Extended Abstract for the
Symposium on Integrating AI and Data Science into School Education Across Disciplines 2025

EXPLORING A US FRAMEWORK OF LEARNING PROGRESSIONS FOR K-12 DATA SCIENCE EDUCATION

Katherine M. Miller[1] & Michelle H. Wilkerson[2]
[1]The Concord Consortium, USA, kmiller@concord.org
[2]University of California Berkeley, USA

Focus Topic: AI and Data Science Competencies

Introduction

While data science education (DSE) is growing quickly as a field within educational research, it is still nascent, especially in regard to K-12 teaching and learning (Rosenberg & Jones, 2024). This creates a troubling dichotomy as uptake of data science education work among K-12 practitioners and developers accelerates (Drozda et al., 2022; National Academies of Sciences Engineering and Medicine [NASEM], 2023). Without coherent research guidance, such implementation occurs with little understanding about what learners need to know and be able to do with data, and when within the course of their learning these aspects are appropriate (Israel-Fishelson et al., 2023; NASEM, 2023; Rosenberg & Jones, 2024). As such, there is a need for a framework that conceptualizes data science learning at the K-12 level which could serve as a guide to policy makers, practitioners and researchers alike. In an attempt to build such a framework, the Concord Consortium and Data Science 4 Everyone joined together, with seed funding from NSF and the Valhalla Foundation to facilitate a series of workshops across the field with the goal of building consensus Learning Progressions (LP) for K-12 DSE.

Background

A recent U. S. National Academies workshop, Foundations of Data Science for Students in Grades K-12 established K-12 DSE as a crucial and growing field (NASEM, 2023). While calls for K-12 data science education research draw from a rich history of statistics education research including learning trajectory development particular to that field (e.g., Franklin & Bargagliotti, 2020), the National Academies workshop highlighted the need for bringing siloed fields of research together to build a more comprehensive, consensus understanding of the scope of data science education. The workshop framed the need for interdisciplinary research to inform areas as far-ranging as identifying the role for computational tools and technologies in data science education, understanding the needs for teacher preparation for data science education, answering questions about what data science education looks like in different school subjects and educational settings, and identifying and encouraging just and ethical approaches to data science education (NASEM, 2023).

Across a variety of international gatherings similar to the Foundations workshop (e.g., International Association for Statistics Education roundtable, International Society for the Learning Sciences workshop) and recent literature, participants, panelists, authors, and co-chairs point to the need for better understanding of research-based LPs, identifying the importance of overall coordination and clarity across data science education research (Miller & Yoon, 2023; NASEM, 2023; Rosenberg & Jones, 2024). As practitioners and providers tackle the challenge of creating materials and preparing students for the skills and understandings necessary for data fluency, the existence of research to guide their work is essential. Such a need hearkens back to the origins of LPs themselves. As the NSF's own Learning Progressions Footprints Conference stated in 2011 (National Science Foundation [NSF], 2011), learning progressions research grew out of the desire to answer the question: "How can we create research-based products that exert a positive influence on large-scale policy and classroom practice?" As that conference's accompanying materials made clear, LPs have sway across multiple contexts including, but not limited to, state and national standards, large-scale assessments, and classroom practice (NSF, 2011).

We adopt the language of LPs because they can inform, and hold sway across multiple contexts including, but not limited to, state and national standards, large-scale assessments, and classroom practice (NSF, 2011). While this work will be sensitive to the strengths and weaknesses of LPs as identified in the research literature, it will be equally sensitive to the needs of the multiple audiences which it will be designed to serve. Given the nascent state of the field of data science education, many

In S. Podworny & S. Schönbrodt (Eds.), *Towards Fostering AI and Data Science Literacy in Schools Across Disciplines. Proceedings of the 1st Symposium on Integrating AI and Data Science into School Education Across Disciplines (AIDEA1 2025)*, Salzburg, Austria.

different groups stand to benefit from the resulting framework, and the potential applications to which it may apply comprise an equally broad set.

