**To: Ta-Tanisha Wallace**

**From: Dhruv Singh**

**Date: 10/8/2021**

**Subject: Data Discrepancies**

**Context:**

Due to the various sources of data used to track DOES’s programmatic performance, having a unified quality control system has been difficult. As part of the ongoing effort to get DataWorks built and rolled out, I would like to point out some ‘data discrepancies’ I came across in DC Networks.

Specifically, the data I have used to identify these errors is the Maryland Custom Report. The Maryland Custom Report pulls from our most up to date data and provides one of the most comprehensive sources of truth for the agency’s programmatic performance. The timeframe for which I looked at data is the past ten years (2011-2021).

**Objective:** Ideally, to remedy historical education data.

1. To have people look into historical education records used to produce the CASAS data in DC Networks, and to overwrite it.
2. To standardize Wagner Peyser education levels entered into DC Networks

**Case:** A lot of DOES’s program enrollment decision are based on customers’ CASAS scores. Therefore, it is imperative to audit the CASAS data for historical consistency as part of a larger drive towards data integration.

**Summary:**

Three types of errors identified:

1. Dates: End Date preceding Begin Date – 3 count, trivial
2. Education Metrics (CASAS): The same column contains data entered on 3 separate scales
   1. Raw scores (eg: 300, 400),
   2. Numbered Grade Categories (eg: 3, 4), and
   3. Text Grade Values (eg. 3rd, 4th)
3. Wagner Peyser Education Level: Unordered categories, categories not listed in data dictionary
   1. these categories could be ordered in a more meaningful way, eg: ‘16’ -> bachelors, and ‘91’ -> associates
   2. There are categories in the data that do not appear in the data dictionary.

