Cleaned DepEd School-Level Datasets, SY 2017–18 to SY 2025–26: Sources, Harmonization and Data Quality

Mark Anthony Llego

NASH Reviewer, Philippines

NASH Reviewer Research Report No. 3 · Version 1.1 · November 2026

DOI: https://doi.org/10.5281/zenodo.24624444

Data set described: https://doi.org/10.5281/zenodo.24624443 (Version 1.1, released November 28, 2026)

Suggested citation: Llego, M. A. (2026). *Cleaned DepEd school-level datasets, SY 2017–18 to SY 2025–26: Sources, harmonization and data quality* (NASH Reviewer Research Report No. 3, Version 1.1). NASH Reviewer. https://doi.org/10.5281/zenodo.24624444

© 2026 Mark Anthony Llego. This report is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0); the data it describes are distributed under DepEd's terms of use (Section 5.1). NASH Reviewer is not affiliated with the Department of Education (DepEd).

Abstract

This paper describes a cleaned, harmonized release of the school-level data that the Philippine Department of Education (DepEd) publishes as machine-ready files. The release draws on 17 DepEd files with 2,216 columns: enrollment of every school in nine school years, SY 2017–18 to SY 2025–26; learner, personnel, facilities, information and communication technology (ICT), and water, sanitation and hygiene (WASH) data for SY 2023–24; Grade 6 National Achievement Test (NAT) and Grade 3 Early Language, Literacy and Numeracy Assessment (ELLNA) results for SY 2023–24; and Comprehensive Rapid Literacy Assessment (CRLA) results for SY 2025–26. It holds five datasets with 343 documented columns: a school master file for SY 2023–24 (60,167 schools), an enrollment panel of 547,425 school-years over 66,967 school IDs, a school spine, CRLA reading levels for 6,121 schools, and empirical Bayes estimates of NAT and ELLNA school means. Two helper tables list 146 likely school-ID changes and assessment coverage by division. Labels follow one scheme across years, regions are also given under SY 2025–26 boundaries, and blank cells follow stated rules; a codebook, a rules file and the README list every rule. No source file has a duplicate school ID or a negative count, the six SY 2023–24 files share one set of 60,167 IDs, and implausible values are flagged, not altered. The assessment files cover a selected set of schools in 44 divisions of five regions, and the input modules cover public schools only. The files are in CSV and Parquet, with SHA-256 checksums, under DepEd's terms.

Keywords: open data; data descriptor; school-level data; Philippines; Department of Education; enrollment panel; data harmonization; data quality

1. Background and summary

The Department of Education publishes school-level planning and assessment data as machine-ready files on its website (Department of Education [DepEd], n.d.-a). Each archive holds one CSV file, with one row per school keyed on a school ID of six digits (one ID in SY 2022–23 has seven), and a README that documents its columns. The planning data come from three information systems: the Learner Information System (LIS), which DepEd names the official source of enrollment; the Basic Education Information System (BEIS), which collects school resources and personnel; and the National School Building Inventory (NSBI), which collects buildings, classrooms and other facilities (DepEd, n.d.-b, Sec. I). The assessment files come from the Bureau of Education Assessment (BEA; DepEd, 2025).

The files are public, but they cannot be analyzed together as published. The SY 2025–26 enrollment file codes sector, management, annex status and the offering flags differently from the eight years before it, and it mixes two text encodings. A blank cell can mean that a school does not offer a level, that a variable was not collected that year, or that a school reported nothing. Region boundaries changed four times between SY 2017–18 and SY 2025–26, a new BARMM division took 111 schools from Region XII, five divisions were renamed, and two division names are each used in two regions. Some schools change ID between years, and no mapping from old to new IDs is published. Each user of the raw files has to settle all of this before the first table, and two users who settle it differently will report different numbers from the same files.

Version 1.1 of the release (Llego, 2026a), published on November 28, 2026, applies one documented set of rules to 17 DepEd files. It keys every file and year on DepEd's school ID and gives every year one set of labels, column names and types. It holds five datasets and two helper tables in CSV and Parquet, 66.6 MB in all, with a codebook of its 343 columns, the label mappings, SHA-256 checksums and a README. The data are DepEd's; the cleaning, the derived columns and the documentation are NASH Reviewer's work.

2. Methods

2.1 Source files

The release draws on 17 archives from DepEd's machine-ready files (Table 1). Each archive holds exactly one CSV file and one README; none holds a workbook. Every file has one row per school, and the 17 files have 2,216 columns between them. The nine enrollment files cover every school in the LIS in each year, in all sectors: public schools, private schools, state and local universities and colleges (SUC/LUC), and Philippine Schools Overseas (PSO), whose enrollment the LIS captures from SY 2018–19 (DepEd, n.d.-b, Sec. VIII.a). The five SY 2023–24 school files list all 60,167 schools of that year, but their personnel, facilities, ICT and WASH data cover public schools only (DepEd, n.d.-b, Sec. VIII.d–f). The three assessment files hold results "for a selected set of schools only," in the words of their README (DepEd, n.d.-a).

