



# Data science for K-12: A design-based research study

Andrew A. Tawfik<sup>1</sup> · Luiz Barboza<sup>2</sup> · Andrew M. Olney<sup>3</sup> · Yeonji Jung<sup>4</sup>

Received: 18 August 2025 / Accepted: 23 September 2026  
© The Author(s) 2026

## Abstract

Data science is a field that merges advancements in computer science, statistics, and machine learning. To date, much data science instruction has been through graduate programs or self-directed course offerings. There are therefore fewer opportunities for formal data science education, especially for K-12 educators and learners who have limited exposure to these fields. While some efforts have emerged to introduce data science in pre-college settings, there is little research regarding learning design strategies that support data science for inservice teachers. As such, this project employs a design-based research (DBR) approach to develop accessible and engaging open-access learning resources as a way to reduce barriers to data science instruction across the educational spectrum. Given the gap regarding data science in K-12, our DBR study identifies the following design elements: (a) revised data science content scope, (b) design standardization for interactive learning, (c) reframing data science for case-based learning, (d) artificial intelligence (AI) towards data personalization, and (e) data science coding through goal-based learning. The project also introduces aspects of the curriculum via a teacher professional development, which provides important insights and challenges regarding data science at the middle and high school levels. Implications for learning design, open-educational resources, data science standards, and others are discussed.

**Keywords** Data Science · Computer Science · Analytics · Learner experience design · STEM

---

✉ Andrew A. Tawfik  
aatawfik@gmail.com

Luiz Barboza  
barbozalc@alma.edu

Andrew M. Olney  
aolney@memphis.edu

Yeonji Jung  
yjjung@tamu.edu

<sup>1</sup> University of Memphis, 3798 Walker Ave, Office: 421D, Memphis, TN 38111, USA

<sup>2</sup> Alma College, 614 W Superior St, Alma, MI 48801, USA

<sup>3</sup> FedEx Institute of Technology, Office: Room 403E, Memphis, TN 38111, USA

<sup>4</sup> Texas A&M University, 517 Harrington Tower, College Station, TX 77843-4226, USA

## Introduction and literature review

### Data science education

Next Generation Science Standards (NGSS) and 21st Century Skills outline a variety of skillsets needed to be successful in science, technology, engineering, and mathematics (STEM). In these instructional approaches, educators align core STEM concepts through problem-solving cases that are representative of ill-structured problems found in practice (Loyens et al., 2023; Schoute & Alexander, 2025). As part of this problem-solving process, learners in NGSS-aligned instruction extend beyond mere retention of declarative knowledge as they engage in inquiry, collaboration, digital literacy, and (Long & Koehler, 2024; Wong et al., 2023). Another important aspect of NGSS problem-solving is how learners engage in observation and sense-making, which requires an understanding of measurement, representation, and data analysis (Rosenberg & Jones, 2024). Given the importance of measurement and evaluation in problem-solving, there have been considerable efforts in terms of data literacy and other analytic skillsets. Data science has also emerged as *“a new interdisciplinary field that synthesizes and builds on statistics, informatics, computing, communication, management, and sociology to study data and its environments (including domains and other contextual aspects, such as organizational and social aspects) in order to transform data to insights and decisions by following a data-to-knowledge-to-wisdom thinking and methodology”* (Cao, 2018, p. 8).

While data science was often considered a byproduct of other STEM fields, there have been increased calls towards a more specialized and standalone area of expertise (Donoho, 2017). The interest became more formalized in the “big data” era, particularly given the emergence in early 2010 around learning analytics, machine learning, artificial intelligence, and natural language processing (McFarland et al., 2021). A corollary response included rapid efforts to train around the unique skillsets that were central to this emerging transdisciplinary field. Given its applied nature among working practitioners, data science education was often targeted for workplace professionals through non-traditional pathways such as informal online learning courses (e.g., MOOCs), YouTube tutorials, and platforms such as Kaggle. These educational approaches sought to upskill workers and supplement specific knowledge deficits (Kuroki, 2023; Romero & Ventura, 2017). In response to the growing demand, a large majority of U.S. degree programs began to focus on data science within postsecondary contexts in the form of courses embedded within statistics, computer programming, and machine learning, reflecting the deep technical prerequisites needed for this interdisciplinary field (Dogucu et al., 2025). For example, computer science courses may require learners to integrate statistical and analytical concepts as they engage in coding tasks. Alternatively, a statistics program might ask learners to apply their understanding through tools such as R or Python, which requires additional expertise beyond the mathematical principles that are inherent within the course. In doing so, learners engage in piecemeal approaches to data science education, navigating disparate disciplinary approaches with limited guidance on how these areas intersect in practice (Israel-Fishelson & Weintrop, 2026; Payne et al., 2021; Rosenberg & Jones, 2024).

To better prepare learners for postsecondary contexts, there have also been recent efforts to formalize K-12 data science, such as the Guidelines for Assessment and Instruction in Statistics Education (GAISE I) (Bargagliotti et al., 2020). Key aspects describe how technology can be used to assist in data collection, variable types, multivariate perspectives, probabilistic reasoning, and others. However, a persistent challenge is that learning data

science is “often through the heavy overlap of statistics and data science concepts and practices in the K–12 curriculum” (Weiland & Engledowl, 2022, p. 3). For example, Israel-Fishelson et al. (2024) found that existing K-12 data science curricula often used relatively small datasets and offered limited student involvement in dataset selection. Similarly, K-12 data science opportunities are often distributed across statistics, computer science, mathematics, STEM inquiry, informal learning platforms, electives, and Career and Technical Education courses, rather than organized through systematic and developmentally appropriate learning progressions (ACM, 2022; Fisher et al., 2019; Rosenberg & Jones, 2024). Consequently, students and teachers may encounter data science through isolated activities, tools, or short-term instructional units rather than through sustained curricula that build conceptual and technical fluency over time (Israel-Fishelson et al., 2024; Lee & Delaney, 2022).

A corollary issue is that there remains a gap for teachers who wish to apply data science within a K-12 setting (Groth, 2025; Lee & Delaney, 2022; Rosenberg & Jones, 2024). Data science requires the coordination of statistical reasoning, computational thinking, machine learning concepts, data visualization, and domain-based interpretation, making it difficult to situate data science within traditional school curricular or teacher preparation pathways (Irizarry, 2020; National Academies of Sciences, Engineering, & Medicine, 2023; Weiland & Engledowl, 2022). K-12 teachers are thus often expected to teach across disciplinary areas in which they may not have received formal preparation (Drozda et al., 2022; Groth, 2025; Lee & Delaney, 2022). Prior research suggests that teachers frequently feel under-prepared to teach data science and must navigate instructional environments in which tools, technologies, and disciplinary expectations continue to evolve (Drozda et al., 2022; Herro & Madison, 2026). However, teacher perception alone does not establish the importance of K-12 data science education. Rather, the need for data science education is grounded in broader standards, policy, and professional efforts that call for learners to engage with data, modeling, measurement, computational representations, and statistical reasoning (Institute of Educational Sciences, 2025; Martinez & LaLonde, 2020; National Academies of Sciences, Engineering, & Medicine, 2023; National Research Council, 2014).

Given this gap, teachers’ perspectives and interactions with curricular materials are important because teachers are the proximal users and classroom enactors of data science resources. If data science learning opportunities are to reach K-12 students in formal educational settings, teachers must be able to understand, facilitate, adapt, and sustain those materials in practice. Prior research outlines various efforts towards design features needed for formalized, open-access data science curriculum in higher education and K-12. In terms of the latter, early studies attempted to address challenges of K-12 data science education through curriculum development and teacher training efforts (Herro & Madison, 2026; Weiland & Engledowl, 2022). While this research has advanced integration efforts, less research has explored data science curriculum from the perspective of K-12 teachers. As such, there is little guidance on how to design data science learning experiences that are comprehensive and engaging, especially given the teaching needs as educators support the cognitive development of younger learners. This gap impacts available curriculum, professional development, and other elements needed for the successful development of data science skills. Accordingly, this design-based research (DBR) study addresses the gap between broad calls for K-12 data science education and the limited empirical guidance available for designing teacher-facing, classroom-ready data science resources (Institute of Educational Sciences, 2025; Lee & Delaney, 2022; Rosenberg & Jones, 2024). Specifically, the study examines how an existing open-access data science curriculum originally developed for higher education can be redesigned for K-12 teacher use. This focus

responds to a practical and underexamined design problem: how to preserve the conceptual integrity of data science while adapting its content, structure, tools, and learning supports for teachers working within K-12 classroom constraints (Barany et al., 2024; Dogucu et al., 2025). Together, this study both considers which data science topics should be included in K-12 education and investigates how complex interdisciplinary content can be scoped, sequenced, scaffolded, contextualized, and made usable for teacher enactment.

## Purpose

This design-based research (DBR) project examines the iterative redesign of open-access data science learning materials for K–12 teacher use. DBR is well suited for this educational problem as it supports the study of interventions in authentic settings, while simultaneously generating practical design knowledge and theoretical insight. Rather than testing a curriculum solely under controlled conditions, DBR emphasizes iterative refinement in naturalistic environments, attending to how learners, teachers, tools, and contexts interact to yield “empirical insights and theoretical advancements that could be used to address real concerns in educational practice” (McKenney & Reeves, 2014; see also Hoadley & Campos, 2022). To achieve this goal, DBR projects often proceed through multiple cycles of design, implementation, analysis, and redesign (Richardson et al., 2025). These cycles may include broader mesocycles and more focused microcycles. The former represent larger phases of design and implementation with distinct participant groups and design goals, while the latter involve targeted revisions to specific curriculum features based on emerging evidence (Adams & Du Preez, 2022; Schmidt & Glaser, 2024).

DBR Meso Cycle 1 focused on a summer program designed to introduce undergraduate students to data science. In line with prior literature (Fisher et al., 2019; Institute of Educational Sciences, 2025), one of the notable challenges identified in DBR Meso Cycle 1 was that the scope of data science is often ill-defined given its interdisciplinary nature. As noted below, learners must integrate various areas (statistics, computer science, machine learning), which impacts one’s ability to gain conceptual mastery and progress in the learning of data science. The findings from that cycle informed the initial curriculum scope and learners’ progression of selected topics (descriptive statistics, data visualizations, measurements of association, clustering), along with strategies to supplement individuals who had undergraduate backgrounds in different subjects.

As the design team discussed ways to further expand data science beyond higher education, DBR Meso Cycle 2 sought to redesign the data science learning resources for the K-12 level, especially for inservice teachers. Although prior work has advanced K-12 data science through tool evaluations (Israel-Fishelson et al., 2025), consideration of curricular needs (Lee & Delaney, 2022), and proposed topic frameworks (ACM, 2022; Fisher et al., 2019; National Academies of Sciences, Engineering, & Medicine, 2023), these efforts largely remain conceptual and lack empirical validation at the K-12 level (Institute of Educational Sciences, 2025). Moreover, very little research has been focused on K-12 teachers who are responsible for instruction of the material. Indeed, Herro et al. (2022) caution that “teachers’ lack of data science preparation hinders their ability to integrate data science practices into their lessons” (p. 496). That is, the interdisciplinary nature may make it especially difficult for K-12 teachers who may be trained in one area (e.g. computer science), but must supplement their knowledge to holistically integrate all areas required for data science education (Fry & Makar, 2021, p. 181). Additional research is needed to understand

how K–12 teachers interact with a data science curriculum and learning resources, particularly given the complexity and interdisciplinary nature of the field (Institute of Educational Sciences, 2025; Irizarry, 2020; Rosenberg & Jones, 2024).

The purpose of DBR Meso Cycle 2 is thus to examine how inservice K–12 teachers progressed through redesigned data science materials and where they encountered challenges. Progress was important to investigate because DBR Meso Cycle 1 considered the scope of the foundational data science topics and how they could be sequenced so the concepts are able to build on one another. Similarly, error patterns are important to understand where individuals encounter challenges in their progress regarding the sequence of data science concepts (Singh et al., 2024; West, 2018). Meso Cycle 2 extends the prior design work with a focus on K-12 by documenting this process and design features, which is important for secondary education data science because the field still lacks sufficient empirical guidance regarding how complex data science content can be scoped, scaffolded, and contextualized in ways that support classroom teachers. By analyzing teacher progression and challenges across the curricular scope and sequence, this study contributes empirically grounded design features for developing teacher-facing and interactive data science learning resources. Specific research questions for DBR Meso Cycle 2 are as follows:

1. How do inservice K-12 teachers progress in a data science curriculum?
2. What are the error patterns of K-12 teachers completing the data science curriculum?

