are summarized in table 2. For some factors one alternative is clearly superior. Two proposed changes had at least one positive and no negative outcomes, and one change had a negative outcome and no positive outcomes. Linking a school's STAR rating directly to a description of how the score is calculated is likely to increase usability relative to placing the description in the top ribbon, and doing so is unlikely to affect other measures. Displaying STAR points earned is likely to increase understanding of how STAR scores work relative to displaying information about floors and targets, and doing so is also unlikely to affect other outcome measures. Displaying the district average for proficiency as its own bar is likely to decrease understanding of whether a school is performing above or below average, and doing so had no positive outcomes relative to displaying the average as a line overlaid on a bar depicting school performance.

Other design choices improved performance on some outcome measures and decreased performance on others. Emphasizing differences in performance between the current and prior year made it easier for participants to correctly identify how school performance changed over time. But participants also rated the report cards that use this design element as less usable and reported being less likely to recommend the site to others. Similarly, listing all possible school offerings increased the proportion of participants who correctly identified schools that had specific offerings without decreasing the proportion who could identify the school with the most offerings. However, when the site listed all offerings, participants also found it more difficult to identify which school had the most offerings and were less willing to recommend the site to others. These tradeoffs are discussed in more detail in the implications section below.

Box 2 summarizes the combinations of design choices that produced the greatest improvement in outcome measures.

Table 2. Effect of proposed design changes on outcome measures

	Proposed design change				
Outcome measure	Change position of STAR link	Display STAR points earned	Put district average proficiency in its own bar	Display year-over-year change in school performance	List all potential offerings
Usability	+ +			– –	
Understanding		+		+	+
Ease of finding specific information			–		–
Willingness to recommend		+		–	–

STAR is School Transparency and Reporting.

+ + indicates that the proposed design change has a substantial (odds ratio > 1.05 or 0.1 standard deviation) positive effect on the measure. + indicates that the proposed design change has a positive effect on the measure. – indicates that the proposed design change has a negative effect on the measure. – – indicates that the proposed design change has a substantial negative effect (odds ratio < .95 or –0.1 standard deviation) on the measure. *Usability* results are based on all 13 items in the instrument that relate to usability. *Understanding* and *ease of finding specific information* results are based on specific items that are conceptually related to the design elements manipulated by each factor. *Willingness to recommend* is a single continuous variable (range of 1, not at all likely, to 10, extremely likely) based on response to a single item.

Source: Authors' analysis of data collected for this study and described in appendix B.

Box 2. Cumulative effects of design factors on outcomes

This report focuses on the effect of individual design choices. Another way to evaluate school report card designs is to examine which combination of design elements produces the largest favorable effect for each outcome measure. This approach is different from looking at the independent effects of each factor (as in the main text) because it estimates the cumulative effect of all design choices and interactions between them. It also forces a choice between one design factor and the other, regardless of whether the effect of that factor by itself is significant. This analysis can help designers select the combination of design choices that, in aggregate, would maximize a particular outcome.

For example, changing the position of the School Transparency and Reporting (STAR) link and displaying STAR points earned (rather than floors and targets) increased the proportion of participants who rated the school report card as usable (that is, above the midpoint of this scale) by 6 percent (see table). There is a 98 percent chance that the true effect of this design combination is superior in usability to the business-as-usual design and a 96 percent chance that the magnitude of the effect exceeds 5 percent.

Best versus worst performing design choices for different outcome measures

	Optimal design choice					Difference between best and worst performing design	Percentage probability that the best design is…	
Outcome measure	Position of STAR link	STAR explanation	District average	Year-over-year change in school performance	School offerings		Better[a]	Substantially better[b]
Usability	Under STAR rating	Display points possible	Overlay on school's bar	Do not display change	Amenities offered by school	+6 percent	98	96
Understanding	Under STAR rating	Display points possible	Separate bar	Display change	Amenities offered by school	+0.8 percent	76	49
Ease of finding specific information	Under STAR rating	Display points possible	Separate bar	Do not display change	Amenities offered by school	+0.6 percent	73	42
Willingness to recommend	Under STAR rating	Display points possible	Separate bar	Do not display change	Amenities offered by school	0.46 scale point	94	73

STAR is School Transparency and Reporting.