2.2 Reading the source files

The archives were extracted to a working copy, and every later step reads that copy; the original archives are never modified. Sixteen of the 17 CSV files are valid UTF-8. In the SY 2025–26 enrollment file, 1,892 lines contain Windows-1252 bytes (0xF1 for "ñ" in "Doña" and "Niño", for example), while 134 other lines contain valid UTF-8 multibyte characters. Each line is decoded as UTF-8 first and as Windows-1252 when that fails, so no row is lost and every school name keeps its letters.

The enrollment files of SY 2017–18 to SY 2024–25 have the same 78 columns in the same order. The SY 2025–26 file adds five. One, `Modified Curricular Offering Classification`, is not documented in its README and is kept as `coc_2025`. The other four count Grade 11 learners by sex in the Strengthened Senior High School Academic and Technical-Professional tracks (`g11_sshs_acad_*`, `g11_sshs_techpro_*`). The CRLA README describes its `*_percent` columns as percentages; their values are proportions from 0 to 1, and the release keeps them as proportions.

2.3 Harmonization of labels, types and school IDs

Labels are mapped to one scheme in code, never by hand, and the published label is kept wherever no rule applies. Table 2 lists all 29 mappings in the release's `harmonisation_rules.csv`. Whitespace is also stripped from every label, since SY 2025–26 writes "Non-Sectarian " and "Sectarian " with trailing spaces.

Table 1. The 17 DepEd source files

Dataset	CSV file	School year	Source	Rows	Columns
Enrollment	`enrollment_2017-18.csv`	2017–18	LIS	61,563	78
Enrollment	`enrollment_2018-19.csv`	2018–19	LIS	61,916	78
Enrollment	`enrollment_2019-20.csv`	2019–20	LIS	61,923	78
Enrollment	`enrollment_2020-21.csv`	2020–21	LIS	60,957	78
Enrollment	`enrollment_2021-22.csv`	2021–22	LIS	60,429	78
Enrollment	`enrollment_2022-23.csv`	2022–23	LIS	60,137	78
Enrollment	`enrollment_2023-24.csv`	2023–24	LIS	60,167	78
Enrollment	`enrollment_2024-25.csv`	2024–25	LIS	60,129	78
Enrollment	`enrollment_2025-26.csv`	2025–26	LIS	60,204	83
Learner data	`learners_2023-24.csv`	2023–24	LIS, BEIS	60,167	316
School personnel	`personnel_2023-24.csv`	2023–24	BEIS	60,167	327
School facilities	`facilities_2023-24.csv`	2023–24	NSBI	60,167	87
ICT data	`ict_2023-24.csv`	2023–24	BEIS	60,167	594
WASH and utilities	`sanitation_utilities_2023-24.csv`	2023–24	BEIS, NSBI	60,167	114
NAT, Grade 6	`nat_2023-24_selected.csv`	2023–24	BEA	6,636	10
ELLNA, Grade 3	`ellna_2023-24_selected.csv`	2023–24	BEA	5,752	10
CRLA, Grades 1–3	`crla_2025-26_selected.csv`	2025–26	BEA	6,121	51
All files					**2,216**

Note. One row per school in every file; `school_id` is unique in every file. LIS = Learner Information System; BEIS = Basic Education Information System; NSBI = National School Building Inventory; BEA = Bureau of Education Assessment. In the WASH file, water, electricity and internet come from the BEIS and toilets from the NSBI (DepEd, n.d.-b, Sec. I). Source: DepEd (n.d.-a).

SY 2025–26 uses a coding scheme that its README does not document: "SUCsLUCs" for SUC/LUC, "DepED Managed" for DepEd, "SCHOOL ABROAD" for PSO, new wording for the four annex categories, and Yes/No for the offering flags that the other eight years code True/False. Each is mapped to the labels of the other years. Two management labels stay as published. "Tahderiyyah" appears only in SY 2018–19 (29 schools). "Local Government" falls from 1,054 schools in SY 2018–19 to 193 in SY 2019–20, because 858 of the SY 2018–19 schools were recoded to "DepEd": a recode, not a transfer of schools.

Four region changes fall within the nine years. ARMM became BARMM in SY 2019–20; Cotabato City moved from Region XII to BARMM in SY 2021–22; the Negros Island Region (NIR) appeared in SY 2024–25 with 21 divisions from Regions VI and VII; and Sulu moved from BARMM to Region IX in SY 2025–26, which raised Region IX from 2,877 to 3,365 schools. Schools also moved without their division: 111 of the 118 schools of BARMM's Special Geographic Area Division, first listed in SY 2021–22, appear under the North Cotabato division of Region XII in SY 2017–18 to SY 2020–21. ARMM is relabeled BARMM in every year. Every school-year row also carries `region_const`, its region under SY 2025–26 boundaries, computed from region and division together; the Special Geographic Area Division's schools, matched by school ID, are BARMM in every year, which moves 422 school-year rows, with 158,366 learners, from Region XII to BARMM. The published `region` stays beside it, so a year's own tables can still be matched.

The five renamed divisions take their current names in every year (Table 2). Divisions created during the period, mostly new city divisions and new PSO countries, are not mapped back, and the number of divisions, counted by region and division, rises from 219 in SY 2017–18 to 242 in SY 2025–26 (217 and 240 distinct names). Division names are not unique: San Carlos City is a division in Region I and in Negros Occidental, and San Fernando City in La Union (Region I) and in Pampanga (Region III). Division-level work therefore keys on region and division together. The division "Lanao del Sur - IB", which appears only in SY 2019–20, keeps its name.