## **Project description**

### **Datawhys project**

#### **Project background**

The overall goal of this DBR project was to understand how to formalize data science education, particularly in terms of design strategies that are instantiated in an open-access curriculum. In DBR Meso Cycle 1, this data science education project focused on upskilling undergraduate learners outside traditional data science fields (e.g., computer science, statistics) through a 6-week summer internship. The project team members include data science faculty, industry experts, and an instructional design team. Across various summers (2020–2025), the project hosted a five-week internship at a university campus whereby undergraduate students from various disciplines would have university faculty lecture on data science (didactic learning) and later engage in hands-on application of their given concept.

### **DBR Meso Cycle 1: original scope and datasets**

In terms of curricula and learning resources to support the formal data science summer instruction, an interdisciplinary group of faculty members from statistics, computer science, education, and cognitive psychology contributed their perspectives and expertise to the development of the training materials. To accomplish this goal, DBR Meso Cycle 1 first conducted a needs assessment that looked at nearly 1,000 different data science resources (e.g., notable higher education textbooks, popular websites). After various

| Analysis & Exploration                                                                                        | Design & Construction                                                                                                                                                                                                 | Evaluation and Reflection                                                                                                                                                                  | Analysis & Exploration                                                                                                                   | Design & Construction                                                                                                                                                                                                                                     | Evaluation and Reflection                                                                                                                                                                                                                                                                                                                   |
|---------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ul style="list-style-type: none"> <li>- Literature Review</li> <li>- Identification of Principles</li> </ul> | <p><b>Design:</b> Jupyter notebooks individually designed from 4 faculty members</p> <p><b>Focus:</b> Interdisciplinary introduction for higher education students</p> <p><b>Assessment:</b> Surveys, LXD studies</p> | <p><b>Key Finding:</b> Students understand the applied nature of statistics, but struggle with coding aspects of data science</p> <p>Relationship between usability and cognitive load</p> | <p>How would data science scale for a K6-12 audience?</p> <p>How can data science learning materials be designed to scale as an OER?</p> | <p><b>Design:</b> Data Science learning materials that were redesigned for scope, usability, and OER</p> <p><b>Focus:</b> K6-12 in-service teacher perspectives on data science education</p> <p><b>Assessment:</b> Log file (execution, error rates)</p> | <ul style="list-style-type: none"> <li>- Revised data science content scope</li> <li>- Design standardization for interactive learning</li> <li>- Reframing data science for case-based learning</li> <li>- Artificial Intelligence (AI) towards data personalization</li> <li>- Data science coding through goal-based learning</li> </ul> |
| Micro Cycle                                                                                                   | Meso Cycle 1                                                                                                                                                                                                          |                                                                                                                                                                                            | Meso Cycle 2                                                                                                                             |                                                                                                                                                                                                                                                           |                                                                                                                                                                                                                                                                                                                                             |

**Fig. 1** Design-based research adapted from McKenney and Reeves (2021)

discussions and iterations, the materials of DBR Meso Cycle 1 consisted of 20 unique topics and 230 hours that progressed from descriptive statistics, data visualizations, measurements of association, clustering techniques, and others. Each module includes two unique aspects: (a) partial worked example and (b) isomorphic problem-solving task. The former is a hands-on activity that included detailed descriptions of data science concepts and methods for coding the models, while the latter is a unique problem-solving activity, but provided no solution steps as it focused on higher-level goals (e.g., conduct a train/test split). This design choice allowed learners an opportunity to fade the scaffold and transfer their lessons learned from the more didactic portions of the class (see Fig. 1).

Another unique aspect of DBR Meso Cycle 1 and the development of open-access learning materials included the use of Jupyter Notebooks as a platform for the data science instruction. Jupyter Notebooks combine text, mathematical equations, code, and graphs. An important feature is that the code is runnable, meaning that the code embedded in the notebooks is interactive and can reproduce the results. In contrast to a traditional paper-based textbook often used for computer science, Jupyter Notebooks can serve as a medium for worked-examples pedagogy in coding (Atkinson et al., 2000), which shows step-by-step problem solutions for computer science (see Fig. 2). Researchers thus leverage Jupyter Notebooks as a learning resource given their ability to be “*read like a journal paper and run like a computer program.*” (O’Hara et al., 2015, p. 263).

In addition to the format of the learning resources, another important design aspect of DBR Meso Cycle 1 is the selection of datasets embedded within the Jupyter Notebook. Showing step-by-step solutions to data science problems requires robust datasets that allow for diverse modelling; however, it is often difficult to find datasets that have optimal properties to demonstrate data science techniques, so instructors often reuse readily available datasets. DBR Meso Cycle 1 followed standard practice in teaching data science by using well-known datasets from the UCI Machine Learning Repository (<https://archive.ics.uci.edu/>). DBR Meso Cycle 1 employed the iris dataset originally introduced by Fisher, which is one of the most widely used dataset in statistics education (Unwin & Kleinman 2021). This prominent dataset includes five variables that detail information about flowers, four of which are width or length measurements of the parts of an iris, and the fifth variable is species. Datasets like iris are frequently used in pedagogy because they are simple yet precise, while instructors are familiar with their properties

### Prepare train/test sets

We need to split the dataframe into training data and testing data, and also separate the predictors from the class labels.

Often we'd drop the label/response and use all the predictors, but in this case, we only want to use two of the predictors. So let's make `X` and `Y` but specify what we want in them, instead of what we don't want (which is what `drop` does):

- Create variable `X` and set to `dataframe[ ]` with a list inside containing
  - "Hits"
  - "Years"
- Create variable `Y` and set to freestyle `np.log1p(dataframe[['Salary']])`

```
]:
```

Import `sklearn.model_selection` so we can split the data into train/test sets:

- `import sklearn.model_selection as model_selection`

```
]:
```

And do the actual split:

- Create `splits` and set to `with model_selection do train_test_split using` a list containing:
  - `X`
  - `Y`
  - freestyle `random_state=1` (this will make your random split the same as mine)

```
]:
```

**Fig. 2** Original data science curriculum for higher education learners (DBR Meso Cycle 1)

and the statistical phenomena that can be employed given the data. This allows the iris dataset to be useful for teaching introductory data science concepts, such as descriptive statistics and plotting, along with advanced modeling topics (Zheng, 2023).

## DBR Meso Cycle 1: learning environment scaffolds

Given the overall goal of design features needed for formalized, open-access data science curriculum, it was important for the design team to consider how to scaffold learners as Meso Cycle 1 introduced data science to postsecondary students. Learning to program and its associated syntax are often considered difficult for novices (Luxton-Reilly et al., 2018; Mergianian et al., 2026), so DBR Meso Cycle 1 designed the Jupyter Notebooks to concurrently work with an optional extension for blocks-based programming, which has been widely used to teach introductory programming over the last decade (Bau et al., 2017; Resnick et al., 2009). Blocks-based programming uses puzzle-piece shaped widgets to represent snippets of code such that learners ‘snap’ blocks together to make programs. All blocks are visually browsable on an interface palette, so learners only need to recognize the blocks they need rather than remember arbitrary code. Moreover, blocks only fit together in syntactically valid ways, eliminating common syntax error mistakes. Blocks-based programming provides a visual representation of coding, often leading to gains in cognitive and affective learning outcomes (Hu et al., 2021). DBR Meso Cycle 1 also introduced a novel ‘blocks to code’ feature, which allows learners to use a blocks-based coding approach and synchronously see how it would align with Python code (see Fig. 3) (Tawfik et al., 2024). From a cognitive load perspective, this support was intended to help reduce demands on working memory associated with element interactivity during coding (Çakiroğlu & Bilgi, 2022).

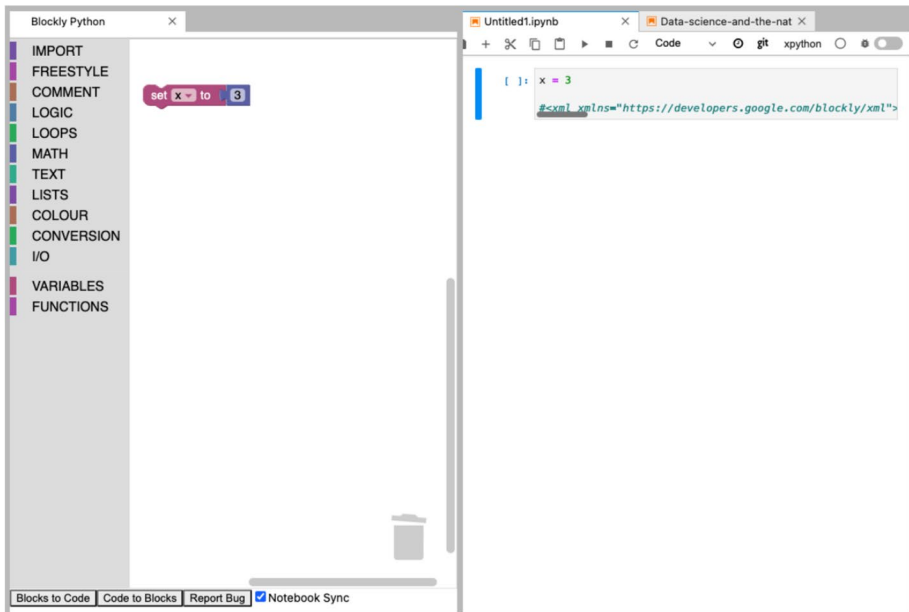


Fig. 3 Blocks-to-code learning environment (DBR Meso Cycle 1)

## DBR Meso Cycle 2: data science design features for a K–12 audience

Research from DBR Meso Cycle 1 yielded important insights about the higher education student perspective. The first design feature focused on the need to consider learners' prior knowledge gaps and appropriate conceptual introduction for novices, including onramping the interdisciplinary nature of data science given its confluence of statistics, computer science, and machine learning (Payne et al., 2021). A second principle also identified the learning modalities and platforms needed to support data science, especially an interactive platform that allowed (a) scaffolds, (b) multiple forms of representation, and (c) contextualized application (Tawfik et al., 2024). The final emergent design feature focused on learner experience design, especially usability challenges given the different scaffolding supports (free-hand coding; block coding) that were part of the data science learning tool.

As the DBR project began to understand important design elements of formalized data science instruction, the team sought to consider design elements beyond higher education (Meso Cycle 1) and how data science might be applied at a K-12 level, (Meso Cycle 2), especially at the middle and high school level. The design features of Meso Cycle 1 led the design team to consider the needs of K-12 teachers who will be at the forefront of data science integration. As such, the design team quickly realized the current higher education version included critical design assumptions in terms of content (mathematics, necessary examples), and also required attention to educational context and learner reading level, which are prominent factors within the classroom. Related research shows K-12 teachers feel ill-suited to integrate data science in the classroom given the intersection of multiple domains and limited opportunities for professional development (Drozda et al., 2022). This is a challenge because Rosenberg and Jones (2024) note that many curriculum

**Table 1** Design features across Meso Cycles

| Meso Cycle 1 Design Features (Higher Education)                                                                                 | Meso Cycle 2 Design Features                                  |
|---------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------|
| Learners prior knowledge given interdisciplinary nature of data science                                                         | Revise data science content scope                             |
| Modalities and platforms that supported (a) scaffolds, (b) multiple forms of representation, and (c) contextualized application | Standardization design for interactive learning               |
| Learning experience design for different approaches to coding                                                                   | Reframe data science for case-based learning                  |
|                                                                                                                                 | Leverage artificial Intelligence (AI) to data personalization |
|                                                                                                                                 | Facilitate data science coding through goal-based learning    |

standards require “*developed strategies, representations, statistics, models, and computational systems to structure the variability in ways that allow us to make observations and inferences that cannot be made by direct observation alone*” (p. 10).

The transition from DBR Meso Cycle 1 to DBR Meso Cycle 2 therefore involved more than a change in audience. It required revisiting assumptions embedded in the higher-education version of the curriculum, including assumptions about mathematical background, programming familiarity, reading level, pacing, and learners’ ability to navigate open-ended analytic tasks. For K-12 settings, these assumptions are especially consequential because teachers must not only understand the data science content themselves, but also determine how to facilitate it for students with diverse prior knowledge, varying levels of coding experience, and limited instructional time. Thus, the redesign process focused on the practical design problem of making data science materials usable for K-12 teacher contexts, while preserving the conceptual integrity of the original curriculum. This required attention to content scope, structural consistency across notebooks, case-based framing, dataset alignment, and goal-based coding supports.

As the project further considered the design elements needed for data science across different learning contexts, Meso Cycle 2 encountered multiple features that were central as the team redesigned the platform as an open-educational resource (OER) for teachers in a K-12 setting. In line with Jung and Mercier (2023), the following section describes an in-depth retrospective reflection of the design features that emerged from Meso Cycle 2 (see Table 1).

