IES Institute of Education Sciences

APPENDIXES
July 2021

Regional Educational Laboratory Mid-Atlantic
At Mathematica

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

See https://go.usa.gov/x6tCt for the full report.

Appendix A. About this study

This appendix provides additional background information about why designing school report cards is difficult and summarizes the existing evidence on the importance of design choices to the user experience.

The challenges of designing information displays

Designing an online school report card is challenging for many reasons. One is that any design is the product of many choices: which data elements to include, whether to use one or more pages to show them, where to place them on each page, and how to organize the pages logically. Each data element in turn requires its own choices, such as what size it should be and whether it should communicate information through numbers, icons, or graphs (and if so, what kind of graphs). Design choices often beget more choices. 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, or have labels that report counts or proportions. As a further complication, these decisions must take into account that people access electronic report cards using devices with different screen sizes and interfaces, including desktop computers and mobile devices such as tablets and smartphones.

A second challenge is that school report card design can be optimized for any number of different outcomes. A central concern is whether users find a website usable (Hornbæk, 2006), which corresponds closely to the topics that states tend to focus on when eliciting feedback about school report cards through community meetings and surveys. Most usability measures assess overall impressions of a site or tool, but they can also focus on specific questions such as the ease of performing a specific task using a website or information display. Usability is important because parents who find report cards hard to use could be less likely to use them. Another consideration is whether people understand the information that they see. Although understanding something can contribute to perceptions of its usability, understanding is distinct from usability in that it is an objective measure rather than an experience. This study also asked participants how willing they would be to recommend the site to others. Willingness to recommend is a widely used measure of customer satisfaction that is associated with growth rates for service and product use (Reichfeld, 2003).

Sometimes, design goals are in tension with each other. Measures of user preferences and user performance (such as being able to understand or report back the information they were shown) are only modestly related (Tractinsky, 2018). In some cases people intuitively prefer designs that lead them to make more factual errors on

tests of understanding (Bundorf & Szrek, 2010; Zacks et al., 1998). For this reason designers of school report cards must sometimes prioritize certain design goals or make tradeoffs to optimize their designs.

A third challenge is that school report cards are often intended to be used by different stakeholders with different levels of ability (literacy, numericity, and digital literacy), subject matter expertise, and reasons for use. Report cards are expected to help school leaders and staff make better policy decisions, help parents understand what is happening in schools, and help community members hold government (at all levels) accountable for school performance. Design choices can have different effects for different users. For example, some evidence suggests that people with less experience reading graphs are more likely than people with more experience to be influenced by design features (Peebles & Ali, 2015). In other domains substantive content knowledge also improves a user's accuracy when interpreting information displays, presumably because people draw on knowledge of the topic to resolve ambiguities about the display (Roth & Bowen, 2001).

A specific concern related to different uses by different stakeholders is that the design of a school report card could exacerbate inequity. Although school report cards can be intended to give under-resourced communities access to valuable information, researchers have expressed concern that parents with higher education levels are better able to take advantage of school report cards and that report cards will exacerbate the tendency for advantaged groups to enroll in high-performing schools (Figlio & Lucas, 2004; Hasan & Kumar, 2019). The design of a school report card could contribute to inequity if design decisions were made with regard for the preferences only of wealthy or highly educated users. Recognizing this challenge, the State Board of Education ESSA Taskforce (2017) recommends testing report card designs on different groups of users.

Design choices influence the usability of information displays

Numerous studies have revealed that many of the decisions required of designers make a difference in the user's experience and understanding of information. These findings illustrate the number and complexity of choices facing designers (for an overview, see Kosslyn, 2006). For example, pie and bar charts can lead to different conclusions about the same underlying data (Spence & Lewandowsky, 1991). However, even within the general category of bar charts, decisions must be made about the axes, whether bars should be horizontal or vertical, and the width and spacing of bars and labels—all of which also matter for usability (Fischer et al., 2005; Talbot et al., 2014).

There is also extensive (albeit not comprehensive) research on the effect of design elements used to support complex information displays in other domains such as health insurance portals (Hibbard et al., 2002; Johnson et al., 2013), food nutrition labels (Campos et al., 2011), e-commerce sites (Ert & Fleischer, 2016), credit card statements (Agarwal et al., 2015), and surveys (Schwarz, 2007). Although the issues faced by schools differ from those faced by health insurance companies, manufacturers of frozen dinners, and banks, these inquiries collectively illustrate that information design plays an important role in people's understanding of and engagement with otherwise objective information.

Even though the unique combination of data elements, audiences, and potential consequences of design choices suggests that it is important to study the effect of design on school report cards, there is scant research on the design elements that matter. Although the underlying design principles are the same, they might play out in different ways. Evidence on which design choices are "best" for school report cards is scarce, but some studies provide nuanced information about how design choices might matter. One study found that parents reported more extreme evaluations of high- and low-performing schools when they saw performance expressed through letter grades than when they saw it expressed through other means (for example, as a numerical performance index, the percentage of students meeting academic goals, or an achievement level; Jacobsen et al. 2014).

A more recent experiment began to address the potential effects of the multitude of design choices by examining how five design features in mock school report cards might affect choice, user satisfaction, and user understanding

(Glazerman et al., 2020). Using an understanding measure similar to the one the current study used, Glazerman and colleagues found that people understood school information better when it was presented using only numbers than when numeric information was supplemented with graphs or icons, but users were less satisfied with numbers alone. Using a usability measure similar to that used by the current study team, Glazerman and colleagues found that sorting schools by distance improved usability. Unlike the current study, Glazerman and colleagues examined how design influenced school choice. They found that parents tended to select schools that scored high on a particular attribute when schools were sorted according to that attribute, when the attribute was displayed as an icon (not as a number or graph), or when it was reported in more granular detail.

Ideally, school report cards are designed by teams of designers who have considerable experience with and knowledge of user experience research. However, sometimes report cards are designed by people who have less experience. Even if designers are experienced enough to avoid some of the most serious design problems, the sheer number of design choices means that many alternatives that seem equally reasonable are untested. The effect of any of these individual design choices are small when examined singularly (otherwise they would not require large-scale testing to find them). But combining a large number of separate design choices to optimize the selected design might lead to large payoffs.

The importance of cumulative effects is easy to underestimate. E-commerce companies run many (sometimes thousands of) experiments a year to optimize user experience. Each experiment tests the effect of changing a single design element—such as making fonts dark gray instead of black or decreasing page load time by a few hundred milliseconds. The usually small but incremental improvements identified by these tests have been used to consistently improve business performance by 10–25 percent a year (Kohavi & Thomke, 2017; Thomke, 2020). Similarly, when Glazerman et al. (2020) examined five design factors in school report cards, the observed effects of each design choice were small, but taken together, the optimal combination produced a 5 percentage point difference in the percentage of factual questions participants answered correctly and a 6 percentage point difference in user satisfaction.

References

Agarwal, S., Chomsisengphet, S., Mahoney, N., & Stroebel, J. (2015). Regulating consumer financial products: Evidence from credit cards. *The Quarterly Journal of Economics*, *130*(1), 111–164.

Bundorf, M. K., & Szrek, H. (2010). Choice set size and decision making: The case of Medicare Part D prescription drug plans. *Medical Decision Making, 30*(5), 582–593. http://doi:10.1177/0272989X09357793.

Campos, S., Doxey, J., & Hammond, D. (2011). Nutrition labels on pre-packaged foods: A systematic review. *Public Health Nutrition, 14*(8), 1496–1506.

Ert, E., & Fleischer, A. (2016). Mere position effect in booking hotels online. *Journal of Travel Research, 55*(3), 311–321.

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.

Fischer, M. H., Dewulf, N., & Hill, R. L. (2005). Designing bar graphs: Orientation matters. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 19*(7), 953–962.

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.

Hibbard, J. H., Slovic, P., Peters, E., & Finucane, M. L. (2002). Strategies for reporting health plan performance information to consumers: Evidence from controlled studies. *Health Services Research, 37*(2), 291–313.

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

Jacobsen, R., Snyder, J. W., & Saultz, A. (2014). Informing or shaping public opinion? The influence of school accountability data format on public perceptions of school quality. *American Journal of Education, 121*(1), 1–27. http://eric.ed.gov/?id=EJ1044003.

Johnson, E. J., Hassin, R., Baker, T., Bajger, A. T., & Treuer, G. (2013). Can consumers make affordable care affordable? The value of choice architecture. *PloS One, 8*(12), e81521.

Kohavi, R., & Thomke, S. (2017, September–October). The surprising power of online experiments. *Harvard Business Review*. https://hbr.org/2017/09/the-surprising-power-of-online-experiments.

Kosslyn, S. M. (2006). *Graph design for the eye and mind*. Oxford University Press USA.

State Board of Education ESSA Taskforce. (2017). *Building a parent-driven school report card.* District of Columbia Office of the State Superintendent of Education. https://osse.dc.gov/sites/default/files/dc/sites/osse/publication/attachments/Jan.%209%2C%202017%20ESSA%20Task%20Force%20Meeting.pdf.

Peebles, D., & Ali, N. (2015). Expert interpretation of bar and line graphs: The role of graphicacy in reducing the effect of graph format. *Frontiers in Psychology, 6*(1), 1673. http://dx.doi.org/10.3389/fpsyg.2015.01673.

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

Roth, W. M., & Bowen, G. M. (2001). Professionals read graphs: A semiotic analysis*. Journal for Research in Mathematics Education, 32*(2), 159–194.

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology*, *21*(2), 277–287.

Spence, I., & Lewandowsky, S. (1991). Displaying proportions and percentages. *Applied Cognitive Psychology, 5*(1), 61–77.

Talbot, J., Setlur, V., & Anand, A. (2014). Four experiments on the perception of bar charts. *IEEE Transactions on Visualization and Computer Graphics, 20*(12), 2152–2160.

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

Tractinsky, N. (2018). The usability construct: A dead end? *Human–Computer Interaction, 33*(2), 131–177.

Zacks, J., Levy, E., Tversky, B., & Schiano, D. J. (1998). Reading bar graphs: Effects of extraneous depth cues and graphical context. *Journal of Experimental Psychology: Applied*, *4*(2), 119–138.

Appendix B. Methods

This study was an online experiment to examine how people responded to changing the design of school report cards. The goal was to determine which design choices participants found easiest to use and most understandable. An initial list of design decisions that could be tested was developed based on observations by the District of Columbia Office of the State Superintendent of Education (OSSE), a literature review conducted by the study team, and the experience of Tembo Inc. (the contractor that partnered with OSSE to develop its current school report cards) in implementing designs for other state education agencies.

Study design

To answer the research questions, the study team tested how people responded to different versions of OSSE's school report card using a survey instrument with three sections. (The full questionnaire is available at https://osf.io/5vw8z/.) The first section was a baseline demographic questionnaire, the second was a task in which participants reviewed seven hypothetical high schools using 1 of 32 randomly assigned information displays (referred to as treatments), and the third was an outcomes survey that included the dependent variables (see below).

Demographic and biographical information. Participants were asked about their experience with District of Columbia schools (whether they are a parent, district resident, student attending a district school, or a school educator in the district, as well as whether they were familiar with any district school report card websites). Parents reported the number of children in their household, the grades children were enrolled in, and whether they have ever applied for the district school lottery.

Participants were also asked to provide information about their demographic characteristics (age, gender, and race/ethnicity), location (zip code), biographical information (education level and how much time they spend on the internet), household income, and language spoken in the home.

School report cards. The study was a randomized factorial experiment that examined five factors simultaneously. The term "factor" refers to a specific design element that could vary across participants. For example, one factor varied whether participants were shown the year-over-year change in school performance measures. Each treatment was constructed out of a combination of display strategies for each factor. To illustrate how a factorial design works, consider that a study with three factors, with each factor consisting of two design decisions, has eight (2 × 2 × 2) treatments, of which each participant would see one (table B1).

Table B1. Illustrative framework for treatments in a 2 × 2 × 2 study design

	A (report card organization) = 1 STAR on top ribbon		A (report card organization) = 2 STAR on main page	
Factor and design choices	B (STAR explanation) = 1 floor/target	B (STAR explanation) = 2 points possible	B (STAR explanation) = 1 floor/target	B (STAR explanation) = 2 points possible
C (proficiency score chart format) = 1 Bar with line indicating district average	$A_1B_1C_1$	$A_1B_2C_1$	$A_2B_1C_1$	$A_2B_2C_1$
C (proficiency score chart format) = 2 School score and district average as stacked bars	$A_1B_1C_2$	$A_1B_2C_2$	$A_2B_1C_2$	$A_2B_2C_2$

STAR is School Transparency and Reporting.
Note: Contrasts include main effects, specifically the effects of factor A (A1 vs. A2), B (B1 vs. B2), and C (C1 vs. C2), as well as two-way interaction effects, specifically AB, BC, and AC. This is a simplified example; the full study design included five factors.
Source: Authors' compilation.

In this study the experiment simultaneously tested five factors, each consisting of two design choices. Consequently, each participant saw 1 of 32 (2 × 2 × 2 × 2 × 2) treatments. Each combination differs in at least one way from all the others. The study team then estimated the effect of each factor (and the interactions between factors) simultaneously while statistically controlling for the effect of other factors. Because random assignment ensured there were no systematic differences between who was exposed to any of the treatments before the study began, any differences between treatment groups is a result of the different design elements.

For each factor one design represented the design used by OSSE in the 2017/18 academic year, and the other represented an alternative design that could be implemented:

- *Report card organization.* A link to the explanation of the STAR rating calculation is either placed in the top ribbon or hyperlinked from the STAR rating on the profile page.
- *Proficiency score chart format.* Graphs displaying academic performance and school environment school metrics indicate the District of Columbia schools average for the metric using a line on the school score bar or a separate bar beneath the school score.
- *Details of STAR rating.* Scores for each metric in the STAR framework report either the floor and target scores or the total points possible that a school could score for the metric.
- *Change over time.* Year-over-year change in academic proficiency metrics is depicted either by using a line graph that reports raw scores for each year or by reporting the difference in scores between years.
- *School offerings.* The display includes only the amenities offered by a school or all possible amenities in the district, with those offered by a school indicated by a check mark and those not offered by a school indicated by an X.

The study team created two sets of school profiles, each of which included seven schools to help ensure that results do not depend on any one set of schools (Judd et al., 2012). Schools were selected from existing District of Columbia high schools so that the covariance of school attributes in the profiles used in the study was similar to what might be observed in the district, but schools were renamed to preempt any potential concerns about why particular schools were included in the study. Schools were selected to ensure variation in values of specific attributes so that factual questions had correct answers. Because participants were also randomly assigned to one of the two sets of schools, each participant was assigned to 1 of 64 unique experimental conditions.

For each school the study team presented a school report card populated with the data elements included in the 2018/19 District of Columbia School Report Card (table B2).

Table B2. Selected data elements on the 2018/19 school report card

Top-level page	Second-level page	Selected data elements on page
Profile page	na	• Picture of school • Contact information • Message from school (a text-based description of the school) • Student population • Offerings
STAR rating	na	• STAR rating score (overall) • STAR framework metric scores for all students • STAR ratings for different student groups
Academic performance	Student achievement	• PARCC 4+/MSAA 3+ performance • PARCC 4+/MSAA 3+ Performance
	College and career readiness	• Advanced Placement/International Baccalaureate participation • District of Columbia SAT percentile
	Graduation rate	• Four-year graduation rate • Five-year graduation rate
School environment	Attendance	• In-seat attendance • Attendance growth
	Student enrollment changes	• Re-enrollment
	School safety and discipline	• Suspension rate • Expulsions
	Teacher and health staff information	• Teacher experience • Teacher qualifications

na is not applicable. MSAA 3+ is the proportion of students who met expectations on the Multi-State Alternate Assessment. PARCC 4+ is the proportion of students who met expectations on the Partnership for Assessment of Readiness for College and Careers assessment. STAR is School Transparency and Reporting.
Note: STAR framework metrics are measures such as SAT percentile, four-year graduation rate, and in-seat attendance that make up the STAR rating. For details, see the STAR Framework Technical Guide (District of Columbia Office of the State Superintendent of Education, 2019).
Source: Authors' compilation.

To motivate participants to explore the school report cards, the study team instructed participants to use the report cards to identify the school that they believed had the strongest academic performance and indicate why. The study team then asked participants to identify the school that they believed had the best school environment and indicate why.

Outcome measures. The study team measured the effect of design choices on four categories of outcomes.

Usability. Participants responded to 13 statements about the usability of the school report cards and about their satisfaction with the site's design using a six-point scale to indicate whether they disagreed strongly (a value of 1), disagreed, disagreed slightly, agreed slightly, agreed, or agreed strongly (a value of 6) with the statement. Five statements were developed with input from OSSE and concerned core functions of a school report card site. Eight were more general evaluative statements about the usefulness and attractiveness of the site that were adapted from two validated measures of usability: the System Usability Scale (Brooke, 1986) and the Standardized User Experience Percentile Rank Questionnaire (Sauro, 2015). Items and means are shown in table B3. Three of the items were negatively worded. Responses to these items were reversed before analysis so that larger values represent more favorable responses.