Note: *Usability* refers to the proportion of respondents who at least slightly agreed that the report card was usable. *Understanding* refers to the percentage of items answered correctly. *Ease of finding specific information* refers to the proportion of respondents who at least slightly agreed that the information was easy to find. *Willingness to recommend* is expressed as scale point difference (range of 1, not at all likely, to 10, extremely likely) based on response to a single item.

a. A difference between the optimal design and business as usual that of more than 0.

b. A greater than 5 percent difference in odds for all outcomes except willingness to recommend, which is defined as a 0.1 difference in standard deviation.

Source: Authors' analysis of data collected for this study and described in appendix B.

As might be expected, differences between the best and worst performing designs are larger than differences between the best performing design and the business-as-usual design (see table C12 in appendix C) because some of OSSE's current design choices were better than the alternatives tested.

The table above also shows the outcome measures for which each design decision is included in the optimal design. For example, changing the position of the STAR link and displaying the STAR points earned on different STAR rating metrics improve outcomes on all measures. There are tradeoffs for some other design changes: displaying year-over-year change maximizes user understanding at the expense of usability, ease of finding specific information, and willingness to recommend the site to others.

Changes to report card design usually had similar effects for different groups of users

All subgroups tended to respond to different report card designs similarly. Of 180 potential differences in how subgroups could respond to outcome measures (5 factors × 4 measures × 9 subgroup comparisons), only 1—a difference in usability ratings between participants of differing education levels—was likely to differ from zero.

Users' average ratings of usability, understanding, ease of finding information, and willingness to recommend the site to others varied widely among subgroups

Although all subgroups responded to different report card designs similarly, there were differences in average outcomes between different subgroups. These differences should be interpreted with caution because they are correlations: any differences could be caused by variables that are correlated with subgroup membership but not controlled for in the analysis, including members of different subgroups selecting into the study for different reasons. Nevertheless, the differences observed are large, and they have important implications for OSSE's refinement of its school report card site and for future studies of report card design.

Mobile device users had more difficulty using the school report card site. They said the site was less usable, they showed less understanding of the materials and had more difficulty finding specific data elements, and they were less willing to recommend it to others. Although the mobile-optimized display used in this study differed from that in the currently deployed version of the website in minor ways, the two displays were close overall, and the test site was developed in a mobile-responsive manner.[2]

Participants who said they had used school report card sites in the past found the site more usable, said it was easier to find specific information, and were more willing to recommend the site to others.

Finally, mean differences between participants from different sample sources were large (table 3). In particular, the community sample recruited directly by OSSE reported more negative attitudes about all school report card designs, saying the designs were less usable and made them less willing to recommend the site than participants from the market research panel did.[3] This sample was made up of people interested enough in school policy to belong to OSSE's email and social media distribution lists and motivated to volunteer feedback about school report cards. The market research panel and Amazon Mechanical Turk samples were also self-selected, so they might not represent the opinions of all school report card users.

Table 3. Differences in average outcomes for selected subgroups

	Outcome measure			
Comparison subgroups	Percentage point difference in usability	Percentage point difference in understanding	Percentage point difference in ease of finding specific information	Scale point difference in willingness to recommend
Mobile vs. desktop	–9 (73 vs. 82)	–9 (26 vs. 35)	–13 (66 vs. 79)	–1.5 (6.4 vs. 7.9)
Used school report card sites vs. did not	+3 (80 vs. 77)	+0.6 (31.6 vs. 31)[a]	+5 (76 vs. 71)	+0.8 (7.8 vs. 7.0)
Community sample vs. research panel	+18 (78 vs. 60)	–4 (29 vs. 33)[b]	+17 (76 vs. 59)	–2.7 (7.4 vs. 4.7)

Note: For usability, understanding, and ease of finding specific elements, the difference in estimated scores between the first and second listed subgroup is reported, and numbers in parentheses are estimated scores listed in the order that the subgroups are listed in the first column. *Usability* refers to the proportion of respondents who at least slightly agreed that the report card was usable. *Understanding* refers to the percentage of items answered correctly. *Ease of finding specific information* refers to the proportion of respondents who at least slightly agreed that the information was easy to find. *Willingness to recommend* is expressed as scale point difference (range of 1, not at all likely, to 10, extremely likely) based on response to a single item. All findings have at least a 95 percent posterior probability of having a substantial effect unless noted otherwise. Complete results are reported in tables C4, C6, C8, and C11 in appendix C.

a. Posterior probability is 41 percent.

b. Posterior probability is 72 percent.