### Design feature 1: revised data science content scope

Given the relative emergence of data science, the guidelines that outline how educators should integrate these interdisciplinary concepts (mathematics, computer science, machine learning) for younger learners are largely ill-defined, which makes it difficult to revise the scope needed for K-12 teachers and learners. In other STEM disciplines, educators are often provided guidance or a learning progression for a particular subject based on learner age and established curricular standards. For example, the ecosystem guidelines in the Next Generation Science Standards (2014) outline a logical progression of relationships in ecosystems (LS2A), cycles of matters (LS2B), ecosystem dynamics (LS2C), group dynamics

The last column in the table titled `Residual` is computed as the Observed HPI minus the Predicted HPI and can be used to diagnose the model fit. Checking the residuals is an important part of model diagnostics. The residuals can be summarized in a single measure called the *root mean square error* (RMSE). It is the square root of the sum of the squared residuals divided by the number of sample points. The smaller the RMSE the better the fit. The formula for the RMSE is

$$RMSE = \sqrt{\frac{\sum_{i=1}^n res_i^2}{n}}$$

where  $res_i$  is the residual for the  $i$ th observation. For the testing set in this example, the RMSE is 56.29. Alone this does not tell us anything about the fit of the model, but we can use it to compare various models. For example, if you wanted to try different values of  $K$  you could use the RMSE to compare the fits.

## Standardization

This example highlights one issue that you may run into when you are running a KNN regression model (or KNN classifier). The magnitude (size) of the Loan variable is much bigger than the magnitude of the Age variable. This difference in magnitude can lead the distance measure to favor differences in the Loan variable and the Age variable can have much less of an impact. One solution to this problem is to standardize each variable by subtracting the sample mean,  $\bar{X}$ , and dividing by the sample standard deviation,  $S = \sqrt{S^2}$ , as

$$Z_i = \frac{(X_i - \bar{X})}{S}$$

and performing the KNN regression on the standardized variables.

## Choosing $K$

Similar to the KNN classifier, the choice of  $K$  has a big impact on the regression. When  $K$  is low, the regression fit will be very choppy and gets smoother as  $K$  increases. The figures below show a progression of values of  $K$  from  $K = 1$  to  $K = n = 506$ . For this data, there is only one predictor on the x axis. The points in blue are the predictor (x-axis) and the associated response (y-axis). The orange line is the KNN regression prediction. As you can see, in Figure 1 where  $K = 1$ , the regression prediction is very choppy and smooths out as  $K$  increases. When  $K = n = 506$  as in Figure 6, the prediction line is completely smooth and horizontal. The choice of  $K$  is a compromise between accurate but choppy predictions (low  $K$ ) and less accurate but less variable predictions (high  $K$ ).

**Fig. 4** Original data science conceptual description for higher education students (DBR Meso Cycle 1)

(LS2D), and so on. Similar guidelines describe a progression regarding the basics of earth (ESS2A) and later biogeology (ESS2E). In doing so, educators have a ‘starting point’, along with guidance about necessary concepts that are important for a learner pathway. These are important considerations for teachers and learning designers because it serves as a progression pathway, while taking into consideration learners’ cognitive development, prior knowledge, scaffolding strategies, etc.

While this initial scope of topics was iteratively refined in DBR Meso Cycle 1 for higher education students, DBR Meso Cycle 2 encountered a recurring issue about how to determine what was foundational to understanding data science at the K-12 level. For example, a topic like K-Nearest Neighbor is a potentially complex mathematical topic that includes detailed equations to determine residuals, data standardization, and others (see Fig. 4). As we attempted to reframe the data science materials for K-12 teachers, the design team frequently questioned whether these topics, equations, or both were needed to understand the concept at the middle and high school level. Given that we had hoped to design this as an open-access resource, the project also could not make assumptions about learners’ prior knowledge in either mathematics or reading levels. A key initial resource was the FORECAST website, a readability assessment tool that assessed the original higher education materials from DBR Meso Cycle 1 at an 11th-grade reading level. There was a constant design tension about what was the appropriate data science scope for a given topic, age level, and what was retained from earlier data science Jupyter Notebooks.

To address this issue, an important aspect of the DBR Meso Cycle 2 included a learner experience design (LXD) component with teachers, which was facilitated by (a) using artificial intelligence (AI) tools to reduce the complexity of the materials, (b) multiple rounds of reviews, and (c) working with multiple stakeholders. In terms of AI tools, the instructional design team conducted an artifact analysis whereby the text-based content was input into AI curriculum tools (Diffit, Twee) to help simplify the data science conceptual description. The selected AI tools each provided similar services; however, they supported DBR Meso Cycle 2 in different ways. For example, the design team inserted the original conceptual text into Diffit, which would allow the team to revise the content for a given grade level. Shortly after clicking ‘Generate Resources,’ a variety of materials populated including the following: student handouts, a summary of the text, vocabulary, and a multiple-choice quiz. For purposes of the redesign, the project used the sections “Summary” and “Key Vocabulary Words.” Whereas Diffit provides a bullet-pointed summary, Twee converted the text in a paragraph format.

Given the different options among multiple tools, comparison among AI-generated passages between Diffit and Twee was an essential part of the redesign process. Upon completion, the design team synthesized the three versions: Diffit, Twee, and the original text to create the final version of the data science conceptual description. An instructional design faculty member would then take this design artifact, meet with a data science subject matter expert, and further refine the scope based on the materials that were generated from the AI revisions. As a final refinement, the project would be handed off to instructional design research assistants for a final review of the formatted content (e.g., scope, headers, placement of multimedia). From an instructional design perspective, these AI tools and design artifacts were advantageous at the outset of DBR Meso Cycle 2 given the complexity of the interdisciplinary concepts (mathematics, statistics, computer science), especially since the instructional designers were not subject matter experts in data science. The AI tools were therefore not summative; rather, they served as a starting point for expert review in creating engaging, grade-level appropriate learning materials that facilitated the discussion between instruction designers and subject matter experts.

## Design feature 2: design standardization for interactive learning

Another important aspect in DBR Meso Cycle 2 for K-12 teachers was the design standardization of the learning resources. In the prior version of the Jupyter Notebooks, the team developed 20 notebooks written by four faculty members from computer science, mathematics, and cognitive psychology. This initial approach used for a higher-education audience in DBR Meso Cycle 1 allowed for an efficient way to generate a large array of topics, while also affording an opportunity for interdisciplinary faculty members to focus on their given areas of expertise. Although the learning materials were largely the same in terms of general aesthetics within the Jupyter Notebooks (see Fig. 4), they included nuanced design differences in terms of tone or multimedia. For example, a mathematics faculty member would often explicate the details of statistical equations. Alternatively, those with a background in computer science would largely focus on the coding syntax and how to streamline analysis using a computational thinking approach. Finally, the psychology faculty would often emphasize the applied nature of data science for various domain problems.

Given our approach to LXD in the K-12 redesign, DBR Meso Cycle 2 sought to engage with various educator stakeholders through additional LXD testing with pre-service teachers, inservice teachers, and informal educators. Initially, the design team shared a low-fidelity prototype with educators to streamline content and reduce complexity in the data science descriptions (see ‘Revised Data Science Content Scope’ section above). While the data provided some positive feedback, the teachers requested additional guidance on how long a unit might take, which highlighted a ‘blindspot’ to the design team in terms of the learning context. Whereas prior Jupyter Notebooks in DBR Meso Cycle 1 were more focused on the concepts and coding, the feedback from teachers also caused the design team to examine additional challenges inherent in a K-12 setting, especially how a task-oriented approach to coding could better address classroom management challenges. The redesign also considered informal learning settings where educators may not be domain experts (e.g., librarians, after-school clubs).

The collective data from the LXD testing necessitated our design to include dedicated sections within our learning materials for ‘anticipated time’, ‘definitions’, ‘learning objectives’, and enable the goals section to be viewed ‘at a glance’. Teachers also wanted a section for the facilitator that was not viewable to the student, which would include the main idea, anticipated issues students encountered (e.g., challenges related to train/test split), and recommended scaffolding questions. Based on the feedback from potential users, the design and development team thus sought to standardize a design in ways that would balance content and classroom dynamics through the following outline:

- Anticipated time
- Learning Objectives
- Definitions
- Multimedia video of the case
- What you need to know (2–3 paragraphs about the data science topic)
- Goals (computer science tasks to complete the data analysis)
- Reflection

As the design progressed and integrated teacher feedback in DBR Meso Cycle 2, the design team conducted an additional LXD expert review among the five members of the instructional design team. The data from the group indicated that the conceptual text was comprehensive, but it was unclear on the transition from the didactic (e.g., ‘What is a Decision Tree’) to the interactive portion (‘How to Code for Decision Tree’) of the Jupyter Notebooks. They therefore suggested the text be more action-oriented, which would foster engagement and cue the learner to progress through a given module.

An additional aspect of the redesign in Meso Cycle 2 also required considerable time to align the learning objectives with the shorter didactic (‘What You Need to Know’) portions of the Jupyter Notebooks, which the team believed helped manage learners’ working memory, intrinsic load, and germane load (see Fig. 5) (Tawfik et al., 2023). This was challenging because the design team was constantly having to revise this text-heavy portion based on the perceived complexity and feedback from the AI, teacher stakeholders, instructional design team, and subject matter experts. If a topic was deemed too complex or not foundational, the design team retroactively removed the definitions and/or revised the text. Again, this was especially challenging given the lack of accepted standards about data science for K-12 audiences.

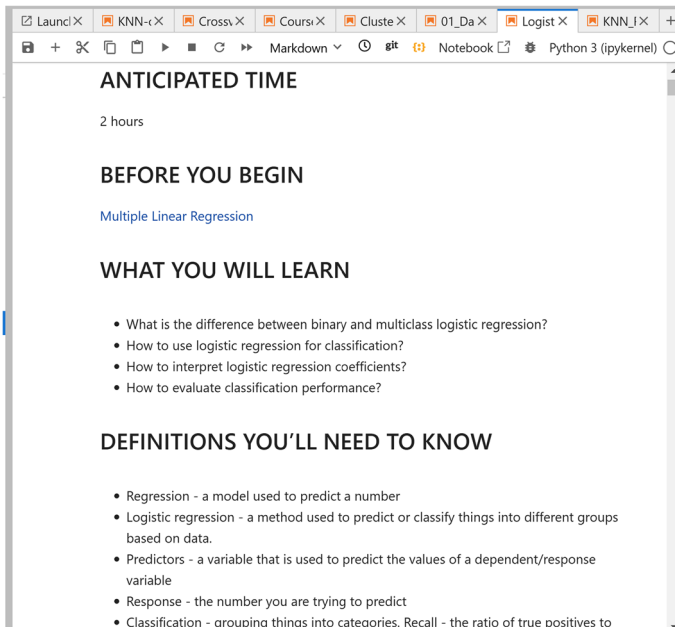


**Fig. 5** Redesigned Jupyter notebook to support cognitive load and classroom management for K-12 (DBR Meso Cycle 2)

### Design feature 3: reframing data science for case-based learning

To facilitate the action-oriented aspect of coding for data science, another key element of DBR Meso Cycle 2 design was the move towards a case-based strategy, which was based on learning theories that knowledge construction can be supported when it is situated within a context (Ertmer et al., 2009; Tawfik & Kolodner, 2016; Zhang & Ma, 2023). As noted earlier, each prior Jupyter Notebook from DBR Meso Cycle 1 was different based on the expertise of the subject matter experts/faculty and used the popular iris dataset often employed in data science education (Ciubotariu & Bosch, 2023). Alternatively, other Jupyter Notebooks from the first round included brief examples, such as how to use a decision tree to predict the weather based on categorical variables (e.g., ‘rainy’ and ‘cloud cover’). Additional subject matter experts decided to illustrate data science examples using different forms of media, such as in the example of KNN below (see Fig. 6).

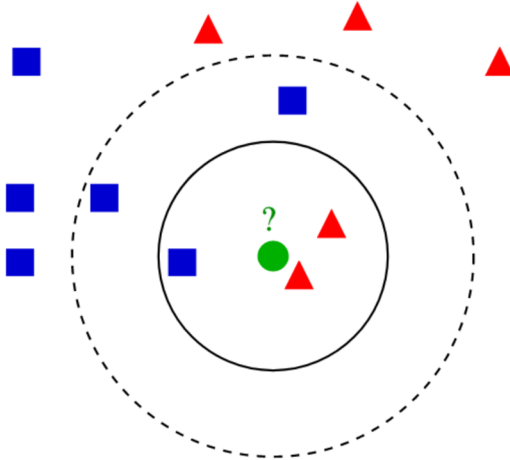
Given the focus on K-12 teachers, DBR Meso Cycle 2 especially wanted to consider how the affective component (motivation, engagement) might serve as a catalyst for data science education. The design team therefore leveraged cases as a way to afford learners agency and a sense of ownership, while also contextualizing the material (Koehler & Vilarinho-Pereira, 2021; Meulenbroeks et al., 2023). A key design decision included the nature and topics of the cases. To that end, we considered multiple different options such as how a community developer might apply data science for a neighborhood beautification project. While this would contextualize the data science concepts, the design team was concerned that cases would be too disparate, along with being far removed from their students’ daily activities. The design team was also mindful of the ‘breadth vs. depth’ challenge of

## When to use classifiers

Classification is useful when you have a set of annotated data. That is, for each set of features the true class is known. In this case you can use the data to train a classifier so that when a new set of features is encountered with unknown class, you are able to infer its class from its features.

