



DATA QUALITY DOCUMENTATION TEMPLATE

sdp.cepr.harvard.edu
sdp@gse.harvard.edu

This tool is part of [The Early Learning Journey](#), a collaboration between the Strategic Data Project and representatives from Boston Public Schools, the California Department of Education, Chicago Public Schools, the D.C. Office of the State Superintendent of Education, and the Washington State Education Research and Data Center. It is supported by the Gates Foundation.



Center for Education Policy Research
HARVARD UNIVERSITY

Data Quality Documentation Template

Use this template to document data quality issues, decisions, and caveats for each data source used in PreK–2 longitudinal analysis. Complete one section per data source. This record supports transparent analysis, honest stakeholder communication, and replicability over time.

Analyst: _____
Agency: _____
Date: _____

How to Use This Template

For each data source, complete the shaded header row with the source name, program or setting, and the date data was received. Then work through each issue type row:

- **Finding:** Describe what you observed in the data, including counts, rates, or patterns.
- **Decision/Action Taken:** Document how you handled the issue, including any exclusions, flags, or follow-up.
- **Analysis Caveat:** Note any limitations or interpretation guidance that should accompany findings from this source.

Leave rows blank if an issue type does not apply. Add rows as needed. Italicized text in each cell is example guidance—replace it with your own notes.



Data Source 1

Data Source 1 Data Source: _____ Program/Setting: _____ Data Received: _____			
Issue Type	Finding	Decision / Action Taken	Analysis Caveat
Record counts	Expected: ____ Actual: ____ Difference: ____	e.g., Followed up with data owner; excluded duplicate records	e.g., Coverage limited to enrolled children only
Missingness (overall)	e.g., 18% of records missing assessment score	e.g., Flagged for sensitivity analysis	e.g., Results may underrepresent children with limited English proficiency
Missingness (by population)	e.g., Missing data concentrated among DLL students	e.g., Analyzed separately; documented in limitations	e.g., Disaggregated results for this group should be interpreted with caution
Assessment administration	e.g., Timing varied across sites by 6–8 weeks	e.g., Restricted analysis to children assessed within standard window	e.g., Findings may not reflect full program population
Assessment fidelity	e.g., Staff in 3 sites reported limited training on tool	e.g., Scores from those sites flagged; excluded from comparative analysis	e.g., Do not compare scores across sites with different fidelity levels
Identifier consistency	e.g., 12% of records could not be matched across systems	e.g., Used probabilistic matching; documented match rate	e.g., Unmatched children excluded; may differ systematically from matched group
Other			

Data Source 2

Data Source 2 Data Source: _____ Program/Setting: _____ Data Received: _____			
Issue Type	Finding	Decision / Action Taken	Analysis Caveat
Record counts	Expected: ____ Actual: ____ Difference: ____	e.g., Followed up with data owner; excluded duplicate records	e.g., Coverage limited to enrolled children only
Missingness (overall)	e.g., 18% of records missing assessment score	e.g., Flagged for sensitivity analysis	e.g., Results may underrepresent children with limited English proficiency
Missingness (by population)	e.g., Missing data concentrated among DLL students	e.g., Analyzed separately; documented in limitations	e.g., Disaggregated results for this group should be interpreted with caution
Assessment administration	e.g., Timing varied across sites by 6–8 weeks	e.g., Restricted analysis to children assessed within standard window	e.g., Findings may not reflect full program population
Assessment fidelity	e.g., Staff in 3 sites reported limited training on tool	e.g., Scores from those sites flagged; excluded from comparative analysis	e.g., Do not compare scores across sites with different fidelity levels
Identifier consistency	e.g., 12% of records could not be matched across systems	e.g., Used probabilistic matching; documented match rate	e.g., Unmatched children excluded; may differ systematically from matched group
Other			

Data Source 3

Data Source 3 Data Source: _____ Program/Setting: _____ Data Received: _____			
Issue Type	Finding	Decision / Action Taken	Analysis Caveat
Record counts	Expected: ____ Actual: ____ Difference: ____	e.g., Followed up with data owner; excluded duplicate records	e.g., Coverage limited to enrolled children only
Missingness (overall)	e.g., 18% of records missing assessment score	e.g., Flagged for sensitivity analysis	e.g., Results may underrepresent children with limited English proficiency
Missingness (by population)	e.g., Missing data concentrated among DLL students	e.g., Analyzed separately; documented in limitations	e.g., Disaggregated results for this group should be interpreted with caution
Assessment administration	e.g., Timing varied across sites by 6–8 weeks	e.g., Restricted analysis to children assessed within standard window	e.g., Findings may not reflect full program population
Assessment fidelity	e.g., Staff in 3 sites reported limited training on tool	e.g., Scores from those sites flagged; excluded from comparative analysis	e.g., Do not compare scores across sites with different fidelity levels
Identifier consistency	e.g., 12% of records could not be matched across systems	e.g., Used probabilistic matching; documented match rate	e.g., Unmatched children excluded; may differ systematically from matched group
Other			

Overall Data Quality Summary

Use this section to summarize cross-cutting quality issues, overarching limitations, and any decisions that apply to the analysis as a whole.

Topic	Notes
Populations with limited coverage	<i>e.g., Children in unlicensed settings not represented in any data source</i>
Cross-source quality concerns	<i>e.g., Identifier inconsistency between PreK and K systems affected match rate across all sources</i>
Decisions affecting generalizability	<i>e.g., Children with fewer than 2 assessment time points were excluded from longitudinal analyses</i>
Recommended caveats for all reporting	<i>e.g., Findings should not be interpreted as causal; selection into PreK programs was not random</i>