The CRLA file spells three of the ELLNA file's languages differently and names two that the ELLNA file lacks. The three take the ELLNA spelling, such as "Ilocano" to "Ilokano/Iloko", and the two take the spelling of DepEd's technical notes, "Bahasa Sug" to "Tausug" and "Chavacano" to "Chavacano/Chabacano" (DepEd, 2025). Counts are stored as nullable integers, flags as Booleans and shares as proportions from 0 to 1, and a column has the same type in its CSV and Parquet files.

Table 2. The label mappings in `harmonisation_rules.csv`

Field	Source label	Released label
`sector`	SUCsLUCs	SUC/LUC
`school_management`	DepED Managed	DepEd
	DepEd Managed	DepEd
	SUC Managed	SUC
	DOST Managed	DOST
	Other GA Managed	Other GA
	SCHOOL ABROAD	PSO
	Local International School	International School
`annex_status`	School with no Annexes	Standalone School
	Mother school	Mother School
	Annex or Extension school(s)	Annex/Extension School
	Mobile School(s)/Center(s)	Mobile School/Center
`region`	ARMM	BARMM
`division`	Compostela Valley	Davao De Oro
	San Jose del Monte City	City of San Jose Del Monte
	Munoz Science City	Science City of Muñoz
	Taguig	Taguig City and Pateros
	Special Geographic Division	Special Geographic Area Division
language (CRLA to ELLNA spelling)	Ilocano	Ilokano/Iloko
	Sinugbuanong Binisaya	Sinugbuanong Binisaya/Cebuano
	Pangasinan	Pangasinan/Pangasinense
	Bahasa Sug	Tausug
	Chavacano	Chavacano/Chabacano
`offers_es`, `offers_jhs`, `offers_shs`	Yes	True
	No	False
	True	True
	False	False
`region_const`	ARMM; Cotabato City (Region XII); Sulu (BARMM)	BARMM; BARMM; Region IX
	21 Negros Island Region divisions under Region VI or VII	NIR

Note. The 29 rows of the release's rules file, in its order. `region_const` is a new column; the published `region` keeps its label apart from the ARMM rule. Labels are also stripped of leading and trailing whitespace. Source: Llego (2026a).

School IDs are taken as published and never re-linked. DepEd's README states that a school's ID can change between school years as the school expands or shrinks its offerings or merges with another, that this is "a minority of cases," and that DepEd intends to publish mapping data in future releases (DepEd, n.d.-a). A likely ID change is a school that leaves the files and reappears the next year under a new ID with the same name, municipality and barangay. There are 146 such cases over the eight transitions between school years (Table 4). The release lists them in `likely_id_changes.csv` and counts each as an exit plus an entry, so a user can link them or not.

2.4 Blank cells, zeros and implausible values

Three kinds of empty or doubtful cell are kept apart. A structural blank is a cell for a variable that does not apply: a level the school does not offer, a variable not collected that year, or a non-public school in a public-only module. An unexpected blank is an empty cell where the variable applies. An implausible value is a reported number that

is possible but doubtful. Five rules handle them.

1. A blank grade cell in an enrollment file is 0 when the school does not offer that level (its `offers_*` flag is False). From SY 2017–18 to SY 2024–25, DepEd codes such cells 0, and the only blank grade cells are in columns that did not exist that year (Table 6).
2. SY 2025–26 writes blanks where earlier years wrote zeros. A school that reports at least one level but leaves an offered level entirely blank is read as 0 at that level. This applies to 128 levels, of which 89 were 0 in SY 2024–25; the rest held 4,354 learners in that year.
3. The 46 schools that left every grade cell blank in SY 2025–26 (35 private, 8 PSO and 3 public; the 31 of them present in SY 2024–25 held 468 learners that year) stay in the panel with `enrol_missing` true and their total, sex and level totals blank, and they are left out of that year's totals. Grade and strand columns of the levels they do not offer read 0.
4. In the personnel, facilities, ICT and WASH modules, non-public rows stay blank, and `personnel_has_data`, `facilities_has_data`, `ict_has_data` and `wash_has_data` record whether a school has any data in each. An unexpected blank stays blank in its own column, but school totals sum the cells that are present, so a blank level counts as 0 in a total. Thus 645 of the 779 public schools with every nationally funded junior or senior high teacher column blank still have a `ptr`, and 42 of the 669 with no elementary classroom count have a `pcr`. An indicator built on the blank cell itself, such as `pcr_es` for those 669 schools, is blank.
5. Implausible values are flagged, not altered (Section 4.3). The one capped value is the derived `comp_acad_w`, which winsorizes academic computer counts at the public-school 99th percentile, 486, because 108 ICT cells exceed 1,000 and one school reports 15,699 tablets. The raw count stays in `comp_acad_total`.

2.5 Linkage into the SY 2023–24 school master file

The school master file starts from the SY 2023–24 learner file (60,167 schools) and joins the personnel, facilities, WASH, ICT, NAT, ELLNA and CRLA files to it on `school_id`, keeping every row of the learner file. The other SY 2023–24 school files and the NAT and ELLNA files match it in every row (Section 4.1). The CRLA file is for SY 2025–26, and 6,090 of its 6,121 schools match; the other 31 first appear in the SY 2024–25 enrollment file and have no SY 2023–24 data. They are kept in `crla_2025_outcomes` (Section 2.7).