## $k$ -Nearest Neighbor

The  $k$ -nearest neighbor (KNN) method of classification based on distance. Consider the picture below.



By Antti Ajanki AnAj, CC BY-SA 3.0

In this picture, we have a new observation represented by a green circle. We want to predict its class, which is either a blue square or a red triangle. KNN does this by looking at the  $K$  closest previous observations; in the picture,  $K$  is 3 (solid line). KNN predicts whatever class is the most common in the  $K$  closest previous observations. In this case, the most common class is red triangle (the three closest neighbors are two red triangles and one blue square), so KNN would predict the new observation (green circle) is a red triangle.

As you might guess, the value of  $K$  can change KNN's behavior a lot. If  $K = 5$  (the dashed line in the picture above), the KNN prediction will change from red triangle to blue square. Imagine in the picture above if  $K = 11$ , which is the total number of previous observations. In that case, KNN will always predict whatever class is most common *overall* (in this case blue square). This is bad because it means that the model will always be wrong for new observations that are red triangles.

**Fig. 6** Original data science multimedia for higher education students (DBR Meso Cycle 1)

case-based learning whereby learners engage in complex problem-solving to solve a contextualized case (depth), but the specifics of the scenario make it difficult to transfer learning to other problem-solving activities (breadth) (Tawfik & Kolodner, 2016).

In light of these challenges, DBR Meso Cycle 2 sought to identify a ‘throughline’ across the cases to establish a broader metanarrative for learners to situate themselves, but allow different problem-solving contexts to apply distinct data science concepts. The revised data science Jupyter Notebooks were thus contextualized as an after-school club that is looking to reduce CO2 emissions within their community (see Fig. 7). The redesign presented the cases through multimedia videos in which characters conveyed the problem and prompted the learner for advice given the available dataset. Examples include using multiple regression to predict emissions levels from multiple ratio variables and using  $k$ -nearest neighbors to classify emissions levels. This contextual approach helps to engender problem-solving, while also serving to frame the ‘Goals’ section and data science steps that follow.



**Fig. 7** Redesigned data science case-based multimedia for K-12 students (DBR Meso Cycle 2)

### Design feature 4: artificial intelligence for data personalization

The move toward a case-based approach required additional design work on how to align the conceptual description, datasets, and coding tasks within the Jupyter Notebooks. In the prior learning materials geared towards higher education audiences (DBR Meso Cycle 1), the design team utilized the iris dataset that is often adopted in higher education data science classes (Ciubotariu & Bosch, 2023). Although this was a key artifact in Meso Cycle 1, the contextualized cases in Meso Cycle 2 focused on pollution within communities required that the design team find a corresponding dataset that aligned with the case details. In line with other research (Kang et al., 2026), the selection of the dataset presented a considerable challenge as we attempted to search across available repositories, such as Google Datasets, Kaggle, and other public repositories.

Rather than identify and integrate a novel dataset, DBR Meso Cycle 2 used large language models (LLMs) to rescale the iris dataset. That is, the design team leveraged LLMs as a way to repurpose existing data, which made the datasets align with new thematic perspectives of the case and thus be more engaging for K-12 audiences. The core design strategy was to prompt-engineer the AI in a way that maintains the underlying structure and statistical properties, while changing the meaning and values to fit the new environmental science theme of the cases. This mapping includes suggesting new column names (e.g., “sepal length” becomes “traffic volume”), specifying data type for columns (numerical, categorical), and defining new data ranges or categories for the new theme. For numerical variables, the original values are rescaled linearly to fit the new ranges provided by the LLM, ensuring that the relative distribution of the data is preserved. For categorical variables, the original categories are replaced with the new categories using a dictionary-based lookup (see below for an example of an LLM-aligned schema). The result is a new dataset with the same number of rows and columns as the original iris dataset, but features new column names, rescaled numerical values, and remapped categorical values that are aligned with the CO<sub>2</sub> emissions theme. For example, the flower themes are now transformed to include illustrative data about air quality levels with columns such as “Traffic Volume”, “Average Speed”, “CO<sub>2</sub> Emissions”, “Noise Level”, and “Traffic Condition” (matching the original ones from the iris dataset, “Sepal Length”, “Sepal Width”, “Petal Length”, “Petal Width”, and “Species”).

**Fig. 8** Redesigned data personalization to align with K-12 case using AI and LLMs (DBR Meso Cycle 2)

Theme. : Arts & Entertainment ▼

Topic. : ✓ Celebrities & Entertainment News

- Comics & Animation
- Entertainment Industry
- Events & Listings
- Fun & Trivia
- Humor
- Movies
- Music & Audio
- Offbeat
- Online Media
- Performing Arts

```
new_ranges = {
    'sepal.length': (0.200, 0.400), # Traffic Volume
    'sepal.width': (0, 50), # Average Speed
    'petal.length': (0, 150), # CO2 Emissions
    'petal.width': (0, 60) # Noise Level
}
```

Given the focus on open-educational resource design in DBR Meso Cycle 2, we also considered how to ensure the newly developed materials were potentially adaptable based on other K-12 learning contexts. To facilitate adaptability in the open-educational resource, the development team also leveraged LLMs to align with popular themes identified through Google Trends to make the system automatically generate diverse datasets based on user selection (see Fig. 8). Each dataset is then organized into a corresponding server folder, effectively pre-populating a directory ready for user access. This method enables the efficient reuse of existing data across different domains, making it easier to create engaging and relevant learning materials or analytical tools that align with the case-based themes adopted in DBR Meso Cycle 2. It also enhances consistency and accuracy because the underlying data structure remains intact, and the transformation process preserves the tabular structure and is automated.

## Design feature 5: data science coding through goal-based learning

Another central design feature of our data science curriculum redesign was the adoption of a goal-based learning approach to coding that provides a scaffolded structure with step-by-step instructions for each notebook (see Fig. 9). In addition to the LXD feedback from teachers, this decision was motivated by research on goal-based scenarios in programming education (Kandin & Şendurur, 2023) and cognitive load principles in scenario-based multimedia learning (Kilic & Yildirim, 2012). Rather than framing the Jupyter Notebooks as broad, open-ended challenges, each module was segmented into approximately 7–8 coding

## Goal 1: Importing the Pandas Library

Need extra tools to help solve this problem? Well, we can bring in extra ' to help us do extra data science stuff. You can think of it like an 'add-on'. case, we bring in pandas, which is a popular library for doing data scienc

### Blockly

+ 4 cells hidden

### Freehand

+ 7 cells hidden

## Goal 2: Bringing in the Dataframe

Let's bring in the data that we want to look at.

### Blockly

+ 5 cells hidden

### Freehand

+ 7 cells hidden

**Fig. 9** Redesigned Goal-based approach to K-12 data science education (DBR Meso Cycle 2)

“goals” that learners could view at a glance. For example, teachers and students now see a Measures of Association notebook as a progression of the following seven tasks: Import the Pandas library, Bring in the dataset, Calculate correlations, Import Plotly for visualization, View a heatmap of correlations, Calculate correlations for ordinal variables, and Analyze categorical correlations. Within each of these goals, learners are given additional instructions to elaborate upon and further segment the tasks. This included a simplified description, along with how to complete the task in the coding language (see Figs. 9 and 10). In lieu of pre-established data science standards, each notebook now includes a scaffolded progression from foundational steps (e.g., data loading) to more advanced tasks, with accompanying reflection questions that encourage learners to consider what they had done and why. Beyond just cognitive processing, the redesign towards a goal-based approach integrates key aspects of educator feedback.

Teacher feedback from LXD testing was treated as formative design evidence about classroom enactment rather than as evidence for whether K-12 data science education was necessary. Teachers' requests for clearer guidance about anticipated time, instructional pacing, common student difficulties, and facilitation prompts highlighted practical classroom constraints that were less visible in the higher-education version of the materials. This feedback helped the design team shift from a primarily content- and coding-centered notebook structure toward a more classroom-aware design that included visible goals, simplified explanations, explicit definitions, facilitator guidance, and reflection

simple command to bring in the file we need (CSV....Comma Separated Values). Let's say we have a file called 'datasets/AirQualityCars.csv' in the folder '**datasets**'. We're telling Python to read the CSV file and store it in a variable called **train**.

From the Variable menu, drag a DO block using the **pd** variable, go ahead with the do operation **read\_csv**. The read\_csv function reads a CSV file and returns a DataFrame object.

In our case, let's bring in the "datasets/AirQualityCars.csv" (use the Quotes from the TEXT menu) because that is what Kiana is working with.

#### Step 4 - Display the variable

Let's see it now by 'displaying' and showing our work.

Drag the **train** variable to the workspace, making it available for further use in our program. This step is more of a visualization step, as it allows us to see the variable in the Blockly workspace

#### Step 5 - Connect the blocks to run the code

Connect the blocks and run the code!



**Fig. 10** Redesigned detailed descriptions for K-12 Coding (DBR Meso Cycle 2)

opportunities. In this way, the goal-based structure served both as a learning scaffold for teachers and students and as a practical tool for classroom planning and management.

The aforementioned design decisions are important in multiple respects in DBR Meso Cycle 2. The redesign was intended to manage processing demands by dividing complex tasks into smaller segments (Rey et al., 2019). The overview introduces upcoming tasks, and the detailed goals break the workflow into manageable steps. Cognitive load was not directly measured in this cycle. In response to the LXD testing, the redesign makes recurring tasks visible across notebooks, including uploading a dataset, importing a library, and splitting data into training and test sets. The goal structure also facilitates scaffolding across programming representations, supporting both block-based and freehand coding within the same learning environment. By anchoring both programming representations to shared goals, DBR Meso Cycle 2 provides learners the flexibility to choose their mode of engagement, while also allowing teachers to support learners at different levels of coding mastery (see Fig. 9). In terms of mastery of coding, novices might start with blocks and then transition to free-hand coding as they become more proficient. By standardizing the structure of the “goals” section in each notebook, the redesign ensured a predictable learner experience, particularly important given the diverse backgrounds of our users, including K-12 teachers, informal educators, and students with limited prior exposure to programming.

## Methods

### Participants

The primary goal of the project was to explore the design features needed for formalized, open-access data science curriculum. As the project transitioned from higher education (DBR 1), DBR Meso Cycle 2 focused on how inservice K-12 teachers progressed in the redesigned learning materials (RQ1), along with the challenges they might encounter as they learn data science (RQ2). To address these questions, the project conducted a professional development program in the summer of 2025 for urban inservice teachers (N=5) situated within the southern region of the United States of America. The school district is the largest in the state and is notably diverse in terms of both demographics (75% African American; 14% Hispanic, 7% White) and socioeconomic status, as 63,000 students in the district self-classify as ‘economically disadvantaged’.

Inservice K-12 teachers were selected as participants as the primary users and intended enactors of the redesigned open-access data science materials in this meso cycle. The broader goal of the DBR project is to support K-12 students’ data science learning; however, this cycle focused on teacher-facing interaction with the open-access curriculum because classroom adoption depends heavily on teachers’ ability to understand, facilitate, and adopt the materials. Teachers were therefore not treated as proxies for students, nor were their perceptions used as the sole evidence for the importance of K-12 data science. Instead, teacher interaction data, including code executions and error rates, provides preliminary evidence about how teachers interact with the redesigned materials. A subsequent meso cycle may extend this work by examining how K-12 students engage directly with the revised materials.

### Procedure

Prior to the professional development, members of the research team worked with a local educational computer science non-profit to discuss the learning materials and clarify any questions. The subsequent professional development consisted of a two-week program, with each day lasting six hours. In this ‘train the trainer’ approach, the facilitators first delivered an introductory, didactic explanation of the data science topics. The stated goals by the facilitator included understanding of data science, Python, Jupyter Notebooks, and pedagogical strategies. Following this introduction, the inservice teachers were directed to engage independently with the interactive Jupyter Notebooks developed in DBR Meso Cycle 2, though they were permitted to collaborate with peers or seek one-on-one support from the instructors as needed. Upon completion of each module, participants were prompted to reflect on the key data science concepts and the corresponding coding elements within the learning materials.