Table B3. Usability items

Item	Item wording	Source	Mean (standard deviation)	Number of responses
USE_1	The site gave me the right information to compare the academic performance of schools.	New	4.3 (1.44)	793
USE_2	The site gave me the information I need to compare the nonacademic characteristics of schools.	New	4.4 (1.27)	789
USE_3	The site gave me the information I need about the programs offered by these schools.	New	4.5 (1.30)	789
USE_4	I could use these school report cards to find a school that is of interest to me and my child.	New	4.4 (1.33)	789
USE_5	I would feel comfortable talking about the information on this site with educators and/or school leaders.	New	4.2 (1.36)	789
USE_6	The school report card site was easy to use.	SPUR-Q	4.7 (1.24)	788
USE_7	The information presented was easy to understand.	New	4.6 (1.26)	786
USE_8	I was able to find the information I was looking for.	SPUR-Q	4.4 (1.33)	784
USE_9	I find the website to be attractive.	SPUR-Q	4.3 (1.31)	785
USE_10	The website has a clean and simple presentation.	SPUR-Q	4.6 (1.19)	784
USE_11[a]	The site is too complex.	SUS	4.2 (1.55)	783
USE_12[a]	I would need someone to help me use the site effectively.	SUS	4.3 (1.61)	783
USE_13[a]	I found the site difficult to navigate.	New	4.4 (1.59)	780

SPUR-Q is the Standardized User Experience Percentile Rank Questionnaire. SUS is the System Usability Scale.
Note: Scale responses ranged from 1 (strongly disagree) to 6 (strongly agree).
a. Items were reversed prior to analysis so that larger values represent more favorable responses.
Source: Data collected for this study and described in this appendix.

The study team explored whether any of the usability items correlated more strongly with others by conducting an exploratory factor analysis. Of the 13 usability items, 10 loaded strongly onto a single factor, and the other three loaded onto a second factor. These results were in line with the Pearson correlations that were calculated between each pair of items (table B4). The study team believes that these 3 items were differentiated from the other 10 by the survey design rather than substantive factors related to the question itself: The three items that loaded onto the second factor were negatively worded questions, whereby the participant's level of attention to the reversed meaning likely affected responses. Accordingly, the analysis of the usability items grouped all 13 usability items into a single hierarchical model (with the negatively worded items reverse scored) but still made separate effect estimates for each item. This approach enabled the study to estimate an overall effect for each report card design choice on usability while allowing for variation in the effects for each individual item.

Table B4. Correlations of usability items

Item	USE_1	USE_2	USE_3	USE_4	USE_5	USE_6	USE_7	USE_8	USE_9	USE_10	USE_11	USE_12
USE_2	**.52**											
USE_3	**.53**	**.51**										
USE_4	**.55**	**.47**	**.61**									
USE_5	**.46**	**.39**	**.50**	**.59**								
USE_6	**.42**	**.32**	**.41**	**.42**	**.39**							
USE_7	**.40**	**.40**	**.47**	**.47**	**.39**	**.67**						
USE_8	**.53**	**.44**	**.51**	**.55**	**.46**	**.55**	**.65**					
USE_9	**.36**	**.30**	**.40**	**.40**	**.39**	**.48**	**.49**	**.52**				
USE_10	**.32**	**.31**	**.38**	**.40**	**.38**	**.57**	**.59**	**.54**	**.60**			
USE_11	.01	–.05	.04	.00	–.07	**.20**	**.21**	.09	.02	**.17**		
USE_12	–.02	.03	.02	–.02	**–.10**	**.16**	**.17**	.05	–.01	**.14**	**.66**	
USE_13	.04	.05	.03	.04	–.06	**.22**	**.25**	**.10**	.05	**.20**	**.72**	**.69**

Note: Bold values indicate correlations greater than 0.1 that are significant at $p < .01$.
Source: Authors' analysis of data collected for this study and described in this appendix.

Understanding. Participants answered six factual questions about the school profiles they saw. Because the two sets of schools had different attributes, the potential responses and correct answers varied depending on the set of schools whose information the participant viewed. These potential differences were accommodated by including the school profile set that participants saw as a factor in the model assessing understanding (see the model described below for more details).

Seven items were designed, and as explained below, each participant was exposed to six of them. Three of the tested design factors were each associated with one understanding item (report card organization, proficiency score chart format, and change over time). The potential effect of change to school offerings was tested with two items because the study team expected that doing so might make one type of information (whether a specific program was offered) easier to find and another (the total number of programs offered) more difficult to find. The study team used two items to measure the change to explanation of the STAR rating because the team expected the change could make one type of information (the relative weight assigned to different metrics when calculating the score) possible to find and another (the effect of metrics with values outside the floor and target score) more difficult to find. This question and the potential responses did not depend on the schools viewed by the participant, but to limit the length of this section, the study team randomly assigned participants to see only one question about STAR ratings.

The items used and the way they map to factors are shown in table B5. For each factor the table shows the factual question asked to assess how the factor affected users' understanding, the percentage of users who answered correctly, the percentage of users who stated that they did not know, and the prediction made before the experiment regarding how the factor would influence users' understanding.

Table B5. Understanding items

Item	Item wording	Percent answering correctly	Percent do not know	Prediction
Treatment factor: Report card organization				
UN_1	Which school has the highest STAR rating?	39	27	Linking from the STAR rating to the STAR rating page will increase the proportion of participants who correctly identify the school with the highest STAR rating because it makes it easier to find this.
Treatment factor: Explanation of the STAR rating calculation				
UN_2	Based on your understanding of the STAR rating page of the school report card, what has a larger impact on a school's STAR total? (Different STAR metric scores)[a]	14	41	Depicting points possible will increase the proportion of people who correctly identify the metric that has a larger effect on the STAR total because it makes the different weightings salient.
UN_3	Based on your understanding of the STAR rating page of the school report card, what change has a larger impact on a school's STAR total? (Changes in rates outside of or inside floor and target score)[a]	17	36	Depicting floors and targets will increase the proportion of people who correctly identify the change in a metric score that has a larger effect on the STAR total because it highlights that some response options are outside the floor and target ranges.
Treatment factor: Proficiency score chart format				
UN_4	Which of the following schools have above average 90% attendance rates? (must select two correct answers)	22	na	Stacked bars will increase the proportion of participants who correctly identify above average schools. The business-as-usual design displays the district average proficiency as a line that is superimposed on the school average and an accompanying legend. The line that appears in the legend is easily confused with the line that indicates the district average.
Treatment factor: Change over time				
UN_5	Which school saw the greatest improvement in students meeting or exceeding grade-level expectations in English language arts from last year to this year?	27	36	Displaying the year-over-year difference will increase the proportion of participants who correctly identify the school with the most improvement because doing so makes change more salient and removes the need to calculate the change score.
Treatment factor: School offerings				
UN_6	Based on the information contained in the school report cards, which school has the most school programs?	37	23	Showing only the offerings at a school will increase the proportion of participants who correctly identify the school with the most school programs because users can estimate the school with the most offerings by the amount of space the offerings list occupies.
UN_7	Which school [does not offer interscholastic sports/offers STEM programs]	48	25	Showing the full list of offerings will make it easier to identify schools that do or do not offer a specific program because the lack of an offering is explicitly stated rather than inferred by its absence.

na is not applicable. STAR is School Transparency and Reporting.
a. Only half of participants saw this question (randomly assigned).
Source: Data collected for this study and described in this appendix.

Ease of finding specific information. Participants answered six questions about the ease of finding specific information contained within design elements that were changed in the study. Four factors had one question that corresponded directly to the contents of an information display affected by the factor. The study team asked two questions about school offerings because design changes could differentially affect the ability to find two potentially important types of information about schools: the number of offerings a school has and whether it has a specific offering. Summary descriptive statistics for each of these outcome measures are in table B6.

Table B6. Ease of finding specific information items

Item	Item wording	Mean (standard deviation)	Number of responses
Factor: Report card organization			
E_1	It is easy to find a school's STAR rating	4.07 (1.44)	737
Factor: Explanation of the STAR rating calculation			
E_2	It is easy to understand how the STAR rating is calculated	4.83 (1.26)	731
Factor: Proficiency score chart format			
E_3	It is easy to figure out which schools have students who score better on state assessments	3.71 (1.53)	732
Factor: Change over time			
E_4	It is easy to see how a school's performance has changed over time	4.55 (1.30)	728
Factor: School offerings			
E_5	It is easy to figure out whether a school has a particular extracurricular activity	4.47 (1.35)	731
E_6	It is easy to figure out which school has listed the most extracurricular activities	3.91 (1.47)	730

STAR is School Transparency and Reporting.
Source: Data collected for this study and described in this appendix.

Willingness to recommend the site to others. Overall satisfaction with the school report card design was assessed with 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?" The mean response to this item was 7.4 out of 10, with a standard deviation of 2.61.

Study sample

Recruitment. Participants were recruited from three sources. The community sample (n = 126) was recruited from social media sites, newsletters, and direct emails sent by OSSE and affiliates. This sample is assumed to consist of people who are engaged with District of Columbia schools and school policy. A second sample of presumed district residents (n = 1,522) was recruited from a market research panel called Prime Panels that oversampled parents. A third sample of U.S. residents (n = 362) was recruited from Amazon Mechanical Turk. The Mechanical Turk sample oversampled parents and district residents by restricting participation to people with one or more children in their household (for n = 189 responses) and people who previously reported living in the district (for n = 31 responses).

One study link was created for each source, and random assignment occurred independently within each instantiation of the survey, leading the sample to be implicitly stratified by sample source. Random assignment occurred independently in each case, and the survey was posted as an open link, so participants could complete the survey more than once. The study team ensured that the market research panel and Mechanical Turk participants completed only one survey by logging the unique participant identifiers used by the vendor and excluding all responses with the same identifier except the first one, which was randomly assigned to a condition.

It is possible that participants recruited by OSSE completed the study more than once. The sample was not statistically representative of any population because participants opted into the survey.

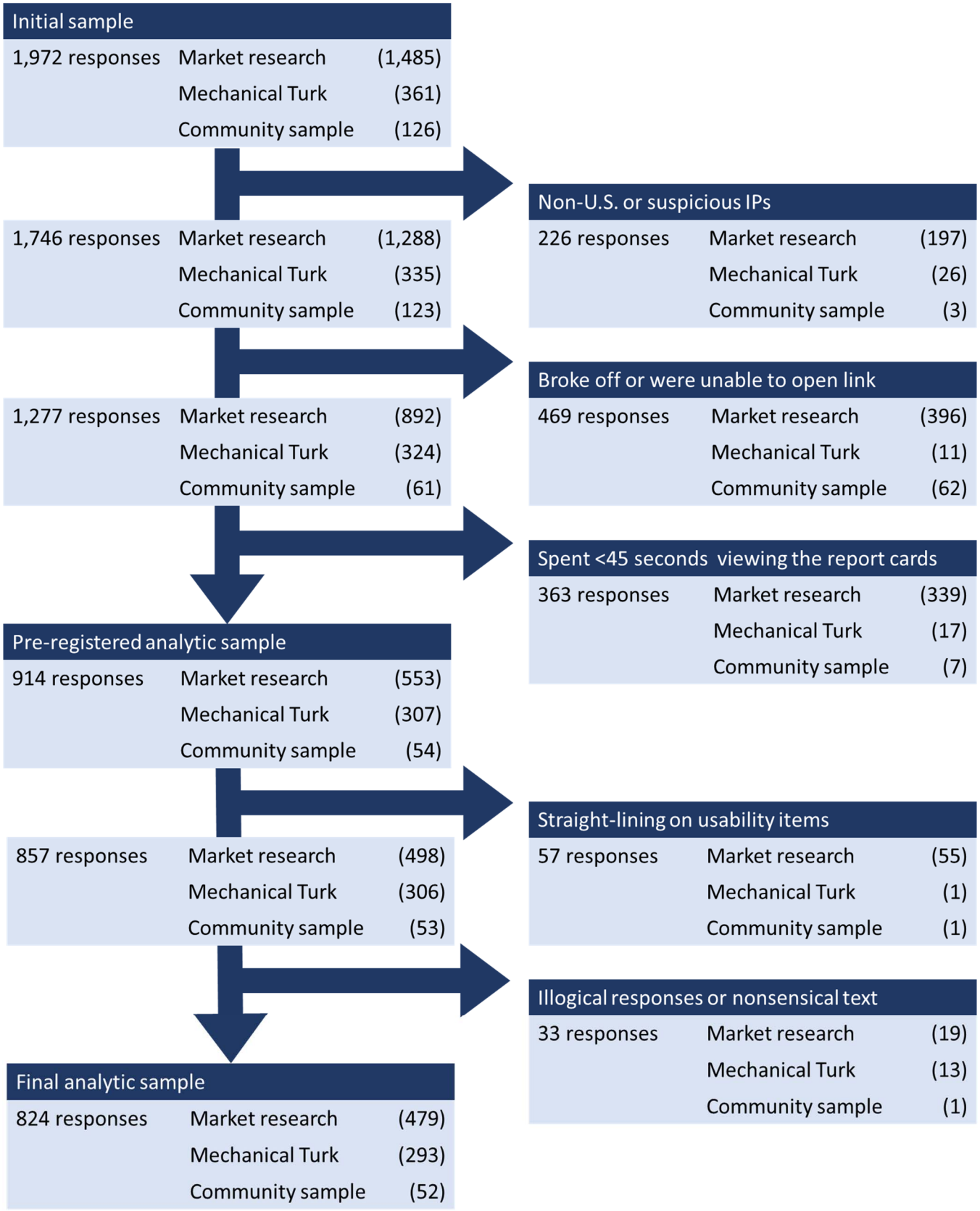
The sample is suitable for testing experimental treatment effects because treatments are randomly assigned and, by design, uncorrelated with the demographic characteristics of survey participants. For this reason, associations between treatments and outcome measures are not influenced by unmeasured characteristics as they are in correlational studies. The only threat to inferences about the general population made from these studies occurs when a factor interacts with individual characteristics (for a discussion, see Coppock et al., 2018). If a subgroup of people is more responsive to a design choice and is substantially more or less common in the sample than in the general population, an experiment will lead to incorrect estimates of treatment effects.

Data cleaning

The data cleaning and analysis plan for this study was registered in the Registry of Efficacy and Effectiveness Studies (REES) and can be accessed at https://sreereg.icpsr.umich.edu/sreereg/subEntry/2604/pdf. Following this plan, the study team excluded participants who completed the survey from non-U.S. IP addresses because they were unlikely to be U.S. residents. It also excluded participants who responded from IP addresses associated with data centers that allow people to run virtual private servers (cloud-based emulations of physical computers) under the assumption that these were either "bots" or non-U.S. residents attempting to circumvent location restrictions set by market research companies and Mechanical Turk (Dennis et al., 2020). Prior studies have demonstrated that participants originating from virtual private servers produce poor-quality data that can falsely decrease or increase observed effect sizes (Chandler et al., 2020). Finally, all participants who spent less than 45 seconds viewing the school report cards were excluded because it was unlikely that they could form a substantive impression of the report cards in so short a time.

After inspecting the distributions of responses in the data but before conducting any analyses, the study team also excluded participants who straight-lined (that is, selected the same response for all questions; Curran, 2016) through the usability questions. These questions were scaled in different directions such that higher or lower values could indicate usability issues, and two items were antonyms ("The school report card site was easy to use" and "I found the site difficult to navigate"). The reason for this exclusion was that anyone selecting the same response for all items was not answering them carefully. Several completed surveys also included improbable responses (people who said they were both District of Columbia public school students and educators, and participants with 10 or more children) or nonsensical responses to open-ended questions, which were also excluded. The effect of these exclusion criteria on sample size is shown in figure B1.

Figure B1. Effect of exclusion criteria on sample size



Source: Data collected for this study and described in this appendix.

Study attrition

The experiment's implementation of random assignment was successful. Study attrition was low (about 23 percent) and virtually identical across report card designs in the analytic sample (table B8).

Table B8. Attrition by report card design

	Completed at least one outcome measure (percent)		Included in the analytic sample (percent)	
Factor	Business-as-usual design	Alternative design	Business-as-usual design	Alternative design
Report card organization	77	78	47	47
Details of the calculation of the STAR rating	77	78	48	47
Proficiency score chart format	78	77	47	47
Change over time	77	77	47	48
School offerings	78	77	47	47

STAR is School Transparency and Reporting.
Note: The analytic sample refers to participants who completed at least one outcome measure and who passed the screening criteria detailed in figure B1. The denominator for all calculations is the number of responses from nonsuspicious U.S. IP addresses.
Source: Data collected for this study and described in this appendix.

