**­­­MPA breakfast analysis processes and terms – for IT team**

This document describes, in detail, the steps needed to take the breakfast and lunch data submitted by NKH partners and state agencies, transform it into a common format, and then use it to report on our results and create useful visualizations.

The document first presents some key terms, and then goes into detail providing guidance on how to move from through this process: Raw Data  Clean Data  Consolidated Data  Consolidated & Enriched Data  Annual School Level Data  Results Data and Visualizations.

**Key terms**

* Raw Data – the original data we receive from the partner/agency. This file is completely untouched by Share Our Strength/SOS (i.e. no formatting changes, et al). The ideal dataset is a plain data table with one row per school (or site) per month per school year, but format varies widely by state agency and some Raw Data requires more cleaning than others. The number of schools is dependent on the geography of the data (e.g. statewide = all schools; single county = just the schools in that county; etc.).
* Partially Clean Data – MPA makes a copy of the Raw Data and pastes as plaintext into a new spreadsheet. In the ideal scenario, cleaning involves basic formatting changes and adding some columns with calculations. If the Raw Data comes in a particularly wonky format, the process to partially clean the data may be more arduous. It is necessary to get the data into a format where there is one row per school (or site) per month per school year. At this stage, data is transformed enough for MPA to conduct analysis in Excel (e.g. for quarterly analysis, planning landscapes, etc.), but additional transformation is needed to standardize data to be ready to ingest into Data Lake.
* Clean Data – data is scrubbed in accordance with details described in the below Clean Data section and is ready to be ingested into the Data Lake.
* Consolidated Data – The Clean Data files for each state and school year are appended into one dataset to allow for analysis across years and across states.
* Consolidated & Enriched Data – The Consolidated Data is enriched to fill in gaps and include data from additional sources, such as NCES, grants or API-generated fields.
* Annual School Level Data: Annual School Level Data is site-level data (like Clean Data), but is now aggregated to the annual level (whereas Clean Data is typically on a monthly level). Annual School Level Data will be the foundation for Results Data and most visualizations. For our purposes, annual refers to a school year period (September through May) not a calendar year.
* Results Data and Visualizations – the data we actually share with the Field team and others. Some of the Results Data fields come directly from the Annual School Level Data (static text fields like School Name), while other results are derived by performing additional calculations on the Annual School Level Data and/or aggregating data on different levels. Examples of Results Data include:
  + **School-level list** that shows key information and metrics on the school level for the current school year, and progress from the previous school year.
  + **Target area-level progress table** that shows changes in breakfast participation, lunch participation, and the breakfast/lunch participation rate from the previous school year to the current school year.
  + **District-level table** showing key metrics at the district level for the current school year, and progress from the previous school year.
  + ***Important note:*** the Field team is the key customer for results and numbers derived from the state agency data. The numbers Communications and CED use in external reports, funder proposals, etc. typically come from our USDA data, which is cleaned and analyzed differently than state agency data.

**Clean Data**

The table below contains a list of the key columns we look for in each breakfast dataset. The Raw Data may contain additional columns to those listed below; additional columns should not be retained in Clean Data. The table shows the formatting changes MPA applies as well as additional calculations and assumptions.

White cells below represent cleaned raw data, while blue cells represent enriched data. In other words:

* White cells below describe columns that are present in state/partner submitted raw data but need to be cleaned/renamed to match the column titles below.
* Blue cells below describe enriched columns, columns that are either generated/calculated based on the white columns, or added from nationwide (non state-specific) datasets.

All columns listed below are needed for an ideal clean dataset, but raw data columns (white cells) most critical to our analysis are noted under “Critical Data.”

Note that content in purple such as Month([Claim Date]) is used internally when creating new recipes, please do not edit them.