The master file sums the source columns into school totals, by level where the source has levels. Teachers are the 36 nationally funded teacher, master teacher and instructor positions, including SPED/SNED and Special Science teachers but not head teachers or principals, plus the 16 locally funded teacher counts. Toilet bowls are counted as male, female, PWD and shared, attached and standalone, with non-functional bowls in columns of their own. Academic computers are the desktops, laptops, all-in-ones, tablets, tablet PCs and virtual terminals that schools report for academic use, from every funding source. The ratios follow DepEd's definitions: total enrollment over teachers, instructional rooms, functional toilet bowls or academically used computers (DepEd, n.d.-b, Sec. VI; Table 3).

Three departures from the technical notes follow from the data. The notes add an annex's learners and teachers to its mother school before the teacher–learner ratio is computed (DepEd, n.d.-b, Sec. VI.6); no mother–annex mapping is published, so annexes stay separate rows, flagged by `is_annex`. The notes also define a seat–learner ratio from Kindergarten chairs, armchairs and desks (Sec. VI.8); the facilities file has no seat columns, so the ratio is not in the release. The proficiency bands of the assessment notes (DepEd, 2025) are applied to each school's mean percentage score (`nat_band`, `ellna_band`), because the files give school means and not learners' scores. A band therefore describes a school's mean, not a share of proficient learners.

2.6 The enrollment panel and the school spine

The panel stacks the nine enrollment files into one long table with one row per school for each school year in which the school appears: 547,425 rows over 66,967 school IDs. Each row has the harmonized labels and `region_const`, and enrollment by grade, level, sex and senior high strand. Its panel columns come from the years in which the ID appears: first and last year present, years absent between them, presence in the previous and the next year, entry and exit flags, the previous row's enrollment, and the year-on-year growth when the previous row is the immediately preceding year. The panel has 69 columns, and no school-year is dropped. The Strengthened Senior High School columns are included in the SY 2025–26 totals (268,830 learners); earlier years have no such columns.

Table 3. Derived indicators in the school master file

Column	Definition	Rows with a value
`ptr`, `ptr_es`	Learners per teacher: `enrol_total` / `teachers_total`; elementary: `enrol_es` / `teachers_es` (elementary positions plus Kindergarten and elementary local teachers)	47,674; 39,334
`pcr`, `pcr_es`	Learners per instructional room, all levels (`rooms_total`) or elementary (`rooms_es`)	46,939; 38,674
`ltr`, `ltr_es`	Learners per functional toilet bowl, all levels or elementary	46,454; 38,162
`no_functional_toilet`	Public school reporting no functional toilet bowl	47,818
`handwash_per_100`	Functional handwashing facilities, group and individual, with or without soap, per 100 learners	44,186
`comp_acad_w`	`comp_acad_total` capped at 486, the public-school 99th percentile	47,817
`computers_per_100`, `clr`	`comp_acad_w` per 100 learners; learners per `comp_acad_w` where the school has at least one academic computer	47,813; 36,040
`has_academic_computer`	Public school reporting at least one academic computer	47,817
`multigrade_share_es`	Elementary learners in multigrade classes / (monograde + multigrade)	49,152
`buildings_bad_share`	Buildings needing major repair, for condemnation or for demolition / all buildings	47,185
`nat_band`, `ellna_band`	Band of the school's mean percentage score: Not Proficient 0–24, Low Proficient 25–49, Nearly Proficient 50–74, Proficient 75–89, Highly Proficient 90–100	6,636; 5,752
`nat_proficient_school`, `ellna_proficient_school`	School mean of 75 or more	6,636; 5,752
`teachers_outnumber_learners`, `computers_outnumber_learners`	Flags for two implausible values (Table 7)	47,818; 47,818
`log_enrol`	Natural logarithm of `enrol_total`, schools with learners	60,052

Note. Input indicators are computed for public schools only. `ptr`, `pcr`, `ltr`, `clr` and their elementary versions are blank where enrollment or the denominator is zero or missing; `computers_per_100` is 0 for the 11,773 schools with learners and no academic computer. Bands follow DepEd (2025). Source: `codebook.csv` (Llego, 2026a).

The school spine has one row per school ID (66,967): its first and last year, the number of years present, gaps, entry and exit flags, a presence flag for each year, and its sector, division and constant-boundary region in its last year. Of the 66,967 IDs, 55,029 appear in all nine years, 5,404 first appear after SY 2017–18, 6,763 last appear before SY 2025–26, and 483 leave and later return. Three schools change sector between their first and last year, all from public to SUC/LUC. Table 4 gives the panel by school year.

2.7 The assessment files

`crla_2025_outcomes` has one row for each of the 6,121 schools in the CRLA file. For each school it keeps the learners assessed, the counts at each of the five reading levels, and the shares at grade level and at Low Emerging as proportions. For each of the six grade-and-language blocks it keeps the learners assessed, those at grade level and those at Low Emerging. It adds the school's SY 2025–26 Grade 1 to 3 enrollment, sector and constant-boundary region, which the SY 2025–26 enrollment file gives for 6,113 of the schools. The block counts at High Emerging, Developing and Transitioning are in DepEd's file but not in the release.