In terms of scope of the professional development, the facilitators shared 15 data science lessons developed in DBR Meso Cycle 2 and published via Jupyter Notebooks (see Table 2). The curriculum’s data science topics were presented in a structured progression, beginning with descriptive analysis concepts such as Data Science and the Nature of Data, Plotting, Descriptive Statistics, and Measures of Association. The training then moved into predictive models, which were segmented into supervised regression techniques, including KNN Regression, Simple Linear Regression, Multiple Linear Regression, and Regression

**Table 2** Professional development data science schedule and topics

| Session Number | Data Science Topic                  |
|----------------|-------------------------------------|
| Session 1      | Data science and the nature of data |
| Session 2      | Plotting                            |
|                | Descriptive statistics              |
| Session 3      | Measures of association             |
| Session 4      | KNN regression                      |
|                | KNN classification                  |
|                | Clustering                          |
| Session 5      | Multiple linear regression          |
|                | Simple linear regression            |
| Session 6      | Decision trees                      |
|                | Logistic regression                 |
| Session 7      | Random forests                      |
|                | Regression tree                     |
| Session 8      | Data cleaning                       |
|                | Cross-validation                    |

Tree, followed by supervised classification models like KNN Classification, Decision Trees, Logistic Regression, and Random Forests. It also included the unsupervised technique of Clustering. The training concluded with foundational data workflow topics: Data Cleaning and Cross-validation. Each Jupyter Notebook included the necessary code, the problem-specific datasets, and the execution environment, thus eliminating any technical effort for installation or configuration by the teachers. The supporting programming language was Python, but the materials employed the blocks-based programming extension noted earlier (see Fig. 3). This process automatically generated the corresponding Python code, enabling teachers to focus on the logical flow and data science goals rather than programming syntax errors.

## Measurements

Given the gap around data science curriculum and the need for K-12 teacher capacity building, the goal of DBR Meso Cycle 2 was to test the redesign of the data science materials that were aimed at the K-12 level. In accordance with other studies (Gobert et al., 2013; Wang et al., 2023) and systematic reviews (Mergianian et al., 2026), Research Question 1 utilized log data (executions) to understand learners' progression with the coding activities. Log data is often used in computer science/data science research *"because the learners' computer programming behavior can be an indicator of their learning performance and, at the same time, their performance as an artifact. Thus, researchers have become increasingly interested in analyzing learners' programming log data as they work on computer programming tasks"* (Song et al., 2021, p. 2).

Research Question 2 further examined error rates to identify where teachers encountered difficulty as they progressed through increasingly complex data science topics. In line with computational thinking and programming education studies focused on error rates (Mergianian et al., 2026; Švábenský et al., 2024), error rates provide insight into potential conceptual or procedural barriers that may not be visible through completion data alone

(Jadud, 2006; Luxton-Reilly et al., 2018; Mergianian et al., 2026; Švábenský et al., 2024; Zhou et al., 2021). Jadud (2006) examined patterns of repeated compilation to characterize the difficulties novice programmers encounter. Similar studies find that error rates in log data are among the measures used in computational thinking for those learning a new programming skill (Luxton-Reilly et al., 2018; Mergianian et al., 2026; Švábenský et al., 2024; Zhou et al., 2021), while error frequency alone does not establish conceptual understanding. In terms of advancing from higher education and supporting K-12 teachers, this measure was particularly important considering the interdisciplinary nature of data science, which can create challenges for teachers who must coordinate statistical reasoning, programming syntax, data interpretation, and model-based thinking. For the purpose of this study, an error is recorded when an inservice teacher triggered a code block that resulted in an incorrect output or a logical failure within the Jupyter Notebook.

## Results

### Research question 1: how do inservice K-12 teachers progress in the data science learning materials?

Figure 11 illustrates the engagement levels of the K-12 instructors throughout the training program by measuring the average number of code executions per Jupyter notebook. To understand participant progress across the learning materials, we aggregated the number of executions per notebook in order of their progression, visualizing the average execution frequency as a measure of the teacher progress along the timeline. Each data point on the line represents the mean frequency with which teachers triggered a code cell within a specific module, while the shaded area indicates the standard deviation that reflects variance in activity levels among the participants. Each Jupyter Notebook consisted of approximately 139.36 (SD = 115.35) executions per cell. A finding is that there was considerable variability in the execution of coding cells across all the Jupyter Notebooks. While the more mathematically complex topics found later in the professional development included more executions, the early stage notebooks did not reveal a consistent pattern, which may speak to the initial complexity and cognitive load of data science. This variability is noteworthy given the feedback of teachers wanting to be able to plan for classroom management, time efforts, and planning.

### Research Question 2: What are the error rates as K-12 teachers progress in the data science learning materials?

To understand the complexity across the notebooks, we considered both absolute errors and the error rate. In particular, we aggregated the number of errors per notebook in order of progress, showing the average error per notebook along the progress timeline. Error counts varied across notebooks, with some larger counts occurring in later modules ( $M = 125.99$ ;  $SD = 117.55$ , see Fig. 12). That said, some of the notebooks ranged considerably in their errors. For example, measures of association ( $M = 29.80$ ;  $STDEV = 20.00$ ) and simple linear regression ( $M = 61.25$ ;  $SD = 35.33$ ) were relatively low, while logistic regression ( $M = 31.83$ ;  $STDEV = 342.20$ ) and decision trees ( $M = 504.71$ ;  $SD = 722.91$ ) were considerably higher. In Fig. 13, error rates were computed by the number of executions resulting in an error divided by the total number of executions; this ratio represents

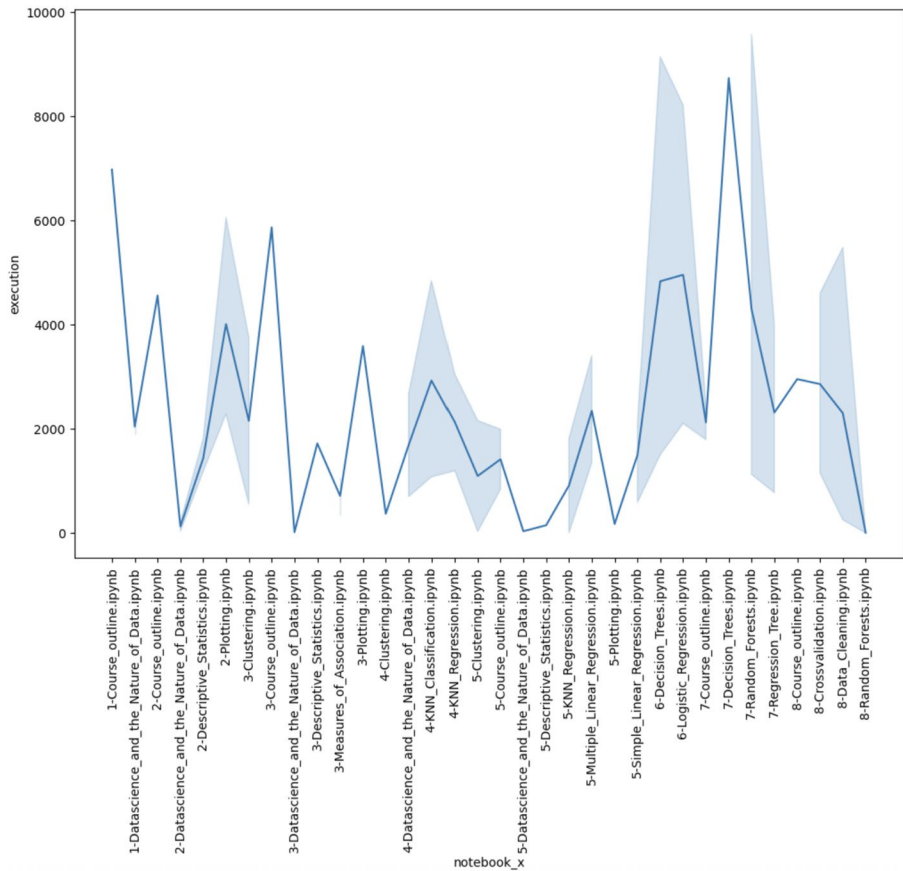


Fig. 11 Overall executions per data science notebook

the proportion of executions that produced errors (Mergianian et al., 2026). Despite the number of executions, the error rate was relatively low ( $M = 0.053$ ).

## Discussion and implications

### Design and the need for K-12 data science standards

Research shows that efforts towards formal approaches to data science are emerging within higher education and K-12 learning settings. This DBR project explores the design elements of open-access curriculum, with a specific exploration of data science design strategies that support K-12 teachers. DBR Meso Cycle 2 is important because concepts of data science education (statistics, coding, and computer science) are often embedded within existing programs, so teachers lack clarity about how these elements should be holistically presented at the K-12 level (Štuikys & Burbaitė, 2024). For example, a variety of studies apply data science via data exploration (Israel-Fishelson et al., 2025), information visualizations (Frischemeier et al., 2021; Sanei et al., 2024), or “informal statistical inference”

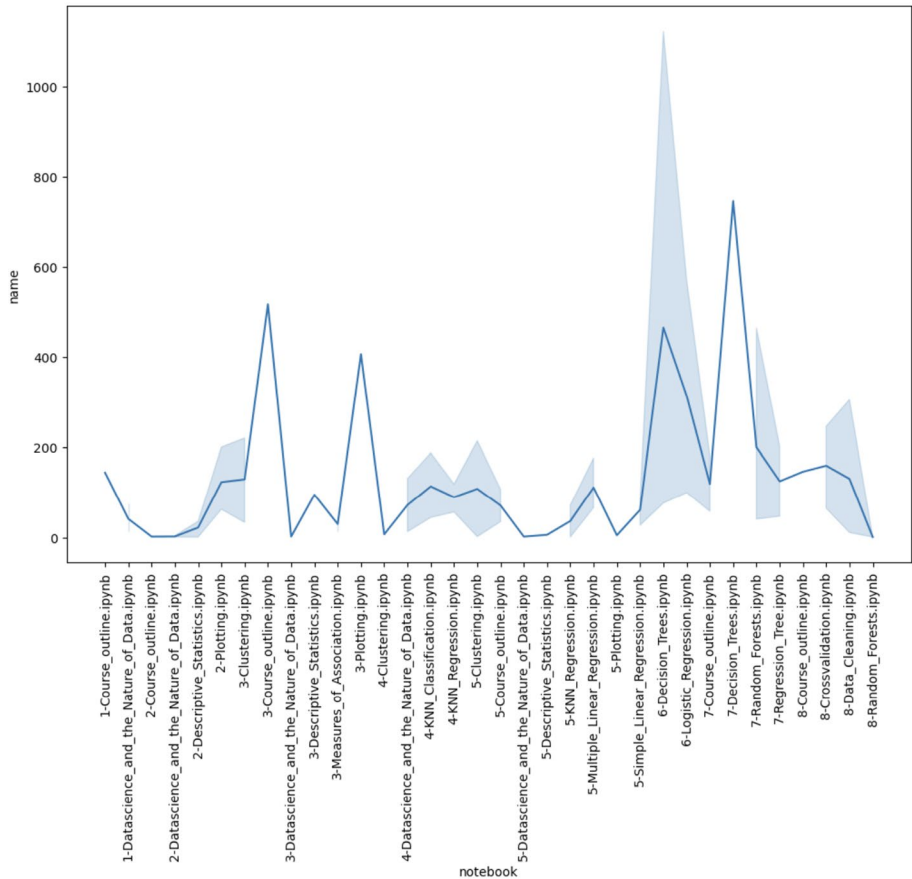


**Fig. 12** Overall errors per data science notebook

(Watson et al., 2024, p. 819) within various STEM classes. One of the related challenges of K-12 data science is the lack of standardized curriculum (Rauber & Gresse von Wangenheim, 2022); instead, the data science resources are largely ad-hoc learning platforms (e.g., MOOCs, Youtube tutorials) (Dogucu et al., 2025). In response, recent work by the American Statistical Association and others (Martinez & LaLonde, 2020; Rauber & Gresse von Wangenheim, 2022) supports more accessible and standardized resources as a way to establish a foundation for data science education. To address this gap, this project presents specific design features that should be considered, along with the empirical evidence of how inservice K-12 teachers progress (RQ1) and error rates as K-12 teachers progress in a data science curriculum (RQ2).