Hiearchical Bayesian analysis

Advantages of Bayesian analysis. For each outcome measure, the study team used a Bayesian hierarchical model to estimate the effects of each design choice on each item. Bayesian inference differs from traditional or "frequentist" statistical analyses by allowing researchers to incorporate a set of prior assumptions about the structure of the data into the analysis. In other words the analysis begins with a set of "priors" about the factors of interest in the experiment and how they relate to each other and updates those priors with experimental data. Importantly, this analysis used weakly informative priors (Gelman, 2006). These priors were centered at the null value (zero) and were wide enough to allow the estimates to take on a broad range of values in a data-driven way. In addition, the priors were hierarchical, meaning that the prior variances on the effects were estimated from the data. The result is that the study's Bayesian models tended to be more conservative than a corresponding frequentist analysis that did not incorporate prior information would have been, while still allowing observed data to drive the analysis (Gelman & Jakulin, 2007).

In the context of this study, a Bayesian approach has three primary benefits over frequentist analysis:

- *Shrinkage.* A major feature of Bayesian models (and Bayesian hierarchical models in particular) is that they shrink parameter estimates. Shrinkage occurs in one of two ways. The first is by shrinking parameters toward zero—this leads to more conservative effect estimates. The second is by shrinking related parameters toward one another. This produces a phenomenon known as borrowing of strength, whereby parameters that are driven by less data (for example, those corresponding to smaller subgroups) will be pulled toward similar parameters that are driven by more data. This is essential for the factorial design, which tests many features simultaneously—for example, five factors, up to 13 items per outcome measure, and many subgroup-specific estimates, as well as interactions. Borrowing strength improves and stabilizes the estimate of each individual parameter in the context of a complicated model, thereby providing more robust effect estimates.
- *Correction for multiple comparisons.* Bayesian analysis avoids multiple comparison problems that arise in traditional frequentist analyses when testing many factors and covariates at the same time. In a conventional analysis of results from a factorial design, each factor tested in an experiment would require its own independent hypothesis test. As the number of hypothesis tests increases, so does the probability that at least one will yield a false positive—a situation referred to as the multiple comparisons problem (Waller & Duncan,

1969). Frequentist corrections for multiple comparisons usually focus on inflating the standard errors around estimates, thereby reducing the likelihood of rejecting any individual null hypothesis. Bayesian hierarchical models take a different approach. Instead of inflating standard errors, estimates themselves are shrunken toward the null. This not only reduces the likelihood of a false positive but also results in better point estimates, as described above, because extreme effect estimates (which are more likely to be high or low due to chance alone) are shrunken toward more plausible values (Gelman et al., 2012).

- *Interpretation.* Bayesian analyses provide posterior probabilities that one design choice is superior to the other, based on a prior distribution updated by the available data. In a medical context a posterior probability is analogous to the number of people who received a false positive test result, divided by the total number of people who received either a true positive or false positive test result. It provides information about how likely someone is to have the condition being tested for, given the test results. Similarly, in a research context Bayesian analysis allows one to assert, "The probability that the alternative report card organization results in more favorable responses for the question on willingness to recommend the site to others is X percent." This type of inference is closely aligned with the needs of policymakers who must determine whether any observed differences are "true."

  This approach differs from (more frequently used) frequentist analyses, which report a *p*-value. In a medical screening test a *p*-value is analogous to the number of people who received a false positive test result, divided by the total number of people who do not have the medical condition. It provides information about the proportion of people who do not have the medical condition who will nevertheless test positive. Similarly, in a research context a *p*-value allows one to assert, "The probability of observing a difference equal to or greater than the observed difference by chance alone is X percent if there were no true effect." Though it is easy to imagine circumstances in which *p*-values provide important information (for example, someone using statistical tools to detect plagiarism might want to know how many people they are likely to falsely accuse), this information is often less relevant for decisions about whether to implement one policy or another.

  One of the most serious difficulties with using *p*-values in research is that even experienced researchers interpret them as posterior probabilities (Cohen, 1994), which often leads evidence to appear stronger than it is. A frequentist test with $p = .05$ (the standard significance threshold in frequentist analysis) has about a 70 percent posterior probability of being true, assuming that there is an equal likelihood that the tested hypothesis is correct or incorrect (Sellke et al., 2001).

The models used for this study. The study team used four (related) models in the analysis, one for each set of outcomes: usability, understanding, ease of finding specific information, and willingness to recommend. The models differ in important ways but share the following features:

- *Multiple items for each outcome measure.* Each model (except the one for willingness to recommend) examined outcomes measured with multiple dependent variables. As described above, the study's Bayesian modeling approach helps mitigate concerns related to the multiple comparisons problem.

- *Five treatment factors.* The experiment defined treatment arms with a set of five factors, described previously. For each item in the model, the models estimated effects of all five factors and the two-way interactions between all possible pairs of factors. Aside from two factors that concerned the design of the STAR rating, each factor influenced distinct design elements. The study team did not expect third-order interactions or higher to matter, so these were omitted from the model (Li et al., 2006).

- *Covariates.* The model included a set of covariates in the model. The covariates served two purposes. First, any residual confounding (differences in the types of people assigned to each treatment) that might have existed despite randomization was controlled for by including main effects of each covariate in the model. Second, the team explored whether the effects of design choices differed for different types of survey

participants by including interactions between covariates and factors. The nine covariates included in each model, all of which are dichotomous, were:

- District of Columbia resident.
- Mobile device user.
- More education (defined as completed a bachelor's, graduate, or professional degree[1]).
- Less education (defined as completed high school or less).
- Speaks a language other than English at home.
- Recruited from Amazon Mechanical Turk.
- Recruited by OSSE.
- Spent less time looking at the site compared to other participants (below median).
- Has used school report card sites before.

- *Participant random effect.* For the models that included multiple items (all but willingness to recommend), participant random effects were included to account for the tendency for individuals' responses to be correlated across similar questions, regardless of which treatment arm they were part of.

The form of the models differed slightly from each other, reflecting the different number and types of variables they included. The next section describes each model in detail.

Outcome measures

Willingness to recommend the site to others. This is the simplest model because it estimates one numeric item. The willingness to recommend item asks survey participants to rate, on a scale of 1 to 10, how likely they would be to recommend the report card website to other potential users. The outcome of interest is their rating on this scale. The scale was treated as continuous and modeled using linear regression. Unlike the other three models, this model pertained to only a single item, so it did not include item–factor interactions.

Let Y_i represent the response for participant *i*, $\{x_{ic}\}$ represent the set of nine covariates ($c \in \{1, \dots, 9\}$) for that participant, and $\{z_{if}\}$ represent the set of five treatment factors ($f \in \{1, \dots, 5\}$) for that participant. The model can then be expressed as:

$$Y_i \sim N(\mu_i, \sigma)$$

$$\mu_i = \alpha + \sum_c \beta_c x_{ic} + \sum_f \theta_f^{trt} z_{if} + \sum_{f<g} \theta_{fg}^{trt,trt} z_{if} z_{ig} + \sum_{f,c} \theta_{fc}^{trt,cov} z_{if} x_{ic}.$$

The model included five sets of terms: α, an overall intercept; $\sum_c \beta_c x_{ic}$, which adjusts for residual confounding of the covariates; $\{x_{ic}\}$; $\sum_f \theta_f^{trt} z_{if}$, which includes main effects of each treatment factor; $\sum_{f<g} \theta_{fg}^{trt,trt} z_{if} z_{ig}$, which accounts for interactions between treatment factors; and $\sum_{f,c} \theta_{fc}^{trt,cov} z_{if} x_{ic}$, which allows effect for each treatment factor to differ for participants with different covariate values.

The cumulative effect of including these five sets of terms is that an effect is estimated for each of the 32 treatments (that is, the possible combinations of the five treatment factors), separately for each person. The effects reported in the main report and in appendixes C and D are marginal effects, defined by averaging these effects across all individuals who responded to the survey item. For example, treatment factor $f = 1$ is report card organization. The study team calculated the marginal effect of report card organization on willingness to

[1] This covariate was treated as a nuisance variable. Its influence was controlled for, but the results were omitted from the main report.

recommend as $\theta_1^{trt} + \frac{1}{2}\sum_{g\neq 1}\theta_{1g}^{trt,trt} z_{ig} + \sum_i \sum_c \theta_{1c}^{trt,cov} x_{ic}$. This marginal effect is referred to as the effect size and interpreted as the average expected increase in willingness to recommend scores, comparing what would happen if all individuals in the study were given the alternative report card design to what would happen if they were all given the business-as-usual design.

A fully Bayesian model provides effect size estimates and posterior probabilities that the effect size exceeds certain meaningful thresholds. The discussion focuses on the posterior probability that the effect size is positive (that is, that the alternative design improves scores on average), as well as the probability that this effect is at least 0.1 standard deviation in the direction of the effect.

Understanding. Participants answered several multiple-choice comprehension questions to assess understanding. The outcome of interest for each factor was how participants answered the questions about the content that was changed by that design factor (though the effect of factors on all understanding questions was estimated as a part of the same model). To model the questions on understanding, the study team first binarized responses into correct and incorrect answers by assigning a value of 1 to participants who selected the correct response and a value of 0 to participants who selected any of the incorrect responses. Whether participants selected the correct response was modeled using a multivariate logistic regression model. Each participant saw one comprehension question to assess the effect of changing report card organization, proficiency score chart format, and depiction of change over time and two comprehension questions to assess the effect of changing the display of school offerings. Participants viewed two sets of schools (with different correct answers), but because the design choices were expected to affect the (a) and (b) versions of a given question in the same direction, both versions were combined into a single correctness indicator. To control for the fact that the set of schools a participant saw might have made each question easier or more difficult to answer, an additional covariate was included in the model to control for whether the participant saw version (a) or (b).

Participants also saw one of two questions related to the explanation of the STAR rating calculation: one about floors and targets for each metric (version a) or one about points possible for each metric (version b). The correct answer to these questions did not depend on which school set the participant saw, so these questions were randomly assigned, independent of school set. Because the design choice associated with the STAR rating calculation displays information to assist either version (a) or (b), these questions were expected to move in opposite directions. To allow for the design choices to have different effects on these two questions, they were not collapsed into a single correctness indicator; rather, the (a) and (b) versions of the question were included in the model as two separate dependent variables.

The study team included all seven dependent variables corresponding to understanding in a single model. Letting $Y_{ij} = 1$ represent a correct response to item j by participant i, and using previous notation to represent the treatment factors (z_{if}) and covariates (x_{ic}) for that participant, the multivariate logistic regression model was expressed as:

$$\Pr(Y_{ij} = 1) = p_{ij}$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_j + a_i + \sum_c \beta_{jc} x_{ic} + \sum_f \theta_{jf}^{trt} z_{if} + \sum_{f<g} \theta_{jfg}^{trt,trt} z_{if} z_{ig} + \sum_{f,c} \theta_{jfc}^{trt,cov} z_{if} x_{ic}.$$

This model expresses the log odds of responding correctly to each dependent variable j as a function of the treatment factors and covariates and allows the effects to vary by item. More specifically, the model included six sets of terms: α_j, a dependent variable–specific intercept; a_i, a participant-specific random effect to control for within-participant correlation of responses; $\sum_c \beta_{jc} x_{ic}$, which adjusts for residual confounding of the covariates, separately for each dependent variable; $\sum_f \theta_{jf}^{trt} z_{if}$, which includes main effects of each treatment factor, separately for each dependent variable; $\sum_{f<g} \theta_{jfg}^{trt,trt} z_{if} z_{ig}$, which accounts for interactions between treatment

factors, separately for each dependent variable; and $\sum_{f,c} \theta_{fc}^{trt,cov} z_{if} x_{ic}$, which allows effects for each treatment factor to differ for participants with different covariate values, separately for each dependent variable.

The study team estimated a different effect for each of the 32 combinations of treatments (as defined by their covariate values) separately for each of the seven dependent variables. A marginal effect was then calculated for each combination of treatment factor and dependent variable, using a calculation analogous to the one used for willingness to recommend. Logistic regression models produce results that are expressed as log odds ratios. The study team exponentiated these to present a more interpretable odds ratio. Odds ratios range from zero to infinity, with greater odds ratios indicating that the alternative design increases the average participant's likelihood of answering a survey item correctly and odds ratios less than 1 indicating that the alternative design decreases the average participant's likelihood of a correct response. This discussion focuses on these marginal odds ratios (effect sizes), the posterior probability that the odds ratio is greater than 1, and the posterior probability that the change in odds is at least 5 percent in the direction of the effect.

Usability and ease of finding specific information. Participants rated both usability and the ease of finding specific information on a Likert response scale (six levels). The outcome of interest for usability was the average effect of design choices across all items. The outcome of interest for each factor was how easy it was to find the information affected by that factor (though the effect of factors on all ease questions was estimated as part of the same model). Although some researchers treat Likert scales as continuous, this approach has been shown to produce biased estimates because it assumes that all levels of the response have an equal amount of space between them (Bürkner & Vuorre, 2019). To avoid this bias, the study team used ordinal logistic regression.

The ordinal logistic regression model is a generalization of logistic regression that can be applied to categorical dependent variables when the categories have a natural ordering (as is the case with dependent variables on a Likert scale). All the ordinal dependent variables in this survey had the same six levels: strongly disagree, disagree, slightly disagree, slightly agree, agree, and strongly agree. The ordinal logistic regression model can be expressed as a series of five binary logistic regressions, one for each possible threshold between adjacent response levels. When an ordinal logistic regression model is fit, the key assumption is that the regression coefficients (other than the intercepts) are identical across the five models. This assumption, known as the proportional odds assumption, implies that a factor that improves one's odds of responding with a level higher than "strongly disagree" improves one's odds of responding with a level higher than "disagree" by the same amount (on the log odds scale) and so forth.

As with the model used to measure understanding, the study team included multiple dependent variables in each regression model, resulting in multivariate ordinal logistic regression models. Two such models were fit: one that included the 13 usability dependent variables and one that included the six dependent variables on the ease of finding specific information.

Let $Y_{ij} \in \{1, \ldots, 6\}$ be the ordinal response by participant i to survey item j. All dependent variables were ordered so that a higher value indicates a more favorable response to the question. As before, $\{x_{ic}\}$ and $\{z_{if}\}$ represent the covariates and treatment factors corresponding to participant i, respectively. The model can be expressed as:

$$\Pr(Y_{ij} > k) = p_{ijk}, k \in \{1, \ldots, 5\}$$

$$\log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = \alpha_{jk} + a_i + \sum_c \beta_{jc} z_{ic} + \sum_f \theta_{jf}^{trt} z_{if} + \sum_{f<g} \theta_{jfg}^{trt,trt} z_{if} z_{ig} + \sum_{f,c} \theta_{jfc}^{trt,cov} z_{if} x_{ic}.$$

The model included five sets of terms: α_{jk}, an intercept that is specific to each dependent variable and threshold; a_i, a participant-specific random effect to control for within-participant correlation of responses; $\sum_c \beta_{jc} z_{ic}$, which adjusts for residual confounding of the covariates, separately for each dependent variable; $\sum_f \theta_{jf}^{trt} z_{if}$, which includes main effects of each treatment factor, separately for each dependent variable; $\sum_{f<g} \theta_{jfg}^{trt,trt} z_{if} z_{ig}$, which

accounted for interactions between treatment factors, separately for each dependent variable; and $\sum_{f,c} \theta_{fc}^{trt,cov} z_{if} x_{ic}$, which allowed effects for each treatment factor to differ for participants with different covariate values, separately for each dependent variable.

Marginal effects were calculated for each combination of treatment factor and dependent variable, using the same calculation as used in the model to measure understanding. As with the logistic regression for understanding, these marginal effects are also interpreted as odds ratios, but in this case (because the dependent variable is ordinal), it is not simply the odds ratio of responding correctly versus incorrectly, as it is with the model used to measure understanding. This odds ratio corresponds to the odds of responding with a higher value versus a lower value for any of the five outcome thresholds. The study team also marginalized over the items to get a marginal effect of each treatment factor across all items in the model, which for convenience is referred to as the average effect. Because the dependent variables were ordered so that a higher number is "better," these are referred to as the odds ratio of a more favorable response to the survey item. An odds ratio greater than 1 indicates that the alternative design increased a participant's odds of responding favorably to the survey item, whereas an odds ratio less than 1 indicates that the alternative design decreased a participant's odds of responding favorably. As with other dependent variables, this discussion focuses on these marginal odds ratios (effect sizes), the posterior probability that the odds ratio is greater than 1, and the posterior probability that the change in odds is at least 5 percent in the direction of the effect.

Prior assumptions

All Bayesian models start out with prior assumptions (priors), which describe the likely values of model parameters in the absence of any data. The prior distributions are similar across the four models, with some differences based on model type. This study's choices of priors were based on recommendations in Gelman (2006) on prior distributions for variance parameters in hierarchical models. Table B9 summarizes the four models, including which models correspond to each of the two prior structures described below.

