Journal of Science Education and Technology (2020) 29:409–420
https://doi.org/10.1007/s10956-020-09826-8

Operationalizing Optimization in a Middle School Virtual Engineering Internship

Ryan Montgomery[1] · Eric Greenwald[1] · Samuel Crane[2] · Ari Krakowski[1] · Jacqueline Barber[1]

Published online: 27 April 2020

Abstract

New national science standards have elevated attention to student performance with a core set of science and engineering practices, yet guidance about how to assess these practices is only just emerging in the literature. This is particularly true for the set of engineering design–focused concepts and practices articulated in the Next Generation Science Standards' (NGSS) Engineering, Technology, and Application of Science (ETS) standards. In this work, we present a model of student cognition for assessing student facility with the engineering design practice of optimization. We operationalize this model of cognition within a set of engineering-focused units for middle school, framed as Virtual Engineering Internships (VEIs). To operationalize the engineering design practice of optimization within our VEIs, we first broke optimization down into two more specific sub-behaviors: exploration and systematicity. We then designed metrics that provide evidence of those behaviors and would be observable given student clickstream data from a digital design tool. We normalized these metrics based on the obtained distributions from a research trial. We discuss the existing correlations between these behaviors and metrics.

Keywords Learning analytics · Engineering design · NGSS · Optimization · Middle school · Virtual internships

Motivation

Traditional approaches to automatically scorable assessments are inadequate to meet the demands of recent education reforms, which call for performance-based demonstrations of understanding (National Research Council [NRC] 2012; NGSS Lead States 2013). The need for more robust assessments is especially salient when attempting to assess science and engineering practices, because these are patterns of behavior that (1) take time to perform and (2) do not lend themselves to discrete correct or incorrect answers but rather a continuum of abilities (Pellegrino et al. 2014). These two aspects of engineering design practices make traditional assessments (e.g., multiple-choice items, which are of short duration and are scored dichotomously) an awkward fit for diagnosing student facility with engineering design practices. Teachers need new approaches to assessment that will enable them to monitor and support student progress in these complex science and engineering practices (Pellegrino et al. 2014).

Extended performance assessments show promise for assessing complex science and engineering practices (Wertheim et al. 2016). Previous work has developed models of student cognition for problem solving within a science context, for example, identifying categories of student performance that show qualitative variation (Baxter et al. 1996). However, performance tasks can be challenging to develop and use at scale: valid and reliable performance tasks are difficult to develop and are often time-consuming to administer and reliably score (Ruiz-Primo and Shavelson 1996; Shavelson et al. 1991). These challenges are particularly problematic for the formative use of performance assessments, when gathering timely information is paramount. Because a teacher is hard-pressed to evaluate and provide feedback on the performance of a full class of students at the same time, teachers must make choices about which students to focus on, which learning goals to monitor, or how much time to allow between a student's performance and formative feedback.

In response to these challenges, a growing number of efforts have emerged to capitalize on the ability of learning analytics technologies to support assessment of complex learning goals, and science and engineering practices in particular (Bennett et al. 2010; Gobert et al. 2013; Quellmalz

✉ Ryan Montgomery
RMontgom@berkeley.edu

[1] Lawrence Hall of Science, University of California, Berkeley, CA, USA

[2] Amplify Education Inc., Brooklyn, NY, USA

et al. 2012). Digital learning environments, when analyzed through the unobtrusive methods of learning analytics, allow students to behave (and be assessed) in more authentic ways that more closely parallel the true behaviors and practices of science and engineering (Shavelson et al. 1991). By enabling automated assessment of complex student performance, learning analytics offers promise to provide a richer and more immediate picture of student understanding than traditional assessments (Quellmalz and Pellegrino 2009; Serrano-Laguna et al. 2012). Already, computer-assisted formative assessments have shown to be good predictors of academic performance and identifiers of underperforming students (Tempelaar et al. 2015). The promise of learning analytics is captured by the Concord Consortium's question, "Ever wish you could be at every student's desk at the same time?" (Online Assessment, n.d.). The hope of learning analytics is that automatically analyzing the massive and varied data generated from student engagement in digital environments will better illuminate student learning progress and provide information in a more timely manner, all without burdening the teacher with the considerable time required to score performance assessments.

For learning analytics to serve in this formative assessment role, they must provide credible, actionable, and timely information to teachers in order to best serve students (Black and Wiliam 2010). However, while important work has been done towards this goal, the classroom implementation of learning analytics technologies is still nascent. Additional research is needed to (1) inform the design of digital environments so that they are well-suited to generating evidence of multidimensional understanding; (2) identify specific approaches to analytics likely to yield useful and credible information for teachers; and (3) determine ways that information may be communicated to teachers so that it can inform instructional decisions (Kuo et al. 2015; Pellegrino et al. 2016).

Background

Engineering Design Practices

Engineering design practices have long been a target of study (e.g., Marples 1961) and have been defined in a multitude of ways (Dubberly 2004). Crismond and Adams (2012) pull from a wide array of engineering design descriptions to propose a set of nine engineering design strategies, including *Understand the Challenge*, *Build Knowledge*, *Generate Ideas*, *Represent Ideas*, *Weigh Options & Make Decisions*, *Conduct Experiments*, *Troubleshoot*, *Revise/Iterate*, and *Reflect on Process*. Student engagement with engineering design has been explored in terms of these strategies within a digital space (Purzer et al. 2015). Mehalik and Schunn's (2007) meta-analysis identified, among the many design process elements, three that are most commonly associated with productive engagement in the design process: *Explore problem representation*, *Use interactive/iterative design methodology*, and *Search the space* (*explore alternatives*). Similarly, the NGSS and NRC Framework describe three primary components of engineering design: *Defining and Delimiting an Engineering Problem*, *Developing Possible Solutions*, and *Optimizing the Design Solution*.

Improved formative assessment of these highlighted elements of the engineering design process could support more effective development of students' abilities. This could lead to significant gains in both engagement (Kolodner 1993; Hammond 1989; Schank 1982, 1999; Kolodner et al. 2003) and associated science learning for a diverse student population (Doppelt et al. 2008). In this work, we will outline a model of cognition and preliminary learning analytics results for one of these primary engineering design elements.