Although interest in K-12 data science has emerged in recent years, the Institute of Educational Science (2025) notes a gap regarding “*the use of evidence-based practices in the design and implementation of educational programs in these areas [data science]—while acknowledging that, as a field, we continue to learn together what works best, for whom, and under what conditions.*”. As the field looks for more evidence-based practices that advance theory and practice, this DBR Meso Cycle 2 project describes the challenges and opportunities as it explores open-access data science education for K-12 teachers. From

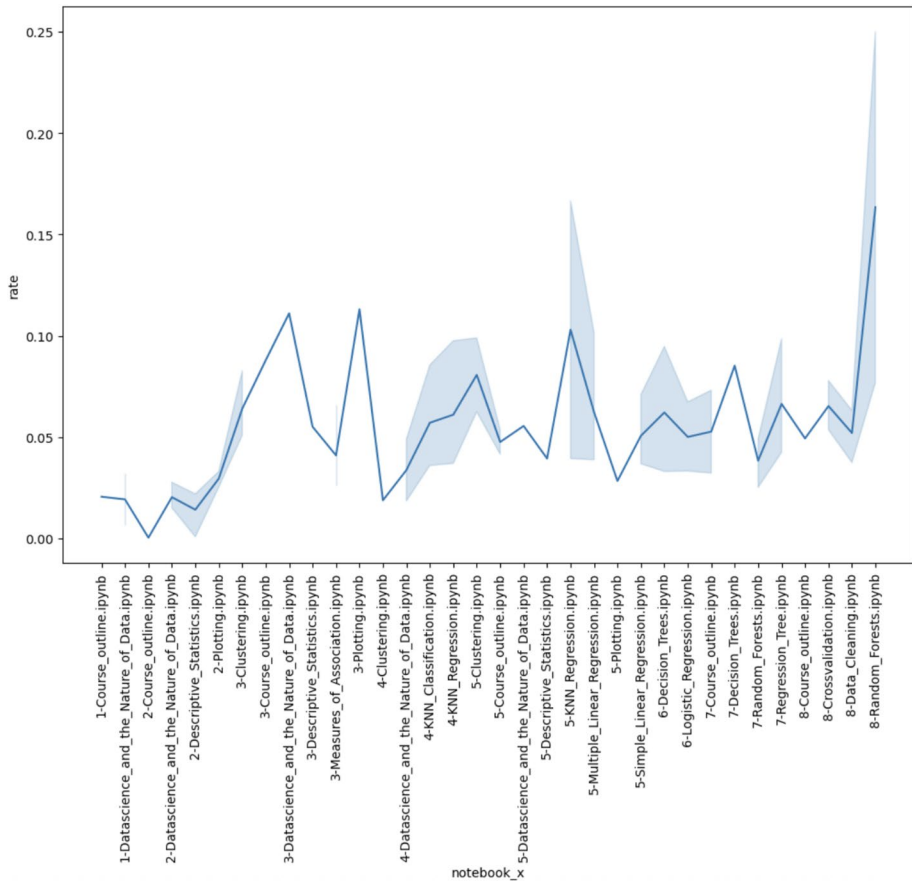


**Fig. 13** Error rates per data science notebook

a design perspective, one of the primary issues encountered in the design cycle was the interdisciplinary nature of data science and the lack of established theory, standards, and design learning features at different educational levels, especially for secondary education contexts. Indeed, this interdisciplinary approach makes it difficult for educators (teachers, learning designers) to identify specific skillsets needed for data science education, as well as how to translate data science for K-12 audiences (Rosenberg & Jones, 2024; Weiland & Engledowl, 2022). In terms of the progression (RQ1) and error rates (RQ2), the results found inconsistency across the data science notebooks, especially as the data science became increasingly complex. An implication for this DBR study is for governing bodies (federal, state, and professional organizations) to identify standards related to data science at various grade levels. Initial efforts towards curriculum frameworks for data science at the undergraduate and high school levels have emerged (ACM, 2022; Fisher et al., 2019; National Academies of Sciences, Engineering, & Medicine, 2023); however, further empirical evaluation is needed to establish how these frameworks can be implemented in K-12 classrooms (Institute of Educational Sciences, 2025; Israel-Fishelson et al., 2025; Lee & Delaney, 2022). For policymakers and standards-setting organizations, findings

underscore the need for K-12 data science frameworks that move beyond lists of topics to specify developmentally appropriate practices, such as data modeling as sense-making, iterative refinement of analytic decisions, and interpretation under uncertainty. Standards that articulate both what students should learn and how data science practices can be enacted in constrained classroom contexts are essential for guiding curriculum development, evaluation, and scaling. In this respect, the teachers' progression and error data were not treated as the determinant for whether K-12 data science education is necessary, which is established by the standards, policy, and framework efforts described above; rather, these data revealed where conceptual understanding of the interdisciplinary content breaks down, information that topic frameworks alone cannot provide because teachers ultimately translate curriculum into instruction.

To further address gaps in data science education, data science at the K-12 level *“requires increased professional development opportunities for teachers to master the practice and instruction of novel data science skills”* (Lee & Delaney, 2022, p. 96). However, the learning opportunities and resources are limited for current educators who wish to engage in data science education. For example, Israel-Fishelson et al. (2024) analyzed the 12 curricula listed by the ‘Data Science for Everyone Initiative’ and identified only four that met their inclusion criteria for a comprehensive, classroom-ready high school curriculum: Bootstrap: Data Science, CodeHS, Introduction to Data Science, and YouCubed. Moreover, these curricula cover only small portions of the International Data Science in Schools Framework (Fisher et al., 2019). A landscape scan presented by Drozda and colleagues reported that 0.1% of high school teachers and 3.6% of high school students participated in data science programs (National Academies of Sciences, Engineering, & Medicine, 2023). In terms of the gap related to learning design for K-12 data science materials needed for data science education, we proffer the following design guidelines that should be considered: (a) scope, (b) situated within a case-based pedagogy, (c) data that is aligned with the primary problem to solve, (d) goal-based, and (e) interactive. For curriculum developers, these guidelines can be enacted as a design sequence that can audit and simplify the content scope with readability and AI curriculum tools before verification by a subject matter expert, standardize a common module template, situate modules within a shared case throughline, align or re-theme datasets to the case, and segment each module into explicit, interactive goals.

## Goal-based learning and classroom management in K-12 data science education

The integration of a goal-based learning approach in DBR Meso Cycle 2 highlights its potential as a meaningful and adaptable design strategy for K-12 data science education. Informed by research on self-regulated learning support in programming (Shin & Song, 2022), the goal-based format segments multi-step processes into manageable coding units. This design is intended to support learners, particularly novices encountering programming and statistical reasoning for the first time, in developing fluency with foundational concepts before progressing to more advanced tasks. Through this segmentation, teachers can guide learners as they organize their knowledge, engage in guided practice, and reflect on their process through a structured learning progression.

Findings from the study suggest that learners were able to progress in the goals in the notebook, particularly for the earlier notebooks. Interpreted through this lens, the activity

patterns inform hypotheses for subsequent design cycles. Variability in executions (RQ1) and errors (RQ2) identifies notebooks for closer examination in the next design cycle. Additional measures are needed to distinguish conceptual difficulty from differences in task length, repeated execution, and available support. Error rates alone do not establish productive engagement or an effect of goal-based segmentation. Consistent with DBR, these patterns serve as conjecture-testing evidence that informs iterative refinement of the design features, not as claims of effectiveness at scale. From the perspective of K-12 classroom instruction, feedback from educators underscored the importance of clearly defined goals to enhance instructional clarity and navigability, especially for younger learners who may be less familiar with the content (Weeda et al., 2023). As each goal maps to a specific data science concept/skill (e.g., create a histogram using a package), teachers can sequence instruction according to classroom constraints or learner needs, without compromising coherence of the overall curriculum.

K-12 teachers in DBR Meso Cycle 2 reported in the LXD testing that visible, modular goals would allow them to better manage key classroom dynamics, such as management of students' progress, and instructional pacing. This includes embedding anticipated time requirements, common error patterns, and facilitation prompts alongside learner-facing content. Design standardization across modules is especially critical in interdisciplinary domains such as data science, where inconsistent representations or task structures can exacerbate cognitive burden and undermine transfer across topics.

Another implication is that the goal-based structure may serve as both a scaffold for learners and a practical tool for classroom management. In a data science field where formal standards and articulated learning progressions are still in development, the goal-based approach offers a principled yet flexible framework for organizing content. The modularity of the goal structure further enhances reusability and adaptability across varying educational contexts. This progressive approach may also be especially beneficial in open-access and informal learning environments, where educators often face diverse instructional conditions and require resources that are both customizable and pedagogically robust. Rather than positioning notebooks as linear coding tasks, teachers need to leverage goal-based structures to (a) make analytic intent explicit, (b) support formative assessment through visible progress markers, and (c) allocate instructional time for interpretation, comparison, and reflection on model outputs. In practice, a teacher could plan notebook activities around available class time: launch with the case video to establish the problem context, preview the 7–8 goals before coding begins, use goal completion as formative checkpoints to identify where students stall, and close with the reflection questions. Given the variability in executions and errors across notebooks (RQ1 and 2), teachers could sequence foundational notebooks first and budget additional time or paired support for the later, more error-prone notebooks.

## **Use of artificial intelligence in design for data science open-educational resources**

To support the redesign process, another outcome relates to how artificial intelligence tools were used in DBR Meso Cycle 2, especially as the design team considered how open-educational resources require adaptation across various learning contexts. Indeed, socio-technical perspectives of learning suggest that open-educational resource content should not employ a 'one-size-fits-all' approach, but instead align with contextual considerations

such as the cultural perspectives and various learner profiles (Kimmons, 2021; Kimmons & Irvine, 2023). While there is considerable discourse on how students can use AI through adaptive scaffolding and other classroom activities (Bai et al., 2025; Matsuda et al., 2020; Umutlu & Gursoy, 2022), this project adds to the growing body of literature on the use of AI and its application to learning design. In the current project, the goal was to redesign data science materials that could be adopted by inservice teachers in urban, suburban, or rural schools. Once the designers considered the various classroom settings, prompt engineering through AI tools (Diffit, Twee) helped refine the reading level and address other contextual learning considerations. Rather than expending considerable time trying to identify key concepts through a think-aloud with a subject matter expert at the beginning of the project, the AI output focused the subject matter expert discussions and helped organize the initial design work (e.g., needs assessment). Additionally, this DBR project describes how AI and strategic prompt engineering can be used as a ‘starting point’ for the design of complex learning topics, especially to (a) repurpose learning materials for diverse audiences, (b) work with the subject matter expert to refine content scope and (c) serve as a baseline for the design team to later develop supplementary resources (e.g., multimedia, scaffolds) that would reinforce key concepts.

Another important implication was how AI was used later in the design process, namely how the tool helped to refine data science materials through prompt engineering that was focused on learner characteristics and potential future use of the open-educational resource (e.g., prior knowledge, reading level). Beyond just support for subject matter expert discussions and scope, the project leveraged AI to personalize the iris dataset to align with the case-based curriculum that was essential in DBR Meso Cycle 2. Researchers in problem-solving note that one challenge of inquiry-based instruction is the effort required to design a detailed case and later align appropriate supplementary resources, especially as learners engage in various pathways during their self-directed problem-solving (Gómez & Suárez, 2020; Long & Koehler, 2024; Tawfik et al., 2021). As it relates to data science education, additional studies suggest challenges for educators as they attempt to identify ‘clean’ datasets that are sufficiently large, accessible, comprehensive, and engaging to diverse learners’ interests (Kang et al., 2026; Rosenberg et al., 2022). The integration of artificial intelligence (AI) into the design presents a significant opportunity for instructional designers who wish to adapt datasets and other supplemental materials. Specifically, this DBR project highlights how AI tools can be used to adapt materials for different instructional contexts, which can aid open-educational resource iterative design and dissemination across different learning contexts and interests. For example, teachers and curriculum developers can reuse the dataset personalization procedure to re-theme the same analytic activities around issues salient to their own communities (e.g., traffic, water quality) or repeat the AI-assisted simplification process to adjust reading levels for a specific grade band, provided that AI-generated adaptations are verified by teachers or subject matter experts before classroom use.

## Limitations and future studies

While the research provides insight into the evolving landscape of K-12 data science education, there are other opportunities to build on this research. The research was an extension of a meso cycle focused on higher education learners, so the transition to K-12 teachers caused the design team to focus on how teachers interacted with concepts (log data;

error rates) in the data science notebooks. However, additional data could be gathered on the teachers themselves. One study could explore how the more advanced topics that resulted in errors could be scaffolded. The results suggest that teachers may be able to progress to the foundational elements of data science (e.g.—nature of data, visualization), but may need differentiated support for topics as teachers progress to the prediction and clustering techniques of data science. In terms of additional studies focused on K-12 teachers, the interdisciplinary nature is a key element of data science, so it is closely aligned with STEM or other teachers who wish to apply elements of computational thinking, statistics, and machine learning. A future study could therefore explore how different teachers engage with data science and how this might differ based on the domain (e.g., economics). Additionally, as this meso cycle involved a small sample of inservice teachers ( $N = 5$ ) within a single district, the findings are not intended to establish the effectiveness of the design or its performance at scale. Consistent with the goals of DBR, the contribution is a set of transferable design features and a design process, surfaced through teachers' interaction with the materials, that can be tested in subsequent cycles and contexts.