| **No.** | **Column** | **Definition** | **Data Type/**  **Storage\*[[1]](#footnote-2)** | **Display Format** | **MPA notes and what to do if this column is missing** | **Critical Data** |
| --- | --- | --- | --- | --- | --- | --- |
|  | School ID | May appear in data as "Site ID" |  | Plaintext (note: will be all numbers but need formatted as text to preserve any leading zeros e.g. 0002368) |  | Yes |
|  | School Name |  |  | Plaintext |  | Yes |
|  | District ID | May appear in data as "Sponsor ID" |  | Plaintext (note: will be all numbers but need formatted as text to preserve any leading zeros e.g. 0002368) |  | Yes |
|  | District Name |  |  | Plaintext |  | Yes |
|  | State-Physical | The state that the site is physically located it. |  | Plaintext, 2-letter state abbreviation | Ok if this is missing.  For mapping purposes only. Filtering and results/reporting roll-up, etc. should use State-Reporting (#45). |  |
|  | County |  |  | Plaintext |  |  |
|  | City |  |  | Plaintext |  |  |
|  | Zip Code |  |  | Plaintext, 5-digit zip code (note: will be all numbers but need formatted as text to preserve any leading zeros e.g. 02368) |  |  |
|  | Street Address-Line 1 |  |  | Plaintext |  |  |
|  | Street Address-Line 2 |  |  | Plaintext | Not applicable for many schools |  |
|  | School Level-Original | Primary, middle, high, other, etc. |  | Plaintext | Preserve original classification of school level. |  |
|  | School Type-Original | Public, private, charter, RCCI, etc. |  | Plaintext | Preserve original classification of school type. |  |
|  | CEP (Y/N) |  |  | Plaintext | In raw data, if #13-17 is in one column (“Provision”), split out each provision into its own Y/N column |  |
|  | Provision 1 (Y/N) |  |  | Plaintext | In raw data, if #13-17 is in one column (“Provision”), split out each provision into its own Y/N column |  |
|  | Provision 2 (Y/N) |  |  | Plaintext | In raw data, if #13-17 is in one column (“Provision”), split out each provision into its own Y/N column |  |
|  | Provision 3 (Y/N) |  |  | Plaintext | In raw data, if #13-17 is in one column (“Provision”), split out each provision into its own Y/N column |  |
|  | Universal Free (Y/N) |  |  | Plaintext | In raw data, if #13-17 is in one column (“Provision”), split out each provision into its own Y/N column |  |
|  | Claim Date | Month and year; may appear in data as "process date", "submitted date", etc. |  | MM/DD/YY | Standard is to use first day of month – e.g. September 2018 should read “09/01/18) | Yes |
|  | Claim Month | Only required if “Claim date” is missing  Month([Claim Date]) |  | Plaintext | Use to generate #18 Claim date, if Claim date is missing | Yes if #18 is missing |
|  | Claim Year | Only required if “Claim date is missing”  YEAR([Claim Date]) |  | Number | Use to generate #18 Claim date, if Claim date is missing | Yes if #18 is missing |
|  | Enrollment-Total | Total number of kids enrolled in the school or district |  | Number, no decimal places, commas separating thousands | Calculate manually if missing. Calc: #22 + #23 + #24  ***Enrollment Note – it is important to be sure this is the total number of kids enrolled in the school. Some datasets report only the number of kids enrolled in the lunch program, which is not what we need.*** | Yes |
|  | Enrollment-Paid | Total number of kids who must pay for school meals (i.e. not eligible for either free meals or reduced meals) |  | Number, no decimal places, commas separating thousands | See “Enrollment Note” in #21 |  |
|  | Enrollment-Free | Total number of kids who are eligible for free meals in the school or district |  | Number, no decimal places, commas separating thousands | See “Enrollment Note” in #21 | Yes |
|  | Enrollment-Reduced | Total number of kids who are eligible for reduced meals in the school or district |  | Number, no decimal places, commas separating thousands | See “Enrollment Note” in #21 | Yes |
|  | Enrollment-Free and Reduced | Total number of kids who are eligible for free and reduced meals in the school or district |  | Number, no decimal places, commas separating thousands | See “Enrollment Note” in #21  Okay if this data is missing, as long as we have #23 and #24. |  |
|  | Lunch Meals-Free |  |  | Number, no decimal places, commas separating thousands |  | Yes |
|  | Lunch Meals-Reduced |  |  | Number, no decimal places, commas separating thousands |  | Yes |
|  | Lunch Meals-Paid |  |  | Number, no decimal places, commas separating thousands |  |  |
|  | Lunch Meals-Free and Reduced |  |  | Number, no decimal places, commas separating thousands | Okay if this data is missing, as long as we have #26 and #27. |  |
|  | Breakfast Meals-Free |  |  | Number, no decimal places, commas separating thousands |  | Yes |
|  | Breakfast Meals-Reduced |  |  | Number, no decimal places, commas separating thousands |  | Yes |
|  | Breakfast Meals-Paid |  |  | Number, no decimal places, commas separating thousands |  |  |
|  | Breakfast Meals-Free and Reduced |  |  | Number, no decimal places, commas separating thousands | Okay if this data is missing, as long as we have #31 and #32 |  |
|  | Operating Days | Track 18.5 rule |  | Number, no decimal places, commas separating thousands | Enter manually if missing. Assume 18.5 service days per month if this, #35 and #36 are all missing. | Yes |
|  | Operating Days-Breakfast Only | Often there is a single column for operating days (also called service days) that will serve for both breakfast and lunch. This is ok. |  | Number, no decimal places, commas separating thousands | Enter manually if missing. Assume 18.5 service days per month if missing.  Okay if this data is missing if we have #34. | Yes if #34 is missing |
|  | Operating Days-Lunch Only | Often there is a single column for operating days (also called service days) that will serve for both breakfast and lunch. This is ok. |  | Number, no decimal places, commas separating thousands | Enter manually if missing. Assume 18.5 service days per month.  Okay if this data is missing if we have #34. | Yes if #34 is missing |
|  | Breakfast Delivery Model from Campaign Tracking Data-Original | Delivery model of breakfast service (Breakfast in the Classroom, Grab and Go, Traditional, etc.)  Present in few states |  | Plaintext | Preserve original classification of breakfast delivery model.  Populate this “Breakfast Delivery Model from…” column if model source is Campaign Tracking data.  May not have this data for all states/schools. |  |
|  | Breakfast Delivery Model from State Agency Tracking-Original | Delivery model of breakfast service (Breakfast in the Classroom, Grab and Go, Traditional, etc.) |  | Plaintext | Preserve original classification of breakfast delivery model.  Populate this “Breakfast Delivery Model from…” column if model source is State Agency Tracking data.  May not have this data for all states/schools. |  |
|  | Breakfast Delivery Model from Other Source-Original | Delivery model of breakfast service (Breakfast in the Classroom, Grab and Go, Traditional, etc.) |  | Plaintext | Preserve original classification of breakfast delivery model.  Populate this “Breakfast Delivery Model from…” column if model source is not listed elsewhere in this document.  May not have this data for all states/schools. |  |
|  | Breakfast Delivery Model from Other Source, Note Source Here | Source of #39 |  | Plaintext | If a source not listed elsewhere in this document is used to determine breakfast delivery model, create plaintext note in each per data dictionary that indicates the source of the data. |  |
|  | BAB Implementation Date from Campaign Tracking Data | Date of model implementation |  | MM/DD/YY | Populate this “BAB Implementation Date from…” column if implementation date source is from Campaign Tracking data.  May not have this data for all state/schools. |  |
|  | BAB Implementation Date from State Agency Tracking | Date of model implementation |  | MM/DD/YY | Populate this “BAB Implementation Date from…” column if implementation date source is from State Agency Tracking data.  May not have this data for all state/schools. |  |
|  | BAB Implementation Date from Other Source | Date of model implementation |  | MM/DD/YY | Populate this “BAB Implementation Date from…” column if implementation date source is from another source.  May not have this data for all state/schools. |  |