Table 4. The enrollment panel by school year

School year	Schools (rows)	Entries	Exits	Likely ID changes	Schools with no enrollment reported	Total enrollment
2017–18	61,563	–	724	–	0	26,311,949
2018–19	61,916	1,077	1,034	19	0	27,018,509
2019–20	61,923	1,041	1,711	27	0	27,030,391
2020–21	60,957	745	1,129	13	0	26,227,022
2021–22	60,429	601	852	49	0	27,560,661
2022–23	60,137	560	699	3	0	27,794,282
2023–24	60,167	729	630	25	0	27,081,292
2024–25	60,129	592	476	2	0	26,400,182
2025–26	60,204	551	–	8	46	25,935,863
All years	**547,425**			**146**	**46**	

Note. Entries: IDs absent in the previous year's file, including IDs that return after a gap. Exits: IDs absent in the next year's file. Likely ID changes: exits from the previous year whose name, municipality and barangay reappear under a new ID in this year. Total enrollment: the sum of `enrol_total` over the year's rows, all sectors; the 46 schools with no enrollment reported in SY 2025–26 are not in it. Source: Llego (2026a).

`school_assessment_shrunk` has one row for each of the 6,840 schools with a NAT or ELLNA result. Beside each published school mean and its number of test-takers, it gives an empirical Bayes estimate (Casella, 1985) that pulls the school mean toward its division mean with weight $n/(n + \sigma^2/\tau^2)$. Here n is the number of test-takers, σ^2 the variance between learners within a school, and τ^2 the variance between schools within a division. The variances are estimated by the method of moments: $\sigma^2 = 54$ and $\tau^2 = 146$ for the NAT, and $\sigma^2 = 98$ and $\tau^2 = 164$ for the ELLNA. A NAT school mean rests on a median of six sampled learners, and the median weight on a school's own NAT mean is 0.94.

`assessment_coverage_by_division.csv` has one row for each of the 63 divisions in the five regions the assessments cover. Each row gives the division's public schools, those with Grade 3 and with Grade 6 learners, the schools with NAT, ELLNA and CRLA results, and the ELLNA and NAT shares of the Grade 3 and Grade 6 schools.

3. Data records

The release is recorded under the DOI https://doi.org/10.5281/zenodo.24624443 (Llego, 2026a). Each file is at https://data.nash.com.ph/deped-data/v1.1/ followed by the file name, and the landing page, https://nash.com.ph/blog/download-cleaned-deped-school-data, links each file and documents every column (Llego, 2026d). The folder name is the version: the data files behind a URL do not change, and a later version is a new folder. Release 1.0 (https://doi.org/10.5281/zenodo.24621194) stays online at https://data.nash.com.ph/deped-data/v1/. Release 1.1 differs from it only in `region_const` for the 422 school-year rows of Section 2.3 and in its documentation (Section 5.3), with the `README.md` and `LICENSE.txt` of November 28, 2026. Table 5 lists the files.

Each dataset comes as gzip-compressed UTF-8 CSV with a header row and as Parquet with the same values and types. The panel's CSV is split into one file per school year, and its Parquet file holds all nine years. In the CSV files a blank cell is a missing value, never a zero.

`codebook.csv` has one row for each of the 343 columns of the five datasets, with six fields: `file`, `column`, `type` (integer, number, text or boolean), `block`, `non_blank_rows` and `description`. The count of non-blank rows shows where a column applies; personnel columns, for example, are filled for public schools only. The school master's 207 columns fall in 14 blocks: identification (2), geography (8), classification (10), enrollment (19), learner types (8), personnel (15), facilities (29), WASH (34), ICT (15), NAT (7), ELLNA (7), CRLA (29), assessment coverage (5) and derived indicators (19).

`CHECKSUMS.sha256` gives the SHA-256 digest of 24 files: the 20 data files, the codebook, the rules file, the README and the license. It is in the format that `sha256sum -c` reads. `manifest.json` lists the same files with their sizes, row counts and digests for scripts.

Table 5. The files of release 1.1

File	One row per	Rows	Columns	Size
`school_master_2023.csv.gz`, `.parquet`	school in DepEd's SY 2023–24 files, all sectors	60,167	207	8.0 MB; 6.6 MB
`enrollment_panel_2017-18.csv.gz` to `enrollment_panel_2025-26.csv.gz` (nine files)	school in that school year (Table 4)	547,425	69	2.7–3.7 MB each
`enrollment_panel_long.parquet`	school per school year present, all nine years	547,425	69	17.2 MB
`school_spine.csv.gz`, `.parquet`	school ID in any enrollment file	66,967	24	882 KB; 1.4 MB
`crla_2025_outcomes.csv.gz`, `.parquet`	school in the CRLA file	6,121	33	228 KB; 273 KB
`school_assessment_shrunk.csv.gz`, `.parquet`	school with a NAT or ELLNA result	6,840	10	261 KB; 255 KB
`likely_id_changes.csv`	likely school-ID change	146	7	13 KB
`assessment_coverage_by_division.csv`	division in the five assessed regions	63	10	4 KB
`codebook.csv`	column of the five datasets	343	6	44 KB
`harmonisation_rules.csv`	label mapping (Table 2)	29	3	1 KB
`README.md`, `LICENSE.txt`	(text)			6 KB; 2 KB
`CHECKSUMS.sha256`, `manifest.json`	(lists of the files above)			2 KB; 8 KB

Note. Sizes as listed in `manifest.json` (1 KB = 1,024 bytes; 1 MB = 1,048,576 bytes), except those of `CHECKSUMS.sha256` (2,231 bytes) and `manifest.json` (8,037 bytes), which are the files' own sizes. The 24 files that `CHECKSUMS.sha256` lists total 69,879,255 bytes (66.6 MB). Source: Llego (2026a).

