Designing Data-informed Intelligent Systems to Create Positive Impact
Design Methods, Questions and Recommendations

Lomas, J.D.; Patel, Nirmal; Forlizzi, Jodi L.

Publication date
2021

Document Version
Final published version

Published in
Proceedings of Relating Systems Thinking and Design (RSD10) Symposium

Citation (APA)
Lomas, J. D., Patel, N., & Forlizzi, J. L. (2021). Designing Data-informed Intelligent Systems to Create Positive Impact: Design Methods, Questions and Recommendations. In M. van der Bijl--Brouwer (Ed.), *Proceedings of Relating Systems Thinking and Design (RSD10) Symposium* (pp. 154-170) https://rsdsymposium.org/towards-data-informed-system-design-for-good-methods-questions-and-recommendations-for-designers/

Designing Data-Informed Intelligent Systems to Create Positive Impact

Design Methods, Questions and Recommendations

J Derek Lomas, Nirmal Patel, Jodi Forlizzi

This paper explores several approaches for designing data-informed intelligent systems to create positive impact. Two contrasting case studies in K12 education are used to illustrate design methods, questions and recommendations. The first case study addresses the poverty achievement gap in America, and shows how product data can be used to identify areas of inequity in digital education. The second case study looks at the unintended consequences of automating data-driven optimization in the context of a digital math game. Together, the two case studies reveal generalizable knowledge that supports the design of intelligent feedback loops to create positive impact. Further, this paper considers both the benefits and limitations of data feedback in complex social-technical systems.

Keywords: DesignX, Intelligent Systems, Smart Systems, Learning Organizations, Cybernetics, Poverty, Education, K12, Artificial Intelligence (AI), Goodhart's Law, Games

Introduction

A Human-Centered Design Perspective on Artificial Intelligence and the Design of Intelligent Systems

Artificial intelligence (AI) has done a marvellous job mastering games like Chess and Go. However, there is a gap in our understanding of how to apply AI for addressing large-scale societal issues such as education. New hardware, better data sources and cutting-edge algorithms are seen as the pathway to the production of more advanced AI systems in education, as this formula has worked in other domains. However, we should be careful — the societal goal is not to create improved educational AI—instead, the goal is to improve education itself (e.g., to enhance the potential for success vis-a-vis students, teachers, administrators and education systems). More advanced AI algorithms do not necessarily lead to better outcomes.

There is a danger in viewing AI as an "algorithm in a box" that can be bought, sold and integrated into an existing system to produce improvements. This leads to a technology-centered solutionism generating hype and deflation. A more nuanced view involves treating AI as a form of distributed intelligence. To explain using an analogy, human intelligence does not just reside in the brain. Instead, human cognition is distributed across the body, across our tools and our social engagements [0].Similarly, AI can be viewed as much more than a disembodied algorithm. Meaningful AI systems should be understood as a distributed intelligence, which necessarily involves people, artifacts, data systems and algorithms.

With these points in mind, we seek to shift perspective from the "Design of Artificial Intelligence in Education" to the "Design of Intelligent Systems in Education". Intelligence can be defined pragmatically through the idea of *success intelligence*, which Robert Sternberg (the most cited authority on human intelligence) defines as "one's ability to achieve success in life in terms of one's personal standards, within one's socio-cultural context"; he further notes that success is to be understood as "a state of well-being within one's cultural context" [0]. If we

extend this notion of human intelligence to systems, then more intelligent systems should necessarily increase the likelihood of success and (stakeholder) wellbeing. Otherwise, as Forrest Gump put it, "Stupid is as Stupid Does" [14]. An expensive "Smart Classroom" that does not improve outcomes is, ultimately, stupid.

Sternberg's idea of success intelligence aligns with the dominant definitions of intelligence in AI, such as the one from Peter Norvig, Google's Director of Research, "the ability to select an action that is expected to maximize a performance measure." [15] This definition has provoked much of the thinking in this paper about the role of feedback loops in helping systems to measure their own performance and to maximize performance measures. In this regard, while a previous paper focused on the question, "How can we translate human values into metrics usable by AI systems?" [14], the present paper addresses data feedback loops in social systems.

Notably, Norvig's definition of intelligence does not entail the use of computers. Although "smart learning environments" typically focus on maximizing exposure to computers [0]), computers by themselves do not make a classroom smart. Instead, at least according to Norvig's definition, systemic intelligence emerges from actions that improve performance measures; what is implicit in his definition is the fact that many intelligent systems (computational and non-computational) use performance measures in a feedback loop, in order to determine their choice of a successful action.

From this perspective, schools and other organizations are filled with non-computational intelligent feedback loops, which use performance measures to inform actions (e.g., see our discussion on formative assessments and mastery learning, below). In this paper, we highlight the role of *data feedback loops* as a key target for designers in enhancing system intelligence, as well as the importance of aligning performance measures to human values. As such, this paper aims to shift conversations from "designing AI in education" (which implies the use of cutting-edge algorithms) to "designing intelligent systems in education" (in order to focus on the efficacy of data feedback loops). Going beyond the domain of education, this paper seeks to offer a response to the following design research question: "How *should* we design data feedback loops in order to improve meaningful outcomes in complex sociotechnical systems?"

Incorporating Cybernetics and Systems Design

This paper, which addresses the Research in Systems Design (RSD) community, does not focus on AI. However, it does broadly deal with cybernetics, which is the conceptual predecessor to AI (Fig. 1). The field of cybernetics is dedicated to the study of feedback loops (circularity) as well as the use of data to achieve goals (teleology) [0]. The word "cybernetics" itself comes from the Greek *kybernētiké*, or governance.

Fig. 1: Representation of cybernetics as the conceptual predecessor of AI

A cybernetic system involves five parts: sensors, actuators, a goal state, a controller (determining how the actuators should respond to differences between the sensor data and the goal state) and the environment itself (which is treated as a causal part of the system [0]). A simple example is a thermostat, which senses the temperature of a room, compares it to the target temperature, and either turns a heater on (if the temperature is too low) or turns it off (if the temperature is too high). A complex example of a cybernetic system is a designer with a pencil [0]; when designers produce certain sketching strokes on paper, they sense and evaluate them with respect to an internal goal, and then act upon the environment again, in order to shift their sketch to a more preferred state.

Artificial Intelligence, in its very name, implies that, an AI system should refer only to the artificial parts of the system — that is, the parts that do not involve human intelligence. In contrast, a cybernetic viewpoint is useful, because it provides designers [14] with a valuable way to view intelligent systems as a whole; that is, where the system can include humans, computers, artifacts and ecologies. This view is broad, but it also focuses on the

effective governance of systems through data feedback loops. Therefore, we find it useful to invoke cybernetics as a starting point for designing intelligent systems in education and other complex sociotechnical domains.