Assessment and Learning Analytics

A valid assessment system requires three components: (1) a model of student cognition (a "student model"); (2) tasks intended to elicit student actions/evidence towards that model of cognition (a "task model"); (3) a set of rules for making sense of that evidence (an "evidence model") (NRC 2001; Mislevy et al. 2003). It has been noted that very few assessments reference an explicit model of student cognition (Brown and Wilson 2011). Thus, we began by developing a clear model of cognition through carefully examining and defining the constructs at play within engineering design, focusing specifically on the practice of *Optimization*. We chose a focus on optimization for both theoretical and practical reasons, described in detail in the "Results" section.

Some work has aimed to quantify various aspects of student facility with engineering design practices through analysis of student log files. For instance, Vieira et al. (2016) analyzed student log files from a computer-aided design environment to determine how many iterations each student performed over the course of their design work, and how systematic (vs confounded) each experiment performed was. They noted that more advanced ("informed") designers would iterate more and be more systematic in their experimentation. Similarly, Xie et al. (2014) used time-series analyses to measure student engagement, detect known gender differences in design practices, and detect iteration from student log files.

Despite these forays into using process data to provide evidence of student ability with engineering design practices, to our knowledge, the literature contains no operationalized definitions of the engineering design practice of optimization. We therefore propose an operationalized definition of the engineering design practice of optimization, in the hopes that this definition will facilitate future discussion, critique, and eventually consensus regarding the classes of student behavioral metrics that contribute to the optimization of an engineering design.

Methods

Curriculum Context: Virtual Engineering Internships

The study took place in the context of six middle school Virtual Engineering Internships (VEIs) that are part of an NGSS-designed K–8 science curriculum, Amplify Science (Regents of the University of California, 2017). The VEIs are a special kind of engineering design unit, in which students take on the role of interns for a fictional engineering firm, using a web-based workspace modeled after an email application, through which the teacher manages students' interactions with a "project manager" from the firm. In each VEI, students apply science ideas to design solutions to a different humanitarian challenge, set in a particular subfield of engineering (described in Table 1).

Each VEI consists of ten (45 min) lessons, composed of three main phases: the research, design, and proposal phases. During the research phase, students gather background information to develop their understanding of the context and science of the humanitarian challenge. During the design phase, students iteratively test design solutions. And during the proposal phase, students write a structured argument/proposal for their chosen optimal design solution. To iteratively design and test solutions, students use a digital design tool custom-built for the VEI. Each VEI's design tool allows students to simulate and test designs, and provides test results, which students can analyze to figure out the effects of different design choices. Student log files enable analysis of all interactions within the design tool. The curriculum elicits both process data (student log files) and product data (students submit preliminary and then final designs at the end of the middle and last design-focused lessons, respectively).

This paper will focus on our findings for one VEI, the Force and Motion VEI. We will highlight notable differences between the VEIs in the "Results" section. In the Force and Motion VEI, students address the unit's humanitarian challenge by designing airdropped disaster relief supply pods. In their design, students must make tradeoffs to minimize cost, minimize cargo damage, and maximize reusability of the pod itself. Students must determine (and argue for) their own definition of "optimal" within the tradeoffs afforded by the design tool. Students utilize the unit-specific digital design tool: SupplyDrop, which allows students to select the type and amount of padding to use for their supply pod design, and add-ons like parachutes or springs (Fig. 1, left). For each design test, SupplyDrop provides results that include physical parameters of the impact: impact force, pod mass, and impact velocity, as well as pod cost, cargo damage, and pod condition (Fig. 1, right).

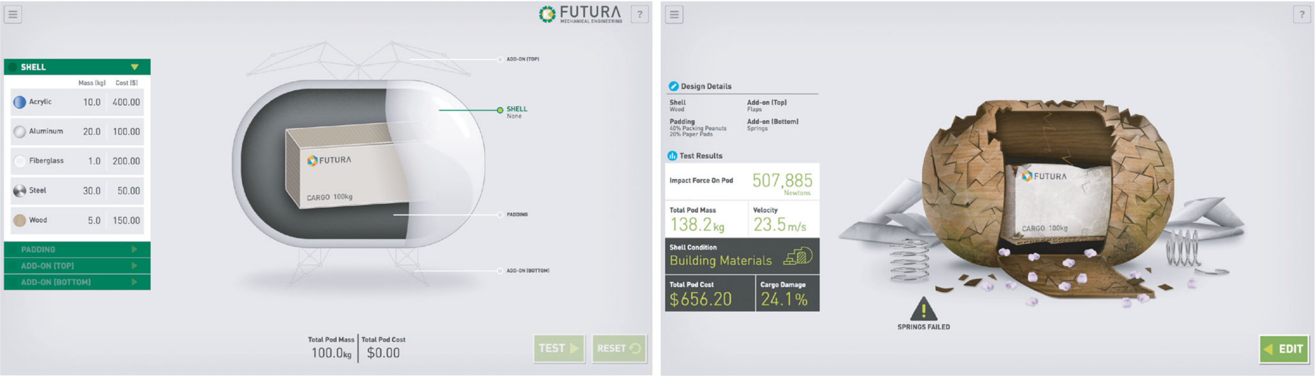
Research Trial Design and Participants

We conducted a series of research trials for each of the VEIs in 2018. Teachers were recruited through a call for participants within a broad national network of districts and schools. This resulted in a large pool of interested and eligible participants. From this large pool of potential teachers, we attempted to select research trial teachers that reflect the diversity of public schools in the USA, in terms of (1) teacher gender, (2) years of teaching experience, (3) teacher expertise in teaching engineering and design, (4) grade level (6th, 7th, or 8th grade), (5) proportion of student English Language Learners, and (6) proportion of students who qualify for free/reduced lunch. One notable exception, reflecting a bias in the larger pool of potential participant teachers, was comfort with technology—in which most participating teachers rated themselves as "very comfortable." Participating teachers were provided with the assistance resources available through the Amplify Science curriculum including training videos, step-by-step lesson guides, and help resources (phone, chat, email, FAQs, forums).