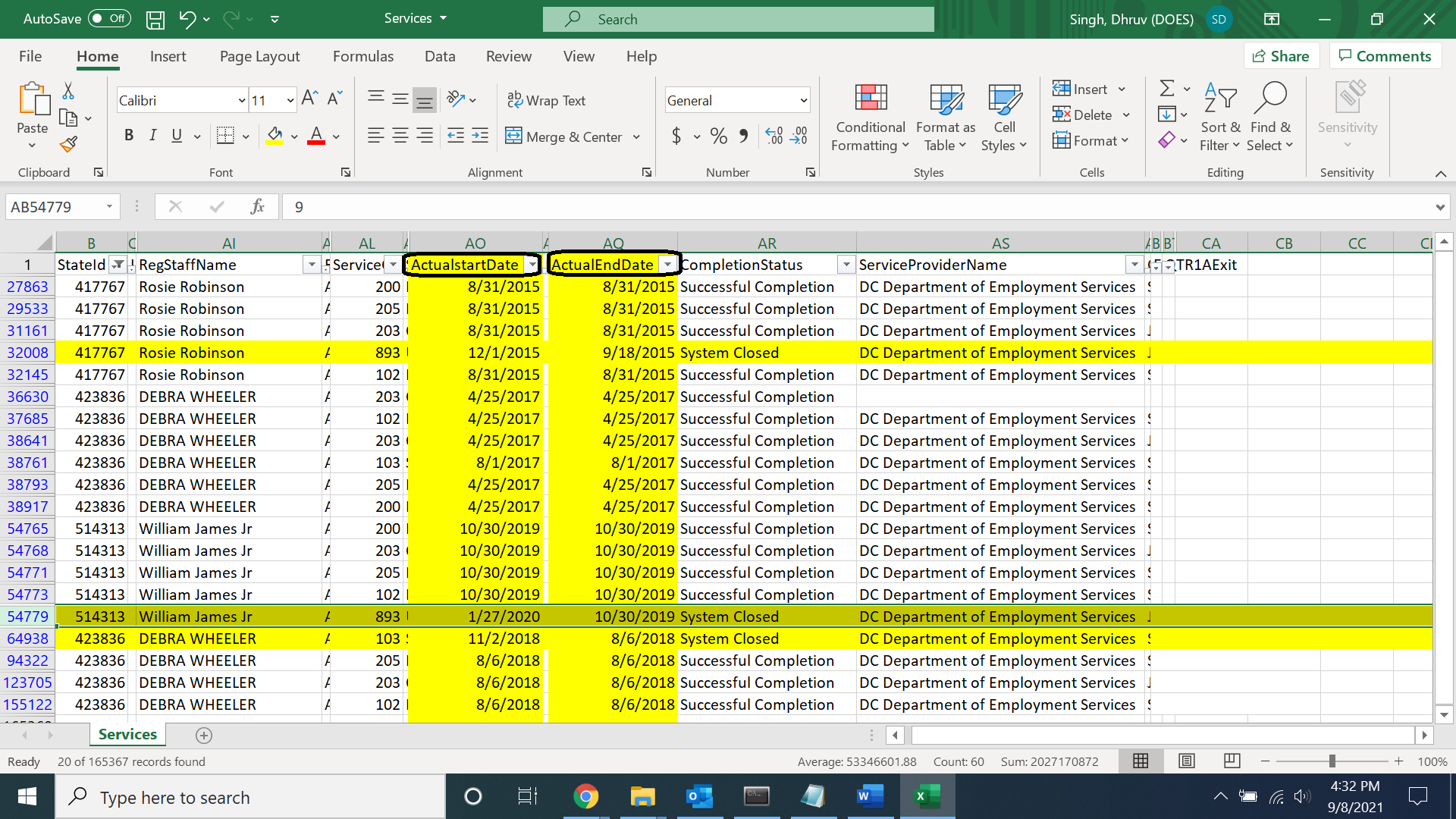
**Appendix A:**

**Problem I:** End Date < Begin Date (3 counts)

**Description:** The first, and most trivial issue is that there are 3 occurrences of end date preceding begin date. The 3 dates and StateId’s for which this error occurs are highlighted in yellow below.

**Why it is a problem:** This presents a logical inconsistency and presents an error when measuring a customer’s time to completion.

**Suggested fix:** A lot of programs start and end on the same day (according to the data). Accordingly, pick either the start date or end date for the 3 rows, and set that to both begin and end date. Possibly look at the other dates for that customer if there are any and base your decision on that.



**Appendix B:**

**Problem II:**

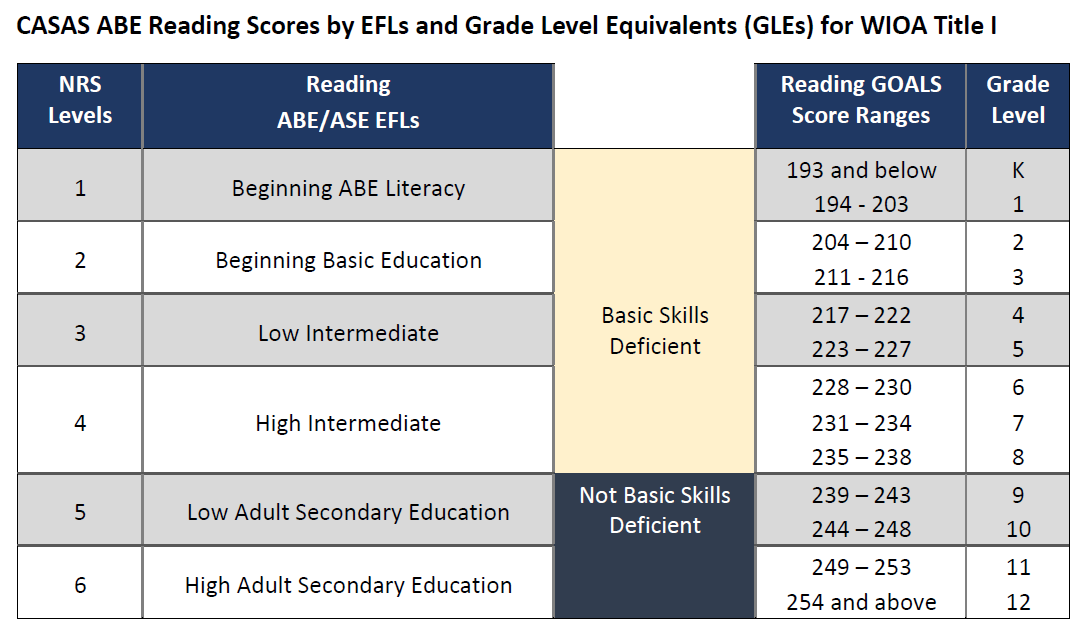
**Description:** CASAS Reporting System: Scales