Mastery Learning as a Cybernetic Loop

Why do students fail in school? One reason is that many educational systems are intentionally designed to ensure that a certain percentage fails (e.g., the lowest 20%). However, most schools do not wish to "weed out the weak"; and in this case, student failure is a failure on the part of the school. How, then, might schools help more students to succeed? Numerous researchers believe that all students can succeed in school, but not within the same period of time: some students will simply require more time and attention than others [0]. In order to investigate this view, educational researcher Benjamin Bloom described a set of controlled experiments [0], which compared conventional classroom instruction to classroom-based *mastery learning*[1] or one-on-one tutored mastery learning. Over 90% of the students receiving personal tutoring and over 70% of the students in the mastery learning class reached a level of performance, which could be attained only by a mere 20% of students in the conventional classroom. How was this achieved? To put it briefly, Mastery Learning uses data feedback loops for informing instruction.

Mastery Learning is based on the use of formative assessments to inform decisions about investing time and effort in the classroom. For instance, after providing an instructional activity in class, a teacher assigns students a quiz, to assess whether the instruction was successful. If the assessment data shows variation in student performance, it can help teachers to understand which students or learning objectives need greater attention. Mastery Learning uses formative assessments to govern decisions about when a student can move to another topic: students are expected to continue to receive additional instruction on each topic, until they succeed in the formative assessment. This, in effect, creates a simple cybernetic loop. For a comparison, see the mastery instruction diagram, which uses the same format as the 1960s TOTE (Test-Operate-Test-Exit) cybernetic system from Miller et al. [0].

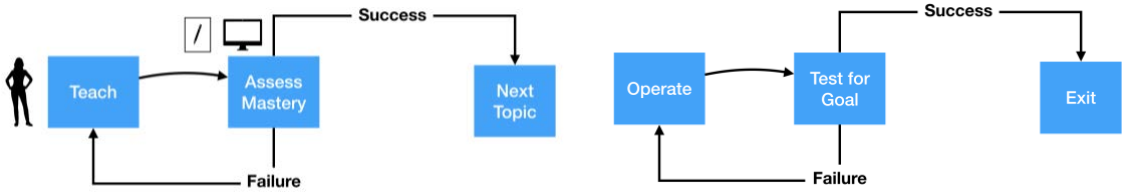

Fig. 2: Mastery Learning as a Cybernetic System: the left part of this diagram shows the TOTE algorithm [0] and the right shows the parallel logic of idealized mastery instruction in a classroom

Mastery Learning is not the only form of cybernetic feedback in education. "Data-Driven Decision Making" involves using formative assessment data to identify student problems and to act upon those problems systematically [0], but it does not require a full commitment to students achieving mastery in each subject. Cybernetic feedback loops can extend beyond individual classrooms, as well, when digital data from formative assessments are aggregated across teachers, in order to provide school administrators with continuous insight (e.g., regarding student or classroom level needs). These data can indicate groups of students or teachers who need additional help, or, it can identify particular learning objectives that create general challenges.

[1] "Teaching under Mastery Learning and under Conventional Instruction is much the same except for the Mastery Learning feedback-corrective process every 2 or 3 weeks in which a formative test is given to students, followed by corrective instruction, and then by a parallel formative test. The first step in the feedback-corrective process typically begins with the teacher's noting the common errors of the majority of the students. Then, the teacher briefly explains the ideas involved using different illustrations or an approach different from what was previously used in teaching these ideas in the class. A second step in this feedback-corrective process is for groups of two or three students (on their own or under the teacher's guidance) to help each other on the items they missed on the test. A third step is for individual students to refer to the instructional material keyed to the test items that they are not confident they fully understand. This three-step process is expected to be used after each 2- or 3-week learning unit, before the students take the parallel formative test" [0].

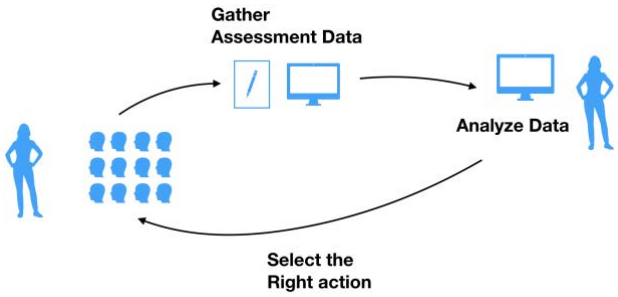

Fig. 3: A model for "Data-Driven Decision-Making" in K12 education [0]: 1. gather assessment data, 2. analyze data to identify problems and their causes, 3. select actions to address those problems

A continuous improvement loop is a data feedback loop, which produces system modifications in response to the measurements of needs. Continuous improvement loops are simple (Fig. 4) and offer a simple heuristic that can guide system designers: *anything that makes it easier to measure or modify a system will facilitate continuous improvement*. The barriers to making changes to a system can be very great – and it can be very painful to try to extract the right data in a timely manner. When it becomes easier to collect outcome data or to make system changes, data is likely to play a bigger role in system improvements.

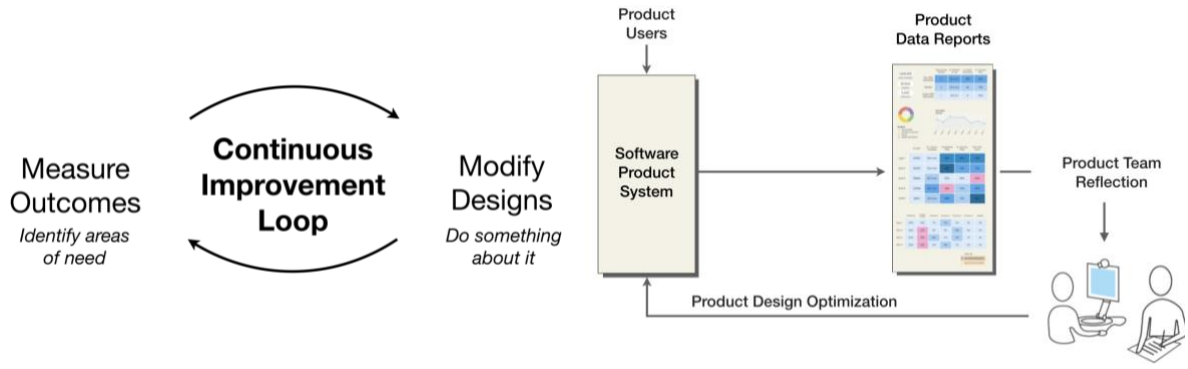

Fig. 4: Left: The basic continuous improvement loop model involves system measurement to gather outcome data about system operation, and the modification of the system with the aim of improving the key outcome measures. Right: A continuous software improvement loop

Case Study #1: Data-Informed System Design for Addressing the Poverty Achievement Gap

How can data-informed designs help produce large-scale social impact? This case study describes the use of digital learning product data, in order to prioritize product improvement goals. It specifically asks: how can we address the poverty achievement gap with data-driven incremental improvements, in a widely used digital learning program? This study has been produced by one of the authors [0].

