# A Tiered Framework for Educational Event Data Documentation: Synthesizing Principles and Addressing Gaps

Wei, X., Prado, Y., Roschelle, J., & Ruiz, P.

data documentation
educational event data
Learning Analytics
learning engineering
Open Science

*Digital learning platforms generate millions of fine-grained behavioral events, yet inadequate documentation restricts equitable research use. We systematically reviewed five documentation traditions (survey/DDI, FAIR, AI transparency, learning-analytics standards, psychometrics) and identified gaps unique to educational event data: temporal complexity, nested structures, platform business logic, scale/granularity, and access heterogeneity. Synthesizing cross-cutting strengths, we propose five principles—Transparency, Accessibility, Usability, Responsibility, Maintainability—operationalized in a 16-item 3-tiered framework designed to enable platforms to self-assess and improve incrementally. Empirical validation through platform piloting and researcher feedback represents essential next steps.*

# Introduction

Digital learning platforms (e.g., ASSISTments, Khan Academy, i-Ready) log millions of student interactions daily, capturing full problem-solving processes: hint requests, errors, time allocation, and help-seeking. Such fine-grained behavioral data can transform how we understand learning and personalize interventions (Baker & Siemens, 2014). Learning engineering leverages learning sciences, cognitive science, computer science, and "the rapid developments in the data available and the learning analytics methods available to analyze it" (Baker et al., 2022, p. 4) to build educational systems that support high-quality learning (Baker et al., 2022). Realizing this promise, however, depends on adequate data documentation.

Yet many educational datasets remain underutilized due to documentation inadequacies. Current practices, inherited from survey research traditions, fail to address temporal event streams, nested data structures, and platform-specific business logic. This creates inequitable access to research opportunities where well-connected researchers navigate data effectively while early-career scholars and under-resourced institutions face substantial learning barriers. This documentation gap thus compounds existing inequities in education research and threatens reproducibility (D'Ignazio & Klein, 2020; Makel & Plucker, 2014).

This paper uses a learning engineering framework to address three questions: (1) What documentation gaps exist when applying established traditions to educational event data? (2)What principles address data documentation gaps? (3) How can these principles be operationalized into a practical, tiered framework?

# Methods

We used a nested learning engineering cycle (Craig et al., 2025; Totino & Kessler, 2024) to address the challenge of making educational data more accessible. Our work follows the iterative Challenge-Creation-Implementation-Investigation cycle (Goodell & Kessler, 2023):

1. Challenge: Understanding documentation inadequacies through systematic review
2. Creation: Synthesizing principles and developing the tiered framework
3. Implementation: Developing platform self-assessment and improvement approaches (future work)
4. Investigation: Evaluating framework impact on accessibility and quality (future work)

This paper documents our work in the Challenge and Creation phases. Our approach is conceptual and synthesis-based, drawing on the research team's collective expertise working with educational platforms and event data. Empirical validation through platform piloting and researcher feedback will be essential next steps.

Our methodology consisted of three phases: documentation tradition review, gap identification, and principle synthesis.

# Phase 1: Review of Five Documentation Traditions

We identified five documentation traditions representing diverse paradigms relevant to educational platforms (social science, computer science, measurement science, open science infrastructure). Selection criteria prioritized traditions with: (1) established standards, (2) research data documentation focus, and (3) different epistemological perspectives. For each tradition, we identified canonical standards and widely cited guidance documents (e.g., specifications, technical notes, flagship papers) through targeted searches of standards organizations' repositories (DDI Alliance, FAIR Data Principles, IMS Global, etc.) and reference chaining from prior reviews. We extracted documentation requirements and assumptions into a structured comparison matrix organized by tradition and documentation focus.

Survey Research and DDI (Data Documentation Initiative): Survey research has developed sophisticated metadata standards for documenting questionnaires, variables, and sampling procedures (Vardigan et al., 2008). DDI provides structured XML schemas for capturing variable-level metadata, including question text, response options, skip logic, and universe statements (DDI Alliance, 2020). This tradition excels at documenting static, rectangular datasets where each row represents a respondent and columns represent variables measured at a single time point (Groves et al., 2009).