|  | BAB Implementation Date Other Source, Note Source Here | Source of #43 |  | Plaintext | If a source not listed elsewhere in this document is used to determine BAB implementation date, create plaintext note in each per data dictionary that indicates the source of the data. |  |
|  | State-Reporting | The state that the site reports its data to (this is almost always the same as the State-Physical (#5), except for some special cases near state borders). |  | Plaintext, 2-letter state abbreviation | Enter manually if missing based on file name and/or based on the person who sends us data.  This is the “state” field that should be used for filtering and results/reporting roll-ups (not State-Physical). |  |
| 64. | ISP | Identified Student Percentage | Percentage, include all decimal places | Percentage, no decimal places. | \*\*We are newly collecting this, starting with SY18-19 data\*\* |  |
|  | FR Lunch Meals | Calc: #26 + #27  IF ISNULL([Lunch Meals-Free]) AND ISNULL([Lunch Meals-Reduced]) THEN  IF ISNULL([Lunch Meals-Free and Reduced])  THEN NULL  ELSE  [Lunch Meals-Free and Reduced]  END  ELSE  IFNULL([Lunch Meals-Free],0) + IFNULL([Lunch Meals-Reduced],0)  END |  | Number, no decimal places, commas separating thousands | This column is always manually added by MPA  If #26 and/or #27 are missing, use #29. |  |
|  | FR Lunch ADP | Calc: IFERROR(#46 / #36,"")  IF ISNULL([FR Lunch Meals]) THEN NULL  ELSEIF ISNULL([Operating Days-Lunch Only]) AND ISNULL([Operating Days]) THEN NULL  ELSE  IF ISNULL([Operating Days-Lunch Only]) AND NOT ISNULL([Operating Days]) THEN  [FR Lunch Meals] /[Operating Days]  ELSE  [FR Lunch Meals]/[Operating Days-Lunch Only]  END  END | Decimal | Number, no decimal places, commas separating thousands | This column is always manually added by MPA  If #36 is missing, use #34 in place |  |
|  | FR Breakfast Meals | Calc: #30 + #31  IF ISNULL([Breakfast Meals-Free]) AND ISNULL([Breakfast Meals-Reduced]) THEN  IF ISNULL([Breakfast Meals-Free and Reduced])  THEN NULL  ELSE  [Breakfast Meals-Free and Reduced]  END  ELSE  IFNULL([Breakfast Meals-Free],0)+IFNULL([Breakfast Meals-Reduced],0)  END |  | Number, no decimal places, commas separating thousands | This column is always manually added by MPA  If #30 and/or #31 are missing, use #33. |  |
|  | FR Breakfast ADP | Calc: IFERROR(#48 / #35,"")  IF ISNULL([FR Breakfast Meals]) THEN NULL  ELSEIF ISNULL([Operating Days-Breakfast Only]) AND ISNULL([Operating Days]) THEN NULL  ELSE  IF ISNULL([Operating Days-Breakfast Only]) AND NOT ISNULL([Operating Days]) THEN  [FR Breakfast Meals] /[Operating Days]  ELSE  [FR Breakfast Meals]/[Operating Days-Breakfast Only]  END  END | Decimal | Number, no decimal places, commas separating thousands | This column is always manually added by MPA  If #35 is missing, use #34 in place |  |
|  | Unique ID | Calc: combine state-reporting, school ID, and district ID, separated in middle with a hyphen (e.g. DE-###-###)  [State-Reporting]+"@"+STR([School ID])+"@"+STR([District ID]) |  | Plaintext (note: will be all numbers but need formatted as text to preserve any leading zeros e.g. 0002368) | This column is always manually added by MPA |  |
|  | NCES ID | National Center for Education Statistics (NCES) School/site ID number  IF(LEN([District ID]) < 4)  THEN  LEFT("0000",4-LEN(STR([District ID])))+STR([District ID])  ELSE  [District ID]  END |  | Plaintext (note: will be all numbers but need formatted as text to preserve any leading zeros e.g. 0002368) | Currently, this column is added by MPA – currently done using VLOOKUP, etc. but perhaps new relational database can easily link-in this info for some states, and we may request it from other states. For some states, it is easy to join the raw data to the NCES ID numbers using the NCES ELSI database (e.g. via vlookup or other methods). In some states, some combination of the “School ID” and “District ID” matches the “State School ID [Public School]” field in the NCES ELSI data. However, in other states, these numbers do not match and there is currently no easy way to link the raw data to NCES data. |  |
|  | School Year |  |  | SY##-## (e.g. SY16-17) | May be present in raw data or may be generated column. Enter manually or calculate using claim date if missing.  Example: July 2016-June 2017 would be SY16-17. |  |
|  | Target Area | Region in which the campaign is focusing its efforts. Varies from state to state but is usually based on school district or county. |  | Plaintext | Not applicable for all states.  Standardize per MPA guidance (in works). |  |
|  | FR Enrollment | #25. If missing, #23 + #24  IF ISNULL([Enrollment-Free and Reduced])  THEN  IF ISNULL([Enrollment-Free]) AND ISNULL([Enrollment-Reduced])  THEN NULL  ELSE  IFNULL([Enrollment-Free],0)+IFNULL([Enrollment-Reduced],0)  END  ELSE  [Enrollment-Free and Reduced]  END |  | Number, no decimal places, commas separating thousands |  |  |
|  | FR Enrollment Percentage | Calc:  If CEP (#13) = “N”:  #54 / #21  If CEP (#13) = “Y”:  #26 / (#26 + #28). If #26 is missing, then #29 / (#29 + #28)  If CEP (#13) is missing from entire data source: #54 / #21  IF([CEP (Y/N)] == "N") THEN  [FR Enrollment] / [Enrollment-Total]  ELSEIF ([CEP (Y/N)] == "Y") THEN  IF ISNULL([Lunch Meals-Free]) AND ISNULL([Lunch Meals-Paid]) THEN  NULL  ELSE  IFNULL([Lunch Meals-Free],0) / (IFNULL([Lunch Meals-Free],0) + IFNULL([Lunch Meals-Paid],0))  END  ELSE  NULL  END | Percentage, include all decimal places | Percentage, no decimal places. | Max value for #55 should be 1 (i.e. 100%). If #55 calculates >1, overwrite as 1. |  |
|  | School Level-Standardized | Standardize to one of the following terms: Primary, Middle, High, Other, Unknown  Use MPA Standardization Guidance-v2.xlsx |  | Plaintext | Using #11, standardize per MPA guidance.  If school level (#11) is missing, we will enter it using NCES data, using formulas to determine school level using the school name, or use other means to determine school level. |  |
|  | School Type-Standardized | Standardize to one of the following terms: Charter, Nonpublic, Public, Other  Use MPA Standardization Guidance-v2.xlsx |  | Plaintext | Using #12, standardize per MPA guidance. |  |
|  | Breakfast Delivery Model from Grants Data-Standardized | Delivery model of breakfast service (BAB, BIC, Cafeteria, GNG, GNG Classroom, GNG Common Area, Other Second Chance, Second Chance Cafeteria, Second Chance GNG) |  | Plaintext | Populated using grants database, joining on NCES ID or unique ID.  The models may need slight standardization before ingestion. Preserve original classification of breakfast delivery model in raw grants data.  Populate this “Breakfast Delivery Model from…” using SOS Grants Data.  May not have this data for all states/schools. |  |
|  | Breakfast Delivery Model from Innovations Data-Standardized | Delivery model of breakfast service (BAB, BIC, Cafeteria, GNG, GNG Classroom, GNG Common Area, Other Second Chance, Second Chance Cafeteria, Second Chance GNG) |  | Plaintext | Populated using grants database, joining on NCES ID or unique ID.  The models may need slight standardization before ingestion. Preserve original classification of breakfast delivery model in raw grants data.  Populate this “Breakfast Delivery Model from…” using SOS Innovations Data.  May not have this data for all states/schools. |  |
|  | Breakfast Delivery Model from Campaign Tracking Data-Standardized | Standardized version of #37. Delivery model of breakfast service (BAB, BIC, Cafeteria, GNG, GNG Classroom, GNG Common Area, Other Second Chance, Second Chance Cafeteria, Second Chance GNG) |  | Plaintext | May not have this data for all states/schools. |  |
|  | Breakfast Delivery Model from State Agency Tracking-Standardized | Standardized version of #38. Delivery model of breakfast service (BAB, BIC, Cafeteria, GNG, GNG Classroom, GNG Common Area, Other Second Chance, Second Chance Cafeteria, Second Chance GNG)  Use MPA Standardization Guidance-v2.xlsx |  | Plaintext | May not have this data for all states/schools. |  |
|  | Breakfast Delivery Model from Other Source-Standardized | Standardized version of #43. Delivery model of breakfast service (BAB, BIC, Cafeteria, GNG, GNG Classroom, GNG Common Area, Other Second Chance, Second Chance Cafeteria, Second Chance GNG) |  | Plaintext | May not have this data for all states/schools. |  |
|  | BAB Implementation Date from Grants Data | Date of model implementation |  | MM/DD/YY | Populated using grants database, joining on NCES ID or unique ID. |  |