The issue of childhood poverty in America is immense. Of the world's richest 35 countries, the United States has the second highest rate of childhood poverty, surpassed only by Romania [0]. Nearly 1 in 5 children grow up below the poverty line, and 24% of all schools in America are considered high-poverty schools [0]. Poverty has an enormous effect on academic achievement; this effect or "the poverty achievement gap" is often defined as the average achievement difference between children from families in the bottom 10% of incomes and children from families in the top 90% of incomes [0]. While poverty and race issues intersect, the poverty achievement gap is roughly twice as large as the achievement gap between white and black students (1.2 standard deviations vs .65 standard deviations). K12 performance affects subsequent educational attainment: only 18% of students in high-poverty schools will complete college within 6 years, while this number is 52% for low-poverty schools [0]. Educational attainment affects future income: as full-time workers without a high school diploma have a median income of $25,636 per year (and an 8% unemployment rate) while workers with a bachelor's degree have a

median income of $59,124 (and a 2.8% unemployment rate) [0]. Thus, poverty is perpetuated: high-poverty students are less likely to succeed in education, which in turn lowers their future income, and escalates chances that their children will suffer the negative effects of poverty, as well.

These issues are complex, and there is no technological "silver bullet" for creating a more equitable society. Political organization is critical to creating social equity. Amidst this reality, there may well be ways to design systems, which can help contribute modestly to improvements. Data-informed systems designs are particularly promising. In a 2017 review of 196 randomized field experiments, several educational interventions failed to improve the poverty achievement gap; yet, Data-Driven Decision Making was one of the most effective approaches [0]. Further studies have found the approach to be more cost-effective, compared to 21 other interventions [0].

Data-informed system design is well-suited to the nature of complex sociotechnical systems. Design theorists Don Norman and PJ Stappers [0] claim that incrementalist approaches are often more successful in addressing human needs within complex sociotechnical systems. In contrast to ambitious or radical reform, which tends to be expensive and failure-prone, data-informed system design entails the continuous implementation of a large number of small incremental changes.

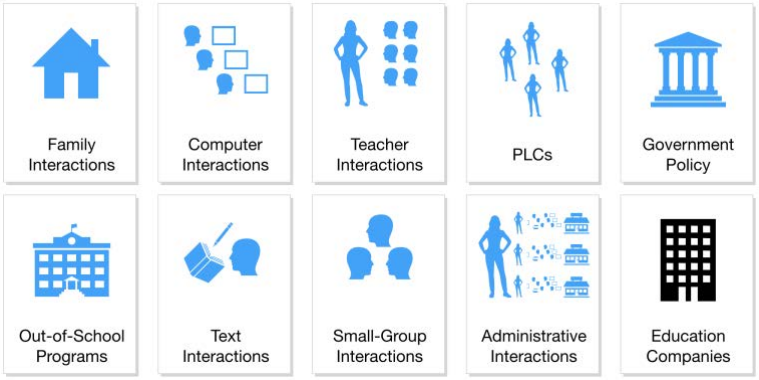

Fig. 5: K12 education is a large, complex, semi-hierarchical sociotechnical system. The icons above represent major components of K12 education, from the out-of-school student experience on the left, to the various kinds of in-school student interactions, to the higher-level interactions of PLCs (teacher groups, often known as "Professional Learning Communities"), school administration and government policy. The focus in this paper is on the role of digital software in K12 education, wherein educational companies have the potential to play a major role in nationwide data-driven improvements.

The Big Impact of Small Improvements on a Large Scale

A world of small, incremental changes may be less satisfying for the ambitions of those who want to "transform" education, but smaller improvements on a larger scale can still produce a major cumulative impact. In order to explore this, we can model the monetary value of new interventions, which improve academic achievement in high-poverty schools.

The total cost of childhood poverty in the USA is believed to be approximately $500 billion per year, as of 2008 [0]. If increased academic achievement enables children to escape a cycle of poverty, then an intervention that successfully cuts the poverty achievement gap in half would be worth $250 billion per year. Improvements in digital curriculum are extremely unlikely to raise student outcomes by 50%—but, they might be able to deliver a smaller improvement to millions of students. Improving the performance of all high-poverty students by 5 percentile points might be worth as much as $25 billion per year. This only goes to show that even small improvements on a large scale could create a big difference.

Fig. 6 shows the effect of poverty (measured in terms of the percentage of students in a school qualifying for a free/reduced-price lunch) on school performance, in a typical online learning system. The higher the poverty level in the school, the lower the average student performance on formative assessments. Reducing the poverty achievement gap means lowering the correlation between school poverty and school performance.

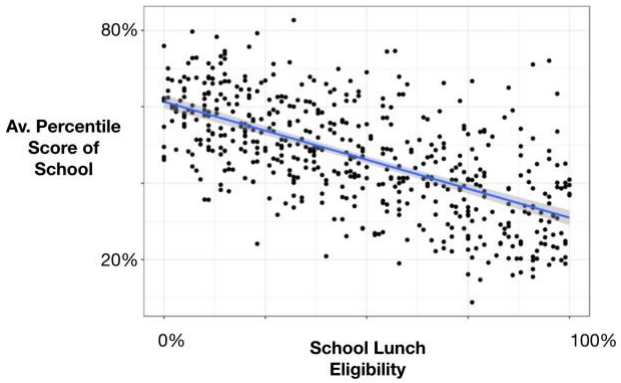

Fig. 6: The data presented here reflect real, observed product data, but are abstracted so as to create anonymity for the products and companies involved. Note the strong linear correlation between the poverty level of a school (percentage of students attending the school, who qualify for a free or reduced-price lunch) and their average percentile achievement on the digital formative assessments in the learning program. Each school is represented by a single dot.

Using Data to Inform Curriculum Improvement

In this case study, our design goal was to help large digital educational companies use data to improve learning products for supporting the needs of high-poverty schools. In order to achieve this goal, we wanted to help companies identify product areas that, if improved, would be likely to have the greatest positive impact on high-poverty schools.

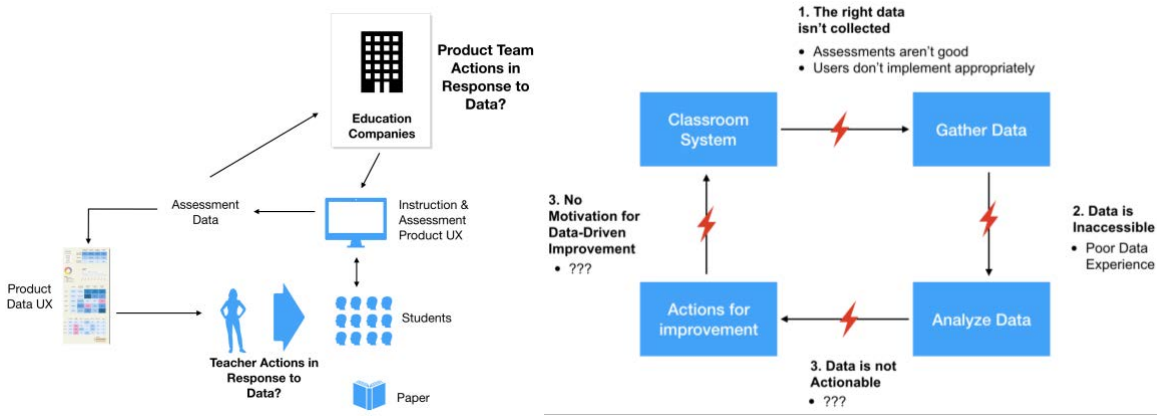

Fig. 7: Left: A data-informed feedback loop exists between teachers and students, via student assessment data. Teachers should have better defined actions in response to data – and likewise, the product teams should have defined actions in response to the data they collect about digital engagement and learning. Note that there is a more rapid feedback loop between students and teachers, which is not mediated by assessment data. This involves processes like empathy. Right: The common barriers to data-informed decision-making in educational products and their root causes are indicated. '???' indicates a root cause that is not understood.