We performed a power analysis to estimate the number of research trial students necessary to detect a small effect size of 0.15. Looking for this effect size in a matched sample *t* test, with 5% false-positive and false-negative rates (alpha = 0.05, beta = 0.05), we needed approximately 580 students. Assuming a 10% teacher attrition rate and a 25% student attrition rate, we expanded this number to a target of 860 students per VEI. To recruit this many students, we advertised

Table 1 Science content areas, engineering subfields, and humanitarian challenges of the six Virtual Engineering Internships developed

VEI science content area	Engineering subfield	Humanitarian challenge
Force and Motion	Mechanical engineering	Designing delivery pods for emergency supplies
Phase Change	Chemical engineering	Designing portable baby incubators
Plate Motion	Geohazard engineering	Designing tsunami warning systems
Natural Selection	Biomedical engineering	Designing antibiotics courses to fight drug-resistant malaria
Earth's Changing Climate	Civil engineering	Designing rooftops for sustainable cities
Metabolism	Food engineering	Designing health bars for disaster relief



SHELL

	Mass (kg)	Cost ($)
Acrylic	10.0	400.00
Aluminum	20.0	100.00
Fiberglass	1.0	200.00
Steel	30.0	50.00
Wood	5.0	150.00

Total Pod Mass 100.0kg | Total Pod Cost $0.00

Design Details

Shell: Wood
Add-on (Top): Flaps
Padding: 40% Packing Peanuts, 20% Paper Pads
Add-on (Bottom): Springs

Test Results

Impact Force On Pod: 507,885 Newtons
Total Pod Mass: 138.2 kg
Velocity: 23.5 m/s
Shell Condition: Building Materials
Total Pod Cost: $656.20
Cargo Damage: 24.1 %

Fig. 1 Images of the Force and Motion VEI's design tool: SupplyDrop, when designing a supply pod (left) and viewing the results of testing a design (right)

for and selected teachers across the USA who could teach the VEIs to one hundred students. For the Force and Motion VEI, we began with 9 teachers and, through attrition, ended with 8 of those teachers successfully completing the research trial. This resulted in a potential pool of 918 students who completed the Force and Motion VEI research trial. Our final sample size was 731—the students from whom we received consent forms. A post hoc power analysis with this larger sample size yields a 98% chance of detecting an effect size of 0.15.

Research Questions and Procedures

The research questions we aimed to answer in this study were the following:

1. Can we discern a model of cognition of the engineering design practice of optimization?
2. What specific metrics, instantiated in our VEIs and tied explicitly to this model of cognition, will allow us to measure student engagement in optimization?
3. Do these metrics capture the variation within the student population?
4. What patterns do we see in student behavior, as expressed through these metrics?

The first of these research questions was motivated by Brown and Wilson's (2011) observation that very few assessments include an explicit model of cognition. Therefore, we began by defining the specific behaviors that seem to comprise the engineering design practice of optimization. We drew from multiple sources to help determine these component behaviors, including the NGSS description, feedback from a panel of experts, and insights from numerical techniques of optimization. We address research question no. 1 in the section "Defining Optimization Behaviors."

The second of these research questions addressed the need to explicitly define the evidence model that will connect student actions in the digital design tool to the component behaviors defined in our model of cognition. We address research question no. 2 in the section "Quantitative Metrics of Optimization Behaviors."

To address research questions 1 and 2, we drew on methods outlined by DeBarger et al. (2013), and Evidence Centered Design's (Mislevy et al. 2003) explicit statement of a student model (model of cognition) linked to an evidence model, to unpack the NGSS' description of optimization "into a coherent association of learning goals, describe the kinds of tasks and situations that would elicit those goals, and demonstrate how particular performances can be interpreted as evidence for students' capabilities" (DeBarger et al. 2013, p. 4).

Once these metrics had been defined, a further concern was that they would not capture substantial variation between students and, thus, would not be useful for identifying variation in students' facility with the practice of optimization. To address this concern, our third research question was aimed at identifying the amount of variation among students on each of the metrics. We address research question no. 3 in the sections "Combining Metric Values" and "Combining Metric Scores into Behaviors."

The fourth of these research questions explored correlations between these student behavior metrics. We address research question no. 4 in the section "Correlations among the Normalized Metric Scores."

Results

Defining Optimization Behaviors

The NGSS breaks down engineering design into three component ideas:

- Defining and Delimiting an Engineering Problem (ETS1.A)
- Developing Possible Solutions (ETS1.B)
- Optimizing the Design Solution (ETS1.C)

We decided to focus on the last of these: *Optimizing the Design Solution*, in part because we were confident that the student log file data would have the depth and variation we would need to investigate metrics for this construct. We also see it as a unique aspect of the VEI learning context, with the rapid test-revise-retest cycles possible within the design tool. In order to discern a model of cognition of the engineering design practice of optimization (research question no. 1), we developed a working/operationalized definition of *Optimizing the Design Solution*, breaking this practice down into a set of more specific behaviors that would count as evidence of student facility with optimization. To determine these specific, constituent behaviors we began with a close reading of the NGSS Appendix I:

> Grades 6–8. At the middle school level, students learn to sharpen the focus of problems by precisely specifying criteria and constraints of successful solutions, taking into account not only what needs the problem is intended to meet but also the larger context within which the problem is defined, including limits to possible solutions. Students can identify elements of different solutions and combine them to create new solutions. Students at this level are expected to use systematic methods to compare different solutions to see which best meet criteria and constraints, and to test and revise solutions a number of times in order to arrive at an optimal design.

From this, we extracted two initial behaviors that contribute to optimization: combining elements of different solutions (combination) and using systematic methods to test and revise solutions (systematicity). Based on the work of Chan et al. (2011) and an analysis of numerical optimization algorithms, we concluded that a third behavior (exploration) was important for the practice of optimization as well. We reason that an engineer must sufficiently explore the possible design space to truly optimize their design.

While attempting to more precisely define each of these behaviors, we had difficulty specifying what the combination feature would look like in our data. Looking at long-form written student responses, we noted that students often did a great deal of intentional combining as they worked on their designs—choosing to use specific materials in future designs to leverage/mitigate the properties of one material with the properties of another. But given the limited number of materials available to students in our design tools, we were unable to determine unambiguously from the digital design tool data that students were intentionally combining elements of different solutions and not just re-using one of the options randomly. Based on this as well as feedback from our panel of experts in engineering education that downplayed the importance of the combination feature, we decided to focus on the behaviors we could appreciably measure. Thus, we focused on two contributing, observable behaviors of optimization: exploration and systematicity.