**Consolidated Data**

The Clean Data files for each state and school year are appended into one dataset to allow for analysis across years and across states.

**Consolidated & Enriched Data**

The Consolidated Data is enriched to fill in gaps (from Clean Data) and add additional information. This information is sourced from nationally based datasets, as opposed to state-specific datasets as is the case with Clean Data information. As such, importing this information can be done on the Consolidated Data file.

Immediate needs to validate and move to Results Data (**high priority for project progress**):

* Target Area, per a living document that defines target areas for each state. These target areas apply to all years of data.
* Schools to exclude: Different states filter out different types of schools (such as private schools) based on campaign priorities and data availability. The rules for filtering out schools differ from state to state and can differ from year to year. In this step, the Consolidated Data will be enriched with a living document specifying which schools we should include. In practice, this could be one column with values such as: “Include” and “Exclude”. These schools to exclude apply to all years of data. Note: we want to filter these schools out, rather than deleting them, as this information may be useful in the future.

Additional needs:

* Using NCES data (or other year data from the same state) to fill in gaps in the Consolidated Data for some states and years (such as enrollment data or school level information).
* Latitude and longitude from NCES and/or API.
* If possible, connecting to the grants database for information such as grant date, implementation date, and breakfast model. (We understand this probably won’t be possible in the near-term).
* State legislative district from API
* *This is not a comprehensive list and additional needs may be added.*

**Annual School Level Data**

Annual School Level Data is site-level data (like Clean Data), but is now collapsed/aggregated to the annual level (whereas Clean Data is typically on a monthly level). Annual School Level Data will be the foundation for Results Data and most visualizations. In other words, for each unique combination of Unique ID and school year, there will be only one row of data. For our purposes, annual refers to the time period reflecting a school year (September through May), not a calendar year.

The table below shows the fields that should be included in Annual School Level Data. You will see that this dataset mirrors Clean Data in many ways. However, some now irrelevant columns (like service days or meals served) no longer appear.

Follow these three steps to collapse the data on the annual level:

1. Drop the months of June, July and August
2. Drop any incomplete months of data: Most non-final datasets we receive will only have partial data for some months since school districts report their claims at different times. To identify and remove incomplete months, make a table that shows the number of claims for each month for each state. Do the same with the sum of F/R Breakfast ADP for each state. An incomplete month will logically have a lower number of claims and a lower sum of F/R Breakfast ADP. Usually you can quickly “eyeball” the data and determine if a month is incomplete. For example, if the dataset contains Sep-Jan data and the values for Sep, Oct, Nov, Dec are all between 1,000-1,200 but Jan is 700, that means January is probably incomplete. If the dataset contains Sep-Jan data and in Sep, Oct, Nov, Dec, about there were about 400 claims in each month, but in January there were only 300 claims, January is probably incomplete. Incomplete months should always be excluded from the analysis. You can expect slight variations between each month (due to differing # school days, holidays, etc.) but overall all months should be within the same range – a dramatic outlier is an indicator of incomplete data.
   1. Note: At current time (Jan 2019), MPA is only sharing Data Dictionaries for final & complete data. During MPA’s next round of mid-year analysis (in Feb & March 2019), we will work to develop a more clear-cut rule around this (for example, if the number of claims in the most recent month are 10% less than the annual average, it’s incomplete, etc.).
3. Data collapse based on column: After finishing these first two steps, the next step is dependent on the particular column. Each column will collapse based on one of the following three rules, as specified in the “Guidance for Collapsing/Aggregating data” column in the below table.
4. **Common/Modal Value:** This is most common for fields that should typically be static (such as School Name, District Name, School Level, etc.). Generally, these values will be identical for all rows of one Unique ID in a given school year. In this case, the Annual School Level Data should show the common value for that column. However, if there are inconsistencies in these values (such as if School Name is “Washington Elementary” 8 times and “Washington Elem” once), use the most common/modal nonmissing value. In this example, it would be “Washington Elementary.” If there are two modes, use the value for October.
5. **October Value:** Instead of using the modal value, populate this field with the value from October. If there is a missing value for October but not other months, use the modal value (for string values) or the average of the nonmissing months (for numeric values, excluding June, July and August).
6. **Average Across Months:** This is mostly used for calculating ADP. First, you sum up all of the values for that Unique ID for a particular school year—this is usually nine months, and it should never exceed nine months. Then you divide by the overall number of months in the dataset for that school year (this is nine for full-year data). *Very important note: per MPA measurement procedures, it is extremely important to divide these sums by the same number of months for all schools in a state. For a full year of data, you should always divide the total by nine, even if a month of data is missing for a particular Unique ID. However, if for an entire state the count of unique months reported is less than 9 (e.g. 8), check with MPA.*