In this context, how did we analyze the curricular data to identify product areas or opportunities for improvement? We first sought to define the product goals and find outcome measures of their achievement. The product goals, simply included mastery on different learning objectives; outcome measures included student performance on the digital assessment items tied to those learning goals. Subsequently, we investigated which learning objectives constituted the biggest needs. That meant looking at the topics, which students struggled with the most (i.e., in which they had the lowest scores). Defining needs as "failed topic performance" would suggest that the worst-performing topics should receive the most attention. However, this is not a sufficient criterion: low-success topics might be viewed as less unimportant. Therefore, we also set a criterion of greatest importance along with that of greatest need. We defined the importance of topics, based on their average correlation with the end-of-year tests. Logically, topics best predicting end-of-year performance are likely to have the greatest impact on end-of-year performance, when improved. Finally, since our goal was to help students from high-poverty schools, we set another criterion: the correlation of school poverty with topic performance. We hypothesize that

the biggest impact would be likely to result from improvements in topics that are: 1. most important for success and 2. disproportionately problematic for high-poverty schools.

In order to validate that this process identified relevant topics, we considered the actual material. “Rounding decimals” topped the list: it both predicted end-of-year performance and was highly associated with school poverty. Why might this be? Logically, if a school has overall poor performance on rounding decimals in the fifth grade, this may indicate a general lack of understanding of decimal place values and number sense. This lack of number sense may impede much other grade-level math learning; while students can continue to learn skills for solving math problems, they will fail to understand the logic of the answer if they lack number sense [14]. If students lack place value number sense in the fifth grade, it does not make sense to move on to other topics that rely on understanding place value.

Table 1: The table presents idealized data showing how average school performance on formative assessments (aggregated items by skills) correlates with school-poverty percentages (percentage of students qualifying for free or reduced-price lunch), and correlates with end-of-year tests. Skills with the greatest magnitude of correlation are *likely* to be those that will have the greatest impact on reducing the poverty achievement gap, when their associated instructional resources are improved.

Skills	# Students	# Schools	Correl. with % Poverty	Correl. with End of Year Test
Rounding Decimals	9033	279	-0.62	0.54
Estimating Sums and Differences of Fractions	7192	235	-0.51	0.53
Estimating Sums and Differences of Mixed Numbers	4825	176	-0.58	0.52
Decimal Place Value	8439	279	-0.57	0.49
Solving Problems Using Division	3492	147	-0.49	0.48

Our analysis aimed to reveal topics that should be prioritized for curriculum-design improvements. This finding is limited, because, only a series of controlled experiments could definitively show that making improvements on these prioritized topics is, in fact, optimal for reducing the poverty achievement gap. There is another more fundamental limitation: these data alone are not enough to create meaningful systemic changes. There would need to be organizational processes in companies or schools that were both capable *and* motivated to take actions in response to the data (e.g., by funding curriculum improvements). This is challenging, in part, because there is no strong financial motivation to improve the outcomes of digital education products. In contrast to typical markets, the educational-curriculum market is not primarily driven by efficacy—while companies and their employees want to help students do better, improving outcomes alone is not likely to drive additional product sales. As no one gets paid more when products work better, no actions are taken. This represents a political and economic challenge as much as a measurement challenge.

In contrast, technology companies have often successfully applied product development methods that use data to inform improvement—after all, this is done in the service of creating more compelling products that make more money. Yet, besides differences in incentives, these data-informed approaches have also been difficult to apply within large legacy systems, in fields such as education and healthcare. Several of the barriers to data-informed decision making are presented in **Error! Reference source not found.**.

Case Study #2: The Dangers of Automatic Data-Driven Optimization

A/B Testing Methods

Companies like Google and Amazon run tens of thousands of A/B tests every day [0]. These are controlled product experiments, which randomly assign users to different versions of the product design. Thus, the organization is able to measure the effects of the design on various outcome metrics that it collects, such as average revenue per user. A/B testing relies on a reflective loop, to respond to the outcomes of the tests and modify the product in response. A/B testing can be a powerful mechanism for causally determining the effects of designs on outcomes, but it can be expensive to produce the multiple designs and inappropriate to run in the absence of large numbers of users. Airbnb suggests pitfalls to avoid [0]; while Google offers an alignment

approach called “Goals-Signals-Metrics”, and categories of common metrics, which they call the HEART (Happiness, Engagement, Adoption, Retention, Task Success) Metrics [0].

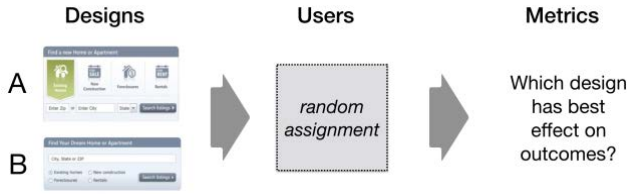

Fig. 8: A representation of the A/B testing method: two different designs, A or B, are randomly assigned to users. Metrics determine which design has the best effects on outcomes; this design is then chosen to be used by everyone.

Success metrics can be used by human teams and AI systems to drive continuous improvement. However, the optimization of metrics can produce unintended consequences when chosen metrics are not fully aligned to intended outcomes and when feedback loops about metric suitability are impoverished. This example shows why systems should be “data informed” but not purely “data-driven.”

In this case study, an online educational game was designed with the goal of motivating students to practice math problems. After being deployed online, the game attracted several thousand students each day; these players were randomly assigned to different game design variations to observe the effects of different designs on key outcome metrics (e.g., duration of voluntary play). To investigate the role of AI in system design optimization, we implemented a multi-armed bandit (a reinforcement learning AI algorithm [0]) to automatically test variations in the existing game parameter space (e.g., time limits, etc). The algorithm was designed to optimally balance the exploration of potential game designs by exploiting the most successful designs; sometimes, it would randomly search the game design space for configurations maximizing metrics (duration of voluntary play time), and sometimes, it would deploy the most successful variations. While the algorithm worked as intended, the system “spun out of control”, and primarily deployed malformed game designs, which maximized the outcome metric, but were misaligned with the original educational intent: the game variations were possibly played for long periods of time, because they were absurdly easy. This shows the pitfalls of having AI systems engage in automatic optimization, without humans as a governing feedback system in the loop. Systemic designers need to design feedback systems for monitoring systemic AI, in order to ensure that the outputs are meaningfully aligned to systemic intentions.