We then compared and contrasted exploration and systematicity with constructs utilized by other researchers seeking to measure engineering design practices. For example, the constructs of Vieira et al. 2017 (generate ideas and conduct experiments) strongly mirror the exploration and systematicity behaviors. We found many similarities between our exploration construct and the metrics employed by other researchers. For instance, the work of Chan et al. (2011) and Shah et al. (2003) on ideation in engineering design, as well as the "breadth of search" metric utilized by Schauble et al. (1991), strongly mirrors our exploration construct.

The behaviors we chose to observe thus seem to be consistent with national standards, insights from a panel of engineering education experts, and previous published work. These behaviors provided the basis of a model of cognition of the engineering design practice of optimization, addressing our first research question (*Can we discern a model of cognition of the engineering design practice of Optimization?*).

Quantitative Metrics of Optimization Behaviors

In order to address research question no. 2, we next developed specific metrics to quantify student engagement with the practice of optimization, through the constituent behaviors of exploration and systematicity. Below, we define each of the behaviors as combinations of these metrics.

Exploration

Exploration of the available design features is important as a means to identify the possibilities and limitations of the design space. Without sufficient exploration, a student's ability to optimize is limited by their lack of understanding of the possibilities available to them. We chose to quantify exploration in several ways. First, we calculated the fraction of possible design modifications made, over the course of an investigation; e.g., they tried adding a parachute, and tried adding flaps, but did not try adding springs to the bottom of the pod (fraction_modified). Second, we calculated the fraction of the design (output) space that a student tested; e.g., they generated designs with high cost and low cargo damage, but never designed anything with low cost but higher cargo damage (fraction_explored). For this fraction_explored to be focused on the most relevant regions of the design space, we performed a k-means grouping on the output variables (cargo damage, shell condition, and cost) of the final designs of students using the commercially available VEIs, and reported the fraction of these groups that students had generated designs within. Third, we calculated the total/cumulative number of tests that each student performed (test_count).

The three exploration metrics described above are cumulative: they reflect all of the designs a student has generated, rather than just the designs made during a specific lesson. We treated exploration as cumulative because, although it is important to engage in, it is not something we think students must repeat anew each lesson—an area of the design space, once explored, remains explored on subsequent design sessions. And tests, once conducted, contribute to the student's understanding of the available design space, largely irrespective of whether those tests occurred during the current or the previous lesson's design time.

Systematicity

Given that there are more possibilities for how to design a solution than can reasonably be tested in the available time, students must make decisions about how to explore the possible designs in order to make progress towards an optimal solution. Being systematic allows students to gain an understanding of how certain types of design decisions will affect their test results, thus enabling design iterations to proceed more efficiently. We chose to quantify systematicity in several ways. First, we calculated the average number of categories of changes (variables_changed) made between consecutive tests (e.g., a student replaced all of their paper padding with metal foam padding and replaced their large parachute with a small parachute [2 categories of changes]). We expect that students exploring systematically will change fewer variables between consecutive tests. And second, we calculated the average distance (Euclidean norm) in the normalized design space between consecutive tests (jump_distance)—the distance between two consecutive tests' normalized results. We expect systematic, iterative exploration of the design space to yield smaller jumps in the design space, compared with wild exploration or random/unprincipled testing.

We calculated both of these systematicity metrics in a non-cumulative (lesson by lesson) manner—a student's behavior is either systematic or not, independent of their behavior in previous lessons.

It is important to note that there is a natural tension between the importance of exploration and the importance of systematicity. Exploring more widely in the available amount of time will often appear as less systematic testing, and vice versa. To navigate this tension and balance the benefits of both behaviors, we envision that students might begin with an initial period of more free exploration aimed at determining the range of possible solutions and generating ideas for promising areas, followed by a period of increased systematicity of testing aimed at honing an acceptable class of design solutions to the students' optimal/final solutions. The system of metric design that we utilize above—using cumulative metrics for the exploration behavior and non-cumulative metrics for the systematicity behavior—allows us to somewhat naturally address this tension. During a later lesson, if students engage in systematic testing behaviors, and if they have also explored the design space previously, then they will score high in both of these behaviors at that point in the VEI.

These metrics operationalized our model of cognition of the engineering design practice of optimization: composed of two behaviors (exploration and systematicity), which are each composed of 2–3 quantitative metrics that are measurable within our VEI design tool. This addressed our second research question (*What specific metrics, instantiated in our VEIs and tied explicitly to this model of cognition, will allow us to measure student engagement in Optimization?*).

Combining Metric Values

To make use of the various quantitative metrics described above, we needed to combine them into the optimization behaviors that they contributed to. Because a simple combination of these values would place undue importance on the largest values, we normalized them prior to combination. We normalized each metric's values by fitting the distribution of student metric values with either a log-normal or a beta distribution (whichever fitted better, as determined by a Kolmogorov-Smirnov test) and then used the corresponding cumulative distribution function (CDF) values. This allowed us to map any metric value to a monotonically increasing score between 0 (worst metric values) and 1 (best metric values) using the fitted distribution's CDF. This parameterization allowed us to effectively differentiate between the student metric values and allowed us to compare and combine scores on a normalized scale. In cases where low values of the metrics were considered optimal (the systematicity metrics), we simply used the survival function (one minus the distribution's CDF) as the normalized metric score instead of the CDF. These scores are shown in Figs. 2 and 3 as overlaid lines.

Figure 2 shows the distributions of student actions in the (cumulative) metrics corresponding to the exploration behavior by the end of the last design-focused lesson: fraction_modified, fraction_explored, and test_count. We see that almost all students have thoroughly explored the range of input elements by this point in time (fraction_modified), modifying (to some degree) all of the possible input options. In reference to our research question no. 3, this metric shows limited variation between students, and so is not particularly useful for differentiating students in this VEI. However, we retain it because, in other VEIs (that have more complex input elements), fraction_modified shows significant variation. Meanwhile, student exploration of the resulting categories of designs (fraction_explored) was much more varied, with some students coming into contact with nearly all categories of designs, other students coming into contact with very few