4. Technical validation

The counts in this section are computed from DepEd's source files (DepEd, n.d.-a) by the cleaning pipeline's audit; Llego (2026c) reports the main ones.

4.1 Keys and matching across files

No school ID is duplicated in any of the 17 files, and no file has more than one row per school; grades and sexes are columns, not rows. In each school year, 130 to 250 rows share a school name, municipality and barangay with another row under a different ID. These are mostly annex schools and public and private schools of the same name, and all are kept as distinct schools. The six SY 2023–24 school files hold the identical set of 60,167 IDs, and the learner file's enrollment columns equal the enrollment file's in every cell. The NAT and ELLNA files match the SY 2023–24 files in every row, and their region and division labels agree with the learner file's in every row. All 6,121 CRLA schools are in the SY 2024–25 enrollment file, and 6,113 are in the SY 2025–26 file.

4.2 Blank cells

In the SY 2023–24 public-only modules, every row of the 12,349 private, SUC/LUC and PSO schools is blank, as the technical notes lead a user to expect (DepEd, n.d.-b, Sec. VIII.d–f). Among the 47,818 public schools, one row is entirely blank in each of the personnel, facilities and ICT files, and none in the WASH file. Unexpected blanks among public schools that offer a level are few except in two places. Of the 39,346 public schools offering elementary, 669 have no elementary classroom count. Every nationally funded teacher column is blank at the junior high level in 595 of the 10,561 schools offering it, at the senior high level in 213 of 7,743, and at the elementary level in 17. Utility fields are blank in one to three schools each.

Table 6 gives the blank grade cells in the enrollment files; Section 2.4 gives the panel's rules for them. Outside the grade cells, `school_district` is blank in every row of SY 2017–18 and `annex_status` in every row before SY 2019–20, because neither variable was collected.

Table 6. Blank grade cells in the enrollment files

School year	Grade cells	Blank cells	Where the blanks are
2017–18	3,816,906	246,252	Grade 11 and 12 columns for tracks unique to some schools, not collected before SY 2024–25
2018–19	3,838,792	371,496	The same columns (247,664) and the junior high non-graded columns (123,832), not collected that year
2019–20	3,839,226	247,692	The unique-track columns
2020–21	3,779,334	243,828	The unique-track columns
2021–22	3,746,598	241,716	The unique-track columns
2022–23	3,728,494	240,548	The unique-track columns
2023–24	3,730,354	240,668	The unique-track columns
2024–25	3,727,998	0	None
2025–26	3,732,648	1,741,150	Blanks written where earlier years wrote zeros; 1,403 schools have every Kindergarten to Grade 10 cell blank (1,347 of them purely senior high schools), and 46 have every grade cell blank

Note. Grade cells are the enrollment cells by grade, sex and strand, in the 62 columns common to all nine files. From SY 2017–18 to SY 2024–25, a grade that a school does not offer is coded 0, not blank. Source: computed from DepEd (n.d.-a).

The NAT file has no blank cell. In the ELLNA file, `numeracy_mps` is blank in 154 rows and `mother_tongue_mps` in 1,925, all of them structural: according to the file's README, schools assessed in English/Filipino take no Numeracy component, and schools assessed in English/Filipino or Tagalog take no Mother Tongue component. The CRLA file has no blank cell.

4.3 Impossible and implausible values

No file has a negative count, a mean percentage score outside 0 to 100, or a share above 1. The values in Table 7 are possible but doubtful. Each is kept as reported, and a flag or the definition of an indicator handles it.

4.4 Sums and internal consistency

In all 6,121 CRLA rows, the five level counts sum to `n_test_takers`, and in each of the six grade-and-language blocks the level counts sum to the block total. DepEd's README does not say how the school-level counts are derived, and they differ from the sum of the three Mother Tongue blocks in 764 schools (Table 7), so the block counts are the safer unit. No learner-type count in the learner file exceeds the enrollment cell it belongs to. In all 49,343 schools with class-structure data, monograde and multigrade enrollment add up to elementary enrollment. In Table 4, each year's rows less its exits plus the next year's entries equal the next year's rows.

4.5 Coverage of the assessment files

All three assessment files cover the same five regions (CAR, Region I, Region III, Region VIII and Region IX) and the same 44 of those regions' 63 divisions (Table 8). None of the other 19 divisions, mostly small city divisions, and none of the other regions has a result. The SY 2023–24 NAT and ELLNA rounds were censuses of schools (DepEd, 2025), so the files are a selection from them. Within the five regions, the SY 2023–24 files list 15,159 public schools, 12,778 of them offering elementary; the ELLNA file covers 5,752 of these (45%) and the NAT file 5,822 public schools (46%). Inside the 44 divisions, the ELLNA file covers a median of 56% of the public schools with Grade 3 learners, from 5% to 100%.