**Note:** In some cases, raw data is submitted to us **already aggregated at the annual level** (example: LA SY16-17) or we may only have **one month of raw data and do not expect to obtain additional months** (example: OH SY16-17 – Oct 16). This will be noted in the “Time Interval” section of the data dictionary. Since the Clean Data for these states is de facto annual data, you should simply import content for appropriate fields from Clean Data (i.e. there is no need look for Oct values or modal values or average across months).

| **No.** | **Column** | **Data Type/**  **Storage\*[[2]](#footnote-3)** | **Display Format** | **Guidance for Collapsing/Aggregating data** |
| --- | --- | --- | --- | --- |
|  | *Long ID* |  | *School Year-Unique ID (note: IT should create this ID by concatenating the mentioned fields)* | *N/A* |
|  | School ID |  | Plaintext (note: will be all numbers but need formatted as text to preserve any leading zeros e.g. 0002368) | N/A |
|  | School Name |  | Plaintext | Common/Modal Value |
|  | District ID |  | Plaintext (note: will be all numbers but need formatted as text to preserve any leading zeros e.g. 0002368) | N/A |
|  | District Name |  | Plaintext | Common/Modal Value |
|  | State-Physical |  | Plaintext, 2-letter state abbreviation | Common/Modal Value |
|  | County |  | Plaintext | Common/Modal Value |
|  | City |  | Plaintext | Common/Modal Value |
|  | Zip Code |  | Plaintext, 5-digit zip code (note: will be all numbers but need formatted as text to preserve any leading zeros e.g. 02368) | Common/Modal Value |
|  | Street Address-Line 1 |  | Plaintext | Common/Modal Value |
|  | Street Address-Line 2 |  | Plaintext | Common/Modal Value |
|  | School Level-Original |  | Plaintext | Common/Modal Value |
|  | School Type-Original |  | Plaintext | Common/Modal Value |
|  | CEP (Y/N) |  | Plaintext | Common/Modal Value |
|  | Provision 1 (Y/N) |  | Plaintext | Common/Modal Value |
|  | Provision 2 (Y/N) |  | Plaintext | Common/Modal Value |
|  | Provision 3 (Y/N) |  | Plaintext | Common/Modal Value |
|  | Universal Free (Y/N) |  | Plaintext | Common/Modal Value |
|  | Enrollment-Total |  | Number, no decimal places, commas separating thousands | October Value |
|  | Enrollment-Paid |  | Number, no decimal places, commas separating thousands | October Value |
|  | Enrollment-Free |  | Number, no decimal places, commas separating thousands | October Value |
|  | Enrollment-Reduced |  | Number, no decimal places, commas separating thousands | October Value |
|  | Enrollment-Free and Reduced |  | Number, no decimal places, commas separating thousands | October Value |
|  | Breakfast Delivery Model from Campaign Tracking Data-Original |  | Plaintext | Common/Modal Value |
|  | Breakfast Delivery Model from State Agency Tracking-Original |  | Plaintext | Common/Modal Value |
|  | Breakfast Delivery Model from Other Source-Original |  | Plaintext | Common/Modal Value |
|  | Breakfast Delivery Model from Other Source, Note Source Here |  | Plaintext | Common/Modal Value |
|  | BAB Implementation Date from Campaign Tracking Data |  | MM/DD/YY | Common/Modal Value |
|  | BAB Implementation Date from State Agency Tracking |  | MM/DD/YY | Common/Modal Value |
|  | BAB Implementation Date from Other Source |  | MM/DD/YY | Common/Modal Value |
|  | BAB Implementation Date Other Source, Note Source Here |  | Plaintext | Common/Modal Value |
|  | State-Reporting |  | Plaintext, 2-letter state abbreviation | Common/Modal Value |
| 56. | ISP | Percentage, include all decimal places | Percentage, no decimal places | October Value |
|  | FR Lunch ADP | Decimal | Number, no decimal places, commas separating thousands | Average Across Months |
|  | FR Breakfast ADP | Decimal | Number, no decimal places, commas separating thousands | Average Across Months |
|  | Unique ID |  | Plaintext (note: will be all numbers but need formatted as text to preserve any leading zeros e.g. 0002368) | N/A |
|  | NCES ID |  | Plaintext (note: will be all numbers but need formatted as text to preserve any leading zeros e.g. 0002368) | Common/Modal Value |
|  | School Year |  | SY##-## (e.g. SY16-17) | N/A |
|  | Target Area |  | Plaintext | Common/Modal Value |
|  | FR Enrollment |  | Number, no decimal places, commas separating thousands | October Value |
|  | FR Enrollment Percentage | Percentage, include all decimal places | Percentage, no decimal places. | October Value |
|  | School Level-Standardized |  | Plaintext | Common/Modal Value |
|  | School Type-Standardized |  | Plaintext | Common/Modal Value |
|  | Breakfast Delivery Model from Grants Data-Standardized |  | Plaintext | Common/Modal Value |
|  | Breakfast Delivery Model from Innovations Data-Standardized |  | Plaintext | Common/Modal Value |
|  | Breakfast Delivery Model from Campaign Tracking Data-Standardized |  | Plaintext | Common/Modal Value |
|  | Breakfast Delivery Model from State Agency Tracking-Standardized |  | Plaintext | Common/Modal Value |
|  | Breakfast Delivery Model from Other Source-Standardized |  | Plaintext | Common/Modal Value |
|  | BAB Implementation Date from Grants Data |  | MM/DD/YY | Common/Modal Value |
|  | F Breakfast ADP | Decimal | Number, no decimal places, commas separating thousands | Average Across Months (see note 1 below) |
|  | R Breakfast ADP | Decimal | Number, no decimal places, commas separating thousands | Average Across Months  (see note 1 below) |
|  | P Breakfast ADP | Decimal | Number, no decimal places, commas separating thousands | Average Across Months  (see note 1 below) |
|  | F Lunch ADP | Decimal | Number, no decimal places, commas separating thousands | Average Across Months (see note 1 below) |
|  | R Lunch ADP | Decimal | Number, no decimal places, commas separating thousands | Average Across Months (see note 1 below) |
|  | P Lunch ADP | Decimal | Number, no decimal places, commas separating thousands | Average Across Months (see note 1 below) |

**Note 1:**

You must first calculate each of these ADPs (F Breakfast ADP, P Breakfast ADP, etc.) for each month for the school, then you average across months. ADP is calculated the same was as FR Breakfast ADP and FR Lunch ADP are, just using the appropriate meal types.