Table B9. Summary of the four models

Outcome measure	Number of items	Model type	Priors shrink estimates for different items toward each other
Willingness to recommend	1	Linear	na
Understanding	7	Logistic	No
Ease of finding specific information	6	Ordinal logit	No
Usability	13	Ordinal logit	Yes

na is not applicable.
Source: Data collected for this study and described in this appendix.

In the models for willingness to recommend, understanding, and ease of finding specific information, weakly informative priors were used on the estimates for the treatment effects. For each type of parameter—the treatment effects for each factor on each dependent variable, the two-way interactions between treatment factors for each dependent variable, and the interaction between treatment factor, dependent variable, and covariates—weakly informative shrinkage priors were used that assumed the parameter is normally distributed with a mean difference of 0 and a variance that might differ from the variances of the other types of parameters. These priors permitted the study team to learn the amount of shrinkage from the data separately for each group of parameters.

The assumption built into these priors is that within each of the three groups of parameters, the parameters are independent from each other and come from the same probability distribution. In addition, the fact that the interactions between treatment factors and between treatment factor, dependent variable, and covariates are

shrunken toward zero has the consequence of shrinking the estimates of treatment effects on each subgroup toward the estimate of the main effect of the treatment factor on each item; that is, the effect is homogenous across subgroups until the data provide evidence otherwise.

For the usability model only, the study's priors on the treatment effects incorporated a slightly stronger assumption, in that they also shrank treatment effects for different dependent variables toward each other. In other words, based on the results of the factor analysis, all usability items were assumed to have measured a single underlying construct of "usability," and differences across items reflected measurement error. Incorporating this structure into the priors leads to more precise inference by enabling the model to borrow strength across different dependent variables. As with all prior assumptions, the model still allows estimates for a treatment factor on different items to diverge if there is evidence for this in the data.

The study team did not shrink estimates of treatment effects on different dependent variables together in the prior structure for the other two models with more than one dependent variable (understanding and ease of finding specific information) because in those models the dependent variables are intended to be affected by different treatment factors. For example, changing the position of the STAR rating link should make it easier to find the STAR rating link but is not expected to make it easier to determine which school had more school offerings.

Concretely, these priors for the usability model that borrow strength across items look like those described above, except that for each group of parameters, instead of shrinking the estimates toward zero, the parameters were shrunken toward a mean across items specific to that group:

- The treatment effects for each factor on each dependent variable were shrunken to an overall treatment effect.
- The two-way interactions between treatment factors for each dependent variable were shrunken toward an overall estimate of the two-way interaction.
- The interactions between treatment factors and moderators for each dependent variable were shrunken toward an overall treatment–moderator interaction.

These means were then shrunken toward zero, with the amount of shrinkage learned from the data, as described above.

For the main effects of covariates on each dependent variable, estimates were shrunken toward each other if they pertained to the same item or the same subgroup. The assumptions behind these priors are that covariates will affect responses to each item similarly and that certain items may be more prone to being affected by covariates.

The mathematical specifications for the priors used in each model include the following:

- Willingness to recommend (dependent variable is standardized for the model to have mean 0, standard deviation 1):
  - $\alpha \sim N(0,1)$.
  - $\beta_c \sim N(0, \sigma^c), \sigma^c \sim N(0,1)$.
  - $\theta_f^{trt} \sim N(0, \tau^{trt}), \tau^{trt} \sim N(0,1)$.
  - $\theta_{fg}^{trt,trt} \sim N(0, \tau^{trt,trt}), \tau^{trt,trt} \sim N(0,1)$.
  - $\theta_{fc}^{trt,cov} \sim N\left(0, \tau_c^{trt,cov}\right), \tau_c^{trt,cov} \sim N\left(0, \tau_0^{trt,cov}\right), \tau_0^{trt,cov} \sim N(0,1)$.

- Understanding and ease of finding specific information:
  - $a_i \sim N(0, \tau_a), \tau_a \sim N(0,1)$.
  - $\beta_{jc} = \beta_j^{item} + \beta_c^{cov} + \beta_{jc}^{resid}$.
    - $\beta_j^{item} \sim N\left(0, \sigma^{item}\right), \beta_c^{cov} \sim N(0, \sigma^{cov}), \beta_{jc}^{resid} \sim N\left(0, \sigma^{resid}\right)$.
    - $\left(\sigma^{item}, \sigma^{cov}, \sigma^{resid}\right) \sim N(0,1)$.
  - $\theta_{jf}^{trt} \sim N(0, \tau^{trt}), \tau^{trt} \sim N(0,1)$.
  - $\theta_{jfg}^{trt,trt} \sim N(0, \tau^{trt,trt}), \tau^{trt,trt} \sim N(0,1)$.
  - $\theta_{jfc}^{trt,cov} \sim N\left(0, \tau_c^{trt,cov}\right), \tau^{trt,cov} \sim N\left(0, \tau_0^{trt,cov}\right), \tau_0^{trt,cov} \sim N(0,1)$.
- Usability (shrink estimates for different dependent variables toward each other):
  - $a_i \sim N(0, \tau_a), \tau_a \sim N(0,1)$.
  - $\beta_{jc} = \beta_j^{item} + \beta_c^{cov} + \beta_{jc}^{resid}$.
    - $\beta_j^{item} \sim N\left(0, \sigma^{item}\right), \beta_c^{cov} \sim N(0, \sigma^{cov}), \beta_{jc}^{resid} \sim N\left(0, \sigma^{resid}\right)$.
    - $\left(\sigma^{item}, \sigma^{cov}, \sigma^{resid}\right) \sim N(0,1)$.
  - $\theta_{jf}^{trt} = \phi_f^{trt} + \phi_{jf}^{trt,resid}$.
    - $\phi_f^{trt} \sim N(0, \tau^{trt}), \tau^{trt} \sim N(0,1)$.
    - $\phi_{jf}^{trt,resid} \sim N\left(0, \tau^{trt,resid}\right), \tau^{trt,resid} \sim N(0,1)$.
  - $\theta_{jfg}^{trt,trt} = \phi_{fg}^{trt,trt} + \phi_{jfg}^{trt,trt,resid}$.
    - $\phi_{jfg}^{trt,trt} \sim N(0, \tau^{trt,trt}), \tau^{trt,trt} \sim N(0,1)$.
    - $\phi_{jfg}^{trt,trt,resid} \sim N\left(0, \tau^{trt,trt,resid}\right), \tau^{trt,trt,resid} \sim N(0,1)$.
  - $\theta_{jfc}^{trt,cov} = \phi_{fc}^{trt,cov} + \phi_{jfc}^{trt,cov,resid}$.
    - $\phi_{fc}^{trt,cov} \sim N\left(0, \tau_c^{trt,cov}\right), \tau_c^{trt,cov} \sim N\left(0, \tau_0^{trt,cov}\right), \tau_0^{trt,cov} \sim N(0,1)$.
    - $\phi_{jfc}^{trt,cov,resid} \sim N\left(0, \tau_c^{trt,cov,resid}\right)$.
    - $\tau_c^{trt,cov,resid} \sim N\left(0, \tau_0^{trt,cov,resid}\right), \tau_0^{trt,cov,resid} \sim N(0,1)$.

Defining "significant" effects

Frequentist analysis usually applies a "bright line" statistical test to define effects as "significant" and "not significant," using a *p*-value of .05 to define the cutoff out of convention. Also, by convention, frequentist analysis tests only whether an effect differs from zero. Bayesian analysis has no such conventions. Decisionmakers can flexibly adjust the criteria they use to determine whether a finding supports a decision. For example, a state considering changing an existing school report card might require strong evidence that doing so will have a large effect if the anticipated effort for making the change is high or weaker evidence that doing so will have a smaller effect if the anticipated effort for making the change is low. A designer developing a new school report card might be willing to accept very weak evidence (such as a 51 percent chance that one design element is superior to the other) in the absence of other considerations.

In this report a 70 percent probability was adopted as the threshold to define "significant" effects for the variables that were most likely to differ across design choices. The study team selected the 70 percent threshold for two reasons. First, this value is consistent with the (arbitrary) threshold used by Glazerman et al. (2020). Second, as discussed earlier, a frequentist p = .05 has at least a 30 percent chance of truly being a null effect, making this conceptually similar to the traditional cutoff used by frequentist statisticians. For all tests exact posterior probabilities are reported in appendix C, and readers who prefer to use a different threshold can do so.

In addition to testing effects against zero, the study team applied a second test to evaluate whether the difference is substantial. All outcomes that have at least a 70 percent probability of a greater than a 5 percent difference (for odds ratios) or 0.1 standard deviation (for the continuous item) were defined as substantial. Readers can adopt this higher threshold when considering whether to make changes to existing report card designs that are likely to incur significant costs.

References

Brooke, J. (1986). *System Usability Scale (SUS): A quick-and-dirty method of system evaluation user information.* Digital Equipment Co. Ltd.

Bürkner, P. C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science, 2*(1), 77–101.

Chandler, J., Sisso, I., & Shapiro, D. (2020). Participant carelessness and fraud: Consequences for clinical research and potential solutions. *Journal of Abnormal Psychology, 129*(1), 49.

Cohen, J. (1994). The earth is round (p <.05). *American Psychologist, 49*(1), 997–1003.

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.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*(1), 4–19.

District of Columbia Office of the State Superintendent of Education (2019). DC School Report Card and STAR Framework Technical Guide. https://osse.dc.gov/sites/default/files/dc/sites/osse/publication/attachments/2019%20DC%20School%20Report%20Card%20and%20STAR%20Framework%20Technical%20Guide%201.6.20.pdf.

Dennis, S. A., Goodson, B. M., & Pearson, C. A. (2020). Online worker fraud and evolving threats to the integrity of MTurk data: A discussion of virtual private servers and the limitations of IP-based screening procedures. *Behavioral Research in Accounting, 32*(1), 119–134.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis, 1*(3), 515–534.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*(2), 189–211. http://eric.ed.gov/?id=EJ961444.

Gelman, A., & Jakulin, A. (2007). Bayes: Radical, liberal, or conservative? *Statisica Sinica, 17*(1), 422–426.

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.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*(1), 54.

Li, X., Sudarsanam, N., & Frey, D. D. (2006). Regularities in data from factorial experiments. *Complexity, 11*(5), 32–45.

Sauro, J. (2015). SUPR-Q: A comprehensive measure of the quality of the website user experience. *Journal of Usability Studies, 10*(2), 68–86.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician, 55*(1), 62–71.

Waller, R. A., & Duncan, D. B. (1969). A Bayes rule for the symmetric multiple comparisons problem. *Journal of the American Statistical Association, 64*(328), 1484–1503.

Appendix C. Supporting analyses

This section describes the results of all analyses on which the conclusions in this report are based. Additional results that follow the preregistered analysis plan are reported in appendix D.

The results of this model are expressed as odds ratios (OR). Odds ratios higher than 1 indicate that the alternative design or subgroup had a more favorable score than the business-as-usual design, which serves as the reference group. Odds ratios lower than 1 indicate that the alternative design or subgroup had a less favorable score than the business-as-usual design. Each odds ratio is accompanied by two posterior probabilities: the probability that the true difference exceeds 0 (PP > 1.00 or PP > 0) and the probability that the true difference exceeds a higher threshold. For the usability, understanding, and ease of finding information measures the higher threshold is a 5 percent difference between conditions or subgroups (PP < 0.95 or PP > 1.05). For the willingness to recommend measure the higher threshold is a difference of 0.1 standard deviation (PP > 0.1 SD), or about 0.26 scale point.

Sample characteristics

Sample characteristics for the full analytic sample and various subsamples are in table C1. Characteristics of the population of the District of Columbia are presented for context. The sample is diverse but deviates from the population substantially in the proportion of parents (as would be expected from the study team's decision to oversample parents). Older adults and people from households with above-average incomes are notably underrepresented, and White participants are overrepresented, which is consistent with other samples collected online (Chandler & Shapiro, 2016).

The subsamples differed from each other in important ways. The community sample can be roughly characterized as older, wealthy, highly educated White female educators. The market research panel can be described as younger, less wealthy, and less educated female District of Columbia residents. The Amazon Mechanical Turk sample falls somewhere between these two groups on most variables, but most people in the sample live outside the district.

Table C1. Demographic characteristics of the analytic sample

Characteristic	Community sample (*n* = 52)	Market research panel sample (*n* = 479)	Mechanical Turk sample (*n* = 293)	Full sample (*n* = 824)	District of Columbia population
District of Columbia resident (%)	60	80	15	55	100
Parent of a child under 18 (%)	27	61	55	57	19
Educator (%)	63	5	1	7	—
High school student (%)	4	14	5	10	4
Age (% of individuals age 13 or older)					
18 or younger	0	2	0	1	6
19–34	34	51	43	47	38
35–54	58	45	46	46	30
55 or older	8	2	11	5	26
Sex (% female)	82	66	50	61	53
At least a bachelor's degree (% of individuals age 25 or older)	98	46	64	56	58
Household income <$35,000 (%)	0	42	18	31	26
Household income <$75,000 (%)	12	63	60	59	47
Language other than English at home (%)	10	25	25	24	17
Hispanic (%)	6	12	10	11	11
Black (%)	20	66	12	44	47
White (%)	70	29	82	51	41
Completed using a mobile device (%)	17	75	2	46	na

— is not available. na is not applicable.
Note: Parents are defined as adults with children younger than age 18 who live with them.
Source: Authors' analysis of data collected for this study and described in appendix B; District of Columbia population estimated using the American Community Survey 2018 five-year estimates (U. S. Census Bureau, 2020).

Effect of report card design on usability

Effects of design choices. Placing the link to the School Transparency and Reporting (STAR) core explanation under the STAR rating increased the odds that a user rated usability at least one category level higher (for example, moved from selecting "slightly agree" to "agree"; OR = 1.12, PP > 1.00 = 86 percent, PP > 1.05 = 73 percent). An examination of the individual items suggests that this difference is driven by parents reporting that the alternative design gives them a better chance to find a school of interest to their child (USE_4), leads them to feel comfortable talking about the information on the site (USE_5) and is simpler (USE_9) and more attractive (USE_10; table C2).

Adding information about change over time in school performance decreased user satisfaction (OR = 0.77, PP < 1.00 = 99 percent, PP < 0.95 = 96 percent), driven by lower ratings on all items in the usability scale.

Of the 10 interactions between factors, 2 had a 70 percent chance of a favorable effect or of an unfavorable effect: 1 indicating that changing school report card organization had less benefit when information about the year-over-year change in school performance was displayed, and 1 indicating that displaying all possible school offerings led to lower usability ratings when the explanation of the STAR rating included floors and targets instead of points possible (table C3). Neither of these interactions was expected, and neither was likely to have an effect greater than 5 percent.

Cumulative effects. Although the effects of each factor on its own might be small, the cumulative effect of changing several factors at once is larger. The combination of changing the location of the STAR rating link and expressing the STAR rating metrics in terms of points possible was superior to the business-as-usual design (OR = 1.17, PP > 1.00 = 85 percent, PP > 1.05 = 76 percent).