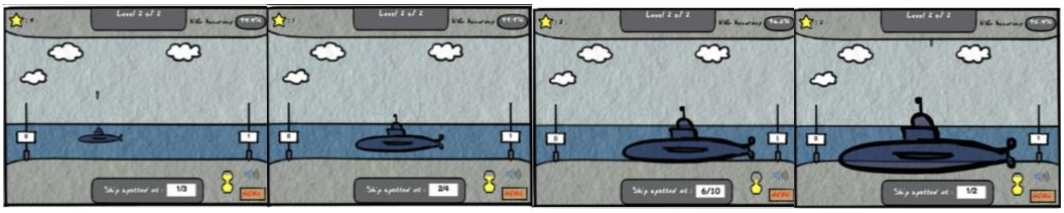

Fig. 9: Several of the game design variations of *Battleship Numberline*

In the first experiment [0], 10,832 players were randomly assigned to 3 different algorithms, each testing 6 different design factors (2x3: target type and target size). This experiment demonstrated that all the algorithms automatically produced design variations, which were more optimal for engagement. However, in a follow-up experiment with 5,849 players, we included some “ridiculously bad” designs. Surprisingly, some of these designs (such as the one with the enormous submarine, above) performed the best. As this resulted in the majority of users receiving the designs, we started to receive phone calls about bugs in the game. We had to explain that it was not a bug — we were just optimizing regarding whatever students were playing the most. It was not that the algorithm failed, rather the metric of success (total trials played) failed — it was not aligned to our “actual” goals of supporting student learning. This example of the runaway algorithm shows the importance of choosing the right overall evaluation criterion (metric for success) for optimization. We might have, for instance, jointly optimized, both for student accuracy in their responses as well as total trials played. However, we wish to make the point that this isn’t merely a matter of choosing better metrics but having a regular and rapid human response to whatever metrics are chosen.

Reflections

The key insight from this case study is the importance of keeping "humans in the loop." Automated optimization is powerful but dangerous. How might we provide a feedback loop about the suitability of success metrics? While we had provided a user experience (UX) to show the performance metrics of different conditions, it was not enough. What we missing was insight into the actual user experience of the designs (i.e., screenshot or links to the different variations). Holistic human judgement would only have been possible by connecting the qualitative experience of the designs and the quantitative metrics.

In addition to UX for system monitoring, data-informed system design needs to attend to ongoing organizational processes for continuing data analysis. It is critical to maintain an alignment between success metrics and the "actual" strategic goals/values underlying success metrics. On the one hand, we accepted a key performance indicator (KPI)_ that was insufficient — we wanted learning, but it was hard to measure; so, we accepted a measure of engagement as a leading indicator. However, we also had *implicit* goals (such as, aesthetic appeal) that would have been violated, when we saw that the submarine was starting to become so large. These implicit intentions only became noticeable while experiencing that version of the game wherein the intentions were violated. We therefore suggest that evaluating metric-goal alignment requires holistic humanistic judgement.

Discussion

These two case studies contribute to our evolving perspective on AI, data-driven design and now "data-informed system design." Our first case study shows the potential for incremental, data-driven design to address complex social problems, but reveals the severe limitations of data, in the absence of organizational processes to respond to data. Our second case study shows the power of fully-automated optimization, reveals the danger of optimizing for the wrong metric, and suggests the need for maintaining a human-in-the-loop for achieving alignment between goals and metrics. By avoiding a blind adherence to quantitative improvement through continuous dialogue with qualitative, humanistic experience, Data-Informed System Design may provide a "second-order" learning loop, which in turn will help avoid the dangerous limitations of a purely quantitative viewpoint.

From Data-Driven Design to Data-Informed Design in K12

K12 education is a large and complex sociotechnical system, with known systemic needs (such as the poverty achievement gap) and known measures of success (such as student assessment performance). As education is resistant to big reforms, there is a need for system-design perspectives that can support the incremental but continuous improvement of positive outcomes.

Data-Driven Design involves *developing* or *improving* a design, based on the measurement of outcome data, particularly indicators of success. For better and worse, data-driven design reduces complex and disputable notions of goal achievement into one or more quantified, numeric outcome measures or metrics. These metrics might emerge from human-generated ratings or rubrics. Alternatively, they might be generated from computational collection methods, such as website analytics or environmental sensors. In any case, it is important that outcome metrics closely align with organizational goals and values. Currently, many in the learning science community are exploring the use of data-driven design to improve K12 educational systems [0]. These "continuous-improvement systems" aim to align strategic goals, outcome metrics and human-computer system processes, for supporting improved learning outcomes. However, some approaches to data-driven instruction have triggered fierce opposition from educators [0], who are concerned about an over-emphasis on test scores, thereby resulting in misleading data and the dehumanization of teaching. These are fair critiques, which should be addressed in a nuanced manner.

If the danger of data-*driven* decisions is that they are inhumane, then one possibility is that decisions should be driven by more than the numbers. Data-*informed* systems semantically leave a place for holistic, qualitative viewpoints. Qualitative insights can help address one of the biggest risks of a quantitative design, namely, a situation wherein the measured outcomes (metrics) do not actually align with the goal, purpose or value of a system, which the metrics intend to measure. If raising students' test scores, for instance, will not actually help them become more successful, then what is the point of such a measure? Stakeholders always need to ask themselves: how well do outcome measures align with our shared intentions for the system?

Thus, we emphasize "data-informed" systems over "data-driven" systems [0,0], so that the opportunities for quantitative improvement do not suppress a more humanistic and holistic view of system design. Data-informed design is presented here in the context of system design, wherein we assume that the data is practically useful, only when functionally integrated into existing organizational systems.

Systemic Alignment as an Innovation Process

A basic heuristic for data-informed design is to ensure that the system is capable of measuring successful outcomes. For instance, **IF** the goal of product X is to develop student mastery on topic A, **THEN** collection of valid measures of student mastery on topic A must be ensured. By defining the data needed, product designers can ensure that their product design is able to collect that data. Subsequently, processes for maintaining systemic *alignment* may help prevent the gaming of quantitative metrics (see section below on Goodhart's Law).

Systemic alignment involves the explicit documentation and alignment of organizational values, strategies, goals, outcomes and instrumented data metrics. For instance, designers may wish to document strategic goals, which do not have measures, or have measures not clearly connected with a strategic goal. They might try to understand whether successful cases are actually "holistically successful", or just a numerical success. They might also follow the Goal-Signal-Metrics approach at Google [0].

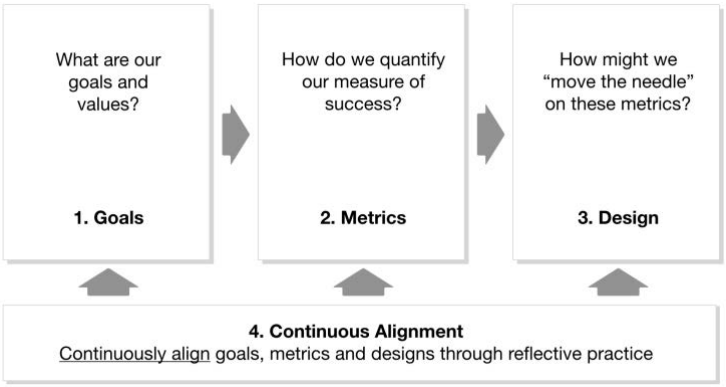

Fig. 10: Systemic alignment is a method for designing new systems to support data-driven design.