Context and Framing

The effort to design consensus LPs for K-12 DSE began with an NSF funded workshop in October 2023. We brought together educational researchers, professional learning coaches, and practitioners to explore possibilities for the structure and functionality of LPs for DSE, as well as an initial attempt to map the content of those progressions. This workshop resulted in an initial high-level outline of strands of learning for DSE. Over the next several months, a series of design focus groups were held to gather expert input from a breadth of groups that represent the full spectrum of data theory and practice, spanning higher-education, industry, K-12 practitioners, and K-12 students themselves. These focus groups identified the knowledge, skills, dispositions, and critical thinking tools students should gain by the time they graduate as relates to data literacy, data analysis techniques, and other data-related technology. Each focus group created a prioritized list of competencies for data science, shared these lists with the larger community, and provided feedback on the initial high-level strands. The work of the focus groups led to a revision of the strands and additional ideas for the content and competencies that belong within each strand. A writing team then convened in November of 2024, to map the strands and content into competencies that built across grade bands. This draft of the full LPs then entered a period of review and feedback across the data science education community.

The K-12 Data Literacy and Data Science Education Model Learning Progressions

While the initial draft of the K-12 Data Literacy and Data Science Education Model Learning Progressions released in July 2025 is intended for public consumption, it was written with specific audience, use-cases and scoping guardrails in mind. This first draft was written as a subject-agnostic guide to data skills and fluency as fundamental learning objectives across multiple school subjects. The primary audience is state education agencies and leaders who are potentially writing or modifying standards or curriculum and seeking to answer the question “what is data science?” While the goal is to eventually make the LPs useful for educators, the current lack of subject articulation and supporting resources makes this initial version most relevant to policymakers and curriculum developers.

The current draft of the model LPs includes five strands representing Data Literacy & Responsibility, Creation & Curation, Data Analysis & Techniques, Interpreting Problems & Results, and Visualization & Communication, each of which contains 4 to 5 substrands. Table 1 depicts the strands, substrands, and descriptions of each.

Table 1: Framework for K-12 Data Literacy and Data Science Education Model Learning Progressions

Strand A Data Literacy and Responsibility This strand focuses on what data is and all of the ways students should think about and frame it as a concept and tool. The nature of data is complex, diverse, and humanistic. When engaging with data you must consider the form it takes, where it can come from, and what it can and should be used for. Working with data is non-linear and often raises new questions while seeking answers to others. Additionally the data process is influenced at all stages by the humans working with it which can lead to biases and concerns about ethics and responsibility. However, data can also be powerful for supporting the advancement of discovery or enactment of change.
A1. Nature of Data: The nature of data is complex, variably, humanistic, and often incomplete. Data can take many forms and may come from many different sources. Additionally, data is integral to the field of AI.
A2. Data Ethics and Responsibilities: The data process is influenced at all stages by the humans working with it which can lead to concerns about ethics and responsibility. It is important when working with data to consider the use risks as well as the benefits. Data can be powerful for supporting the advancement of discovery or enactment of change.
A3. Investigative Dispositions: Working with data is non-linear and often requires cycling between phases in various orders multiple times. The process of investigating with data often raises new questions while seeking answers to others. Additionally, data is influenced by the humans working with it and the contexts within which they work.

Strand B Creation and Curation

This strand focuses on where data comes from and how it should be collected, organized, and formatted in order to make it useful. Data collected from real world scenarios is often complex and messy, and whether it is collected firsthand, or retrieved second hand from an external source, it requires curation and cleaning before analysis. The context of data collection matters and affects the nature of errors in data collection. The methods and decisions made during data collection affect the usefulness of the data and its ability to answer different questions.

B1. Organization & Processing: In order for data to be useful for analysis and visualization, it often needs to be organized and formatted in particular ways. Organization can include both procedural cleaning up of errors or mistakes and processing or transforming the data through calculations and logic statements to create new or summative measures.

B2. Designing for Data Collection: The design of a data investigation is as important as the data collection process. Framing a data-based investigation requires identifying a problem or question to be explored. Additionally, the methods must be carefully chosen and the values and tradeoffs considered.

B3. Measurement & Datafication: The methods and decisions made during data collection affect the usefulness of the data and its ability to answer different questions. It is important to consider the potential effects of methodological decisions when collecting data and to determine the methodological decisions made by others when using secondary data. It is also important to consider ethical practices of using other's data.