Table C2. Effect of design decisions on overall usability and on specific usability items

Item		Odds ratio [95 percent credible interval]	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: Report card organization					
Overall		1.12 [0.92, 1.39]	86	73	na
USE_1	The site gave me the right information to compare the academic performance of schools.	1.13 [0.90, 1.44]	85	72	793
USE_2	The site gave me the information I need to compare the nonacademic characteristics of schools.	1.07 [0.84, 1.36]	70	54	789
USE_3	The site gave me the information I need about the programs offered by these schools.	1.14 [0.92, 1.44]	86	74	789
USE_4	I could use these school report cards to find a school that is of interest to me and my child.	1.20 [0.94, 1.53]	93	86	789
USE_5	I would feel comfortable talking about the information on this site with educators and/or school leaders.	1.15 [0.91, 1.47]	87	77	789
USE_6	The school report card site was easy to use.	1.11 [0.88, 1.42]	82	69	788
USE_7	The information presented was easy to understand.	1.10 [0.87, 1.39]	79	64	786
USE_8	I was able to find the information I was looking for.	1.09 [0.87, 1.39]	78	63	784
USE_9	I find the website to be attractive.	1.15 [0.91, 1.48]	88	77	785
USE_10	The website has a clean and simple presentation.	1.16 [0.92, 1.49]	88	78	784
USE_11	The site is too complex.	1.07 [0.84, 1.35]	73	57	783
USE_12	I would need someone to help me use the site effectively.	1.10 [0.88, 1.39]	77	63	783
USE_13	I found the site difficult to navigate.	1.08 [0.85, 1.35]	73	59	780

Item		Odds ratio [95 percent credible interval]	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: Explanation of the STAR rating calculation					
Overall		1.02 [0.84, 1.24]	60	39	na
USE_1	The site gave me the right information to compare the academic performance of schools.	1.08 [0.86, 1.35]	73	58	793
USE_2	The site gave me the information I need to compare the nonacademic characteristics of schools.	1.01 [0.80, 1.26]	55	39	789
USE_3	The site gave me the information I need about the programs offered by these schools.	1.02 [0.81, 1.28]	56	40	789
USE_4	I could use these school report cards to find a school that is of interest to me and my child.	1.07 [0.85, 1.34]	71	56	789
USE_5	I would feel comfortable talking about the information on this site with educators and/or school leaders.	1.08 [0.85, 1.35]	74	58	789
USE_6	The school report card site was easy to use.	1.03 [0.83, 1.30]	59	43	788
USE_7	The information presented was easy to understand.	1.01 [0.80, 1.27]	53	37	786
USE_8	I was able to find the information I was looking for.	0.99 [0.78, 1.23]	45	37	784
USE_9	I find the website to be attractive.	1.06 [0.84, 1.34]	69	53	785
USE_10	The website has a clean and simple presentation.	1.02 [0.81, 1.27]	56	39	784
USE_11	The site is too complex.	0.97 [0.77, 1.22]	40	43	783
USE_12	I would need someone to help me use the site effectively.	0.97 [0.77, 1.22]	40	44	783
USE_13	I found the site difficult to navigate.	1.00 [0.79, 1.26]	50	34	780

Item		Odds ratio [95 percent credible interval]	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: Proficiency score chart format					
Overall		0.98 [0.81, 1.17]	42	38	na
USE_1	The site gave me the right information to compare the academic performance of schools.	0.97 [0.77, 1.20]	37	44	793
USE_2	The site gave me the information I need to compare the nonacademic characteristics of schools.	0.96 [0.77, 1.19]	36	46	789
USE_3	The site gave me the information I need about the programs offered by these schools.	0.97 [0.77, 1.21]	39	45	789
USE_4	I could use these school report cards to find a school that is of interest to me and my child.	1.01 [0.81, 1.25]	53	35	789
USE_5	I would feel comfortable talking about the information on this site with educators and/or school leaders.	1.03 [0.83, 1.30]	60	45	789
USE_6	The school report card site was easy to use.	0.96 [0.77, 1.19]	36	47	788
USE_7	The information presented was easy to understand.	1.00 [0.80, 1.26]	53	33	786
USE_8	I was able to find the information I was looking for.	0.94 [0.75, 1.16]	31	53	784
USE_9	I find the website to be attractive.	0.97 [0.78, 1.21]	41	42	785
USE_10	The website has a clean and simple presentation.	1.01 [0.81, 1.27]	55	37	784
USE_11	The site is too complex.	0.93 [0.74, 1.15]	28	55	783
USE_12	I would need someone to help me use the site effectively.	0.99 [0.79, 1.25]	47	37	783
USE_13	I found the site difficult to navigate.	0.98 [0.79, 1.23]	44	38	780

Item		Odds ratio [95 percent credible interval]	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: Change over time					
Overall		0.77 [0.61, 0.97]	1	96	na
USE_1	The site gave me the right information to compare the academic performance of schools.	0.80 [0.62, 1.04]	5	89	793
USE_2	The site gave me the information I need to compare the nonacademic characteristics of schools.	0.74 [0.57, 0.97]	1	97	789
USE_3	The site gave me the information I need about the programs offered by these schools.	0.80 [0.62, 1.06]	6	88	789
USE_4	I could use these school report cards to find a school that is of interest to me and my child.	0.79 [0.61, 1.03]	5	91	789
USE_5	I would feel comfortable talking about the information on this site with educators and/or school leaders.	0.81 [0.62, 1.05]	7	87	789
USE_6	The school report card site was easy to use.	0.76 [0.59, 0.99]	2	95	788
USE_7	The information presented was easy to understand.	0.74 [0.57, 0.96]	1	97	786
USE_8	I was able to find the information I was looking for.	0.73 [0.56, 0.94]	1	98	784
USE_9	I find the website to be attractive.	0.75 [0.58, 0.97]	1	96	785
USE_10	The website has a clean and simple presentation.	0.76 [0.60, 1.00]	3	95	784
USE_11	The site is too complex.	0.77 [0.59, 1.00]	2	95	783
USE_12	I would need someone to help me use the site effectively.	0.74 [0.57, 0.97]	1	97	783
USE_13	I found the site difficult to navigate.	0.77 [0.60, 1.01]	3	94	780

Item		Odds ratio [95 percent credible interval]	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: School offerings					
Overall		0.96 [0.79, 1.15]	33	48	na
USE_1	The site gave me the right information to compare the academic performance of schools.	0.97 [0.78, 1.21]	38	45	793
USE_2	The site gave me the information I need to compare the nonacademic characteristics of schools.	0.97 [0.77, 1.21]	39	42	789
USE_3	The site gave me the information I need about the programs offered by these schools.	0.93 [0.74, 1.15]	26	59	789
USE_4	I could use these school report cards to find a school that is of interest to me and my child.	0.99 [0.79, 1.24]	49	35	789
USE_5	I would feel comfortable talking about the information on this site with educators and/or school leaders.	0.96 [0.76, 1.21]	39	45	789
USE_6	The school report card site was easy to use.	0.95 [0.77, 1.19]	32	51	788
USE_7	The information presented was easy to understand.	0.91 [0.73, 1.14]	22	62	786
USE_8	I was able to find the information I was looking for.	0.95 [0.74, 1.18]	32	51	784
USE_9	I find the website to be attractive.	0.98 [0.78, 1.23]	44	39	785
USE_10	The website has a clean and simple presentation.	1.01 [0.80, 1.29]	53	36	784
USE_11	The site is too complex.	0.91 [0.72, 1.13]	22	64	783
USE_12	I would need someone to help me use the site effectively.	0.94 [0.75, 1.17]	29	55	783
USE_13	I found the site difficult to navigate.	0.96 [0.76, 1.21]	36	47	780

na is not applicable.
Note: Coefficients in light blue shaded rows have a 70 percent chance of differing from zero. Dark blue shaded rows have a 70 percent chance of having a substantial effect. The effects presented are marginal effects averaged over the demographics of the sample (the expected effect if all participants saw the alternative design compared with the effect if all participants saw the business-as-usual design).
Source: Authors' analysis of data collected for this study and described in appendix B.

Table C3. Effect of two-way interactions between design decisions on overall usability

Treatment factor	Item	Effect size (odds ratio) [95 percent credible interval]	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)
Proficiency score chart format × school offerings	Overall	0.99 [0.81, 1.20]	45	25
Proficiency score chart format × explanation of the STAR rating	Overall	1.00 [0.82, 1.22]	49	22
Proficiency score chart format × report card organization	Overall	1.01 [0.84, 1.22]	53	24
Proficiency score chart format × change over time	Overall	0.98 [0.80, 1.19]	44	29
School offerings × explanation of the STAR rating	Overall	1.08 [0.92, 1.43]	76	47
School offerings × report card organization	Overall	0.96 [0.74, 1.13]	35	34
School offerings × change over time	Overall	1.00 [0.82, 1.25]	48	22
Explanation of the STAR rating × report card organization	Overall	1.03 [0.86, 1.28]	63	33
Explanation of the STAR rating × change over time	Overall	0.98 [0.79, 1.19]	44	29
Report card organization × change over time	Overall	0.93 [0.70, 1.08]	27	46

STAR is School Transparency and Reporting.
Note: The effects presented are coefficients on the interactions between pairs of treatment factors. Coefficients in shaded rows have a 70 percent chance of differing from zero.
Source: Authors' analysis of data collected for this study and described in appendix B.

Subgroup differences. The study team observed large differences in reported usability across subgroups (table C4). These differences are observed in the model, which simultaneously controls for the effects of all other subgroups. One complication with this measure is that the final three items are reversed so that more positive responses are reported as disagreement with the item. For this reason, mean differences can emerge, either because of true differences in usability or because of how much attention participants paid to the questions. Unless noted otherwise, all reported differences remain when reversed items are excluded.

As might be expected, participants who had used school report cards before found the site to be more usable, probably because they were more familiar with the basic concepts that were reported or (in the case of participants who have seen the OSSE site before) with the layout and reporting metrics contained in the report cards. Participants who speak a language other than English at home (and are likely non-native English speakers) reported that the site was less usable, but this difference probably was a result of their endorsing the reversed items.

Participants using mobile devices instead of computers were more likely to say the site was less usable. The OSSE school report card contains an enormous amount of information that is more difficult to display effectively on a mobile device than on a personal computer. This finding should be interpreted with caution, however, because the mobile version of sites created for this study differed from the mobile version of the currently deployed version of the site.

The relationship between education and usability was not linear. Compared with participants with some college or an associate degree, those with either more education or less education rated the site as less usable (OR = 0.68, PP < 0.95 = 98 percent, and OR = 0.72, PP < 0.95 = 91 percent, respectively; see table C4).

By far, the largest differences in usability stemmed from the sample source. The community sample rated the report card site as substantially (0.75 scale point, OR = 0.33, PP < 0.95 = 99 percent) less usable than did people in the other samples (see table C4).

Table C4. Subgroup differences in perceived usability

Subgroup	Regression-adjusted mean score		Difference between subgroup and reference level (odds ratio) [95 percent credible interval]	Percentage probability of subgroup having higher ratings (odds ratio > 1)	Percentage probability of subgroup having substantially different rating from that of reference group (odds ratio < 0.95 or > 1.05 in direction of difference)
	Reference group	Subgroup			
District of Columbia resident	4.40	4.43	1.04 [0.76, 1.43]	59	46
Mobile device user	4.59	4.21	0.47 [0.33, 0.67]	0	100
More education	4.56	4.34	0.68 [0.49, 0.95]	2	98
Less education	4.56	4.44	0.72 [0.42, 1.07]	5	91
Speaks language other than English	4.44	4.35	0.87 [0.65, 1.16]	18	72
Mechanical Turk sample	4.44	4.47	1.06 [0.71, 1.59]	61	52
Community sample	4.44	3.76	0.33 [0.16, 0.74]	1	99
Spent less than median amount of time looking at site	4.53	4.29	0.64 [0.46, 0.85]	0	100
Used school report cards before	4.37	4.53	1.39 [1.01, 1.91]	98	96

Note: The effects presented are marginal effects averaged over the demographics of the sample (the expected difference in responses if all participants were in the subgroup as opposed to the reference group). The reference group for education is moderate education (some college or an associate degree), and the reference group for sample is the market research sample. Coefficients in shaded rows have a 70 percent chance of having a substantial effect.
Source: Authors' analysis of data collected for this study and described in appendix B.

Interactions between treatment factor and subgroup. Despite large differences in average ratings of usability across subgroups, subgroups were affected by design choices in the same way. In fact, across the 45 possible factor-by-subgroup interactions (5 factors × 9 subgroups), only 1 was likely to differ from zero. For brevity these results are not presented. The interaction between the low education subgroup and the explanation of the STAR rating was in the positive direction (OR 1.12, PP > 1.05 = 42 percent), meaning that the explanation of the STAR rating expressed as points possible was particularly helpful for the subgroup with lower education levels (OR 1.14, PP > 1.00 = 76 percent, PP > 1.05 = 60 percent) compared with those with a moderate (OR 1.00) or high (OR 1.02) education level.

Effect on understanding

Effects of design choices. Changes in school report card design influenced understanding, but the effects were small (table C5). Three of the seven theoretically relevant factor-comprehension question combinations had effects in the expected direction. Providing information about the points possible in the explanation of the STAR rating increased the odds that participants correctly identified the metric that was the biggest driver of the overall STAR rating (UN_2, OR = 1.08, PP > 1.00 = 74 percent, PP > 1.05 = 55 percent) and, notably, did not harm understanding of floors and targets. Also as expected, providing information about change over time in school performance made it easier to identify the schools that improved the most (UN_5, OR = 1.09, PP > 1.00 = 79 percent, PP > 1.05 = 58 percent). Finally, displaying all possible school offerings made it easier to identify schools that did or did not have specific offerings (UN_7, OR = 1.07, PP > 1.00 = 74 percent, PP > 1.05 = 52 percent), notably

without making it harder to identify schools with more offerings. Changing the location of the STAR link did not affect participants' ability to identify the school with the highest STAR rating, and changing the format of the proficiency score charts did not affect participants' ability to identify the schools that performed above the district average.

In all, 6 of the 28 treatment factor–question combinations that were not expected to be affected showed evidence of being affected by the design choice. Three of these differences were on questions about whether schools did or did not offer extracurricular activities (see table C5).

None of the interactions between treatment factors had more than a 70 percent chance of a favorable effect or an unfavorable effect. For brevity these estimates are not displayed.

Cumulative effects. The combination of design choices that led to the greatest understanding of report card contents changed the location of the STAR rating link, expressed the STAR rating metrics in terms of points possible, changed the format of the bar chart, and included information about change over time in school performance. It was superior to the business-as-usual design (OR = 1.06, PP > 1.00 = 73 percent, PP > 1.05 = 47 percent).

Table C5. Effect of design choices on understanding

Item		Effect size (odds ratio) [95 percent credible interval]	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: Report card organization					
UN_1	Which school has the highest STAR rating?	1.04 [0.86, 1.26]	65	42	751
UN_2	Based on your understanding of the STAR rating page of the school report card, what has a larger impact on a school's STAR total? (Different STAR metric scores	1.04 [0.84, 1.37]	62	43	371
UN_3	Based on your understanding of the STAR rating page of the school report card, what change has a larger impact on a school's STAR total? (Changes in rates outside of or inside floor and target score)	1.04 [0.85, 1.32]	66	44	374
UN_4	Which of the following schools have above average 90% attendance rates? (must select two correct answers)	1.02 [0.84, 1.28]	60	36	739
UN_5	Which school saw the greatest improvement in students meeting or exceeding grade-level expectations in English language arts from last year to this year?	1.03 [0.85, 1.28]	60	37	751
UN_6	Based on the information contained in the school report cards, which school has the most school programs?	1.02 [0.85, 1.25]	58	35	752
UN_7	Which school [does not offer interscholastic sports/offers STEM programs]	0.92 [0.74, 1.11]	22	60	742
Treatment factor: Explanation of the STAR rating calculation					
UN_1	Which school has the highest STAR rating?	1.02 [0.85, 1.23]	57	35	751
UN_2	Based on your understanding of the STAR rating page of the school report card, what has a larger impact on a school's STAR total? (Different STAR metric scores	1.08 [0.88, 1.45]	74	55	371
UN_3	Based on your understanding of the STAR rating page of the school report card, what change has a larger impact on a school's STAR total? (Changes in rates outside of or inside floor and target score)	1.03 [0.84, 1.33]	59	39	374
UN_4	Which of the following schools have above average 90% attendance rates? (must select two correct answers)	1.02 [0.82, 1.27]	59	36	739
UN_5	Which school saw the greatest improvement in students meeting or exceeding grade-level expectations in English language arts from last year to this year?	1.01 [0.83, 1.22]	53	29	751
UN_6	Based on the information contained in the school report cards, which school has the most school programs?	1.07 [0.88, 1.33]	73	52	752
UN_7	Which school [does not offer interscholastic sports/offers STEM programs]	0.94 [0.75, 1.14]	28	50	742

Item		Effect size (odds ratio) [95 percent credible interval]	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: Proficiency score chart format					
UN_1	Which school has the highest STAR rating?	0.91 [0.73, 1.10]	19	62	751
UN_2	Based on your understanding of the STAR rating page of the school report card, what has a larger impact on a school's STAR total? (Different STAR metric scores	0.98 [0.76, 1.20]	42	36	371
UN_3	Based on your understanding of the STAR rating page of the school report card, what change has a larger impact on a school's STAR total? (Changes in rates outside of or inside floor and target score)	1.01 [0.82, 1.28]	55	35	374
UN_4	Which of the following schools have above average 90% attendance rates? (must select two correct answers)	1.03 [0.84, 1.29]	63	40	739
UN_5	Which school saw the greatest improvement in students meeting or exceeding grade-level expectations in English language arts from last year to this year?	1.03 [0.86, 1.29]	62	40	751
UN_6	Based on the information contained in the school report cards, which school has the most school programs?	0.98 [0.79, 1.19]	42	36	752
UN_7	Which school [does not offer interscholastic sports/offers STEM programs]	1.06 [0.89, 1.31]	71	49	742
Treatment factor: Change over time					
UN_1	Which school has the highest STAR rating?	0.97 [0.78, 1.16]	37	42	751
UN_2	Based on your understanding of the STAR rating page of the school report card, what has a larger impact on a school's STAR total? (Different STAR metric scores	0.99 [0.79, 1.25]	48	32	371
UN_3	Based on your understanding of the STAR rating page of the school report card, what change has a larger impact on a school's STAR total? (Changes in rates outside of or inside floor and target score)	1.06 [0.86, 1.36]	69	49	374
UN_4	Which of the following schools have above average 90% attendance rates? (must select two correct answers)	1.04 [0.86, 1.30]	63	41	739
UN_5	Which school saw the greatest improvement in students meeting or exceeding grade-level expectations in English language arts from last year to this year?	1.09 [0.90, 1.39]	79	58	751
UN_6	Based on the information contained in the school report cards, which school has the most school programs?	1.02 [0.84, 1.23]	60	35	752
UN_7	Which school [does not offer interscholastic sports/offers STEM programs]	1.01 [0.84, 1.23]	54	31	742

Item		Effect size (odds ratio) [95 percent credible interval]	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: School offerings					
UN_1	Which school has the highest STAR rating?	0.98 [0.80, 1.18]	42	35	751
UN_2	Based on your understanding of the STAR rating page of the school report card, what has a larger impact on a school's STAR total? (Different STAR metric scores	1.01 [0.81, 1.29]	53	33	371
UN_3	Based on your understanding of the STAR rating page of the school report card, what change has a larger impact on a school's STAR total? (Changes in rates outside of or inside floor and target score)	0.94 [0.72, 1.15]	28	51	374
UN_4	Which of the following schools have above average 90% attendance rates? (must select two correct answers)	0.98 [0.78, 1.21]	44	36	739
UN_5	Which school saw the greatest improvement in students meeting or exceeding grade-level expectations in English language arts from last year to this year?	0.96 [0.78, 1.17]	36	42	751
UN_6	Based on the information contained in the school report cards, which school has the most school programs?	1.04 [0.86, 1.28]	65	43	752
UN_7	Which school [does not offer interscholastic sports/offers STEM programs]	1.07 [0.89, 1.35]	74	52	742

STAR is School Transparency and Reporting.
Note: The effects presented are marginal effects averaged over the demographics of the sample (the expected effect if all participants saw the alternative design compared with the effect if all participants saw the business-as-usual design). Coefficients in shaded rows have a 70 percent chance of differing from zero.
Source: Authors' analysis of data collected for this study and described in appendix B.