Future research could also explore other supplemental learning resources used in data science education. In the prior meso cycle, the goal was to extend an open-ended invitation for teachers to take part in the professional development. Because the research suggests a fragmented approach to data science education is often common (Lewis & Stoyanovich, 2022), a future study could extend this DBR project via artifact analysis or other qualitative means that explore the specific lessons, materials, and learning resources that are currently used for data science. This would provide important insight into what teachers adopt and their rationale, which could help address the resource gap outlined in prior studies (Drozda et al., 2022). This would also inform the design for open-education data science resources that wish to scale to a broader audience and across diverse learning contexts.

## Conclusion

Despite the rapid expansion of data science education in recent years, research and instructional resources remain concentrated at the higher education level, leaving K-12 teachers with limited empirical guidance for integrating data science into their classrooms (Barany et al., 2024; Dogucu et al., 2025; Olney & Fleming, 2021; Rosenberg & Jones, 2024). This gap in learning design and curriculum likely stems from the inherently interdisciplinary nature of data science, which draws together statistics, computer science, machine learning, data visualization, and domain-specific reasoning. Such breadth demands that content be deliberately scoped, carefully scaffolded, and intentionally rendered teachable for K-12 contexts.

Through a design-based research cycle, this study re-leveled an existing data science open-educational resource for K-12 teacher use. Its central contribution lies not in proposing an additional set of data science design features in the abstract, but in empirically documenting how such features emerged through the iterative redesign process. By positioning teachers as the proximal users and enactors of the resource, the study surfaced design constraints central to classroom implementation, including content scope, instructional pacing, scaffolding, coding difficulty, case alignment, dataset relevance, and classroom management. The findings suggest that teacher-facing data science OERs should be scoped for K-12 contexts, standardized in structure across interactive notebooks, situated within case-based learning, supported by adaptable and contextually relevant datasets, and organized through goal-based structures that render analytic intent visible.

Beyond these design features, the study also demonstrates that artificial intelligence can meaningfully support the instructional design process, reducing reading complexity, focusing subject-matter-expert review, and adapting datasets for illustrative instructional scenarios. In doing so, this study extends the role of AI in education beyond student-facing scaffolding to encompass the work of curriculum and resource development. Taken together, these findings indicate that progress in K-12 data science education will depend less on the continual addition of topics and tools than on the establishment of coherent standards, sustained investment in teacher professional development, and the cultivation of empirically grounded, teacher-facing design features. Future design cycles can build on this work by examining how K-12 students engage with these materials directly, and how teachers adapt them across subjects, grade levels, and instructional settings.

**Funding** This material is based upon work supported by the National Science Foundation under Grant No. 1918751. Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## Declarations

**Competing interests** The authors report there are no competing interests to declare.

**Ethical approval** This study was approved by the Institutional Review Board of the University of Memphis.

**Informed Consent** Informed consent was obtained from all individual participants included in the study.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- ACM. (2022). Computing competencies for undergraduate data science curricula. *Association for Computing Machinery (ACM)*. <https://doi.org/10.1145/3453538>
- Adams, S. P., & Du Preez, R. (2022). Supporting student engagement through the gamification of learning activities: A design-based research approach. *Technology, Knowledge and Learning*, 27(1), 119–138. <https://doi.org/10.1007/s10758-021-09500-x>
- Armoni, M., Meerbaum-Salant, O., & Ben-Ari, M. (2015). From scratch to “real” programming. *ACM Transactions on Computing Education*, 14(4), 1–15. <https://doi.org/10.1145/2677087>
- Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, 70(2), 181–214.
- Barany, A., Scarola, A. D., Acquah, A., Reza, S. M., Johnson, M. A., & Walker, J. (2024). Learning designs that empower: Navigating sandbox data science at the intersection of computing, big data and social media. *Information and Learning Science*, 125(10), 794–812. <https://doi.org/10.1108/ils-12-2023-0211>
- Bai, S., Lo, C. K., & Yang, C. (2025). Enhancing instructional design learning: A comparative study of scaffolding by a 5E instructional model-informed artificial intelligence chatbot and a human teacher. *Interactive Learning Environments*, 33(3), 2738–2757. <https://doi.org/10.1080/10494820.2024.2420184>
- Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). *Pre-K–12 guidelines for assessment and instruction in statistics education II (GAISE II): A framework for statistics and data science education* (Vol. Vol. 2). American Statistical Association.

- Bau, D., Gray, J., Kelleher, C., Sheldon, J., & Turbak, F. (2017). Learnable programming: Blocks and beyond. *Communications of the ACM*, 60(6), 72–80. <https://doi.org/10.1145/3015455>
- Çakiroğlu, Ü., & Bilgi, Ş. (2022). Exploring intrinsic cognitive load in the programming process: A two dimensional approach based on element interactivity. *Interactive Learning Environments*, 32(5), 2045–2063. <https://doi.org/10.1080/10494820.2022.2137527>
- Cao, L. (2018). Data science: A comprehensive overview. *ACM Computing Surveys*, 50(3), 1–42. <https://doi.org/10.1145/3076253>
- Ciubotariu, I. I., & Bosch, G. (2023). Teaching students to R3ason, not merely to solve problem sets: The role of philosophy and visual data communication in accessible data science education. *PLoS Computational Biology*, 19(6), Article e1011160. <https://doi.org/10.1371/journal.pcbi.1011160>
- Dogucu, M., Demirci, S., Bendekgey, H., Ricci, F. Z., & Medina, C. M. (2025). A systematic literature review of undergraduate data science education research. *Journal of Statistics and Data Science Education*, 33(4), 459–471. <https://doi.org/10.1080/26939169.2025.2486656>
- Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4), 745–766. <https://doi.org/10.1080/10618600.2017.1384734>
- Drozda, Z., Johnstone, D., & Van Horne, B. (2022). Previewing the national landscape of K-12 data science implementation. *Workshop on Foundations of Data Science for Students in Grades K-12]. National Academies*.
- Ertmer, P., Glazewski, K. D., Jones, D., Ottenbreit-Leftwich, A., Goktas, Y., Collins, K., & Kocaman, A. (2009). Facilitating technology-enhanced problem-based learning (PBL) in the middle school classroom: An examination of how and why teachers adapt. *Journal of Interactive Learning Research*, 20(1), 35–54.
- Fisher, N., Anand, A., Bailer, J., Bailey, J., & Burr, W. (2019). *Curriculum frameworks for introductory data science*. IDSSP. <http://www.idssp.org/files>
- Frischmeier, D., Biehler, R., Podworny, S., & Budde, L. (2021). A first introduction to data science education in secondary schools: Teaching and learning about data exploration with CODAP using survey data. *Teaching Statistics*, 43(S1), S182–S189. <https://doi.org/10.1111/test.12283>
- Fry, K., & Makar, K. (2021). How could we teach data science in primary school? *Teaching Statistics*, 43, S173–S181. <https://doi.org/10.1111/test.12259>
- Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, 22(4), 521–563. <https://doi.org/10.1080/10508406.2013.837391>
- Gómez, R. L., & Suárez, A. M. (2020). Do inquiry-based teaching and school climate influence science achievement and critical thinking? Evidence from PISA 2015. *International Journal of STEM Education*, 7(1), 1–11. <https://doi.org/10.1186/s40594-020-00240-5>
- Groth, R. E. (2025). Data science teacher education goals: Essential elements of pre-K-12 data science curriculum implementation. *Journal of Statistics and Data Science Education: An Official Journal of the of the American Statistical Association*. <https://doi.org/10.1080/26939169.2025.2526627>
- Herro, D., Madison, M., Irgens, G. A., Hirsch, S., Abimbade, O., & Adisa, O. (2022). Exploring elementary teachers' perceptions of data science and curriculum co-design through professional development. *Journal of Technology and Teacher Education*, 30(4), 493–525.
- Herro, D., & Madison, M. (2026). The impact of professional development on elementary teachers' self-efficacy and interest in data science education. *Professional Development in Education*. <https://doi.org/10.1080/19415257.2026.2630864>
- Hoadley, C., & Campos, F. C. (2022). Design-based research: What it is and why it matters to studying online learning. *Educational Psychologist*, 57(3), 207–220. <https://doi.org/10.1080/00461520.2022.2079128>
- Hu, Y., Chen, C.-H., & Su, C.-Y. (2021). Exploring the effectiveness and moderators of block-based visual programming on student learning: A meta-analysis. *Journal of Educational Computing Research*, 58(8), 1467–1493. <https://doi.org/10.1177/0735633120945935>
- Institute of Educational Sciences. (2025). *Data science education*. <https://ies.ed.gov/use-work/supporting-recovery-with-evidence-based-practices/data-science-education#step-4-implementation>
- Institute of Educational Sciences. (2025). *Postsecondary education*. Postsecondary Education. <https://ies.ed.gov/postsecondary-education>
- Irizarry, R. A. (2020). The role of academia in data science education. *Harvard Data Science Review*. <https://doi.org/10.1162/99608f92.dd363929>
- Israel-Fishelson, R., & Weintrop, D. (2026). Authenticity in a high school data science curriculum. *Educational Technology Research and Development*. <https://doi.org/10.1007/s11423-026-10622-1>

- Israel-Fishelson, R., Moon, P. F., Tabak, R., & Weintrop, D. (2025). Preparing students to meet their data: An evaluation of K-12 data science tools. *Behaviour & Information Technology*, 44(5), 934–953. <https://doi.org/10.1080/0144929x.2023.2295956>
- Israel-Fishelson, R., Moon, P., Tabak, R., & Weintrop, D. (2024). Understanding the data in K-12 data science. *Harvard Data Science Review*. <https://doi.org/10.1162/99608f92.4f3ac3da>
- Jadud, M. C. (2006, September 9). Methods and tools for exploring novice compilation behaviour. *Proceedings of the Second International Workshop on Computing Education Research*. ICER06: International Computing Education Research Workshop 2006, Canterbury United Kingdom, 2006/9/9–2006/9/10. <https://doi.org/10.1145/1151588.1151600>
- Jung, J., & Mercier, E. (2023). Design-based implementation research: Milestones and trade-offs in designing a collaborative representation tool for engineering classrooms. *Educational Technology Research and Development: ETR & D*, 71(6), 2457–2481. <https://doi.org/10.1007/s11423-023-10288-z>
- Kang, J., Na, C., Diederich, M., Swanson, H., & Yan, L. (2026). Cultivating science data literacy in K-16 science education through data practices: A systematic review. *Instructional Science*, 54(2), 50. <https://doi.org/10.1007/s11251-026-09803-5>
- Kilic, E., & Yildirim, Z. (2012). Cognitive load and goal based scenario centered 3D multimedia learning environment: Learners' motivation, satisfaction and mental effort. *Journal of Educational Computing Research*, 47(3), 329–349.
- Kimmons, R. (2021). Shifting to open education in our digital learning. *Educational Technology Research and Development*, 69(1), 379–381. <https://doi.org/10.1007/s11423-021-09949-8>
- Kimmons, R., & Irvine, J. (2023). Future directions in OER. In D. Otto, G. Scharnberg, M. Kerres, & O. Zawacki-Richter (Eds.), *Distributed learning ecosystems: Concepts, resources, and repositories* (pp. 183–199). Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-38703-7>
- Koehler, A. A., & Vilarinho-Pereira, D. R. (2021). Using social media affordances to support ill-structured problem-solving skills: Considering possibilities and challenges. *Educational Technology Research and Development*. <https://doi.org/10.1007/s11423-021-10060-1>
- Kuroki, M. (2023). Integrating data science into an econometrics course with a Kaggle competition. *The Journal of Economic Education*. <https://doi.org/10.1080/00220485.2023.2220695>
- Lee, V. R., & Delaney, V. (2022). Identifying the content, lesson structure, and data use within pre-collegiate data science curricula. *Journal of Science Education and Technology*, 31(1), 81–98. <https://doi.org/10.1007/s10956-021-09932-1>
- Lewis, A., & Stoyanovich, J. (2022). Teaching responsible data science: Charting new pedagogical territory. *International Journal of Artificial Intelligence in Education*, 32(3), 783–807. <https://doi.org/10.1007/s40593-021-00241-7>
- Long, Y., & Koehler, A. A. (2024). Exploring expert instructors' conceptualization and teaching practices in asynchronous online discussions during case-based learning: A multiple case study. *Journal of Computing in Higher Education*. <https://doi.org/10.1007/s12528-024-09405-5>
- Loyens, S., van Meerten, J. E., Schaap, L., & Wijnia, L. (2023). Situating higher-order, critical, and critical-analytic thinking in problem- and project-based learning environments: A systematic review. *Educational Psychology Review*. <https://doi.org/10.1007/s10648-023-09757-x>
- Luxton-Reilly, A., Simon, Albluwi, I., Becker, B. A., Giannakos, M., Kumar, A. N., Ott, L., Paterson, J., Scott, M. J., Sheard, J., & Szabo, C. (2018). Introductory programming: a systematic literature review. *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, 55–106. <https://doi.org/10.1145/3293881.3295779>
- Martinez, W., & LaLonde, D. (2020). Data science for everyone starts in kindergarten: Strategies and initiatives from the American statistical association. *Harvard Data Science Review*, 2(3), 1–7. <https://doi.org/10.1162/99608f92.7a9f2f4d>
- Matsuda, N., Weng, W., & Wall, N. (2020). The effect of metacognitive scaffolding for learning by teaching a teachable agent. *International Journal of Artificial Intelligence in Education*, 30(1), 1–37. <https://doi.org/10.1007/s40593-019-00190-2>
- McFarland, D. A., Khanna, S., Domingue, B. W., & Pardos, Z. A. (2021). Education data science: Past, present, future. *AERA Open*, 7(23328584211052056), 233285842110520. <https://doi.org/10.1177/23328584211052055>
- McKenney, S., & Reeves, T. (2014). Educational design research. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (4th ed., pp. 131–140). Springer.
- Mergianian, C., Hienonen, N., Dindar, M., & Vainikainen, M.-P. (2026). Assessing computational thinking using log data: A scoping review. *Lecture notes in computer science* (pp. 306–320). Springer Nature Switzerland. [https://doi.org/10.1007/978-3-032-03870-8\\_21](https://doi.org/10.1007/978-3-032-03870-8_21)