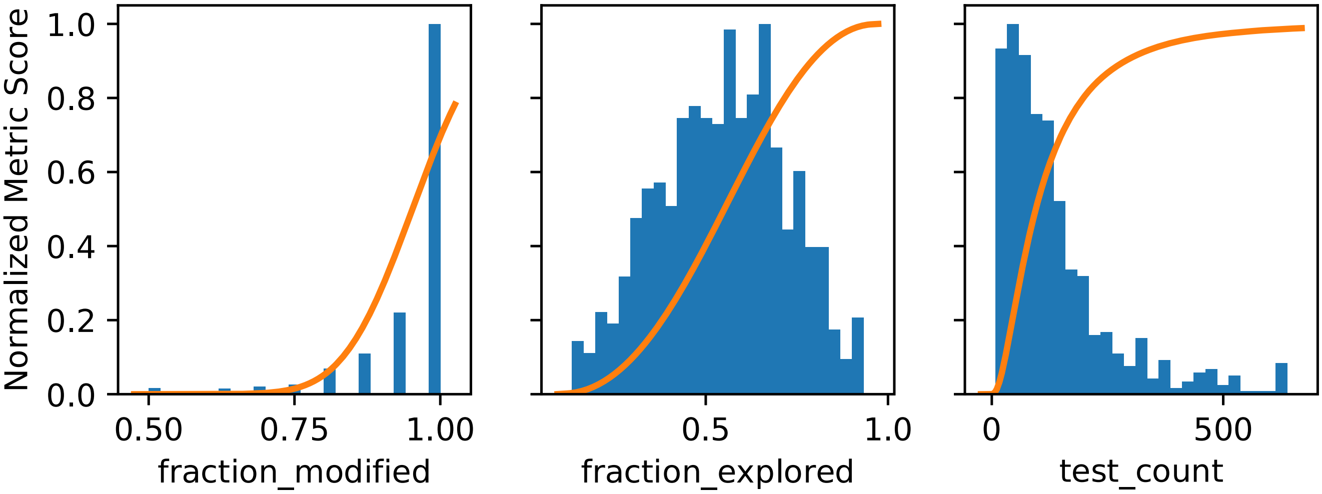


Fig. 2 Distribution of students' fraction of the input parameters modified (left), fraction of the design space (output parameter space) tested (center), and cumulative number of tests (right) by the end of the last design-focused lesson in the Force and Motion VEI's design tool: SupplyDrop. Normalized metric scores are shown by the fitted lines, with higher values of all metrics tending towards more complete exploration

categories of designs, and the majority of students coming into contact with just over half of the categories of designs. Looking at the last of these metrics, we see that most students have performed around 100 design tests by the end of the last design-focused lesson (test_count), with a few students having tested many more times than that.

Turning to the systematicity behavior of students during only (non-cumulatively) the last design-focused lesson, we show the student metric value distributions in Fig. 3: variables_changed and jump_distance. In variables_changed, we see that most students changed between one and three variables (on average) between tests—distributed around the common behavior of "undoing" a previous change in one variable and then changing another variable (two variables changed). Looking at jump_distance, students seem to demonstrate bimodality in how much their changes altered their resulting designs: one group of students made fairly small changes to their designs resulting in smaller "jumps" in the output space (peaking around 0.5), while another group of students made larger changes to their designs resulting in larger "jumps" in the output space (peaking around 1.3). Further, over the course of the VEI, students' average jump_distance values declined significantly. In the first design-focused lesson (not shown), the majority of students made larger changes and 51% of students were in the jump_distance ≈ 1.3 peak region. By the last design-focused lesson (shown in Fig. 3), the fraction of students who made larger changes had fallen to 34%, and the majority of students were instead in the jump_distance ≈ 0.5 peak region. This trend extended to other VEIs, but was not universal. For instance, in the Phase Change VEI, the distribution of student jump_distance scores was only weakly bimodal, and the movement to lower values over time was not statistically significant.

Combining Metric Scores into Behaviors

Each of these sets of metric scores was then combined (geometrically averaged together) to generate an exploration and

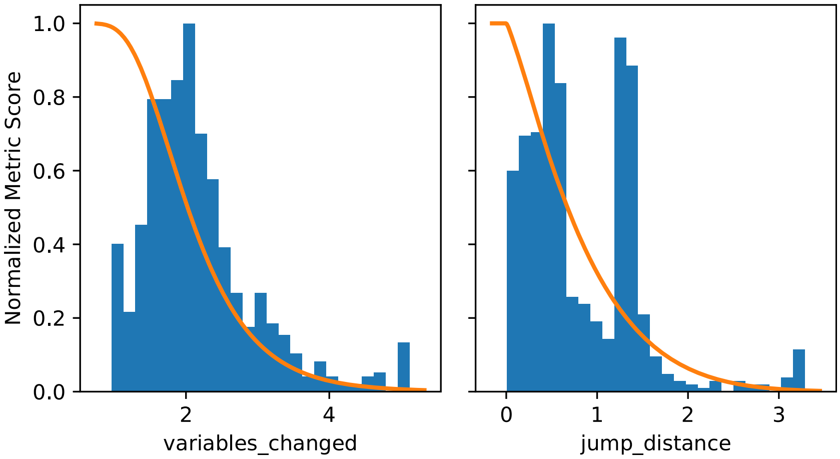


Fig. 3 Distribution of students' average number of variables changed between tests (left), and average distance between tests in the design space/output parameter space (right) during the last design-focused lesson in the Force and Motion VEI's design tool: SupplyDrop. Normalized metric scores are shown by the fitted lines, with fewer variables changed and smaller jumps between tests tending towards being more systematic

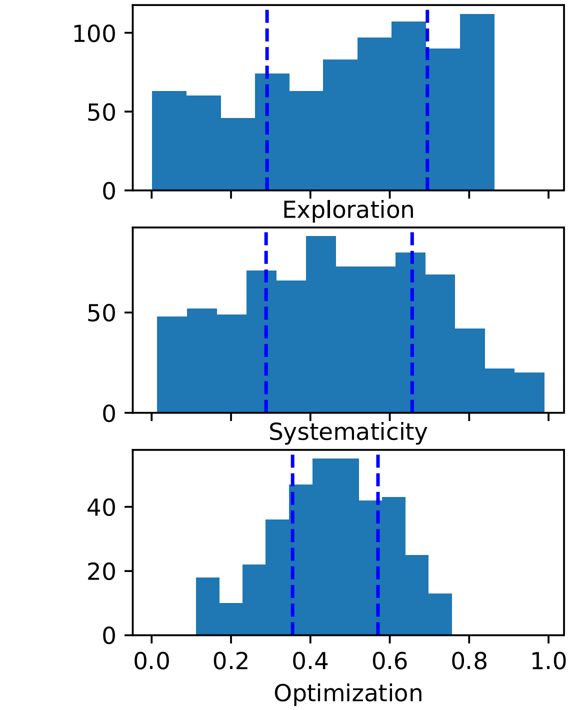


Fig. 4 Distributions of students' constituent behaviors (exploration and systematicity) and overall optimization scores during the last design-focused lesson in the Force and Motion VEI's design tool: SupplyDrop. Vertical lines mark the 25th and 75th percentiles of the distributions of student performance for each behavior