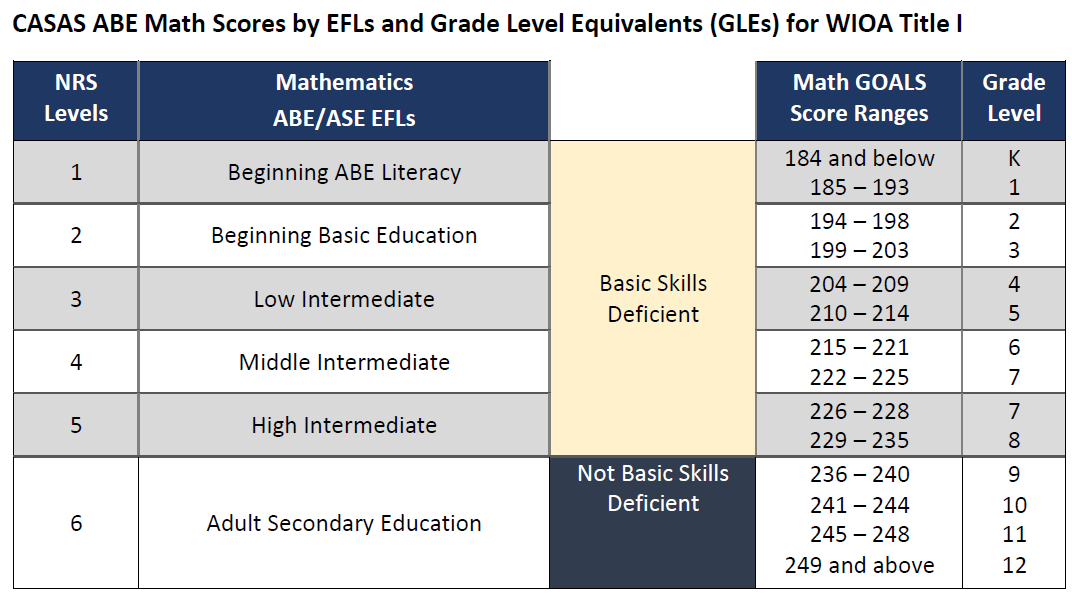
The problem here is that the reading and writing scores are reported on multiple scales.

1. The first scale: Actual score range as per test
2. The coded grade level: this scale has the highest frequency in the data. And while the conversion scale says to convert scores to integers, there are far more decimals at around the grade levels.
3. Lastly, text-based grade levels

**Why it is a problem:** Far too inconsistent to be of any use when conducting any analysis linking customer scores to their success with obtaining employment.

1. Screenshot 1: Below are the conversion scales for CASS





1. Screenshot 2: Numeric Scores

**Description:**

While a majority of scores are coded on the K-12 reading and writing range, a few remain coded as the raw scores

Moreover, for the scores that coded on the K-12 range, the data dictionary lists integer values as the categories. But the data mostly features decimal values.

Lastly, the raw scores are to range from 180-250 for reading and writing scores. But the data contains values such as 1010, 1300, 7000, 9999 that are wildly off the CASAS scoring range.