FAIR (Findable, Accessible, Interoperable, Reusable) Principles: Emerging from open science, FAIR principles provide high-level guidance for making research data discoverable and reusable (Wilkinson et al., 2016). FAIR emphasizes consistent identifiers, machine-readable metadata, standardized vocabularies, and clear licensing (Wilkinson et al., 2016). These principles tend to prioritize general infrastructure and cross-repository interoperability over domain-specific documentation needs (Jacobsen et al., 2020).

AI Transparency Frameworks: Model cards (Mitchell et al., 2019) and datasheets for datasets (Gebru et al., 2018) document machine learning training data and model behavior. These frameworks address algorithmic transparency, training data provenance, performance metrics across demographic groups, and intended use cases (Raji et al., 2020). They emphasize ethical considerations, bias documentation, and limitations statements (Arnold et al., 2019).

Learning Analytics Standards: Technical specifications like xAPI (Experience API) and IMS Caliper define standardized event formats for capturing learning activities (ADL, 2013; IMS Global, 2015). These standards provide controlled vocabularies for common educational events (e.g., “viewed,” “completed,” “scored”) and specify JSON schemas for event structure. They enable cross-platform event aggregation but provide limited guidance on platform-specific business logic documentation (Siemens & Baker, 2012).

Psychometrics Documentation: Measurement research emphasizes construct validity, multiple forms of reliability evidence, and appropriate score interpretation (AERA et al., 2014). This tradition documents measurement instruments, scoring

algorithms, validation studies, and known limitations (Kane, 2013). It focuses on what data means substantively rather than technical access mechanisms (Messick, 1995).

This comparative analysis revealed that while each tradition contributes valuable documentation practices, none adequately addresses the unique characteristics of educational event data, as documented in Phase 2.

# Phase 2: Gap Identification for Event Data

Two members of the research team independently applied each tradition's documentation guidance to educational event data characteristics (temporal sequences, nested structures, platform-generated processes), identifying areas where guidance proved insufficient or silent. Through iterative discussion and cross-tradition thematic coding (Braun & Clarke, 2006), we synthesized recurring breakdowns into five critical gaps:

Temporal complexity gap: DDI was developed for documenting survey-based datasets and their instruments, where time is typically represented as discrete collection periods (Vardigan et al., 2008) rather than continuous event streams. Educational platforms generate continuous event streams where temporal ordering, duration, session boundaries, and time-on-task patterns carry substantive meaning for understanding learning. While learning analytics standards (xAPI, Caliper) specify individual event formats, they provide limited guidance on documenting temporal dependencies, sequence patterns, or longitudinal trajectories.

Nested structure gap: Survey documentation assumes flat, rectangular datasets (rows = respondents, columns = variables). Event data are hierarchically nested (e.g., events within attempts within problems within sessions within students within classrooms) with many-to-many relationships. No existing tradition adequately addresses how to document these complex relational structures or specify aggregation rules across levels.

Platform logic gap: AI frameworks document training algorithms but not production business logic generating behavioral data. Psychometric documentation describes instruments but not software systems mediating collection. Educational platforms embed complex rules (e.g., hint sequencing, mastery thresholds, adaptive pathways) that researchers must understand to interpret observed events, yet no tradition provides guidance for documenting these systems.

Scale and granularity gap: Survey research documents hundreds of curated variables; platforms generate millions of events. Researchers need guidance for event subset selection matched to research questions, not exhaustive documentation of every event type. Existing traditions lack frameworks for tiered documentation at multiple aggregation levels (e.g., raw events, session summaries, student trajectories) or decision trees guiding researchers to appropriate granularity.

Access heterogeneity gap: Survey data is released as single files with codebooks. Event data exists at multiple levels accessible through different mechanisms: bulk downloads, APIs, derived datasets, or restricted-access portals. FAIR principles address repository-level access but not the tradeoffs between different data representations, computational requirements for event-level versus aggregate analysis, or how access method shapes analytic possibilities.