For example, to calculated “F Breakfast ADP” for Annual School Level data, follow the following steps:

1. Calculate “F Breakfast ADP” = iferror(“Breakfast Meals-Free” / “Operating Days-Breakfast Only”,””). If “Operating Days-Breakfast Only” is null, then use “Operating Days”
2. Average these values across months, as defined above.

Likewise, to calculated “P Lunch ADP” for Annual School Level data, follow the following steps:

1. Calculate “P Lunch ADP” = iferror(“Lunch Meals-Paid” / “Operating Days-Lunch Only”,””). If “Operating Days-Lunch Only” is null, then use “Operating Days”
2. Average these values across months, as defined above.

**Results Data & Visualizations**

There are typically five levels of Results Data we can pull from Annual School Level Data: school level results, district level, county level, Target Area level, and state level.

Important note regarding filtering for appropriate site types – in general, we only look at data from public schools and public charter schools and we typically remove other school types where possible (such as RCCIs, juvenile detention centers, daycares, etc.). However, not all state data includes site type information and some of the analysis to-date may inadvertently include the unneeded school types. Because of this variability from state to state, we added the “Schools to exclude” column during the “Consolidated & Enriched Data” stage. **All results data should filter out the schools marked as “exclude” in the “Schools to exclude” column.**

A. Target Area Level Numbers **(MPA can use this table for data validation)**

| **No.** | **Column** | **How calculated** | **Format** | **Grand Total Calculation** |
| --- | --- | --- | --- | --- |
|  | Target Area | From Annual School Level Data for selected school year |  | N/A |
|  | Total Number of Schools | Count of Unique ID from Annual School Level Data for selected school year | Number, no decimal places, commas separating thousands | Sum of column |
|  | Total Enrollment | Sum of Total Enrollment for all schools in Target Area from Annual School Level Data for selected school year. | Number, no decimal places, commas separating thousands | Sum of column |
|  | FR Enrollment | Sum of FR Enrollment for all schools in Target Area from Annual School Level Data for selected school year. | Number, no decimal places, commas separating thousands | Sum of column |
|  | FR % | Calc. #4 / #3 | Percent, no decimal places | Row calculation |
|  | FR Lunch ADP (current school year) | Sum of FR Lunch ADP for all schools in Target Area from Annual School Level Data for selected school year. | Number, no decimal places, commas separating thousands | Sum of column |
|  | FR Breakfast ADP (current school year) | Sum of FR Breakfast ADP for all schools in Target Area from Annual School Level Data for selected school year. | Number, no decimal places, commas separating thousands | Sum of column |
|  | FR Breakfast as a % of FR Lunch | Calc: #7 / #6 | Percent, no decimal places | Row calculation |
|  | Gap to Guidepost | Calc: MAX(#6\*.7 - #7, 0) | Number, no decimal places, commas separating thousands | Row calculation |

*Include a Grand Total Row at the bottom. In this Grand Total row, please follow the calculation method detailed in the final row above.*

B. School-level numbers – School-level are mostly used by the Field team for planning and targeting.

| **No.** | **Column** | **How calculated** | **Format** |
| --- | --- | --- | --- |
|  | Unique ID | From Annual School Level Data for selected school year | Plaintext (note: will be all numbers but need formatted as text to preserve any leading zeros e.g. 0002368)  See definition in “Clean Data” |
|  | Target Area | From Annual School Level Data for selected school year |  |
|  | School Name | From Annual School Level Data for selected school year | Plaintext |
|  | District Name | From Annual School Level Data for selected school year | Plaintext |
|  | State | From Annual School Level Data for selected school year | Plaintext, 2-letter state abbreviation |
|  | County | From Annual School Level Data for selected school year | Plaintext |
|  | School Level | From Annual School Level Data for selected school year | Plaintext |
|  | School Type | From Annual School Level Data for selected school year | Plaintext |
|  | Total Enrollment | From Annual School Level Data for selected school year | Number, no decimal places, commas separating thousands |
|  | FR Enrollment | From Annual School Level Data for selected school year | Number, no decimal places, commas separating thousands |
|  | FR % | From Annual School Level Data for selected school year | Percent, no decimal places |
|  | Eligibility Level | Write a formula to populate cells as follows:  If #11 is >= 60%, “High Eligibility”  If #11 is >=40% but <60%, “Medium Eligibility”  If #11 is <40%, “Low Eligibility”  If #11 is Null, Null | Plaintext |
|  | CEP (Y/N) | From Annual School Level Data for selected school year | Plaintext |
|  | Provision 2 (Y/N) | From Annual School Level Data for selected school year | Plaintext |
|  | FR Lunch ADP (selected school year) | From Annual School Level Data for selected school year | Number, no decimal places, commas separating thousands |
|  | FR Breakfast ADP (selected school year) | From Annual School Level Data for selected school year | Number, no decimal places, commas separating thousands |
|  | FR Breakfast as a % of FR Lunch | Calc: #16 / #15 | Percent, no decimal places |
|  | Gap to Guidepost | Calc: MAX(#15\*.7 - #16, 0) | Number, no decimal places, commas separating thousands |
|  | Target? (Y/N) | User-entered field (ideally) | Only three values allowed: blank cell, Y, and N |
|  | Grant Program Start Date | Joined from grants database (ideally), otherwise, a user-entered field. | Valid date, MM/DD/YYYY |
| 34. | ISP | From Annual School Level Data for selected school year | Percent, no decimal places |
|  | Notes | User-entered field | Plaintext |
| **Targeting and Projections Section** | | | |
|  | Breakfast Model for Projection | User-entered field from dropdown of 5 options: BIC, GNG Classroom, 2nd Chance (GNG), 2nd Chance (Cafeteria), GNG Common Area | Plaintext |
|  | Implementation Date for Projection | User-entered field | Valid date, MM/DD/YYYY |
|  | Implementation Complete? Y/N | User-entered field | Only three values allowed: blank cell, Y, and N |
|  | Projected kids added full SY | *Calculation based on the information included in #20, which determines the projection participation rate. Projected participation rate =*   * *88% if #22 = “BIC”* * *59% if #22 = “GNG Classroom” or “GNG Common Area”* * *58% if #22 = “2nd Chance (GNG)” or “2nd Chance (Cafeteria)*   Calculation = MAX(#15\**projected participation rate* - #16, 0) | Number, no decimal places, commas separating thousands |
|  | Projected kids added in current SY | *Calculation based on the information included in #23 and #25.*  Calculation:   * If #23 is in a previous school year or August or September of current school year: 1 \* #25 * If #23 is in October of current school year: (8/9) \* #25 * If #23 is in November of current school year: (7/9) \* #25 * If #23 is in December of current school year: (6/9) \* #25 * If #23 is in January of current school year: (5/9) \* #25 * If #23 is in February of current school year: (4/9) \* #25 * If #23 is in March of current school year: (3/9) \* #25 * If #23 is in April of current school year: (2/9) \* #25 * If #23 is in May of current school year: (1/9) \* #25 * If #23 is in the following school year: 0 * If #23 is blank: blank | Number, no decimal places, commas separating thousands |
| **Historical Data Section** | | | |
|  | FR Lunch ADP (previous school year) | FR Lunch ADP from Annual School Level Data for *previous school year.* | Number, no decimal places, commas separating thousands |
|  | FR Breakfast ADP (previous school year) | FR Lunch ADP from Annual School Level Data for *previous school year.* | Number, no decimal places, commas separating thousands |
|  | FR Breakfast as a % of FR Lunch | #28/#26 | Percent, no decimal places |
|  | Change in FR Lunch ADP | #27 - #15 | Number, no decimal places, commas separating thousands |
|  | Change in FR Breakfast ADP | #28 - #16 | Number, no decimal places, commas separating thousands |
|  | Change in FR Breakfast %age | #29 - #17 | Percent, no decimal places |