Subgroup differences. There were large differences in reported usability across subgroups (table C6). The model used to analyze the data statistically estimates all parameters simultaneously while controlling for the effects of other subgroups, so the results are not explainable by a single true subgroup difference and correlation between the subgroup variables. As might be expected, after all other subgroup differences were controlled for, people who spent less time looking at the site and people who speak a language other than English at home (and are likely non-native English speakers) answered fewer comprehension questions correctly. Participants with a bachelor's degree or graduate degree (the high education group) were more likely than those with less education to answer comprehension questions correctly. Perhaps surprisingly, District of Columbia residents were less likely than nondistrict residents to answer questions correctly.

Participants using mobile devices answered fewer questions correctly than participants using computers. The OSSE school report card contains an enormous amount of information that is more difficult to display effectively on a mobile device than on a personal computer. However, this finding should be interpreted with caution because the mobile version of the sites created for this study differed from the mobile version of the production version of the site. Again, the differences across samples were quite large, with participants from the community sample and from the Mechanical Turk sample much more likely than those from the market research sample to answer questions correctly.

Table C6. Subgroup differences in overall understanding

Subgroup	Regression-adjusted mean score		Difference between subgroup and reference group (odds ratio) [95 percent credible interval]	Percentage probability of subgroup having higher ratings (odds ratio > 1)	Percentage probability of subgroup having substantially different rating from that of reference group (odds ratio < 0.95 or > 1.05 in direction of difference)
	Reference group	Subgroup			
District of Columbia resident	0.33	0.30	0.83 [0.65, 1.05]	6	87
Mobile device user	0.35	0.26	0.60 [0.44, 0.79]	0	100
More education	0.29	0.33	1.32 [1.04, 1.68]	99	97
Less education	0.29	0.29	1.03 [0.77, 1.38]	59	46
Speaks language other than English	0.33	0.27	0.75 [0.59, 0.97]	1	97
Mechanical Turk sample	0.29	0.34	1.34 [0.98, 1.82]	97	94
Community sample	0.29	0.33	1.20 [0.75, 1.85]	79	72
Spent less than median amount of time looking at site	0.38	0.22	0.43 [0.33, 0.55]	0	100
Used school report cards before	0.32	0.31	0.98 [0.75, 1.25]	42	41

Note: The effects presented are marginal effects averaged over the demographics of the sample (the expected difference in responses if all participants were in the subgroup as opposed to the reference group). The reference group for education is moderate education (some college or an associate degree), and the reference group for sample is the market research sample. Coefficients in shaded rows have a 70 percent chance of having a substantial effect.
Source: Authors' analysis of data collected for this study and described in appendix B.

Interactions between treatment factor and subgroup. There were no interactions between treatment factor and subgroup that were likely to differ from zero for any of the comprehension questions or overall understanding.

Effect on ease of finding specific information

Effects of design choices. Changes in school report card design influenced how easy it was to find specific items, but the effects were small. Changing the format of the proficiency score charts led participants to report more difficulty in figuring out which schools performed better than average (E_1, OR = 0.96, PP < 1.00 = 72 percent, PP

< 0.95 = 58 percent). Displaying all possible school offerings made it harder to figure out which schools had more extracurricular activities (E_5 OR = 0.95, PP < 1.00 = 75 percent, PP < 0.95 = 53 percent).

Of the 28 treatment factor–question combinations that were not expected to be affected, 5 showed evidence of being affected by the design choice. Among these was the finding that changing the position of the STAR link made it easier to figure out how the STAR rating was calculated, which might reflect that the change made this information easier to find. Three of the differences were for questions about whether schools did or did not offer extracurricular activities (table C7).

None of the interactions between treatment factors had more than a 70 percent chance of a favorable effect or an unfavorable effect. For brevity, these estimates are not shown.

Cumulative effects. The combination of design choices that led to the greatest reported ease in finding specific data elements changed the location of the STAR rating link, expressed the STAR rating metrics in terms of points possible, and changed the format of the bar chart. It was superior to the business-as-usual design (OR = 1.03, PP > 1.00 = 70 percent, PP > 1.05 = 32 percent).

Table C7. Effect of design choices on ease of finding information

Item		Effect size (odds ratio) [95 percent credible interval]	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: Report card organization					
E_1	It is easy to figure out which schools have students who score better on state assessments (C9_1)	1.03 [0.89, 1.20]	64	34	737
E_2	It is easy to find a school's STAR rating (C9_2)	0.99 [0.86, 1.15]	46	25	731
E_3	It is easy to understand how the STAR rating is calculated (C9_3)	1.05 [0.91, 1.24]	72	43	732
E_4	It is easy to figure out whether a school has a particular extracurricular activity (C9_4)	1.01 [0.87, 1.17]	56	24	728
E_5	It is easy to figure out which school has listed the most extracurricular activities (C9_5)	0.99 [0.85, 1.16]	43	26	731
E_6	It is easy to see how a school's performance has changed over time (C9_6)	1.01 [0.87, 1.19]	57	28	730
Treatment factor: Explanation of the STAR rating calculation					
E_1	It is easy to figure out which schools have students who score better on state assessments (C9_1)	1.05 [0.92, 1.25]	78	47	737
E_2	It is easy to find a school's STAR rating (C9_2)	0.98 [0.84, 1.13]	38	31	731
E_3	It is easy to understand how the STAR rating is calculated (C9_3)	1.03 [0.89, 1.21]	67	37	732
E_4	It is easy to figure out whether a school has a particular extracurricular activity (C9_4)	0.99 [0.84, 1.15]	44	24	728
E_5	It is easy to figure out which school has listed the most extracurricular activities (C9_5)	0.97 [0.83, 1.13]	35	36	731
E_6	It is easy to see how a school's performance has changed over time (C9_6)	1.01 [0.87, 1.18]	55	26	730
Treatment factor: Proficiency score chart format					
E_1	It is easy to figure out which schools have students who score better on state assessments (C9_1)	0.96 [0.81, 1.09]	28	42	737
E_2	It is easy to find a school's STAR rating (C9_2)	1.04 [0.90, 1.23]	68	38	731
E_3	It is easy to understand how the STAR rating is calculated (C9_3)	1.05 [0.92, 1.23]	73	42	732
E_4	It is easy to figure out whether a school has a particular extracurricular activity (C9_4)	1.07 [0.93, 1.29]	80	52	728
E_5	It is easy to figure out which school has listed the most extracurricular activities (C9_5)	0.97 [0.82, 1.10]	33	36	731
E_6	It is easy to see how a school's performance has changed over time (C9_6)	0.99 [0.85, 1.16]	45	25	730

Item		Effect size (odds ratio) [95 percent credible interval]	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: Change over time					
E_1	It is easy to figure out which schools have students who score better on state assessments (C9_1)	1.02 [0.88, 1.18]	57	29	737
E_2	It is easy to find a school's STAR rating (C9_2)	1.01 [0.87, 1.18]	52	26	731
E_3	It is easy to understand how the STAR rating is calculated (C9_3)	1.00 [0.87, 1.16]	51	22	732
E_4	It is easy to figure out whether a school has a particular extracurricular activity (C9_4)	1.01 [0.88, 1.18]	56	25	728
E_5	It is easy to figure out which school has listed the most extracurricular activities (C9_5)	0.93 [0.78, 1.07]	17	57	731
E_6	It is easy to see how a school's performance has changed over time (C9_6)	1.01 [0.88, 1.18]	56	28	730
Treatment factor: School offerings					
E_1	It is easy to figure out which schools have students who score better on state assessments (C9_1)	1.02 [0.88, 1.19]	60	31	737
E_2	It is easy to find a school's STAR rating (C9_2)	0.94 [0.76, 1.07]	22	52	731
E_3	It is easy to understand how the STAR rating is calculated (C9_3)	1.01 [0.86, 1.18]	55	25	732
E_4	It is easy to figure out whether a school has a particular extracurricular activity (C9_4)	0.99 [0.85, 1.14]	46	25	728
E_5	It is easy to figure out which school has listed the most extracurricular activities (C9_5)	0.95 [0.80, 1.09]	25	47	731
E_6	It is easy to see how a school's performance has changed over time (C9_6)	1.02 [0.88, 1.17]	60	28	730

STAR is School Transparency and Reporting.

Note: The effects presented are marginal effects averaged over the demographics of the sample (the expected effect if all participants saw the alternative design compared with the effect if all participants saw the business-as-usual design). Coefficients in shaded rows have a 70 percent chance of differing from zero. Authors' analysis of data collected for this study and described in appendix B

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

Subgroup differences. There were large differences across subgroups in the reported ease of finding information (table C8). These differences are observed in the model, which simultaneously controls for the effects of all other subgroups. As might be expected, after all other subgroup differences are controlled for, participants who had used school report card sites before reported that it was easier to find specific data elements, whereas people using mobile devices found it harder to find specific data elements. Participants who spent less time looking at the site also said it was harder to find specific data elements. Perhaps unexpectedly, participants who spoke a language other than English at home reported that it was easier to find specific data elements, while participants with more education reported that it was harder to find specific data elements.

Again, the differences between samples were large, with participants from the community sample and from the Mechanical Turk sample more likely than those from the market research sample to say that it was harder to find specific data elements.

Table C8. Subgroup differences in reported ease of finding specific information

Subgroup	Regression-adjusted mean score: Reference group	Regression-adjusted mean score: Subgroup	Difference between subgroup and reference group (odds ratio) [95 percent credible interval]	Percentage probability of subgroup having higher ratings (odds ratio > 1)	Percentage probability of subgroup having substantially different rating from that of reference group (odds ratio < 0.95 or > 1.05 in direction of difference)
District of Columbia resident	4.25	4.25	1.00 [0.72, 1.39]	49	39
Mobile device user	4.47	3.96	0.38 [0.25, 0.59]	0	100
More education	4.29	4.22	0.86 [0.64, 1.16]	17	74
Less education	4.29	4.29	0.98 [0.66, 1.46]	46	43
Speaks language other than English	4.22	4.35	1.25 [0.92, 1.70]	93	88
Mechanical Turk sample	4.36	4.14	0.66 [0.43, 1.06]	4	94
Community sample	4.36	3.70	0.30 [0.13, 0.68]	0	100
Spent less than median amount of time looking at site	4.29	4.21	0.79 [0.58, 1.09]	8	86
Used school report cards before	4.19	4.39	1.49 [1.10, 2.08]	99	98

Note: The effects presented are marginal effects averaged over the demographics of the sample (the expected difference in responses if all participants were in the subgroup as opposed to the reference group. The reference group for education is moderate education (some college or an associate degree), and the reference group for sample is the market research sample. Coefficients in shaded rows have a 70 percent chance of having a substantial effect.
Source: Authors' analysis of data collected for this study and described in appendix B.

Interactions between treatment factor and subgroup. There were no interactions between treatment factor and subgroup that were likely to differ from zero for any of the ratings of the ease of finding specific items or the average ease of finding all items.

Effect on willingness to recommend the site to others

Effects of design choices. Three changes in the design of school report cards influenced people's willingness to recommend the site to others (table C9). Providing information about the change over time in school performance made participants less willing to recommend the site to others (–0.19 scale point, PP > 0 = 90 percent, PP > 0.1 SD = 31 percent). Displaying all possible school offerings also made participants less willing to recommend the site to others (–0.11 scale point, PP > 0 = 79 percent, PP > 0.1 SD = 14 percent). Finally, providing information about points possible in the explanation of the STAR rating calculation increased willingness to recommend (0.1 scale

point, PP > 0 = 79 percent, PP > 0.1 SD = 12 percent. Changing the location of the STAR link and the format of the proficiency score graphs did not influence willingness to recommend.

Of the 10 interactions between factors, 1 had a 70 percent chance of an unfavorable effect. This negative interaction between school offerings and report card organization suggests that moving the STAR rating link increased willingness to recommend when only the offerings particular to a school were listed, instead of all possible offerings (table C10).

Cumulative effects. The combination of design choices that led to a greater willingness to recommend changed the location of the STAR rating link, expressed the STAR rating metrics in terms of points possible, and changed the format of the bar. It was superior to the business-as-usual approach (0.21 score point, PP > 0 = 79 percent, PP > 0.1 SD = 41 percent).

Table C9. Effect of design choices on willingness to recommend the site to others

Treatment factor	Effect size (scale points) [95 percent credible interval]	Percentage probability of a favorable effect (effect size > 0)	Percentage probability of substantial effect (at least 0.1 standard deviation in direction of effect)	Number of responses
Report card organization	0.06 [–0.22, 0.39]	66	10	791
Explanation of the STAR rating calculation	0.10 [–0.17, 0.38]	79	12	791
Proficiency score chart format	0.00 [–0.27, 0.27]	49	3	791
Change over time	–0.19 [–0.53, 0.08]	10	31	791
School offerings	–0.11 [–0.40, 0.16]	21	14	791

STAR is School Transparency and Reporting.
Note: The effects presented are marginal effects averaged over the demographics of the sample (the expected effect if all participants saw the alternative design compared with the effect if all participants saw the business-as-usual design). Coefficients in shaded rows have a 70 percent chance of differing from zero.
Source: Authors' analysis of data collected for this study and described in appendix B.

Table C10. Effect of two-way interactions between design choices on willingness to recommend site to others

Treatment factor	Item	Effect size (scale points) [95 percent credible interval]	Percentage probability of favorable effect (effect size > 0)	Percentage probability of substantial effect (at least 0.1 standard deviation in direction of effect)
Bar chart × offerings	Overall	–0.01 [–0.27, 0.24]	47	3
Bar chart × STAR points earned	Overall	0.01 [–0.23, 0.27]	55	3
Bar chart × STAR rating link	Overall	–0.03 [–0.32, 0.20]	42	4
Bar chart × trend chart	Overall	–0.01 [–0.25, 0.24]	46	2
Offerings × STAR points earned	Overall	0.05 [–0.17, 0.39]	62	7
Offerings × STAR rating link	Overall	–0.08 [–0.46, 0.10]	28	10
Offerings × trend chart	Overall	–0.04 [–0.35, 0.18]	38	5
STAR points earned × STAR rating link	Overall	0.02 [–0.23, 0.27]	57	3
STAR points earned × trend chart	Overall	–0.02 [–0.31, 0.20]	43	3
STAR rating link × trend chart	Overall	–0.03 [–0.31, 0.20]	40	4

STAR is School Transparency and Reporting.
Note: The effects presented are coefficients on the interactions between pairs of treatment factors. Coefficients in shaded rows have a 70 percent chance of differing from zero.
Source: Authors' analysis of data collected for this study and described in appendix B.

Subgroup differences. There were large differences in reported usability across subgroups (table C11). These differences are observed in the model, which simultaneously controls for the effects of all other subgroups. As might be expected, after all other subgroup differences are controlled for, participants who had used school report card sites before were more willing to recommend the site to others, and people using the site on mobile devices were less willing to recommend it. Participants who spent less time looking at the site were also less willing to recommend it. Perhaps unexpectedly, participants who spoke a language other than English at home (and are likely non-native English speakers) were more willing to recommend the site.