- Meulenbroeks, R., van Rijn, R., & Reijkerkerk, M. (2023). Fostering secondary school science students' intrinsic motivation by inquiry-based learning. *Research in Science Education*, 54(3), 339–358. <https://doi.org/10.1007/s11165-023-10139-0>
- National Academies of Sciences, Engineering, and Medicine. (2023). *Foundations of data science for students in grades K-12: Proceedings of a workshop*. National Academies Press. <https://doi.org/10.17226/26852>
- National Research Council. (2014). *Developing assessments for the Next Generation Science Standards* (J. Pellegrino, M. Wilson, J. Koenig, & A. Beatty, Eds.). The National Academies Press. <https://doi.org/10.17226/18409>
- O'Hara, K. J., Blank, D. S., & Marshall, J. B. (2015). Computational notebooks for AI education. *The Florida AI Research Society*. <https://doi.org/10.13140/2.1.2434.5928>
- Olney, A. M., & Fleming, S. D. (2021). JupyterLab extensions for blocks programming, self-explanations, and HTML injection. In T. W. Price & S. San Pedro (Eds.), *Joint Proceedings of the Workshops at the 14th International Conference on Educational Data Mining*, Vol. 3051, CSEDM-8. <https://CEUR-Ws.org>
- Payne, L. A., Tawfik, A., & Olney, A. (2021). Datawhys phase 1: Problem solving to facilitate data science & STEM learning among summer interns. *International Journal of Designs for Learning*, 12(3), 102–117. <https://doi.org/10.14434/ijdl.v12i3.31555>
- Rauber, M. F., & Gresse von Wangenheim, C. (2022). Assessing the learning of machine learning in K-12: A ten-year systematic mapping. *Informatics in Education*, 22(2), 295–328. <https://doi.org/10.15388/infedu.2023.11>
- Resnick, M., Maloney, J., Monroy-Hernández, A., Rusk, N., Eastmond, E., Brennan, K., Millner, A., Rosenbaum, E., Silver, J., Silverman, B., & Kafai, Y. (2009). Scratch: Programming for all. *Communications of the ACM*, 52(11), 60–67. <https://doi.org/10.1145/1592761.1592779>
- Richardson, J. C., Castellanos-Reyes, D., Janakiraman, S., & Duha, M. S. U. (2025). Using design-based research to gauge the impact of an open educational resource repository to support online teaching and learning. *Journal of Computing in Higher Education*, 38(3), 1076–1106. <https://doi.org/10.1007/s12528-025-09483-z>
- Romero, C., & Ventura, S. (2017). Educational data science in massive open online courses: Educational data science in massive open online courses. *Data Mining and Knowledge Discovery*, 7(1), Article e1187. <https://doi.org/10.1002/widm.1187>
- Rosenberg, J. M., & Jones, R. S. (2024). Data science learning in grades K-12: Synthesizing research across divides. *Harvard Data Science Review*, 6(3), 2–30. <https://doi.org/10.1162/99608f92.b1233596>
- Rosenberg, J. M., Schultheis, E. H., Kjelvik, M. K., Reedy, A., & Sultana, O. (2022). Big data, big changes? The technologies and sources of data used in science classrooms. *British Journal of Educational Technology*, 53(5), 1179–1201. <https://doi.org/10.1111/bjet.13245>
- Sanci, H., Kahn, J. B., Yalcinkaya, R., Jiang, S., & Wang, C. (2024). Examining how students code with socioscientific data to tell stories about climate change. *Journal of Science Education and Technology*, 33(2), 161–177. <https://doi.org/10.1007/s10956-023-10054-z>
- Schmidt, M., & Glaser, N. (2024). Design principles for participatory XR design with autistic users. *The Journal of Applied Instructional Design*, 10(59668/2033), 18643.
- Schoute, E. C., & Alexander, P. A. (2025). A critical analysis of critical thinking interventions in higher education. *Review of Educational Research*. <https://doi.org/10.3102/00346543251352539>
- Shin, Y., & Song, D. (2022). The effects of self-regulated learning support on learners' task performance and cognitive load in computer programming. *Journal of Educational Computing Research*, 60(6), 1490–1513. <https://doi.org/10.1177/07356331211052632>
- Singh, A., Fariha, A., Brooks, C., Soares, G., Henley, A. Z., Tiwari, A., Chethan, Choi, H., & Gulwani, S. (2024). Investigating student mistakes in introductory data science programming. *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, 1258–1264. <https://doi.org/10.1145/3626252.3630884>
- Song, D., Hong, H., & Oh, E. Y. (2021). Applying computational analysis of novice learners' computer programming patterns to reveal self-regulated learning, computational thinking, and learning performance. *Computers in Human Behavior*, 120(106746), Article 106746. <https://doi.org/10.1016/j.chb.2021.106746>
- Štuitkys, V., & Burbaitė, R. (2024). Introducing data science concepts into STEM-driven computer science education. *Evolution of STEM-driven computer science education* (pp. 217–245). Springer Nature Switzerland. [https://doi.org/10.1007/978-3-031-48235-9\\_8](https://doi.org/10.1007/978-3-031-48235-9_8)
- Švábenský, V., Pankiewicz, M., Zhang, J., Cloude, E. B., Baker, R. S., & Fouh, E. (2024). Comparison of three programming error measures for explaining variability in CS1 grades. *Proceedings of the 2024*

- on *Innovation and Technology in Computer Science Education V. 1*, 87–93. <https://doi.org/10.1145/3649217.3653563>
- Tawfik, A., & Kolodner, J. (2016). Systematizing scaffolding for problem-based learning: A view from case-based reasoning. *Interdisciplinary Journal of Problem-Based Learning*. <https://doi.org/10.7771/1541-5015.1608>
- Tawfik, A., Gishbaugher, J. J., Gatewood, J., & Arrington, T. L. (2021). How K-12 teachers adapt problem-based learning over time. *Interdisciplinary Journal of Problem-Based Learning*, 15(1), Article n1. <https://doi.org/10.14434/ijpbl.v15i1.29662>
- Tawfik, A., Payne, L., & Olney, A. M. (2024). Scaffolding computational thinking through block coding: A learner experience design study. *Technology, Knowledge and Learning*, 29(1), 21–43. <https://doi.org/10.1007/s10758-022-09636-4>
- Tawfik, A., Payne, L., Olney, A., & Ketter, H. (2023). Exploring the relationship between usability and cognitive load in data science education. *Journal of Applied Instructional Design*, 12(3), 165–178. <https://doi.org/10.59668/515.13078>
- Umutlu, D., & Gursay, M. E. (2022). Leveraging artificial intelligence techniques for effective scaffolding of personalized learning in workplaces. *Advances in analytics for learning and teaching* (pp. 59–76). Springer International Publishing. [https://doi.org/10.1007/978-3-031-14489-9\\_4](https://doi.org/10.1007/978-3-031-14489-9_4)
- Wang, K. D., Cock, J. M., Käser, T., & Bumbacher, E. (2023). A systematic review of empirical studies using log data from open-ended learning environments to measure science and engineering practices. *British Journal of Educational Technology*, 54(1), 192–221. <https://doi.org/10.1111/bjet.13289>
- Watson, J., Fitzallen, N., & Kelly, B. (2024). Interpolation and extrapolation in year 5 STEM activities: Exploring data about viscosity without advanced statistics. *Mathematics Education Research Journal*, 36(4), 815–848. <https://doi.org/10.1007/s13394-023-00473-x>
- Weeda, R., Smetsers, S., & Barendsen, E. (2023). Unraveling novices' code composition difficulties. *Computer Science Education*. <https://doi.org/10.1080/08993408.2023.2169067>
- Weiland, T., & Engledow, C. (2022). Transforming curriculum and building capacity in K–12 data science education. *Harvard Data Science Review*, 4(4), 2–20. <https://doi.org/10.1162/99608f92.7fea779a>
- West, J. (2018). Teaching data science: An objective approach to curriculum validation. *Computer Science Education*. <https://doi.org/10.1080/08993408.2018.1486120>
- Wong, G.K.-W., Reichert, F., & Law, N. (2023). Reorienting the assessment of digital literacy in the twenty-first century: A product-lifecycle and experience dependence perspective. *Educational Technology Research and Development: ETR & D*, 71(6), 2389–2412. <https://doi.org/10.1007/s11423-023-10278-1>
- Xu, Z., Ritzhaupt, A. D., Tian, F., & Umapathy, K. (2019). Block-based versus text-based programming environments on novice student learning outcomes: A meta-analysis study. *Computer Science Education*, 29(2–3), 177–204. <https://doi.org/10.1080/08993408.2019.1565233>
- Zhang, L., & Ma, Y. (2023). A study of the impact of project-based learning on student learning effects: A meta-analysis study. *Frontiers in Psychology*, 14, Article 1202728. <https://doi.org/10.3389/fpsyg.2023.1202728>
- Zheng, Y. (2023). ChatGPT for teaching and learning: An experience from data science education. *The 24th Annual Conference on Information Technology Education*, 66–72. <https://doi.org/10.1145/3585059.3611431>
- Zhou, Z., Wang, S., & Qian, Y. (2021). Learning from errors: Exploring the effectiveness of enhanced error messages in learning to program. *Frontiers in Psychology*, 12, Article 768962. <https://doi.org/10.3389/fpsyg.2021.768962>

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

**Andrew A. Tawfik** Ph.D., is an Associate Professor of Instructional Design & Technology at the University of Memphis, where he also serves as the director for the Instructional Design & Technology Studio. His research interests include problem-based learning, case-based reasoning, case library instructional design, and learning experience design (LXD).

**Dr. Luiz Barboza** holds a bachelor's degree in computer science from the Federal University of Pernambuco (UFPE), a master's degree from University of Sao Paulo (IPT/USP), and a Ph.D. from CESAR School, specializing in Data Science and AI education. He has over 20 years of technology industry experience, including leadership roles at Google and managing cloud and AI/ML projects across AWS, GCP, and Azure. Teaching in higher education since 2008, he served as a Visiting Professor at Florida International

University (FIU), collaborating with the Break-Through Tech nonprofit, followed by a postdoctoral appointment at the University of Memphis. He is currently a Professor at Alma College in Michigan. His research focuses on Data Science education, particularly for non-STEM students, with growing interests in GenAI, LLMs, and platforms such as ChatGPT, Claude, and Gemini.

**Andrew M. Olney** is a former Professor (retired 2025) in both the Institute for Intelligent Systems and Department of Psychology at the University of Memphis. Dr. Olney received a B.A. in Linguistics with Cognitive Science from University College London in 1998, an M.S. in Evolutionary and Adaptive Systems from the University of Sussex in 2001, and a Ph.D. in Computer Science from the University of Memphis in 2006. He is the former editor of the *Journal of Educational Data Mining* (2017–2022) and former Director/Associate Director of the Institute for Intelligent Systems at the University of Memphis (2006–2017), where he oversaw tenure track faculty, staff, and approximately 40 faculty affiliates. Dr. Olney has over 150 publications in the fields of artificial intelligence, education, and psychology, with a particular focus on using natural language processing to assess and enhance learning from text and learning through conversation. He has been an investigator on \$18M in federal grants (\$6M as PI) in these fields. His primary research interests are in natural language interfaces. Specific interests include language and knowledge acquisition (both human and AI) with applications in learning and performance support systems.

**Yeonji Jung** Ph.D., is an Assistant Professor of Learning Technologies & Performance Systems at Texas A&M University. Her research focuses on translating educational data into actionable insights and strategies that promote learner agency, meaningful collaboration, and responsive instructional practice.