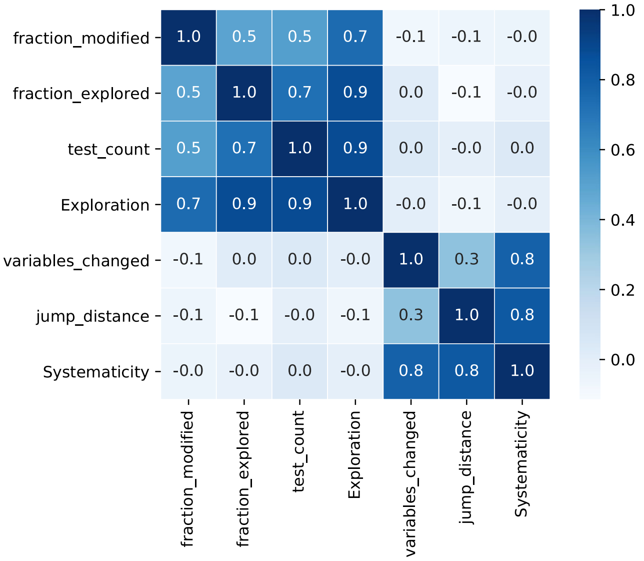


Fig. 5 Correlation matrix of the *normalized* metric and behavior scores. Note that higher normalized scores correspond to better performance on each of these metrics—corresponding to the fitted lines in Figs. 2 and 3

systematicity behavior score for each student. Figure 4 shows the resulting distributions of student behavior scores, along with the resulting combined student optimization scores. The overall optimization scores are a geometric average of the two behavior scores. Note that, for these calculations, we chose the geometric average (product-based) over the arithmetic average (sum-based) to capture our intuition that serious deficiency in any of the contributing metrics should trigger a lower combined behavior score, to better highlight room for improvement. We see that there is a significant spread in the distributions of each of the behavior scores, and the resulting optimization score distribution. This implies that these scores capture some aspects of the diversity of student performance, and so likely contain useful information for differentiating student performances, addressing research question no. 3 (*Do these metrics capture the variation within the student population?*).

Correlations among the Normalized Metric Scores

These normalized metric scores allowed us to more easily identify patterns among the metrics, addressing research question no. 4 (*What patterns do we see in student behavior, as expressed through these metrics?*). We first looked for relationships among these normalized metric scores and behaviors by constructing a correlation matrix (Fig. 5).

Looking at the correlation matrix, we see several significant positive correlations. For example, we see the (mathematically necessary) high positive correlations between the metric scores and the behaviors that they contribute directly to (e.g., jump_distance and systematicity). We see moderate-high correlation between the three exploration-related metric scores: fraction_modified, fraction_explored, and test_count. These exploration metrics show the understandable correlation that students who conducted many tests were more likely to vary more of the possible inputs ("turn all the knobs") and also to explore more of the possible output space. We also see a moderate correlation between the two systematicity metric scores: variables_changed and jump_distance. However, beyond these expected correlations, we note the independence of the metric scores within the exploration and systematicity behaviors. We address these independences in the "Discussion" section.

Discussion

In order to assess the performance of a practice, it is critical to clearly define the behaviors that represent mastery of that practice (Mislevy et al. 2003). Accordingly, while "optimization" is broadly understood to be a core practice in engineering design (Mehalik and Schunn 2007; Crismond and Adams 2012; NRC 2012; NGSS Lead States 2013), monitoring and supporting student development of this practice require a concrete picture of what student action optimization entails. This challenge is magnified with the practice of optimization in particular in that one must not only know what to look for, but also when and how to look for it—from an assessment perspective, optimization is less an observable feature of a particular design in isolation, and more a feature of the changes between design iterations, and in relation to the criteria and constraints of the design problem. Our study advances the field's capacity to meet this challenge by presenting an approach to operationalizing optimization precisely in terms of the changes between iterations and within the context of the criteria and constraints of the design problem. Moreover, through applying this assessment evidence model to a rich student data set, we have been able to shed light on how observable sub-components of optimization, such as exploration and systematicity, plays out across a series of student design iterations.

Our exploration metrics are similar to those of Shah et al. (2003), who defined ideation within engineering design as composed of four metrics: *novelty*, *variety*, *quantity*, and *quality*. However, we struggled to determine any satisfactory metrics that correspond to *quality* within our VEIs, being limited by ambiguity about how to define a single measure of *quality* from a multidimensional optimization problem. Any design change entails a set of tradeoffs between the different priorities (e.g., cost versus performance) which one student may consider an improved design but another student might consider a worse design. Past work has measured the overall *quality* of a multidimensional design solution through expert determined holistic ratings (Kruger and Cross 2006), but the variance of such an assessment is substantial and such an approach is difficult to apply broadly. In cases where the multidimensional nature of optimization can be reduced to a unidimensional problem (plus design constraints), the *quality* of a design solution can be more easily defined. For example, in their study of the quality of students' designs of earthquake-resistant structures, Apedoe and Schunn (2013) were able to unambiguously quantify the *quality* of students' design solutions as the achieved height of a structure, constrained by the requirement that the structure did not collapse in a test earthquake. In cases where such a unidimensional reduction is intractable, the *quality* of a design remains difficult to objectively assess.

Despite the difficulty to apply Apedoe and Schunn's (2013) definition of quality to more multidimensional problems, their work gives us insight into which metrics are helpful to improvement in students' design work, and which are less relevant. Apedoe and Schunn (2013) found that some metrics are significantly related to success in a design project (metrics similar to our test_count and fraction_explored), but that other metrics are uncorrelated with student success (metrics akin to our variables_changed and fraction_modified). It is difficult to compare our results with these findings, since our design challenges lack a unique definition of success, and a design solution that is optimal for one student will likely be suboptimal for another student given their different prioritizations of the design criteria. Utilizing a generic (priority agnostic) measure for design quality, we see no significant correlations between any of our metric scores and this measure of quality/success. We retain these additional metrics in our analysis, despite Apedoe and Schunn's (2013) findings that they are uncorrelated with student success, because they play an important role in the scientific mode of experimentation and may be of interest to teachers and students within the science classroom.