A key part of the alignment work is similar to the process of operationalization of constructs in psychological research, to the constructive alignment of learning objectives and assessments in education, and to the operationalization of values in design [0]. Specific measures of quality in the system can be collected through a specific instrumentation of data metrics. Yet, there is a need to explicitly document the alignment between instrumented metrics, the outcomes they are trying to measure, and the relationship between those outcomes and the "real objective." Positive change in quantitative metrics should result in improved "meaningful outcomes", which are to be holistically assessed.

When metrics are defined for the success of systemic processes, a system can evolve to make the improvement of the metrics as easy as possible. Designers need to make sure that this facilitation effect is not due to the corruption of systemic goals and processes that the metrics are designed to support. Designers should aim to develop an integrated viewpoint on system needs and opportunities, based on their holistic understanding of the system and the data it produces. System designers should anticipate the organizational barriers to investing in a continuous explicit alignment process, over time. In order to overcome these challenges, designers may wish to connect outcome improvements to specific economic metrics for demonstrating bottom-line value (i.e., to quantify the estimated economic impact of the improvements).

Towards Visionary Incrementalism

Design theorists Don Norman and PJ Stappers refer [0] to economist Charles Lindblom's theory of large-scale change, which he calls "muddling through." As large-scale plans are highly prone to failure in large and complex sociotechnical systems, making big plans and trying to execute them is somewhat irrational. Instead, Lindblom advocates for an incrementalist approach of small, continuous improvements. This is fully compatible with approaches [0,0] that drive change by setting a common *vision* for an organization's future: when a future vision

can be realized through an incremental strategy, we call this *Visionary Incrementalism*. For instance, visionary incrementalism may be appropriate when considering how design might address entrenched socio-economic disparities. Designers can play a role in developing and communicating vision, and in helping facilitate the instrumentation of incremental data-informed feedback loops that can help system stakeholders move towards a common and compelling vision, one small step at a time.

Limitations: What Can Go Wrong?

There are a number of important limitations in data-driven design and the use of quantitative measures to define value. When outcome metrics are established in complex systems, there is a tendency for the metrics to be "gamed", producing unintended consequences.

1. **Poor goals (misalignment with values):** Case study #1 is oriented around improving end-of-year test scores, but perhaps in high-poverty schools we should focus more on improving basic factors of student well-being.

2. **Poor metrics (misalignment with goals):** In Case study #2, we maximized the metric of voluntary time on task, but we actually wanted students to be learning fractions during their voluntary time. Measuring learning is much harder than measuring time, but we might have chosen a metric that jointly optimized time and student performance.

3. **Unspecified actions for responding to data:** In case study #1, we found that there is a lack of organizational processes to respond to product data.

4. **Limited incentive for improvement:** Case study #1 refers to the lack of financial incentive to improve educational products.

5. **Mis*leading* data (invalid data) and mis*reading* data (inappropriate analyses):** Much of the data available for data-driven design needs to be validated (e.g., formative assessment scores). Organizations that do not have a strong data culture may treat it inappropriately (e.g., treating the data as more valid than it might actually be) [0].

6. **Unintended consequences:** Accountability can create perverse incentives (e.g., schools encourage low-performing students to drop out or to cheat) [0].

Goodhart's Law

"When a measure becomes a target, it ceases to be a good measure" [0]. This is known as Goodhart's Law, and it represents an important limitation of data-driven design in education and other social systems. Since this law is becoming more well-known, a larger quote is provided below, in order to share the substance of his critique of education and then existing command economies.

"Achievement tests may well be valuable indicators of general school achievement under conditions of normal teaching aimed at general competence. But when test scores become the goal of the teaching process, they both lose their value as indicators of educational status and distort the educational process in undesirable ways… [such as] administering pretests in a way designed to make scores as low as possible so that larger gains will be shown on the post test, or limiting treatment to those scoring lowest on the pretest so that regression to the mean will provide apparent gains."

Goodhart then turns to discuss the *"harmful effects of setting quantitative industrial production goals* [in the USSR, which] *created dysfunctional distortions of production when used as the official goal in terms of which factory production was evaluated. If monetary value, then factories would tool up for and produce only one product to avoid the production interruptions of retooling. If weight, then factories would produce only their heaviest item (e.g., the largest nails in a nail factory). If number of items, then only their easiest item to produce (e.g., the smallest nails). All these distortions led to overproduction of unneeded items and underproduction of much needed ones."*

Goodhart additionally provides his most resonant depiction of the dangers of metrics: the actual use of a "body bag count" to assess success in Vietnam. Goodhart suggests it is "the worship of a quantitative indicator" that causes these sorts of corruptions of intent (e.g., the intent to educate, to produce goods or to win a war).

In the end, Goodhart does not actually advocate against the use of quantitative metrics. Instead, he encourages his peers to "develop ways to avoid the problem". For instance, "the use of multiple indicators, all recognized as imperfect." Further, to avoid the subversion of a measure, he recommended that we "study the social processes through which corruption is being uncovered and try to design social systems that incorporate these features."

Caution about Continuous Improvement Loops

One such social process that can be easily corrupted is "single-loop" learning. Chris Argyris and Donald Schön originally presented the idea of single- and double-loop learning in organizations [0,0,0]. Data-driven continuous improvement loops, e.g., optimizing KPIs in a business, can be dangerous when they only involve a single loop of learning. To give an example, a single-loop learning thermostat will turn on the heat whenever the temperature in a room drops below 68 degrees. In contrast, a thermostat with double-loop learning will first investigate whether 68 degrees is an appropriate goal temperature (e.g., based on, for instance, the season and whether a person is home or on vacation) and, only then, set a goal temperature [0]. Case study 2 demonstrates the dangers of single loop learning, when it leads to blind optimization. For this reason, we recommend that complex sociotechnical systems always use double-loop learning to avoid the corruption of metrics and to promote broader organizational learning [0].

Evaluating System Outcomes: Towards a Humanistic Use of Data

Evaluating holistic alignment requires humanistic judgement. No computer program alone can replace the human social capacity for holistic evaluation. Only humans can sit around and ask: "Should *this* really be our goal?" But neither are computers necessary for systems to be metric-driven in a dehumanizing manner – as shown by the brutality of the body-bag count described by Goodhart. We need ways to ensure that humanistic values and human sensibilities can be in dialogue with our quantitative urge. Humanistic aspirations can be vague—certainly in comparison to the goal of improving a number—but they are an important check on whether goals are being met in an appropriate way. This becomes all the more important in case of computational systems of evaluation and automated optimization.

Humanistic control systems seem critical for overseeing automated system optimizations. Humans in the loop can help mitigate unintended consequences resulting from improving outcome metrics. As designers, we need to consider how to negotiate and balance the role of the qualitative and quantitative in system evaluations. Organizational processes should ideally create a dialogue for successfully negotiating metrics — the holistic values of the organization should be in alignment with their quantitative measures.

To support this, goals and vision also need alignment. Whereas goals need to be measurable, technical, reductive and specific, design visions should convey a quality of experience —they should be holistic, intuitive and usefully vague. The tension between these domains of reason and sense, of quantitative and qualitative, can be extraordinarily valuable when ensuring that human values are being met in earnest. If math scores are increasing but students are feeling alienated by their education, this something may be wrong in the alignment of goals, vision, values and metrics. Developing and sharing a vision can help ensure that data-informed feedback systems stay on track.