# Phase 3: Principle Synthesis

We synthesized five documentation principles by extracting underlying goals from the five traditions and adapting them to address the identified gaps. To operationalize these principles into the tiered framework, we: (1) mapped each principle to observable documentation artifacts (e.g., variable definitions, API documentation, example code), (2) ordered artifacts from minimal to comprehensive, and (3) defined three capability thresholds (Baseline, Enhanced, Advanced). The team refined item definitions and tier boundaries, ultimately producing a 16-item framework. External validation with platform developers and researchers remains essential future work.

The five principles are:

## Transparency (what data exists and what it means):

Documents not just individual event types but also temporal relationships, business logic generating events, and how raw events relate to derived measures. Synthesizes variable-level metadata (DDI), provenance (FAIR), training data descriptions (AI frameworks), event semantics (learning analytics), and construct definitions (psychometrics)—adapted to address temporal complexity, nested structures, and platform logic gaps.

## Accessibility (how data can be obtained and used):

Addresses multiple access pathways (raw events vs. aggregated datasets), technical prerequisites (API authentication, query languages), and computational requirements for large-scale temporal data. Builds on persistent identifiers and machine-readable formats (FAIR), codebook distribution (DDI), API specifications (learning analytics), and dataset availability statements (AI frameworks)—adapted for access heterogeneity.

## Usability (ease of understanding and working with data):

Provides onboarding materials for navigating temporal complexity, example analyses for common research questions, and guidance matching events to questions. Draws from questionnaire documentation (survey research), interoperability standards (FAIR), datasheets with use cases (AI frameworks), reference implementations (learning analytics), and scoring guidance (psychometrics)—addressing scale, granularity, and nested structure challenges.

## Responsibility (ethical considerations and limitations):

Addresses temporal privacy risks (behavioral patterns may identify individuals despite demographic anonymization), algorithmic fairness concerns (platform logic influences observed behavior), and limitations of observational vs. experimental inference. Integrates consent documentation (survey research), licensing requirements (FAIR), bias and limitation statements (AI frameworks), privacy compliance (learning analytics), and appropriate use guidelines (psychometrics).

## Maintainability (keeping documentation current as platforms evolve):

Addresses rapid platform evolution where new features continuously change event generation. Synthesizes version control practices (survey research), metadata versioning (FAIR), dataset update cycles (AI frameworks), schema versioning (learning analytics), and measurement revision tracking (psychometrics).

We operationalized each principle into assessable criteria across three tiers (Baseline, Enhanced, Advanced), recognizing that platforms have varying resources and maturity while researchers have differing documentation needs. This operationalization involved: (1) identifying documentation artifacts or practices demonstrating each principle, (2) arranging these by increasing comprehensiveness, and (3) defining three capability thresholds.

# Results

The resulting 16-item framework (Table 1) emerged by decomposing each of the five principles into 2-4 assessable questions; for example, Transparency encompasses variable clarity (Item 1), provenance documentation (Item 2), and limitations disclosure (Item 3). Each item is assessed at three capability levels:

1. Baseline: Minimum documentation enabling basic understanding and appropriate use. Provides essential information but may require substantial researcher effort.

2. Enhanced: Additional documentation supporting efficient research use through structured metadata, example analyses, and clearer guidance. Reduces time-to-insight for researchers.
3. Advanced: Comprehensive documentation supporting reproducibility, equity (lowering barriers for early-career researchers), and sophisticated secondary analyses. Represents documentation best practices.

This structure enables platforms to self-assess documentation maturity and pursue incremental improvement. For example, a platform might achieve the Enhanced tier for most Accessibility questions while remaining at Baseline for Transparency question 2 (e.g., lacking processing pipeline documentation). This granular assessment helps researchers evaluate whether documentation meets their research needs.