B4. Complexity of Data: Data collected from real world scenarios is often complex across many dimensions including messiness, size, and structure. In order to be able to work with authentic real-world datasets of high complexity, these dimensions must be scaffolded such that increasingly higher levels of complexity are encountered as one approaches mastery.

Strand C Analysis & Modeling Techniques

This strand focuses on the process of analyzing data. Analyzing data includes many different techniques such as examining single and multi-variable patterns, measures of centrality, variability, and uncertainty. Knowing which techniques to use on which types of data to answer which questions is as important as the skills to conduct analysis techniques. Additionally, understanding simulation and the relational nature of data is important to the analysis process, as is the use of technological tools for analysis and modeling.

C1. Summarizing Data: Raw data often is not useful for answering questions, making claims, or telling a story. In order to derive understanding it is usually useful to have a summary of the data which provides measures of the centrality, spread, and shape of the dataset.

C2. Identifying Patterns and Relationships in Data: A primary use of data is in understanding patterns and relationships across different variables and scenarios. As all data contains variability it is important to understand and analyze distributions both within and across variables.

C3. Variability in Data: Variability is omnipresent within data and datasets. Working with data depends on understanding, explaining, and quantifying variability of all forms (variability within a group, between different groups, or between samples).

C4. Digital Tools of Data Analysis: While some datasets can be explored by hand, as they get bigger and more complex it becomes necessary to use digital tools for analyzing data. It is important to understand which tools to use for which application or scenario, the affordances and tradeoffs, and the ethical considerations of using certain tools.

C5. Models of Data: Interpreting, creating, and using models is a central component of working with data. Models are both a way to analyze data and a source of data.

Strand D Interpreting Problems and Results

This strand focuses on justification and explanation of reasoning when making inferences, claims, or suggestions from data within the context and processes of the dataset collection and analysis. An important component of interpreting results is understanding the relationship between questions, problems and datasets. Formulating a strong question or identifying a problem that can be addressed with data affects the opportunities for interpretation and results from the data. Additionally, the applicability of inferences and claims that are made are constrained by the sample, population, and context of the data.

D1. Making and Justifying Claims: As all data contains variability, it is important to use probabilistic thinking and language when making claims from data. This requires paying attention not only to patterns and comparisons within and across variables but also such things as expected and prior values, sample sizes, and significance.

D2. Problem Identification & Question Formation: Formulating a question or identifying a problem that can be addressed with data affects the opportunities for interpretation and results from the data. The ability to make and justify strong claims relies on identifying questions that are testable and can be answered with data. Additionally, identifying the uncertainty or limitations within the problem space is an important component of formulating conclusions.

D3. Generalization: Though there is often an instinct to use data to make large generalized claims, the applicability of inferences and claims that are made are constrained by the sample, population, and context of the data.

Strand E Visualization and Communication

This strand focuses on how to communicate about data through the creation and examination of visualizations. Visualizations are a vital component of the sensemaking process when working with data. Being able to communicate with and about data using visualizations that are clear and tailored to a purpose and audience are an important step for creating action and impact through data. Additionally important are skills and habits for how to read, interpret, and critique other's data communication, paying attention to context, audience and purpose.
E1. Representations and Dynamic Visualizations: The creation and interpretation of graphic and interactive visualizations are vital components of the sensemaking process when working with data. Working with data visualizations requires an understanding of conventional components and best practices along with graphical literacy and representational fluency.
E2. Data Storytelling: Being able to communicate with and about data using visualizations connected to a narrative is an important step for creating action and impact through data. Understanding the audience for the narrative is vital to clear communication.
E3. Acting on Data to Benefit Society: One of the ultimate goals of working with data is applying interpretation and conclusions to real-world problems and scenarios in order to engage in civic practice and enact positive change on the world.

Each of the substrands encompasses between three and eight core data concepts. Across the five strands, there are 82 data concepts that articulate the learning goals for data within K-12 education. Some examples of the data concepts included are, explaining significance, understanding modeling, measures of center, dynamic inferences, the investigative process, data cleaning, graphical literacy, and civic data practices. Each concept is divided into competencies across the five grade bands. Table 2 depicts an example of a single concept broken into its grade-relevant competencies.