*Add bars to the column for “Gap to Guidepost”*

C. School District Level Numbers

| **No.** | **Column** | **How calculated** | **Format** |
| --- | --- | --- | --- |
|  | State & District ID | State + “-“ + District ID from Annual School Level Data for selected school year. |  |
|  | Target Area | Include only if all schools within each school district are in the same Target Area. |  |
|  | Total Number of Schools | Count of Unique ID from Annual School Level Data for selected school year |  |
|  | District Name | Common/Modal Value for District Name for the State and District ID combination in #1 (from Annual School Level Data for selected school year). |  |
|  | Total Enrollment | Sum of Total Enrollment for all schools in State & District ID from Annual School Level Data for selected school year. | Number, no decimal places, commas separating thousands |
|  | FR Enrollment | Sum of FR Enrollment for all schools in State & District ID from Annual School Level Data for selected school year. | Number, no decimal places, commas separating thousands |
|  | FR % | Calc. #6 / #5 | Percent, no decimal places |
|  | FR Lunch ADP (current school year) | Sum of FR Lunch ADP for all schools in State & District ID from Annual School Level Data for selected school year. | Number, no decimal places, commas separating thousands |
|  | FR Breakfast ADP (current school year) | Sum of FR Breakfast ADP for all schools in State & District ID from Annual School Level Data for selected school year. | Number, no decimal places, commas separating thousands |
|  | FR Breakfast as a % of FR Lunch | Calc: #9 / #8 | Percent, no decimal places |
|  | Gap to Guidepost | Calc: MAX(#8\*.7 - #9, 0) | Number, no decimal places, commas separating thousands |

*Add bars to the column for “Gap to Guidepost”*

D. County Level Numbers

| **No.** | **Column** | **How calculated** | **Format** |
| --- | --- | --- | --- |
|  | State & County | State + “-“ + County from Annual School Level Data for selected school year. |  |
|  | Target Area | Include only if all schools within each county are in the same Target Area. |  |
|  | Total Number of Schools | Count of Unique ID from Annual School Level Data for selected school year |  |
|  | Total Enrollment | Sum of Total Enrollment for all schools in State & County from Annual School Level Data for selected school year. | Number, no decimal places, commas separating thousands |
|  | FR Enrollment | Sum of FR Enrollment for all schools in State & County from Annual School Level Data for selected school year. | Number, no decimal places, commas separating thousands |
|  | FR % | Calc. #5 / #4 | Percent, no decimal places |
|  | FR Lunch ADP (current school year) | Sum of FR Lunch ADP for all schools in State & District ID from Annual School Level Data for selected school year. | Number, no decimal places, commas separating thousands |
|  | FR Breakfast ADP (current school year) | Sum of FR Breakfast ADP for all schools in State & District ID from Annual School Level Data for selected school year. | Number, no decimal places, commas separating thousands |
|  | FR Breakfast as a % of FR Lunch | Calc: #8 / #7 | Percent, no decimal places |
|  | Gap to Guidepost | Calc: MAX(#7\*.7 - #8, 0) | Number, no decimal places, commas separating thousands |