Our metrics give substantial insight into behavioral patterns among students in our VEI experiences. Among the correlations shown in Fig. 5, we see some interesting lack of correlation. We also see the independence between behavior groups, which might suggest that the behaviors we have delineated (systematicity, exploration) and their contributing metric scores are relatively independent skills from each other. This is somewhat surprising, as one might imagine that students skilled in one of these behaviors might also be preferentially skilled at employing the other. We hypothesize that middle school students have so little experience with optimization that such a correlation has not yet developed. Further study of individuals with intermediate and advanced experience in engineering design would be useful to determine if this proposed explanation is accurate.

There is very little previous work investigating the relationship between exploration and systematicity. The natural tension between these two optimization behaviors, described in the "Quantitative Metrics of Optimization Behaviors" section, is particularly important to consider in assessments that encompass both of these aspects of engineering design. This tension must be addressed in some way in order to properly assess student performance. In this work, we have mitigated this tension by treating exploration as being composed of cumulative metrics, while treating systematicity as being composed of non-cumulative metrics. This allows exploration that occurred earlier in the VEI to still "count" on subsequent days of the investigation, allowing students to transition to a more systematic style of optimization later in their investigation. But this initial work invites future studies to delve deeper into how development of the sub-constructs may best be supported in students and, importantly, *when* in the design process to expect exploratory and/or systematic behaviors. Other approaches that reconcile the necessarily serial nature of these two behaviors that contribute to optimization should be devised.

We would also expect some significant relationships to exist between the metric scores. For example, looking at the metric scores that focus specifically on the input parameters within the correlation matrix, it is somewhat surprising that the correlation between variables_changed and fraction_modified is not more negative. We might expect that changing more variables between tests (larger variables_changed metric values, renormalized to lower variables_changed scores) would lead to a larger fraction of the input parameters modified (higher fraction_modified scores). Instead, we find that, among the students who are changing many variables between tests, many of those changes involve changing the same, limited set of variables between tests. Similarly, looking at the metric scores that focus on the results/outputs of design tests, we would expect that larger shifts between tested designs (larger jump_distance metric values, renormalized to lower jump_distance scores) would lead to a larger fraction of the output/design space explored (larger fraction_explored scores). We see some evidence of this, as this pair of metric scores has a small negative correlation, but it is somewhat surprising that this correlation is not more negative. We find that, among the students who make more drastic design changes to their designs between tests, there is a significant group that are consistently making jumps to previously explored designs (previously explored regions of the output parameter space) and failing to fully explore the design space.

Future Directions

Utilizing and Extending This Work

To leverage this methodology and extend the results presented here, researchers could measure the above proposed metrics to compare their observed student performances and cross-metric correlations in relation to those presented here, highlighting relevant differences in their engineering design experiences. Researchers could also propose additional metrics (either specific to their educational situations or more general) and check for correlations between those metrics and the metrics presented here. In particular, developing objective and automatically calculable metrics for the quality of a multidimensional design solution would be highly valuable.

Teachers could leverage these results by making time in their engineering/design lesson-plans to provide students with explicit structures and learning experiences focused on useful strategies highlighted by the model of cognition presented here, including exploring the design space and systematic techniques for goal-oriented improvement of a design.

Product developers could leverage these results by making sure that, as they develop digital engineering design experiences, they design their log files to allow easy capture/analysis of student behaviors similar to those highlighted by the behaviors and metrics presented in this study.

Task Design

We note that giving students explicit freedom to individually define what type of design they think is optimal is an important feature of our units; however, diagnosing student facility with the practice of optimization is significantly complicated by the ambiguity about what a student defines as their optimal design. Therefore, having more explicit and timely information about what type of design students are aiming towards would be valuable for quantifying student success in the practices of engineering design and optimization. This kind of information could be gathered through an in-app priority setting exercise, done periodically throughout the design experimentation phase, in which students rate or prioritize the design criteria as they submit their designs. Another potential source of this student-specific goal information (for research purposes, as opposed to teacher formative assessment) could come from student think-aloud interviews as they work within the design tools.

The development of these metrics and the underlying access to digital clickstream data has given us significant insight into student engagement with the engineering design practice of optimization. There might be differences in the behaviors incentivized in different types of engineering experiences. For instance, one could imagine that in design spaces where the number of input variables is small and those variables only have a few options each, then the best way to explore the options is simply to test all of the possible permutations. Conversely, in design spaces where complete exploration is obviously impossible, students will need to limit the amount of exploration they pursue. These two modes of exploration might benefit from very different types of systematic behavior. In this vein, future work could focus on the features of the VEI experiences (such as volume of the design space) that affect student engagement with optimization behaviors.

Validation

Additional validation work on these metrics and behaviors is needed to further clarify the relationships between them and to identify and explore any additional factors not explored here. For instance, initial discussions with our panel of engineering education experts led us to investigate an additional contributing behavior beyond exploration and systematicity: student responsiveness to feedback. Future work will explore this and other potential additional contributing behaviors of optimization.

We are in the process of obtaining a set of expert practitioner ratings of optimization for sets of student behavior, so that we can determine the level of correlation with the metrics and behavior scores that we have discussed above. This will help us build a stronger validity argument for the metrics that correlate most strongly. Of equal importance, this process of obtaining expert ratings, and discussing the reasons for those ratings, will allow us to surface important implicit justifications that experts are applying, and lead us to a stronger set of operationalized metrics.

Beyond the metrics presented and explored here, there is important work that needs to be done to explore additional metrics' contributions to optimization. These explorations could proceed from a theoretical foundation as pursued in this work, or from an empirical foundation based on discovery of predictively powerful patterns in the data. Ideally, future work would combine these two approaches, leveraging theoretical foundations to construct information-dense features and then leveraging machine learning algorithms to search for patterns among the combination of the raw data and these information-dense features for unintuitive but predictively powerful new combinations of features.

Funding Information This research is based upon work supported by the National Science Foundation under grant no. 1417939.