When designers consider both visionary experiences ("what will it feel like?") and goal-driven metrics ("what will move the needle on the metrics of success?"), mismatches can help inform revisions. This holistic reflection leans on both intuition and reason. Such a reflective process can help keep goals aligned to values, which is essential for ensuring that data feedback loops produce beneficial effects [0].

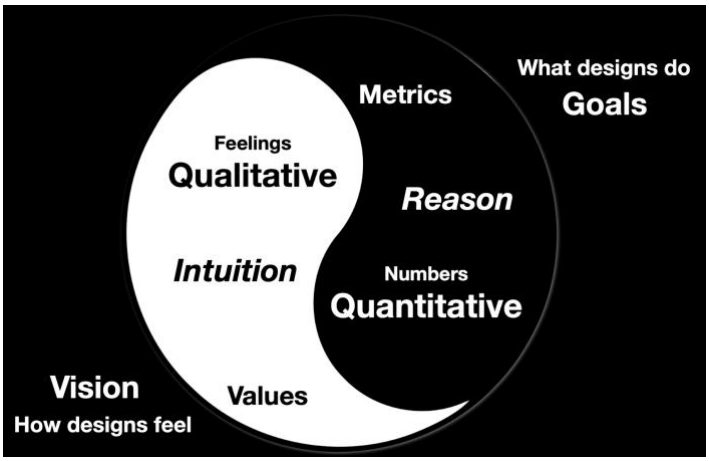

Fig. 11: Data-Informed System Design can benefit from the tension between goals and vision.

Conclusion

This paper contributes to a new perspective on intelligent system design. It describes the potential for data feedback loops in informing continuous system improvement. It reviews several benefits and limitations of data feedback loops in complex sociotechnical systems, such as the risk of an over-reliance on misaligned metrics. Smart systems involve the use of data to inform actions contributing to system success and wellbeing. Cybernetics offers a conceptual framework for a more graceful integration of artificial intelligence and systemic design.

This paper presents two case studies that illustrate the design of data feedback loops in educational systems. Our first case study walks through a potential approach to reducing the poverty achievement gap through incremental, data-informed design. Our second case study shows the power of fully-automated optimization and also reveals the danger of optimizing for the wrong metric.

These case studies show the potential benefits of data feedback loops in educational systems and also illustrate the challenges facing designers. The first case study shows how data alone is insufficient, in the absence of organizational actions to respond to the data. The second case study highlights the need for humanistic judgement to maintain alignment between goals (even implicit goals) and metrics.

Designers should play an important role in ensuring that system outcomes align with sustainable, humanistic values, which can often be expressed well in a design vision. Designers need to be prepared to define and negotiate meaningful metrics of success so that a system can measure the achievement of outcomes, to facilitate human-in-the-loop governance of AI systems, and to map existing system activity in order to understand where to best intervene.

REFERENCES

Adamson, P. (2012). *Measuring child poverty: New league tables of child poverty in the world's rich countries.* Technical Report. UNICEF.

Anderson, S. A. (1994). *Synthesis of Research on Mastery Learning*. Technical Report. IES. ERIC.

Argyris C. (1990). *Overcoming Organisational Defences: Facilitating Organisational Learning*. Prentice-Hall. (p.94).

Argyris, C. (1991). Teaching Smart People How to Learn. *Harvard Business Review*, *69(3),* 99-109.

Argyris, C., & Schon, D. A. (1983). *Organizational Learning*.

Beardow, C., van der Maden, W., & Lomas, J. (2020). Designing Smart Systems: Reframing Artificial Intelligence for Human-Centered Designers. In *TMCE 2020: 13th International Tools and Methods of Competitive Engineering Symposium* (pp. 143-154).

Bloom, B. S. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher*, *13*(6), 4-16.

Bloom, B. S. (1987). A Response to Slavin's Mastery Learning Reconsidered. *Review of Educational Research*, *57(4)*, 507.

https://nces.ed.gov/programs/coe/indicator_clb.asp

https://www.userzoom.com/blog/what-is-data-driven-and-where-do-i-start/

Clark, K. B., & Fujimoto, T. (1990). The Power of Product Integrity. *Harvard Business Review*, *68*(6), 107-118.

Coley, R. and Baker, B. (2013). *Poverty and Education: Finding the Way Forward.* Educational Testing Service Center for Research on Human Capital and Education.

Dubberly, H., & Pangaro, P. (2010). *Introduction to Cybernetics and the Design of Systems*.

Fryer Jr, R. G. (2017). The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments. In A.V. Banerjee & E. Duflo (Eds.), *Handbook of Economic Field Experiments* (Vol. 2). pp. 95-322. North-Holland.

Glanville, R. (2007). Try again. Fail again. Fail better: the cybernetics in design and the design in cybernetics. *Kybernetes*, *36 (9/10)*, **https://doi.org/10.1108/k.2007.06736iad.001**

Graven, M., & Venkat, H. (2017). Advocating Linked Research and Development in the Primary Mathematics Education Landscape in Contexts of Poverty. In *Improving Primary Mathematics Education, Teaching and Learning* (pp. 11-23). Palgrave Macmillan,.

Hamilton, L., Halverson, R., Jackson, S. S., Mandinach, E., Supovitz, J. A., Wayman, J. C., Pickens, C., Martin E.S., & Steele, J. L. (2009). *Using Student Achievement Data to Support Instructional Decision Making*. *IES Practice Guide*. NCEE 2009-4067. U.S. Department of Education.

Hekkert, P., & van Dijk, M. (2011). *Vision in Product Design: A Guidebook for Innovators.* BIS.

Holzer, H., Schanzenbach, D., Duncan, G., Ludwig, J. (2008). The economic costs of childhood poverty in the United States. *Journal of Children and Poverty*, *14, (1)*. 41–61.

Hutchins, E. (1995). *Cognition in the Wild*. MIT Press.

Hwang, G. J. (2014). Definition, framework and research issues of smart learning environments - a context-aware ubiquitous learning perspective. *Smart Learning Environments*, *1*(*1*), 4.

Kohavi, R., Deng, A., Frasca, B., Longbotham, R., Walker, T., and Xu, Y. (2012) *Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained*. KDD

Kroeger, T., & Gould, E. (2017). The Class of 2017. *Economic Policy Institute.*

Kroes, P., & van de Poel, I. (2015). Design for Values and the Definition, Specification, and Operationalization of Values. *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, 151-178.

Legg, S., & Hutter, M. (2007). A Collection of Definitions of Intelligence. *Frontiers in Artificial Intelligence and applications*, *157*, 17.

Legg, S., Hutter, M. (2007). Universal Intelligence: A Definition of Machine Intelligence. *Minds and Machines, 17(4)*,391–444.

Lomas, J. D., Forlizzi, J., Poonwala, N., Patel, N., Shodhan, S., Patel, K., Koedinger, Brunskill, E. (2016). Interface Design Optimization as a Multi-Armed Bandit Problem. In *Proceedings of 34th Conference on Human Factors in Computing Systems,2016.*

Lomas, J. D. (2020). How might data-informed design help reduce the poverty achievement gap? *Neuroscientific Perspectives on Poverty*, 267.