Table 1.
Tiered Documentation Framework Rubric

| Dimensions | Checklist Questions | Baseline | Enhanced | Advanced |
|---|---|---|---|---|
| Transparency | 1. Are all variables clearly named and described? | Variable names listed with brief labels. | Full definitions with units and value meanings. | Complete semantic mapping or ontology crosswalks with usage examples. |
| | 2. Is the dataset provenance documented? | Data collection dates or versions stated. | Sources and processing steps described. | Full lineage (raw → processed → analytic) with version-controlled scripts. |
| | 3. Are assumptions and limitations disclosed? | Generic caveats noted. | Specific limitations and known data issues described. | Quantified uncertainty or bias linked to validation results. |
| Accessibility | 4. Is there a human-readable documentation package? | README / PDF with summary info. | Structured manual with cross-references and examples. | Interactive or searchable web documentation portal. |
| | 5. Is structured (machine-readable) metadata provided? | None. | CSV or JSON variable dictionary with types and domains. | Metadata validated against a standard (DDI, Caliper, xAPI) and registered via API. |
| | 6. Are example analyses or domain-appropriate demonstrations provided? | None. | One or more static examples appropriate to dataset's main use (e.g., joining tables, simulation scripts). | Full reproducible notebook or API workflow with narrative interpretation. |
| | 7. Is the dataset publicly findable and citable? | No DOI / repository record. | DOI or repository record exists but limited indexing. | Indexed in major catalogs (e.g., Google Dataset Search, Dataverse) with persistent identifiers. |

| | | | | |
|---|---|---|---|---|
| Usability | 8. Are data tables organized and relationships clear? | Basic list of tables. | Join keys / relationships explained textually or diagrammed. | Machine-readable schema or entity-relationship diagram with validation. |
| | 9. Are there guides for common research tasks? | None. | Text instructions for typical joins or aggregations. | Step-by-step workflow or template scripts covering end-to-end analysis. |
| | 10. Are data quality checks or diagnostics available? | None. | Missingness / range summaries provided. | Automated validation scripts or dashboards released with data. |
| Responsibility | 11. Are privacy protections or ethical safeguards documented? | De-identification mentioned. | Detailed anonymization / aggregation procedures. | Formal privacy-impact or bias audit (for sensitive data); for open data, clear attribution & ethical use guidance. |
| | 12. Are sampling or bias considerations disclosed? | None. | Narrative description of representativeness limits. | Quantitative subgroup or equity analysis with interpretation. |
| | 13. Is appropriate-use guidance provided? | Generic license text. | Clear do’s / don’ts for intended uses. | Ethical-use policy with community or IRB oversight. |
| Maintainability | 14. Is versioning and change history available? | Single version or date label. | Changelog / release notes maintained. | Semantic versioning with archived prior releases and linked documentation. |
| | 15. Is there a support or feedback channel? | Contact email. | Form or issue log for user feedback. | Public issue tracker with transparent resolution workflow. |
| | 16. Is there a plan or cadence for future updates? | None stated (snapshots acceptable if clearly archived). | Informal commitment to periodic review or metadata refresh. | Formal maintenance schedule or automated continuous integration for documentation. |

# Discussion

This framework advances educational data documentation in five ways. First, it provides systematic synthesis of documentation traditions explicitly adapted for temporal, nested, platform-generated event data. Second, it offers actionable guidance through 16 checklist questions enabling platforms to self-assess maturity, identify improvement priorities, and progress incrementally without complete re-engineering. Third, it promotes equity by setting transparent expectations that help

researchers evaluate whether documentation meets their needs before substantial time investment, reducing advantages currently enjoyed by well-networked researchers. Fourth, it supplies shared stakeholder language. Funders can specify tier requirements in data management plans. Repositories can provide structured feedback using the rubric. Finally, it provides essential R&D infrastructure for learning engineering by reducing time spent deciphering undocumented data, enabling cross-platform comparisons, and establishing shared standards that accelerate the iterative improvement cycles central to learning engineering (Baker et al., 2022).