Compliance with Ethical Standards

Conflict of Interest Samuel Crane provided data and assisted in the development of the digital tools used in this study, as part of his role as Director of Data Science at Amplify Education.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Informed consent and parental consent were obtained from all individual participants included in the study.

Disclaimer Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Apedoe, X., & Schunn, C. D. (2013). Strategies for success: uncovering what makes students successful in design and learning. *Instructional Science, 41*(4), 773–791.

Baxter, G. P., Elder, A. D., & Glaser, R. (1996). Knowledge-based cognition and performance assessment in the science classroom. *Educational Psychologist, 31*(2), 133–140.

Bennett, R. E., Persky, H., Weiss, A., & Jenkins, F. (2010). Measuring problem solving with technology: a demonstration study for NAEP. *The Journal of Technology, Learning and Assessment, 8*(8).

Black, P., & Wiliam, D. (2010). Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan, 92*(1), 81–90.

Brown, N. J. S., & Wilson, M. (2011). A model of cognition: the missing cornerstone of assessment. *Educational Psychological Review, 23*(2), 221–234.

Chan, J., Fu, K., Schunn, C. D., Cagan, J., Wood, K., & Kotovsky, K. (2011). On the benefits and pitfalls of analogies for innovative design: ideation performance based on analogical distance, commonness, and modality of examples. *Journal of Mechanical Design, 133* 081004-1-11.

Crismond, D. P., & Adams, R. S. (2012). The informed design teaching and learning matrix. *Journal of Engineering Education, 101*(4), 738–797.

DeBarger, A. H., Penuel, W. R., & Harris, C. J. (2013). Designing NGSS assessments to evaluate the efficacy of curriculum interventions. In Invitational Research Symposium on Science Assessment, Washington, DC. http://www.k12center.org/rsc/pdf/debarger-penuel-harris.pdf. Accessed 15 Aug 2019.

Doppelt, Y., Mehalik, M. M., Schunn, C. D., Silk, E., & Krysinski, D. (2008). Engagement and achievements: a case study of design-based learning in a science context. *Journal of Technology Education, 19*(2), 22–39.

Dubberly, H. (2004). How do you design? A compendium of models. http://www.dubberly.com/wp-content/uploads/2008/06/ddo_designprocess.pdf. .

Gobert, J., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics for science inquiry using educational data mining. *The Journal of the Learning Sciences, 22*(4), 521–563.

Hammond, K. J. (1989). *Case-based planning: Viewing planning as a memory task*. Academic Press.

Kolodner, J. (1993). *Case-based reasoning*. San Mateo: Morgan Kaufmann Publishers.

Kolodner, J. L., Camp, P. J., Crismond, D., Fasse, B., Gray, J., Holbrook, J., Puntambekar, S., & Ryan, M. (2003). Problem-based learning meets case-based reasoning in the middle-school science classroom: putting learning by design (tm) into practice. *The Journal of the Learning Sciences, 12*(4), 495–547.

Kruger, C., & Cross, N. (2006). Solution driven versus problem driven design: strategies and outcomes. *Design Studies, 27*(5), 527–548.

Kuo, C. Y., Wu, H. K., Jen, T. H., & Hsu, Y. S. (2015). Development and validation of a multimedia-based assessment of scientific inquiry abilities. *International Journal of Science Education, 37*(14), 2326–2357.

Marples, D. L. (1961). The decisions of engineering design. *IRE Transactions on Engineering Management, 2*, 55–71.

Mehalik, M., & Schunn, C. (2007). What constitutes good design? A review of empirical studies of design processes. *International Journal of Engineering Education, 22*(3), 519.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–62.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (2014). *Developing assessments for the next generation science standards*. National Academies Press.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59–81.

Purzer, Ş., Goldstein, M. H., Adams, R. S., Xie, C., & Nourian, S. (2015). An exploratory study of informed engineering design behaviors associated with scientific explanations. *International Journal of STEM Education, 2*(1), 9.

Quellmalz, E. S., & Pellegrino, J. W. (2009). Technology and testing. *Science, 323*(5910), 75–79.

Quellmalz, E. S., Timms, M. J., Silberglitt, M. D., & Buckley, B. C. (2012). Science assessments for all: integrating science simulations into balanced state science assessment systems. *Journal of Research in Science Teaching, 49*(3), 363–393.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching, 33*(6), 569–600.

Schank, R. C. (1982). *Dynamic memory: A theory of reminding and learning in computers and people*. New York: Cambridge University Press.

Schank, R. C. (1999). *Dynamic memory revisited*. New York: Cambridge University Press.

Schauble, L., Klopfer, L. E., & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching, 28*(9), 859–882.

Serrano-Laguna, Á., Torrente, J., Moreno-Ger, P., & Fernández-Manjón, B. (2012). Tracing a little for big improvements: application of learning analytics and videogames for student assessment. *Procedia Computer Science, 15*, 203–209.

Shah, J. J., Vargas-Hernandez, N., & Smith, S. M. (2003). Metrics for measuring ideation effectiveness. *Design Studies, 24*(2), 111–134. https://doi.org/10.1016/S0142-694X(02)00034-0.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education, 4*(4), 347–362.

Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most informative data for feedback generation: learning analytics in a data-rich context. *Computers in Human Behavior, 47*, 157–167.

Vieira, C., Goldstein, M. H., Purzer, Ş., & Magana, A. J. (2016). Using learning analytics to characterize student experimentation strategies in the context of engineering design. *Journal of Learning Analytics, 3*(3), 291–317.

Vieira, C., Magana, A. J., & Purzer, S. (2017). Identifying engineering students' design practices using process data. In Proceedings of 2017 research in engineering education symposium (REES). Bogotá-Colombia.

Wertheim, J., Osborne, J., Quinn, H., Pecheone, R., Schultz, S., Holthuis, N., & Martin, P. (2016). An analysis of existing science assessments and the implications for developing assessment tasks for the NGSS. https://snapgse.stanford.edu/sites/default/files/snap_landscape_analysis_of_assessments_for_ngss_0.pdf. Accessed 15 Aug 2019.

Xie, C., Zhang, Z., Nourian, S., Pallant, A., & Hazzard, E. (2014). Time series analysis method for assessing engineering design processes using a CAD tool. *International Journal of Engineering Education, 30*, 218–230.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.