Source: Authors' analysis of data collected for this study and described in appendix B.

2. The test site used bootstrap (https://getbootstrap.com), a web development framework that is widely used and regarded by Tembo (the contractor that partnered with OSSE to develop its current school report card) as following best practices for user experience.
3. The Mechanical Turk sample was essentially identical to the market research panel (see tables C4, C6, C8, and C11 in appendix C).

Implications

Design choices matter

The study team identified small design changes that produce measurable effects on the user experience. Each design choice might have a small effect on its own, but when the changes are combined, they have an additive effect (see box 2). Important to note, the design choices that were compared are all reasonable, being both acceptable to experts and used in different report cards. The design elements and choices examined in any one study are not exhaustive, and the effect of changing design elements examined in one study is likely to be additive with design choices studied elsewhere (for example, Glazerman et al., 2020) or yet to be studied, leaving room for further incremental improvements in performance.

The available data suggest that OSSE's school report card should link the STAR metric score directly to the explanation of how it is calculated and prioritize the explanation of points possible over the floors and targets used for the metric. These two changes have positive effects on some outcomes, do not have negative effects on any outcomes, and are always included in the optimal combination of design elements. The effect of changing the position of the STAR link suggests that report card designers can benefit from thinking not only about how information is displayed but also about how different design elements can be organized to clarify the relationship between them. The effect of points possible on user outcomes relative to floors and targets can inform state education agencies' decisions about which details of their accountability score calculation to emphasize.

The decision about whether to display year-over-year change is less straightforward and depends on which outcomes are most important to OSSE. Users were better able to identify how much schools have changed over time when year-over-year differences are calculated for them, but they also disliked school report cards that display differences between current and past performance.

Users' apparent dislike for year-over-year change scores is surprising. In previous surveys parents have said that they want information about schools' historical performance (District of Columbia Office of the State Superintendent of Education, 2018; Mikulecky & Christie, 2014). One possible explanation is that users want this information but disliked the way the designs in this study presented it. Another possibility is that users believe that they would like to see this information but find it confusing in practice. Year-over-year changes can be large (because the underlying data are imprecise; see Kane & Staiger, 2002, for a discussion) and can seem inconsistent with how schools perform relative to peer institutions. Users might be unsure how to evaluate a school that has shown large improvement but still performs below the district average. Additional qualitative research could clarify why participants disliked this design.

The decision about how to display schools' programmatic offerings involves a similar set of tradeoffs, and the best course of action depends on the outcome that is most important for policymakers. Listing all possible offerings increased the proportion of participants who correctly identified schools that had specific offerings. However, when offerings were listed this way, users felt that it was harder to figure out which schools had the most offerings and were less willing to recommend the site to others.

School report card designers could further investigate how to improve the experience of the school report card site for mobile device users. About 40 percent of visitors to OSSE's school report card page use mobile devices. The differences in outcomes for mobile device users and other users were large. There are many potential causes of these differences. They could reflect individual differences between mobile device users and computer users (other than those measured and controlled for). Or simultaneously using the report card and completing a survey might be more difficult on a mobile device than on a computer, while only browsing school report cards (without having to answer survey questions about them) is equally difficult across platforms. More substantively, it is

possible that there are user experience issues when accessing the site using a mobile device, either when trying to compare multiple schools (as users were asked to do) or in general.

Randomized controlled trials as a means of collecting feedback on report card designs

Most education policymakers recognize the importance of gathering feedback while developing school report cards, but the cost of collecting input from representative samples often makes such samples impractical to use for testing design choices before deploying them. Despite the limitations of convenience samples collected from community feedback meetings, email lists, or other means, they are frequently used as an affordable method of collecting input. Convenience samples can provide outside perspectives to designers that can help prioritize design elements, specific design choices, and outcome measures to use in a more rigorous design on a more representative sample.

The quality of input received from convenience samples can be improved through randomized trials that focus on comparing the relative performance of alternative designs. Relative performance scores for different displays are much less likely than absolute performance scores to be correlated with individual differences between respondents, making relative performance scores less sensitive to sample composition. Important and unaccounted for differences would have to be large to reverse the relative performance of design alternatives.