The selection rule is not documented in the files. In the 44 divisions, covered and uncovered schools have similar Grade 3 enrollment (a median of 25 against 23 learners) and a similar mix of annex status, and the standardized mean differences between the 5,752 assessed and 6,094 unassessed schools are below 0.25 on every input. The outcomes of the unassessed schools are unknown.

Table 7. Implausible values and their handling in the release

Value	Count	Handling
Public schools with learners and no teacher, national or local	139	`ptr` blank
Public schools with more teachers than learners	7	Flagged in `teachers_outnumber_learners`
Public schools with learners and no instructional classroom	267	`pcr` blank
Public schools with no functional toilet bowl	1,362	Kept as 0, a real shortage; `no_functional_toilet` true, `ltr` blank
Public schools with more academic computers than learners (24 with more than twice as many)	316	Flagged in `computers_outnumber_learners`
ICT cells above 1,000 (largest 15,699 tablets in one school)	108	Raw counts kept; `comp_acad_w` capped at 486
Public schools with no learners in SY 2023–24	4	Ratios blank
Schools with more than 10,000 learners in some year (largest 25,148, SY 2021–22)	192	Kept; large urban schools exist
Schools with more ELLNA test-takers than Grade 3 enrollment	544	Kept; the test was given in May 2024, enrollment is counted at the start of the year
Schools with more NAT test-takers than Grade 6 enrollment	39	Kept
Schools with more CRLA test-takers than Grade 1 to 3 enrollment, SY 2025–26	2,129	Kept
CRLA schools whose school-level counts differ from the sum of the three Mother Tongue blocks	764	Both kept as published

Note. Counts of schools unless stated; school inputs are for SY 2023–24. Source: computed from DepEd (n.d.-a).

Table 8. The three assessment files

Assessment	School year, grade	Design of the round (DepEd, 2025)	Schools in the file	Matched to SY 2023–24 files
NAT	SY 2023–24, Grade 6	Census of schools, sample of learners; April 2024	6,636 (5,822 public, 814 private)	6,636
ELLNA	SY 2023–24, Grade 3	Census of schools and of learners; May 2024	5,752 public	5,752
CRLA	SY 2025–26, Grades 1–3	Not described in the technical notes	6,121 public	6,090

Note. Each file covers 44 divisions in CAR and Regions I, III, VIII and IX. Sources: DepEd (2025); computed from DepEd (n.d.-a).

4.6 Checks in the release build

The script that writes the release stops if any released column lacks a codebook description, so every column in the files is documented. It writes the gzip files without a timestamp, so a run on the same inputs reproduces the same bytes and checksums. A downloaded copy can be checked with `sha256sum -c CHECKSUMS.sha256` in the folder that holds the files.

5. Usage notes

5.1 Terms of use

Two sets of terms apply to the release (Llego, 2026d; `LICENSE.txt`). The five cleaned datasets and the two helper tables are built from DepEd's public datasets, which DepEd publishes under its Terms of Use and Fair Use Policy (DepEd, n.d.-a). Those terms require anything built on DepEd's data to be distributed under the same terms, and they do not allow added restrictions, so the data files are distributed under DepEd's terms, not NASH Reviewer's. The release states the conditions as follows. If you use the data files:

- credit the Department of Education as the source of the data;
- do not suggest that DepEd is involved in, associated with, or endorses your work;
- follow DepEd's guidelines on assessment data: do not compare test scores across school years, and do not rank schools;
- read DepEd's full terms before you publish anything built on the files.

DepEd's terms also say its materials may not be modified beyond editorial changes without its approval. The original DepEd files were never edited, and every change in the release is listed in the harmonization rules, the codebook and the README. NASH Reviewer will change or withdraw the release if DepEd asks.

The cleaning pipeline, the codebook, the harmonization rules, the README and the documentation are © 2026 Mark Anthony Llego, NASH Reviewer, all rights reserved. Reusing, republishing or adapting any of them needs written permission first, asked for through the contact form at https://nash.com.ph/contact or, as `LICENSE.txt` gives it, by email to support@nash.com.ph, saying who is asking, what they want to use and how.

5.2 Limits to read before analysis

- **Divisions by school ID across the boundary change.** The Special Geographic Area Division's schools carry the division North Cotabato before SY 2021–22; match them by school ID for a division series, since grouping by `region_const` and division gives a "BARMM / North Cotabato" division in SY 2017–18 to SY 2020–21.
- **Assessment results describe 44 divisions, not the country.** They cover the assessed schools in 44 divisions of five regions, chosen by a rule DepEd has not published (Section 4.5).
- **Test scores cannot be compared across years.** DepEd's notes say that "the difficulty of the tests may vary from one year to another despite having the same test specification," so that "percent correct may not be comparable from year to year, hence, not valid to describe trends" (DepEd, 2025, ELLNA Sec. VI). Each file holds one round.
- **A NAT school mean rests on a median of six learners.** School-level work should use `school_assessment_shrunk` and report the number of test-takers. DepEd's terms rule out ranking schools.
- **Every file is a school aggregate.** Nothing describes an individual learner, so shares of proficient learners cannot be computed; a band describes a school's mean.
- **The counts come from different points in time.** Enrollment is counted at the start of the school year, while the NAT was given in April 2024 and the ELLNA in May 2024 (Tables 7 and 8). The school inputs are for SY 2023–24 and the CRLA results for SY 2025–26; the 31 CRLA schools first seen in SY 2024–25 lack inputs.
- **Inputs describe public schools only**, and annexes are separate rows, so ratios are not adjusted for annexes as the technical notes prescribe.
- **Division names repeat and regions change.** Group by region and division together; use `region_const` for series across years and `region` to match a year's published tables.
- **School IDs are not re-linked.** A likely ID change appears as an exit and an entry; `present_all_years` in the spine selects the 55,029 schools present in all nine years.
- **Two DepEd indicators cannot be computed.** Gross and net enrollment rates need the Philippine Statistics Authority's population projections (DepEd, n.d.-b, Sec. I.d), which are not in the files, and the seat–learner ratio needs seat counts.
- **Totals are sums over the released rows, not DepEd's official statistics.** For official figures, cite DepEd's own releases.