Lomas, J. D., Van der Maden, W., Hekkert, P. (2022, under review). Designing Tools to Support University Governance for Wellbeing.

Lomas, J. D., Van der Maden. (2021). *My Wellness Check: Designing a student and staff wellbeing feedback loop to inform university policy and governance.*

Lomas, J. D., Matzat, U., Stevens, T., Pei, L., Rouwenhorst, C., den Brok, P., Klaassen, R. (2021*). The impact of COVID-19 on university teaching and learning: Evidence for the central importance of student and staff well-being*. 4TU Centre for Engineering Education White Paper.

Van der Maden, W., Shah, J., Lomas, J.D., (2021) Student Wellbeing at TU Delft in 2020. *TU Delft Technical Report*

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the Structure of Behavior*. Henry Holt and Co.

Nelson, H. G., & Stolterman, E. (2003). *The Design Way: Intentional Change in an Unpredictable World: Foundations and Fundamentals of Design Competence*. Educational Technology.

Norman, D., and Stappers, P.J. (2015). DesignX: Complex Sociotechnical Systems. *She Ji: The Journal of Design, Economics, and Innovation 1*, *2*, 83–106

Russell, S. J., Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Prentice Hall, NJ, USA.

Overgoor, J. (2014, May 28). *Experiments at Airbnb*. **https://medium.com/airbnb-engineering/experiments-at-airbnb-e2db3abf39e7**

Patel, N., Sharma, A., Sellman, C., & Lomas, D. (2018). Curriculum Pacing: A New Approach to Discover Instructional Practices in Classrooms. In *International Conference on Intelligent Tutoring Systems* (pp. 345-351). Springer, Cham.

Pratt, D. (1982). A Cybernetic Model For Curriculum Development. *Instructional Science*, *11*(1), 1-12.

Ramage, M. (2009). Norbert and Gregory: Two strands of cybernetics. *Information, Communication & Society*, *12*(5), 735-749.

Ravich, D. (2017, August 28). *Why You Must Not Be "Data Driven"*. **https://dianeravitch.net/2017/08/28/why-you-must-not-be-data-driven/**

Rodden, K. (2015, December 2). *How to Choose the Right UX Metrics for Your Product*. **https://library.gv.com/how-to-choose-the-right-ux-metrics-for-your-product-5f46359ab5be**

Schildkamp, K., Lai, M. K., & Earl, L. (Eds.). (2012). *Data-based Decision Making in Education: Challenges and Opportunities* (Vol. 17). Springer Science & Business Media.

Simpson, A. (2015, July 29). *Data-driven vs. data-informed design in enterprise products*. **https://medium.com/designing-atlassian/data-driven-vs-data-informed-design-in-enterprise-products-538749b1b4eb**

Singer, S. (2018, September 25). *The Six Biggest Problems with Data-Driven Instruction*. https://gadflyonthewallblog.com/2018/09/25/the-six-biggest-problems-with-data-driven-instruction/

Slavin, R. E., Cheung, A., Holmes, G., Madden, N. A., & Chamberlain, A. (2013). Effects of a Data-Driven District Reform Model on State Assessment Outcomes. *American Educational Research Journal*, *50(2)*, 371-396.

Sternberg, R. J., & Grigorenko, E. L. (2004). Intelligence and culture: how culture shapes what intelligence means, and the implications for a science of well–being. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *359(1449)*, 1427-1434.

Goodhart's Law. Wikipedia. Retrieved September 17, 2021 https://en.wikipedia.org/wiki/Goodhart%27s_law

Van Geel, M., Keuning, T., Visscher, A. J., & Fox, J. P. (2016). Assessing the Effects of a School-Wide Data-Based Decision-Making Intervention on Student Achievement Growth in Primary Schools. *American Educational Research Journal*, *53(2)*, 360-394.

Yeh, S. S. (2010). The Cost Effectiveness of 22 Approaches for Raising Student Achievement. *Journal of Education Finance*, 38-75.

Zimmerman, J., Forlizzi, J., & Evenson, S. (2007, April). Research Through Design as a Method for Interaction Design Research in HCI. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems* (pp. 493-502). ACM.

Appendix

Questions and Recommendations for Designers

The following questions and recommendations may help designers support intelligent data feedback loops and overall system integrity/alignment.

1. **Goals: What are we trying to accomplish?**
   a. Define product strategy in terms of specific values and goals. Additionally, consider articulating an emotionally compelling vision for success (e.g., using metaphor) that offers a more holistic complement to measurable goals.
   b. Define specific measures of goal achievement. Which needles are you trying to move?
   c. Explicitly align goals and measures to show how a particular measure relates to a particular goal and how particular goals relate to core values or the strategic purpose.
   d. Document and characterize other stakeholder needs.
2. **Metrics: How can we measure success in the system?**
   a. Use brainstorming techniques to identify signals (observable behaviors or self-reported perceptions), which serve as indicators when a system is doing well and when there are unmet needs.
   b. Take stock of existing data, asking "what of this data might serve as a measure of system success or system need?"
   c. Document system instrumentation needs so that future data will adequately capture indicators of system success and needs.
   d. What tools, talent or resources are available for analyzing or making sense of system data? Is there a budget to support this?
   e. How might known system needs manifest in the data? How widespread are different needs?
   f. Identify key performance indicators (KPIs) of the system as a whole. What other metrics can serve as leading indicators of core performance?
   g. Which metric would be appropriate as an overall evaluation criteria metric for evaluating the effects of different conditions in a controlled experiment? (e.g., what would be the outcome criterion for A/B testing)
3. **Design: What can and should be done in response to data?**
   a. What would help "move the needle" on specific metrics of success?
   b. Brainstorm and develop potential helpful responses to specific needs.
   c. What are the affordances for action in the system? What *could* be done in response to data? Define the action space and leverage points.
   d. Which existing processes are already in place, which are well-positioned to respond to data?
   e. Are there opportunities to automate processes, in order to make data access faster, broader, or more compelling?
   f. Brainstorm potential actions that would serve as responses to different data scenarios.
   g. Continuous Alignment: "Humans-in-the-Loop"
   h. Ensure regular organizational meetings to review system data and metrics and their alignment to values/goals/strategy.

i. Identify the parts of the system or organization that would be well-positioned to respond to success metrics.
j. Identify potential risks of misaligned values/goals/metrics.

Table 2: Example of Metric-Goals-Strategy alignment chart to support Alignment-Driven Design in a training program for a Digital Curricula

Strategic Objectives	Operational Goals	Outcome Metrics
Successfully transition from paper to digital	"Increase digital platform usage"	• Percent of days average teacher uses platform • % of total teachers accessing platform
	"increase teacher confidence in using digital curriculum"	• Difference between pre-training survey and 3 month follow-up survey to teachers "how confident are you in using digital curriculum"
Enable continuous improvement of curriculum efficacy	Implement measures of curriculum efficacy	• % of instructional resources with post-tests to check mastery
Support data-driven instruction in the classroom	"Increase teacher use of student data"	• % of teachers accessing data tab in digital platform