Although the current study's sample was not compared with a representative sample of potential report card users, design effects were similar across participants recruited through different methods and across participants with different demographic and biographical characteristics (including whether they were or were not from the District of Columbia). These similarities exist despite large differences in how users with different characteristics think about report cards, as demonstrated by the large subgroup differences observed in this study. This finding is consistent with the observation by other researchers that experiments produce results that are often similar for probability samples or convenience samples, even when studying topics such as moral judgment or political beliefs that should depend on participants' lived experiences and cultural backgrounds (Coppock, 2019; Coppock et al., 2018).

State education agencies collecting feedback about school report cards through nonexperimental designs such as surveys, interviews, or focus groups should carefully consider how large differences between subgroups might affect their research findings. Although subgroups tended to show consistent preferences for one design over another, these preferences can be obscured by large differences in outcomes between them. These differences underscore the importance of focusing on whether individual designs perform well relative to a standard of comparison such as an alternative design. Without this standard of comparison, it is difficult to detect the performance of specific design choices from participants' general preexisting beliefs and abilities. In particular, when collecting feedback about different designs from convenience samples without using random assignment (such as iterating on a design over several community meetings), differences in sample characteristics for each data collection could easily be interpreted as differences in design performance rather than changes in sample compositions.

References

Coppock, A. (2019). Generalizing from survey experiments conducted on Mechanical Turk: A replication approach. *Political Science Research and Methods, 7*(3), 613–628.

Coppock, A., Leeper, T. J., & Mullinix, K. J. (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences, 115*(49), 12441–12446.

District of Columbia Office of the State Superintendent of Education. (2018). *DC school report card community engagement: Feedback on layout, navigation, and terminology and definitions.* https://osse.dc.gov/sites/default/files/dc/sites/osse/publication/attachments/DC%20School%20Report%20Cards%20Spring%20Summer%202018%20Engagement%20Report%20vFinal.pdf.

Figlio, D. N., & Lucas, M. E. (2004). What's in a grade? School report cards and the housing market. *American Economic Review, 94*(3), 591–604.

Gentner, D., & Markman, A. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*(1), 45–56. http://eric.ed.gov/?id= EJ538687.

Glazerman, S., Nichols-Barrer, I., Valant, J., Chandler, J., & Burnett, A. (2020). The choice architecture of school choice websites. *Journal of Research on Educational Effectiveness, 13*(2), 322–350. http://eric.ed.gov/?id=EJ1254365.

Hasan, S., & Kumar, A. (2019). *Digitization and divergence: Online school ratings and segregation in America.* SSRN. http://dx.doi.org/10.2139/ssrn.3265316.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies, 64*(2), 79–102.

Hornbæk, K., & Law, E. L. C. (2007). Meta-analysis of correlations among usability measures. In Rosson, M. B. (Ed.), *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 617–626). Association for Computing Machinery.

Kane, T. J., & Staiger, D. O. (2002). Volatility in school test scores: Implications for test-based accountability systems. *Brookings Papers on Education Policy*, *5*(1), 235–283. http://eric.ed.gov/?id=EJ898074.

Mikulecky, M., & Christie, K. (2014). *Rating states, grading schools: What parents and experts say states should consider to make school accountability systems meaningful.* Education Commission of the States. http://eric.ed.gov/?id=ED561935.

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review, 81*(12), 46–55.

Schkade, D. A., & Kleinmuntz, D. N. (1994). Information displays and choice processes: Differential effects of organization, form, and sequence. *Organizational Behavior and Human Decision Processes*, *57*(3), 319–337.

Thomke, S. H. (2020). *Experimentation works: The surprising power of business experiments*. Harvard Business Press.

Vessey, I. (1991). Cognitive fit: A theory-based analysis of the graphs versus tables literature. *Decision Sciences, 22*(2), 219–240

Zhang, S., & Markman, A. B. (2001). Processing product unique features: Alignability and involvement in preference construction. *Journal of Consumer Psychology*, *11*(1), 13–27.

REL 2021–101

July 2021

This report was prepared for the Institute of Education Sciences (IES) under Contract ED-IES-17-C-0006 by the Regional Educational Laboratory Mid-Atlantic administered by Mathematica. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Chandler, J., Hartog, J., Lipman, E., & Gellar, J. (2021). *The effect of school report card design on usability, understanding, and satisfaction* (REL 2021–101). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. http://ies.ed.gov/ncee/edlabs.

This report is available on the Regional Educational Laboratory website at http://ies.ed.gov/ncee/edlabs.