Again, the differences across samples were large, with participants from the community sample and from the Mechanical Turk sample less willing to recommend the site than participants from the market research sample. The community sample differed from the other two samples by about three scale points, which is equivalent to 1.06 standard deviation.

Table C11. Subgroup differences in willingness to recommend the site to others

Subgroup	Regression-adjusted mean score		Difference between subgroup and reference group (scale points) [95% CI]	Percentage probability of subgroup having higher ratings (difference > 0)	Percentage probability of subgroup differing by at least 0.1 standard deviation in direction of difference
	Reference	Subgroup			
District of Columbia resident	7.19	7.28	0.09 [–0.39, 0.57]	63	24
Mobile device user	7.92	6.43	–1.48 [–2.05, –0.92]	0	100
More education	7.32	7.21	–0.13 [–0.62, 0.38]	31	30
Less education	7.32	7.2	–0.14 [–0.72, 0.42]	33	34
Speaks language other than English	7.13	7.55	0.43 [–0.03, 0.87]	97	77
Mechanical Turk sample	7.44	7.24	–0.20 [–0.80, 0.45]	27	42
Community sample	7.44	4.68	–2.76 [–3.71, –1.80]	0	100
Spent less than median amount of time looking at site	7.40	7.05	–0.38 [–0.86, 0.09]	6	68
Used school report cards before	7.00	7.76	0.75 [0.25, 1.21]	100	97

Note: The effects presented are marginal effects averaged over the demographics of the sample (the expected difference in responses if all participants were in the subgroup as opposed to the reference group). The reference group for education is moderate education (some college or an associate degree), and the reference group for sample is the market research sample. Coefficients in light blue shaded rows have a 70 percent chance of differing from zero. Dark blue shaded rows have a 70 percent chance of having a substantial effect.
Source: Authors' analysis of data collected for this study and described in appendix B.

Interactions between treatment factor and subgroup. None of the interactions between treatment factor and subgroup for willingness to recommend had coefficients that were likely to differ from zero.

Cumulative effect of design choices

The cumulative effect of design choices is best illustrated through a comparison of the best versus business-as-usual design choices (table C12).

Table C12. Changes to make for different outcomes

Outcome measure	Optimal design includes this design change					Difference between optimal and current design	Percentage probability that the optimal design is…	
	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		Better[a]	Substantially better[b]
Usability	Yes	Yes	No	No	No	+4 percent	85	76
Understanding	Yes	Yes	Yes	Yes	No	0.08 percent	73	47
Ease of finding specific information	Yes	Yes	Yes	No	No	0.02 percent	70	32
Willingness to recommend	Yes	Yes	Yes	No	No	0.21 scale point	79	41

STAR is School Transparency and Reporting. YoY is year over year.
Note: All differences are expressed as percentage difference in odds ratios, except the difference for willingness to recommend (a continuous variable), which is expressed in scale points (range of 1–10).
a. A difference between the optimal design and business-as-usual design that is greater than zero.
b. A difference that is more than 5 percent (for odds ratios) or 0.1 standard deviation (for willingness to recommend).
Source: Authors' analysis of data collected for this study and described in appendix B.

References

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology*, *12*(1), 53–81.

U.S. Census Bureau (2020). *2014-2018 American Community Survey 5 year estimates, Table DP05*. Retrieved February 1, 2021, from https://data.census.gov/cedsci/.

Appendix D. Sensitivity analyses

This appendix contains tables that report results of all analyses using only the preregistered exclusion criteria described in the data cleaning section in appendix B to define the analytic sample. This is a larger sample that does not exclude participants based on analytic decisions made after the data were collected. All results reported as having a 70 percent probability of differing from zero in the main report also have a 70 percent probability of differing from zero when using the preregistered exclusion criteria, with the following exceptions:

- The probability that listing all school offerings decreases understanding of which schools offer a specific program decreased from 74 percent (see table C5 in appendix C) to 65 percent (table D5).
- The probability that changing the proficiency score chart format makes it harder to find state assessment scores decreased from 72 percent (see table C7) to 69 percent (tables D5–D7).
- Explaining the STAR rating calculation in terms of points possible decreases willingness to recommend from 79 percent (see table C9) to 69 percent (table D9).

The results of analyses using the preregistered exclusion criteria also differed in the following ways that would have been included in the main report if this sample had been used:

- Explaining the STAR rating calculation in terms of points possible increased usability (table D2).
- Moving the link to the explanation of the STAR rating link from the top ribbon to the STAR rating score increased willingness to recommend the site to others (table D9).

Table D1. Attrition by treatment factor (percent)

	Completed at least one outcome measure		Included in the preregistered sample	
Factor	Business as usual	Alternative design	Business as usual	Alternative design
Report card organization	77	78	52	53
Details of the calculation of the STAR rating	77	78	53	52
Proficiency score chart format	78	77	54	51
Change over time	77	77	52	52
School offerings	78	77	53	52

Note: The analytic sample refers to participants who completed at least one outcome measure and who passed the preregistered screening criteria detailed in figure B1. The denominator for all calculations is the number of responses from nonsuspicious U.S. IP addresses.
Source: Authors' analysis of data collected for this study and described in appendix B.

Effect of report card design on usability

Table D2. Effect of design choices on overall usability and on specific usability items (preregistered analysis)

Item		Odds ratio [95 percent credible interval]	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: Report card organization					
Overall		1.13 [0.94, 1.39]	90	79	na
USE_1	The site gave me the right information to compare the academic performance of schools.	1.14 [0.92, 1.45]	89	78	883
USE_2	The site gave me the information I need to compare the nonacademic characteristics of schools.	1.08 [0.87, 1.36]	76	60	876
USE_3	The site gave me the information I need about the programs offered by these schools.	1.14 [0.92, 1.46]	88	76	879
USE_4	I could use these school report cards to find a school that is of interest to me and my child.	1.22 [0.96, 1.59]	95	90	879
USE_5	I would feel comfortable talking about the information on this site with educators and/or school leaders.	1.16 [0.94, 1.47]	90	80	879
USE_6	The school report card site was easy to use.	1.14 [0.92, 1.44]	88	78	877
USE_7	The information presented was easy to understand.	1.12 [0.89, 1.40]	84	71	875
USE_8	I was able to find the information I was looking for.	1.13 [0.90, 1.42]	85	73	873
USE_9	I find the website to be attractive.	1.17 [0.94, 1.48]	91	82	874
USE_10	The website has a clean and simple presentation.	1.19 [0.96, 1.52]	94	85	873
USE_11	The site is too complex.	1.08 [0.85, 1.35]	74	58	871
USE_12	I would need someone to help me use the site effectively.	1.10 [0.87, 1.39]	81	66	872
USE_13	I found the site difficult to navigate.	1.08 [0.86, 1.36]	75	59	868

Item		Odds ratio [95 percent credible interval]	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: Explanation of the STAR rating calculation					
Overall		1.02 [0.84, 1.24]	60	39	na
USE_1	The site gave me the right information to compare the academic performance of schools.	1.08 [0.86, 1.35]	73	58	793
USE_2	The site gave me the information I need to compare the nonacademic characteristics of schools.	1.01 [0.80, 1.26]	55	39	789
USE_3	The site gave me the information I need about the programs offered by these schools.	1.02 [0.81, 1.28]	56	40	789
USE_4	I could use these school report cards to find a school that is of interest to me and my child.	1.07 [0.85, 1.34]	71	56	789
USE_5	I would feel comfortable talking about the information on this site with educators and/or school leaders.	1.08 [0.85, 1.35]	74	58	789
USE_6	The school report card site was easy to use.	1.03 [0.83, 1.30]	59	43	788
USE_7	The information presented was easy to understand.	1.01 [0.80, 1.27]	53	37	786
USE_8	I was able to find the information I was looking for.	0.99 [0.78, 1.23]	45	37	784
USE_9	I find the website to be attractive.	1.06 [0.84, 1.34]	69	53	785
USE_10	The website has a clean and simple presentation.	1.02 [0.81, 1.27]	56	39	784
USE_11	The site is too complex.	0.97 [0.77, 1.22]	40	43	783
USE_12	I would need someone to help me use the site effectively.	0.97 [0.77, 1.22]	40	44	783
USE_13	I found the site difficult to navigate.	1.00 [0.79, 1.26]	50	34	780

Item		Odds ratio [95 percent credible interval]	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: Proficiency score chart format					
Overall		1.01 [0.84, 1.21]	53	32	na
USE_1	The site gave me the right information to compare the academic performance of schools.	1.00 [0.81, 1.25]	51	33	883
USE_2	The site gave me the information I need to compare the nonacademic characteristics of schools.	0.98 [0.79, 1.22]	44	40	876
USE_3	The site gave me the information I need about the programs offered by these schools.	0.98 [0.79, 1.22]	43	38	879
USE_4	I could use these school report cards to find a school that is of interest to me and my child.	1.05 [0.85, 1.31]	66	49	879
USE_5	I would feel comfortable talking about the information on this site with educators and/or school leaders.	1.06 [0.85, 1.35]	70	53	879
USE_6	The school report card site was easy to use.	1.00 [0.81, 1.23]	51	33	877
USE_7	The information presented was easy to understand.	1.04 [0.84, 1.32]	64	46	875
USE_8	I was able to find the information I was looking for.	0.98 [0.79, 1.23]	44	37	873
USE_9	I find the website to be attractive.	1.02 [0.82, 1.26]	55	38	874
USE_10	The website has a clean and simple presentation.	1.05 [0.85, 1.31]	68	52	873
USE_11	The site is too complex.	0.94 [0.76, 1.18]	29	52	871
USE_12	I would need someone to help me use the site effectively.	1.00 [0.81, 1.24]	51	32	872
USE_13	I found the site difficult to navigate.	0.98 [0.80, 1.22]	44	38	868

Item		Odds ratio [95 percent credible interval]	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: Change over time					
Overall		0.77 [0.61, 0.97]	1	96	na
USE_1	The site gave me the right information to compare the academic performance of schools.	0.80 [0.62, 1.04]	5	89	793
USE_2	The site gave me the information I need to compare the nonacademic characteristics of schools.	0.74 [0.57, 0.97]	1	97	789
USE_3	The site gave me the information I need about the programs offered by these schools.	0.80 [0.62, 1.06]	6	88	789
USE_4	I could use these school report cards to find a school that is of interest to me and my child.	0.79 [0.61, 1.03]	5	91	789
USE_5	I would feel comfortable talking about the information on this site with educators and/or school leaders.	0.81 [0.62, 1.05]	7	87	789
USE_6	The school report card site was easy to use.	0.76 [0.59, 0.99]	2	95	788
USE_7	The information presented was easy to understand.	0.74 [0.57, 0.96]	1	97	786
USE_8	I was able to find the information I was looking for.	0.73 [0.56, 0.94]	1	98	784
USE_9	I find the website to be attractive.	0.75 [0.58, 0.97]	1	96	785
USE_10	The website has a clean and simple presentation.	0.76 [0.60, 1.00]	3	95	784
USE_11	The site is too complex.	0.77 [0.59, 1.00]	2	95	783
USE_12	I would need someone to help me use the site effectively.	0.74 [0.57, 0.97]	1	97	783
USE_13	I found the site difficult to navigate.	0.77 [0.60, 1.01]	3	94	780

Item		Odds ratio [95 percent credible interval]	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: School offerings					
Overall		0.96 [0.80, 1.15]	35	45	na
USE_1	The site gave me the right information to compare the academic performance of schools.	0.97 [0.78, 1.21]	38	43	793
USE_2	The site gave me the information I need to compare the nonacademic characteristics of schools.	0.97 [0.77, 1.21]	45	38	789
USE_3	The site gave me the information I need about the programs offered by these schools.	0.93 [0.74, 1.15]	28	53	789
USE_4	I could use these school report cards to find a school that is of interest to me and my child.	0.99 [0.79, 1.24]	54	36	789
USE_5	I would feel comfortable talking about the information on this site with educators and/or school leaders.	0.96 [0.76, 1.21]	40	42	789
USE_6	The school report card site was easy to use.	0.95 [0.77, 1.19]	33	49	788
USE_7	The information presented was easy to understand.	0.91 [0.73, 1.14]	25	59	786
USE_8	I was able to find the information I was looking for.	0.95 [0.74, 1.18]	32	49	784
USE_9	I find the website to be attractive.	0.98 [0.78, 1.23]	45	38	785
USE_10	The website has a clean and simple presentation.	1.01 [0.80, 1.29]	52	34	784
USE_11	The site is too complex.	0.91 [0.72, 1.13]	21	66	783
USE_12	I would need someone to help me use the site effectively.	0.94 [0.75, 1.17]	32	53	783
USE_13	I found the site difficult to navigate.	0.96 [0.76, 1.21]	37	45	780

na is not applicable. STAR is School Transparency and Reporting.
Note: The effects presented are marginal effects averaged over the demographics of the sample (the expected effect if all participants saw the alternative design compared with the effect if all participants saw the business-as-usual design). Coefficients in light blue shaded rows have a 70 percent chance of differing from zero. Dark blue shaded rows have a 70 percent chance of having a substantial effect.
Source: Authors' analysis of data collected for this study and described in appendix B.

Table D3. Effect of two-way interactions between design choices on overall usability (preregistered analysis)

Treatment factor	Item	Effect size (odds ratio) [95 percent credible interval]	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)
Proficiency score chart format × school offerings	Overall	1.00 [0.84, 1.19]	49	19
Proficiency score chart format × explanation of the STAR rating	Overall	1.01 [0.86, 1.22]	51	21
Proficiency score chart format × report card organization	Overall	1.03 [0.89, 1.28]	60	28
Proficiency score chart format × change over time	Overall	0.98 [0.79, 1.17]	44	25
School offerings × explanation of the STAR rating	Overall	1.05 [0.92, 1.32]	69	35
School offerings × report card organization	Overall	0.96 [0.74, 1.09]	33	32
School offerings × Change over time	Overall	1.00 [0.85, 1.24]	51	19
Explanation of the STAR rating × report card organization	Overall	1.02 [0.88, 1.24]	57	24
Explanation of the STAR rating × change over time	Overall	0.97 [0.79, 1.12]	39	28
Report card organization × change over time	Overall	0.96 [0.76, 1.10]	34	32

STAR is School Transparency and Reporting.
Note: The effects presented are coefficients on the interactions between pairs of treatment factors.
Source: Authors' analysis of data collected for this study and described in appendix B.

Subgroup differences

Table D4. Subgroup differences in perceived usability (preregistered analysis)

Subgroup	Regression-adjusted mean score		Difference between subgroup and reference group (odds ratio) [95 percent credible interval]	Percentage probability of subgroup having higher ratings (odds ratio > 1)	Percentage probability of subgroup having substantially different rating from that of reference group (odds ratio < 0.95 or > 1.05 in direction of difference)
	Reference group	Subgroup			
District of Columbia resident	4.37	4.41	1.05 [0.78, 1.41]	63	49
Mobile device user	4.58	4.17	0.44 [0.31, 0.63]	0	100
More education	4.50	4.33	0.76 [0.57, 1.05]	4	92
Less education	4.50	4.43	0.80 [0.49, 1.13]	12	81
Speaks language other than English	4.42	4.32	0.87 [0.66, 1.15]	18	73
Mechanical Turk sample	4.42	4.43	1.02 [0.69, 1.49]	53	42
Community sample	4.42	3.67	0.30 [0.15, 0.70]	0	100
Spent less than median amount of time looking at site	4.52	4.26	0.64 [0.48, 0.84]	0	100
Used school report cards before	4.32	4.54	1.55 [1.13, 2.06]	100	99

Note: Coefficients in shaded rows have a 70 percent chance of having a substantial effect.
Source: Authors' analysis of data collected for this study and described in appendix B.