5.3 Documentation errors of release 1.0, fixed in release 1.1

Release 1.1 fixes six entries that were wrong in the release 1.0 documentation; the data files were not affected. The 1.0 README describes `assessment_coverage_by_division.csv` as covering "the 44 covered divisions," but the file has 63 rows, one for each division in the five regions, 19 of them with no assessed school. The 1.0 `codebook.csv` describes `shs_sshs_acad` and `shs_sshs_techpro` as special senior high school programs; DepEd's SY 2025–26 README names them the Grade 11 enrollment in the Strengthened Senior High School Academic and Technical-Professional tracks. It also gives the type of `is_public` and `is_annex` in

`school_master_2023` and of `sector_changed` in `school_spine` as text; all three hold true and false values and are Boolean in the Parquet files. One gap remains in release 1.1: the Special Geographic Area Division rule for `region_const` is in its README but not in `harmonisation_rules.csv` or the codebook.

5.4 Reading, citing and correcting the files

Every dataset reads with `pd.read_parquet` in Python or `arrow::read_parquet` in R, and each CSV file fits in a spreadsheet once unzipped, where `school_id` is best kept as text. Summing `enrol_total` by `sy_label` in the panel gives the totals in Table 4, the series behind NASH Reviewer's report on enrollment from SY 2017–18 to SY 2025–26 (Llego, 2026b).

A work that uses the files credits DepEd as the source of the data and cites the release as given under Data availability, followed by "Source data: Department of Education (Philippines) public datasets." A work that relies on the cleaning rules or on the counts in Section 4 can cite this report as well. An error in the files can be reported through https://nash.com.ph/contact with the file, column and `school_id`; corrections go into the next version.

Data availability

The release is Llego, M. A. (2026). *Cleaned DepEd school-level datasets, SY 2017-18 to SY 2025-26* (Version 1.1) [Data set]. NASH Reviewer. https://doi.org/10.5281/zenodo.24624443. Its files are free to download, without sign-up, at https://data.nash.com.ph/deped-data/v1.1/ followed by the file name (Section 3). DepEd's source files and technical notes are on DepEd's Machine Ready Files page (DepEd, n.d.-a).

Version history. Version 1.0 (November 28, 2026; https://doi.org/10.5281/zenodo.24622267) described release 1.0. Version 1.1 (November 28, 2026) describes release 1.1 and corrects the SY 2024–25 enrollment of the 46 schools with every SY 2025–26 grade cell blank: 468 learners, not about 465.

Code availability

The code of the cleaning pipeline and its documentation are © 2026 Mark Anthony Llego, NASH Reviewer, all rights reserved, and are not publicly released. Every rule the pipeline applies is listed in the release's `harmonisation_rules.csv`, `codebook.csv` and `README.md` and in Section 2 of this report.

Funding and competing interests

This work received no funding. The author runs NASH Reviewer, a review service for the National Assessment for School Heads, and has no affiliation with DepEd.

References

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician, 39*(2), 83–87. https://doi.org/10.1080/00031305.1985.10479400

Department of Education. (n.d.-a). *Machine ready files* [Data sets]. https://www.deped.gov.ph/machine-ready-files/

Department of Education. (n.d.-b). *Technical notes on the school characteristics data* [Technical notes]. Policy and Planning Service. https://www.deped.gov.ph/machine-ready-files/

Department of Education. (2025). *Technical notes on learning outcomes data* [Technical notes]. Bureau of Education Assessment. https://www.deped.gov.ph/machine-ready-files/

Llego, M. A. (2026a). *Cleaned DepEd school-level datasets, SY 2017-18 to SY 2025-26* (Version 1.1) [Data set]. NASH Reviewer. https://doi.org/10.5281/zenodo.24624443

Llego, M. A. (2026b, October 25). DepEd enrollment trends SY 2017-2018 to SY 2025-2026. *NASH Reviewer Blog*. https://nash.com.ph/blog/deped-enrollment-trends-2017-2026

Llego, M. A. (2026c, October 25). DepEd open data quality: Errors, blanks and coverage limits. *NASH Reviewer Blog*. https://nash.com.ph/blog/deped-open-data-quality

Llego, M. A. (2026d, October 25). Download cleaned DepEd school data, SY 2017-18 to 2025-26. *NASH Reviewer Blog*. https://nash.com.ph/blog/download-cleaned-deped-school-data