*Add bars to the column for “Gap to Guidepost”*

E. Target Area Year Over Year Numbers

| **No.** | **Column** | **How calculated** | **Format** | **Grand Total Calculation** |
| --- | --- | --- | --- | --- |
|  | Target Area | From Annual School Level Data for selected school year |  | N/A |
| 2. | FR Lunch ADP (previous school year) | Sum of FR Lunch ADP from Annual School Level Data for *previous school year.* | Number, no decimal places, commas separating thousands  Ideally there would be a header above columns #2 - #4 indicating the school year (e.g. SY16-17) | Sum of column |
| 3. | FR Breakfast ADP (previous school year) | Sum of FR Breakfast ADP from Annual School Level Data for *previous school year.* | Number, no decimal places, commas separating thousands  Ideally there would be a header above columns #2 - #4 indicating the school year (e.g. SY16-17) | Sum of column |
| 4. | FR Breakfast as a % of FR Lunch (previous school year) | #3 / #2 | Percent, no decimal places  Ideally there would be a header above columns #2 - #4 indicating the school year (e.g. SY16-17) | Row calculation |
| 5. | FR Lunch ADP (selected school year) | Sum of FR Lunch ADP for all schools in Target Area from Annual School Level Data for selected school year. | Number, no decimal places, commas separating thousands  Ideally there would be a header above columns #5 - #7 indicating the school year (e.g. SY17-18) | Sum of column |
| 6. | FR Breakfast ADP (selected school year) | Sum of FR Breakfast ADP for all schools in Target Area from Annual School Level Data for selected school year. | Number, no decimal places, commas separating thousands  Ideally there would be a header above columns #5 - #7 indicating the school year (e.g. SY17-18) | Sum of column |
| 7. | FR Breakfast as a % of FR Lunch (selected school year) | Calc: #6 / #5 | Percent, no decimal places  Ideally there would be a header above columns #5 - #7 indicating the school year (e.g. SY17-18) | Row calculation |
| 8. | Change in FR Lunch ADP | #5 - #2 | Number, no decimal places, commas separating thousands | Row calculation |
| 9. | Change in FR Breakfast ADP | #6 - #3 | Number, no decimal places, commas separating thousands | Row calculation |
| 10. | Change in FR Breakfast %age | #7 - #4 | Percent, no decimal places | Row calculation |
| 11. | Gap to Guidepost | Calc: MAX(#5\*.7 - #6, 0) | Number, no decimal places, commas separating thousands | Row calculation |

*Include a Grand Total Row at the bottom. In this Grand Total row, please follow the calculation method detailed in the final row above.*

F. Other Examples of Results Data and Visualizations

A-E above are the basic forms of results data that MPA uses on a regular basis for breakfast and lunch data. We have created and used a variety of other tables and visualizations that could be useful to have automated as part of this system. MPA will populate a shared folder with examples of these additional tables and visualizations to serve as a guide for IT regarding what additional visualizations and tables to create.

G. *Important notes about Results Data*

* State-level USDA results differ from state-level state agency results due to timing and other factors. We typically never use state agency results in external reports, funder reports, the dashboard, etc. unless otherwise determined appropriate by Katheryn and Eliza; we always use USDA state-level results. USDA is our “single source of truth” and all state agency data is considered preliminary.
* State-level results from the state agency data are most often used by the Field team for more “micro-level” campaign planning and management.
* National results are derived using the USDA data, which has separate cleaning/analysis process than described herein.

**Analysis Pitfalls and Other Important Nuances**

Since the state agency data is typically only used by the Field team for campaign performance management, we do not perform a great amount of QA on datasets received throughout the year unless a large issue is identified. The following “quick checks” and “how to handle” rules are usually enough to give us a good level of confidence.

* Raw Data
  + School lunch enrollment vs total school enrollments – as mentioned in the notes for #21 in Clean Data section, it is important to be sure you are looking at total enrollment information and not lunch program enrollment information. Lunch program enrollment only captures a subset of kids within a school but we need to know the total kids in a school. If working with unfamiliar data from a new state, it is best to spot check and corroborate statistics with NCES ELSI enrollment and eligibility data.
  + Look at F/R enrollment rates among CEP schools – when a school participates in CEP, all of the kids in that school can eat breakfast for free regardless of whether they actually qualify for free or reduced-price meals. This leads to many inconsistencies in what CEP schools are reporting in terms of free enrollments and reduced enrollments. In the ideal scenario, F/R enrollments will reflect the number of kids *actually* eligible for F/R meals. However, many schools will override this number or not report it at all. When this happens, the F/R enrollment data is not accurate and this causes issues in our Results Data. Two common examples of how to spot issues with CEP data are:
    - *Schools report 100% F/R enrollment* – when the free enrollments in a CEP school equal the total enrollments, this may indicate the F/R numbers are not accurate. However, in high-poverty areas it is entirely possible for 100% of the children in that school to legitimately qualify for F/R meals regardless of CEP status.
    - *Schools report 0% F/R enrollment* – when total enrollments are reported for a school but there are zero free enrollments and zero reduced-price enrollments, this may indicate the F/R numbers are not accurate. However, in low-need areas it is entirely possible for none of the children in that school to qualify for F/R meals.

There are other quality issues we have found with state agency data that we do not typically check for on a regular basis. Though not a priority at the moment, it may be wise to tackle the following issues at some point in the future:

* Clean Data
  + Outliers –
    - In our in-depth analysis project, we have noticed a very small number of clearly erroneous values in the data, even though the data comes from trusted state agency sources. For example, some schools have reported breakfast meals in the lunch category and vice-versa, and in other schools enrollment and meals numbers spike in just one month. Often times, this appears to be caused by an accidental additional keystroke (an additional digit added to a number)—this is more prevalent in enrollment numbers than meal numbers since reimbursements are tied to meals and not enrollment.
    - We do not perform outlier checks during our regular data analysis and thus do not have a recommendation for how to deal with outliers on a regular basis.
    - However, as a low-priority item, it would be nice to include a check in our data system to identify potential outliers. This would allow us understand how common outliers are and correct them for specific, detailed analysis projects, if necessary. We have a developed a draft methodology for identifying potential outliers that we can share.
* Clean Data, Consolidated Data, Consolidated & Enriched Data
  + Check for missing values – review the amount of missing values in each column, specifically looking for:
    - Missing values for District ID, District Name, County or Target Area are especially problematic since they will affect the Results Data if the data field of the level of aggregation is missing. Missing fields should be inputted if it is possible to determine their true value (for example, county can often be inputted using zip code).
    - If the clean dataset is compiled from multiple data sources, checking for missing values is a good way to ensure that the data merging/joining process worked for most of the rows.
  + Check for duplicate entries – two rows with the same unique ID and claim date should be reviewed to determine if one row is erroneous or if it is an adjustment to the original claim. We do not do this systematically now but have noticed some issues with duplicate values in state agency data. Adding in such a check would be helpful.

1. This is the same as “Display Format” unless otherwise noted. [↑](#footnote-ref-2)
2. This is the same as “Display Format” unless otherwise noted. [↑](#footnote-ref-3)