Effect on understanding

Table D5. Effect of design decisions on understanding (preregistered analysis)

Item		Effect size (odds ratio) [95 percent credible interval]	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: Report card organization					
UN_1	Which school has the highest STAR rating?	1.04 [0.88, 1.26]	67	41	838
UN_2	Based on your understanding of the STAR rating page of the school report card, what has a larger impact on a school's STAR total? (Different STAR metric scores)[a]	1.02 [0.86, 1.25]	58	34	417
UN_3	Based on your understanding of the STAR rating page of the school report card, what change has a larger impact on a school's STAR total? (Changes in rates outside of or inside floor and target score)[a]	1.02 [0.85, 1.24]	55	30	416
UN_4	Which of the following schools have above average 90% attendance rates? (must select two correct answers)	1.01 [0.85, 1.21]	56	30	828
UN_5	Which school saw the greatest improvement in students meeting or exceeding grade-level expectations in English language arts from last year to this year?	1.02 [0.87, 1.22]	58	32	840
UN_6	Based on the information contained in the school report cards, which school has the most school programs?	1.01 [0.87, 1.19]	56	31	839
UN_7	Which school [does not offer interscholastic sports/offers STEM programs]	0.97 [0.81, 1.15]	37	36	830
Treatment factor: Explanation of the STAR rating calculation					
UN_1	Which school has the highest STAR rating?	0.99 [0.84, 1.15]	46	28	838
UN_2	Based on your understanding of the STAR rating page of the school report card, what has a larger impact on a school's STAR total? (Different STAR metric scores)[a]	1.07 [0.90, 1.38]	72	50	417
UN_3	Based on your understanding of the STAR rating page of the school report card, what change has a larger impact on a school's STAR total? (Changes in rates outside of or inside floor and target score)[a]	1.01 [0.84, 1.24]	55	31	416
UN_4	Which of the following schools have above average 90% attendance rates? (must select two correct answers)	1.02 [0.86, 1.24]	59	34	828
UN_5	Which school saw the greatest improvement in students meeting or exceeding grade-level expectations in English language arts from last year to this year?	0.99 [0.83, 1.17]	46	30	840
UN_6	Based on the information contained in the school report cards, which school has the most school programs?	1.03 [0.88, 1.25]	64	39	839
UN_7	Which school [does not offer interscholastic sports/offers STEM programs]	0.98 [0.83, 1.15]	41	31	830

Item		Effect size (odds ratio) [95 percent credible interval]	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: Proficiency score chart format					
UN_1	Which school has the highest STAR rating?	0.94 [0.76, 1.08]	23	51	838
UN_2	Based on your understanding of the STAR rating page of the school report card, what has a larger impact on a school's STAR total? (Different STAR metric scores)[a]	0.98 [0.81, 1.17]	42	34	417
UN_3	Based on your understanding of the STAR rating page of the school report card, what change has a larger impact on a school's STAR total? (Changes in rates outside of or inside floor and target score)[a]	0.99 [0.82, 1.20]	48	29	416
UN_4	Which of the following schools have above average 90% attendance rates? (must select two correct answers)	1.02 [0.86, 1.25]	56	31	828
UN_5	Which school saw the greatest improvement in students meeting or exceeding grade-level expectations in English language arts from last year to this year?	1.03 [0.88, 1.24]	63	35	840
UN_6	Based on the information contained in the school report cards, which school has the most school programs?	1.00 [0.83, 1.18]	48	23	839
UN_7	Which school [does not offer interscholastic sports/offers STEM programs]	1.05 [0.90, 1.27]	71	44	830
Treatment factor: Change over time					
UN_1	Which school has the highest STAR rating?	0.97 [0.81, 1.13]	36	37	838
UN_2	Based on your understanding of the STAR rating page of the school report card, what has a larger impact on a school's STAR total? (Different STAR metric scores)[a]	0.99 [0.81, 1.19]	46	30	417
UN_3	Based on your understanding of the STAR rating page of the school report card, what change has a larger impact on a school's STAR total? (Changes in rates outside of or inside floor and target score)[a]	1.04 [0.88, 1.29]	65	40	416
UN_4	Which of the following schools have above average 90% attendance rates? (must select two correct answers)	1.02 [0.86, 1.24]	61	34	828
UN_5	Which school saw the greatest improvement in students meeting or exceeding grade-level expectations in English language arts from last year to this year?	1.07 [0.92, 1.34]	77	54	840
UN_6	Based on the information contained in the school report cards, which school has the most school programs?	1.02 [0.86, 1.21]	58	32	839
UN_7	Which school [does not offer interscholastic sports/offers STEM programs]	1.01 [0.87, 1.20]	57	29	830

Item		Effect size (odds ratio) [95 percent credible interval]	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: School offerings					
UN_1	Which school has the highest STAR rating?	0.95 [0.79, 1.10]	27	47	838
UN_2	Based on your understanding of the STAR rating page of the school report card, what has a larger impact on a school's STAR total? (Different STAR metric scores)[a]	1.00 [0.83, 1.21]	50	27	417
UN_3	Based on your understanding of the STAR rating page of the school report card, what change has a larger impact on a school's STAR total? (Changes in rates outside of or inside floor and target score)[a]	0.97 [0.79, 1.17]	38	36	416
UN_4	Which of the following schools have above average 90% attendance rates? (must select two correct answers)	0.98 [0.82, 1.18]	42	32	828
UN_5	Which school saw the greatest improvement in students meeting or exceeding grade-level expectations in English language arts from last year to this year?	0.98 [0.83, 1.16]	43	32	840
UN_6	Based on the information contained in the school report cards, which school has the most school programs?	1.01 [0.87, 1.21]	57	29	839
UN_7	Which school [does not offer interscholastic sports/offers STEM programs]	1.03 [0.88, 1.23]	64	35	830

STAR is School Transparency and Reporting.

Note: The effects presented are marginal effects averaged over the demographics of the sample (the expected effect if all participants saw the alternative design compared with the effect if all participants saw the business-as-usual design). Coefficients in shaded rows have a 70 percent chance of differing from zero.

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

Table D6. Subgroup differences in overall understanding (preregistered analysis)

Subgroup	Regression-adjusted mean score: Reference group	Regression-adjusted mean score: Subgroup	Difference between subgroup and reference group (odds ratio) [95 percent credible interval]	Percentage probability of subgroup having higher ratings (odds ratio > 1)	Percentage probability of subgroup having substantially different rating from that of reference group (odds ratio < 0.95 or > 1.05 in direction of difference)
District of Columbia resident	0.31	0.29	0.87 [0.69, 1.10]	11	77
Mobile device user	0.34	0.25	0.64 [0.49, 0.83]	0	100
More education	0.28	0.32	1.31 [1.05, 1.63]	99	98
Less education	0.28	0.27	0.95 [0.71, 1.27]	36	51
Speaks language other than English	0.31	0.27	0.82 [0.65, 1.01]	3	91
Mechanical Turk sample	0.28	0.34	1.43 [1.09, 1.88]	100	98
Community sample	0.28	0.3	1.16 [0.76, 1.79]	74	65
Spent less than median amount of time looking at site	0.37	0.21	0.42 [0.34, 0.53]	0	100
Used school report cards before	0.30	0.30	1.02 [0.81, 1.28]	55	38

Note: The effects presented in this table are marginal effects averaged over the demographics of the sample (the expected difference in responses if all participants were in the subgroup as opposed to the reference group). The reference group for education is moderate education (some college or an associate degree), and the reference group for sample is the market research sample. Coefficients in light blue shaded rows have a 70 percent chance of differing from zero. Dark blue shaded rows have a 70 percent chance of having a substantial effect.
Source: Authors' analysis of data collected for this study and described in appendix B.

Effect on ease of finding specific information

Table D7. Effect of design decisions on ease of finding information (preregistered analysis)

Item	Wording	Effect size (odds ratio) [95 percent credible interval	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: Report card organization					
E_1	It is easy to figure out which schools have students who score better on state assessments (C9_1)	1.03 [0.91, 1.22]	67	36	824
E_2	It is easy to find a school's STAR rating (C9_2)	0.99 [0.85, 1.15]	45	24	816
E_3	It is easy to understand how the STAR rating is calculated (C9_3)	1.04 [0.92, 1.23]	71	41	816
E_4	It is easy to figure out whether a school has a particular extracurricular activity (C9_4)	1.00 [0.87, 1.16]	52	24	813
E_5	It is easy to figure out which school has listed the most extracurricular activities (C9_5)	0.99 [0.86, 1.14]	44	25	816
E_6	It is easy to see how a school's performance has changed over time (C9_6)	1.01 [0.88, 1.19]	56	27	816
Treatment factor: Explanation of the STAR rating calculation					
E_1	It is easy to figure out which schools have students who score better on state assessments (C9_1)	1.05 [0.92, 1.24]	75	47	824
E_2	It is easy to find a school's STAR rating (C9_2)	0.99 [0.86, 1.15]	45	24	816
E_3	It is easy to understand how the STAR rating is calculated (C9_3)	1.04 [0.91, 1.22]	69	39	816
E_4	It is easy to figure out whether a school has a particular extracurricular activity (C9_4)	0.99 [0.84, 1.14]	42	28	813
E_5	It is easy to figure out which school has listed the most extracurricular activities (C9_5)	0.96 [0.82, 1.10]	27	43	816
E_6	It is easy to see how a school's performance has changed over time (C9_6)	1.01 [0.88, 1.17]	55	26	816

Item	Wording	Effect size (odds ratio) [95 percent credible interval	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: Proficiency score chart format					
E_1	It is easy to figure out which schools have students who score better on state assessments (C9_1)	0.96 [0.81, 1.10]	31	40	824
E_2	It is easy to find a school's STAR rating (C9_2)	1.04 [0.91, 1.23]	68	38	816
E_3	It is easy to understand how the STAR rating is calculated (C9_3)	1.05 [0.92, 1.23]	72	43	816
E_4	It is easy to figure out whether a school has a particular extracurricular activity (C9_4)	1.08 [0.94, 1.33]	85	60	813
E_5	It is easy to figure out which school has listed the most extracurricular activities (C9_5)	0.98 [0.85, 1.12]	39	31	816
E_6	It is easy to see how a school's performance has changed over time (C9_6)	0.99 [0.85, 1.13]	42	28	816
Treatment factor: Change over time					
E_1	It is easy to figure out which schools have students who score better on state assessments (C9_1)	1.01 [0.88, 1.18]	57	26	824
E_2	It is easy to find a school's STAR rating (C9_2)	1.01 [0.87, 1.17]	54	25	816
E_3	It is easy to understand how the STAR rating is calculated (C9_3)	1.00 [0.86, 1.15]	49	21	816
E_4	It is easy to figure out whether a school has a particular extracurricular activity (C9_4)	1.01 [0.87, 1.18]	57	28	813
E_5	It is easy to figure out which school has listed the most extracurricular activities (C9_5)	0.92 [0.77, 1.05]	16	60	816
E_6	It is easy to see how a school's performance has changed over time (C9_6)	1.01 [0.89, 1.18]	56	26	816

Item	Wording	Effect size (odds ratio) [95 percent credible interval	Percentage probability of favorable effect (odds ratio > 1)	Percentage probability of substantial effect (odds ratio < 0.95 or > 1.05 in direction of effect)	Number of responses
Treatment factor: School offerings					
E_1	It is easy to figure out which schools have students who score better on state assessments (C9_1)	1.02 [0.88, 1.18]	61	31	824
E_2	It is easy to find a school's STAR rating (C9_2)	0.95 [0.79, 1.09]	26	43	816
E_3	It is easy to understand how the STAR rating is calculated (C9_3)	1.01 [0.87, 1.17]	55	26	816
E_4	It is easy to figure out whether a school has a particular extracurricular activity (C9_4)	0.99 [0.85, 1.14]	45	27	813
E_5	It is easy to figure out which school has listed the most extracurricular activities (C9_5)	0.94 [0.79, 1.07]	22	50	816
E_6	It is easy to see how a school's performance has changed over time (C9_6)	1.02 [0.89, 1.19]	58	28	816

STAR is School Transparency and Reporting.
Note: The effects presented are marginal effects averaged over the demographics of the sample (the expected effect if all participants saw the alternative design compared with the effect if all participants saw the business-as-usual design). Coefficients in shaded rows have a 70 percent chance of differing from zero.
Source: Authors' analysis of data collected for this study and described in appendix B.

Table D8. Subgroup differences in reported ease of finding specific information (preregistered analysis)

Subgroup	Regression-adjusted mean score: Reference group	Regression-adjusted mean score: Subgroup	Difference between subgroup and reference group (odds ratio) [95 percent credible interval]	Percentage probability of subgroup having higher ratings (odds ratio > 1)	Percentage probability of subgroup having substantially different rating from that of reference group (odds ratio < 0.95 or > 1.05 in direction of difference)
District of Columbia resident	4.30	4.29	0.96 [0.67, 1.35]	43	47
Mobile device user	4.54	3.98	0.33 [0.22, 0.49]	0	100
More education	4.28	4.28	0.99 [0.72, 1.35]	49	39
Less education	4.28	4.36	1.16 [0.78, 1.72]	78	69
Speaks language other than English	4.26	4.39	1.27 [0.94, 1.74]	94	88
Mechanical Turk sample	4.42	4.14	0.57 [0.36, 0.89]	1	99
Community sample	4.42	3.61	0.22 [0.09, 0.49]	0	100
Spent less than median amount of time looking at site	4.33	4.26	0.82 [0.59, 1.12]	11	82
Used school report cards before	4.20	4.50	1.83 [1.33, 2.53]	100	100

Note: The effects presented are marginal effects averaged over the demographics of the sample (the expected difference in responses if all participants were in the subgroup as opposed to the reference group). The reference group for education is moderate education (some college or an associate degree), and the reference group for sample is the market research sample. Coefficients in light blue shaded rows have a 70 percent chance of differing from zero. Dark blue shaded rows have a 70 percent chance of having a substantial effect.
Source: Authors' analysis of data collected for this study and described in appendix B.

Effect on willingness to recommend the site to others

Table D9. Effect of design decisions on willingness to recommend the site to others (preregistered analysis)

Treatment factor	Effect size (scale points) [95 percent credible interval]	Percentage probability of a favorable effect (effect size > 0)	Percentage probability of substantial effect (at least 0.1 standard deviation in direction of effect)	Number of responses
Report card organization	0.08 [–0.19, 0.38]	71	9	881
Explanation of the STAR rating calculation	0.07 [–0.17, 0.33]	69	7	881
Proficiency score chart format	0.05 [–0.21, 0.32]	64	5	881
Change over time	–0.20 [–0.50, 0.06]	8	31	881
School offerings	–0.12 [–0.40, 0.14]	20	15	881

STAR is School Transparency and Reporting.
Note: The effects presented are marginal effects averaged over the demographics of the sample (the expected effect if all participants saw the alternative design compared with the effect if all participants saw the business-as-usual design). Coefficients in shaded rows have a 70 percent chance of differing from zero.
Source: Authors' analysis of data collected for this study and described in appendix B.

Table D10. Effect of two-way interactions between design decisions on willingness to recommend the site to others (preregistered analysis)

Treatment factor	Item	Effect size (scale points) [95 percent credible interval]	Percentage probability of favorable effect (effect size > 0)	Percentage probability of substantial effect (at least 0.1 standard deviation in direction of effect)
Bar chart × offerings	Overall	–0.01 [–0.26, 0.23]	48	2
Bar chart × STAR points earned	Overall	0.03 [–0.19, 0.31]	61	4
Bar chart × STAR rating link	Overall	0.00 [–0.23, 0.24]	50	2
Bar chart × trend chart	Overall	–0.03 [–0.28, 0.16]	39	3
Offerings × STAR points earned	Overall	0.02 [–0.20, 0.29]	54	3
Offerings × STAR rating link	Overall	–0.08 [–0.45, 0.11]	29	10
Offerings × trend chart	Overall	–0.05 [–0.35, 0.15]	35	6
STAR points earned × STAR rating link	Overall	0.00 [–0.22, 0.24]	51	1
STAR points earned × trend chart	Overall	–0.03 [–0.30, 0.18]	41	4
STAR rating link × trend chart	Overall	–0.01 [–0.24, 0.21]	46	2

STAR is School Transparency and Reporting.
Note: The effects presented are coefficients on the interactions between pairs of treatment factors. Coefficients in shaded rows have a 70 percent chance of differing from zero.
Source: Authors' analysis of data collected for this study and described in appendix B.

Table D11. Subgroup differences in willingness to recommend the site to others

Subgroup	Regression-adjusted mean score		Difference between subgroup and reference group (scale points) [95 percent credible interval]	Percentage probability of subgroup having higher ratings (difference > 0)	Percentage probability of subgroup differing by at least 0.1 standard deviation in direction of difference
	Reference group	Subgroup			
District of Columbia resident	7.16	7.35	0.18 [–0.28, 0.61]	78	34
Mobile device user	7.95	6.50	–1.44 [–1.96, –0.92]	0	100
More education	7.38	7.28	–0.10 [–0.58, 0.37]	32	24
Less education	7.38	7.09	–0.32 [–0.87, 0.23]	13	58
Speaks language other than English	7.16	7.58	0.43 [0.03, 0.85]	98	78
Mechanical Turk sample	7.43	7.32	–0.11 [–0.71, 0.48]	38	31
Community sample	7.43	4.71	–2.72 [–3.62, –1.76]	0	100
Spent less than median amount of time looking at site	7.42	7.11	–0.34 [–0.79, 0.10]	7	63
Used school report cards before	6.99	7.87	0.87 [0.37, 1.36]	100	99

Note: The effects presented are marginal effects averaged over the demographics of the sample (the expected difference in responses if all participants were in the subgroup as opposed to the reference group). The reference group for education is moderate education (some college or an associate degree), and the reference group for sample is the market research sample. Coefficients in light blue shaded rows have a 70 percent chance of differing from zero. Dark blue shaded rows have a 70 percent chance of having a substantial effect.
Source: Authors' analysis of data collected for this study and described in appendix B.