Table 2: Example concept divided into competencies across grade bands

Substrand A1. Nature of Data **The nature of data is complex, variably, humanistic, and often incomplete. Data can take many forms and may come from many different sources. Additionally, data is integral to the field of AI.**					
Concepts	**K-2**	**3-5**	**6-8**	**9-10**	**11-12**
A1.2 Data are produced by people Recognize that data represent decisions about measurement and inclusion involving people who *are* and *are not* immediately present.	K-2.A.1.2a Recognize the importance of asking questions about how data were collected.	3-5.A.1.2a Ask questions about how data are collected or considered .	6-8.A.1.2a Ask questions regarding the origins of specific automated measures (e.g., webtracking, email metadata, Spotify).	9-10.A.1.2a Recognize that decisions about data are revisited over time and as need arises (e.g., blood pressure cut-off numbers, dietary guidance, medical benchmarks).	11-12.A.1.2a Explore the origins of some standardized unit measurements (e.g., horsepower, mole, scores on AP Exams).
		3-5.A.1.2b Understand that data are generated by people who make decisions about what and how to measure.	6-8.A.1.2b Recognize the limits of the information the data can provide and the story it can tell.	9-10.A.1.2b Understand that models based on data need to have their data updated to remain current.	11-12.A.1.2b Identify the risks and tradeoffs of using traditional measurements (i.e., IQ, BMI, etc.).
			6-8.A.1.2c Recognize that conclusions may need to be revised in the future as more knowledge and data become available.		

While the full set of LPs (accessed at https://teachingdatak12.org) include hundreds of competencies, the intention is not for any one subject matter to hold the responsibility for enacting all of them, but rather for an interdisciplinary approach which sees data practices infused into all subject matter teaching across the current subject-siloed structure of education. The next phase of development for the LPs will expand on the current model to create subject-articulated alignment to help guide this interdisciplinary approach.

Conclusion and Next Steps

The next step of this project is to convene multiple subject-specific groups of experts to map alignment between the subject-agnostic LPs and current standards across the different subjects of mathematics, science, social studies, and computer science through 2025 and into 2026. These subject specific LPs will help educators guide the work within their specific classrooms. Additionally, there

will multiple ongoing opportunities for different parties and communities of interest, as well as the wider public, to provide insights and feedback on the development, production and interactive structure of the framework. The work will proceed as an iterative design cycle, engaging interest groups in a sustained way going forward with an open amendment suggestion platform and a convening every two years to assess and implement proposed amendments, with the first refinement happening in February 2027. In this way, the proposed activities will generate a framework clear and specific enough to inform work across both research and development, yet flexible enough to evolve and incorporate the many new findings certain to arise from each.

References

Drozda, Z., Johnstone, D., & Van Horne, B. (2022). *Previewing the national landscape of K-12 data science implementation*. In Workshop on Foundations of Data Science for Students in Grades K-12.

Franklin, C., & Bargagliotti, A. (2020). Introducing GAISE II: A guideline for precollege statistics and data science education. *Harvard Data Science Review, 2*(4), 1-9.

Israel-Fishelson, R., Moon, P. F., Tabak, R., & Weintrop, D. (2023). Preparing students to meet their data: an evaluation of K-12 data science tools. *Behaviour & Information Technology*, 1-20.

Miller, K., & Yoon, S. (June 2023). *Supporting High School Science Teachers in Developing Pedagogical Content Knowledge for Data Literacy*. Paper presented at The Annual Conference of the International Society of the Learning Sciences (ISLS), Montreal, QC.

National Academies of Sciences, Engineering, and Medicine. (2023). *Foundations of Data Science for Students in Grades K-12: Proceedings of a Workshop*. Washington D.C.: The National Academies Press.

National Science Foundation. (2011, July 26). *Learning Progressions Footprint Conference. Learning Progressions Footprint Conference Final Report*. Learning Progressions Footprint Conference, Washington, DC.

Rosenberg, J., & Jones, R. S. (2024). Data science learning in grades K–12: Synthesizing research across divides. *Harvard Data Science Review, 6*(3). DOI: 10.1162/99608f92.b1233596

Witte, V., Schwering, A., & Frischemeier, D. (2024). Strengthening Data Literacy in K-12 Education: A Scoping Review. *Education Sciences, 15*(1), 25.