There is no indication as to what these values mean.

**Why it is a problem:** This presents a significant challenge to conducting any meaningful analysis using CASAS scores

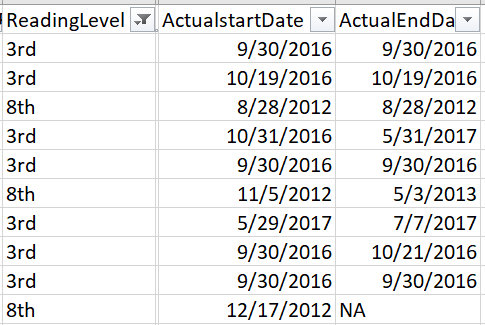
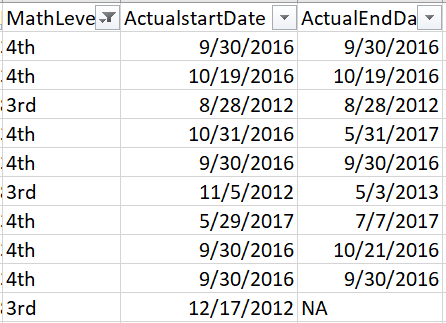
**Proposed solution:**

* Bring all scores on the K-12 range.
* If decimals are acceptable, say so in the documentation.
* Account for the values that are far above the acceptable range, as well as for 0 values., in the documentation.

1. Screenshot 3:

**Proposed solution:** This is far easier to remedy and is an issue for only the rows listed below, far fewer in number. All we have to do is to convert these rows to the corresponding numeric values for grade level in reading and writing.

**Note:** Additionally, workforce specialists should be given instruction on entering CASAS data only on the grade level scale moving forward.

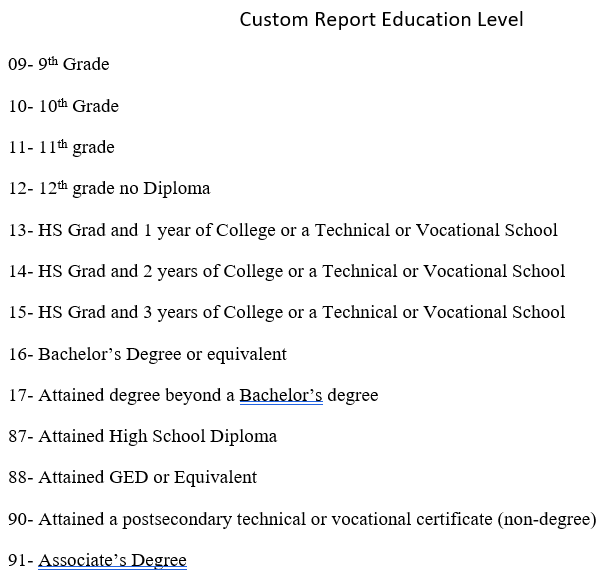
**Problem III:** Education level categories.

**Description:**

The education level categories (Wagner Peyser) that are included in the Maryland custom report, feature data outside the categories listed in the data dictionary [specifically - 89].

Moreover, the data dictionary lists the data beginning from the 9th grade. But the actual data contains values from 1 to 8 as well, and significant counts of the same. Of course, it can be inferred that these values correspond directly to years of schooling (1=> 1st grade, 8=> 8th grade), but this should be made clear in the documentation. And if missing, should be updated to reflect the same.

1. Screenshot 1: Counts of education levels entered into DC Networks by customers
2. Screenshot 2: Current coding scale



**Description:**

A variable like education, where more of something can typically be tied to other outcomes, needs to be ordered in order to be useful. However, the current progression of categories is

(a.) unordered according to progression

(b.) spaced unevenly

**Proposed solution:**

For the education levels variable to be used in producing any meaningful analytics, it must be recoded to be ordered, and spaced evenly. See below for a proposed recode solution.

1. Screenshot 3: Suggested coding scale