This framework represents a conceptual synthesis requiring empirical validation. We have not tested whether platforms can reliably self-assess using these items, whether tier distinctions meaningfully differentiate documentation quality, or whether improved documentation reduces researcher barriers. The framework should be viewed as a theoretically-grounded starting point, not a validated instrument ready for high-stakes deployment.

Following learning engineering practice (Baker et al., 2022; Goodell et al., 2023), the next Implementation and Investigation phases involve: (1) piloting with platforms to assess reliability, (2) gathering researcher feedback on which elements reduce time-to-first-analysis, (3) examining relationships between documentation quality and research productivity, and (4) testing generalizability beyond formal learning platforms. These investigations will inform framework refinement, automated assessment tools, and documentation templates. The framework focuses on content and standards; complementary technical infrastructure (e.g., scalable APIs, privacy-preserving access) and policy challenges (e.g., consent models) represent parallel work.

# Acknowledgments

The research reported here was supported by the Institute of Education Sciences (IES), U.S. Department of Education, through SEERNet Grant R305N210034 and National Science Foundation (NSF), through Grant 2153481. The opinions expressed are those of the authors and do not represent the views of the IES, the U.S. Department of Education, or NSF.

# References

ADL (Advanced Distributed Learning). (2013). Experience API (xAPI) specification, Version 1.0.0. https://github.com/adlnet/xAPI-Spec

AERA, APA, & NCME. (2014). Standards for educational and psychological testing. American Educational Research Association.

Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., ... & Varshney, K. R. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. IBM Journal of Research and Development, 63(4/5), 6:1-6:13. https://doi.org/10.1147/JRD.2019.2942288

Baker, R. S., & Siemens, G. (2014). Educational data mining and learning analytics. In R. K. Sawyer (Ed.), The Cambridge handbook of the learning sciences (pp. 253-272). Cambridge University Press.

Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. Technology, Mind, and Behavior, 3(1), 1-23. https://doi.org/10.1037/tmb0000058

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101.

Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training

Environment. In International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale (pp. 115-132). https://doi.org/10.59668/2109.21735

D'Ignazio, C., & Klein, L. F. (2020). Data feminism. MIT Press.

DDI Alliance. (2020). DDI-Codebook 2.5. https://ddialliance.org/Specification/DDI-Codebook/2.5/

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. Communications of the ACM, 64(12), 86-92. https://doi.org/10.1145/3287560.3287561

Goodell, J., Kessler, A., & Schatz, S. (2023). Learning engineering at a glance. Journal of Military Learning, 7(1), 46-59.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). Survey methodology (2nd ed.). Wiley.

IMS Global Learning Consortium. (2015). IMS Caliper Analytics® specification, Version 1.1. https://www.imsglobal.org/spec/caliper/v1p1

Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., ... & Goble, C. (2020). FAIR principles: Interpretations and implementation considerations. Data Intelligence, 2(1-2), 10-29. https://doi.org/10.1162/dint_a_00024

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1-73. https://doi.org/10.1111/jedm.12000

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. Educational Researcher, 43(6), 304-316. https://doi.org/10.3102/0013189X14545513

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220-229. https://doi.org/10.1145/3287560.3287577

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33-44. https://doi.org/10.1145/3351095.3372873

Siemens, G., & Baker, R. S. (2012). Learning analytics and educational data mining: Towards communication and collaboration. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, 252-254. https://doi.org/10.1145/2330601.2330661

Totino, L., & Kessler, A. (2024). "Why did we do that?" A systematic approach to tracking decisions in the design and iteration of learning experiences. The Journal of Applied Instructional Design, 13(2), 277-281. https://doi.org/10.59668/1269.15630

Vardigan, M., Heus, P., & Thomas, W. (2008). Data documentation initiative: Toward a standard for the social sciences. International Journal of Digital Curation, 3(1), 107-113. https://doi.org/10.2218/ijdc.v3i1.45

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), 1-9. https://doi.org/10.1038/sdata.2016.18