## 

[Amira Interaction Data - High Level](#_13ahazjcugzn)

[March 23 Dataset](#_4o4g5dpalwed)

## 

## Amira Interaction Data - High Level

For each session, where a session could either be a student engaging in foundational skill building (Early Reader Skills Scaffold, or ERSS) or reading a connected text passage, we have:

* Session date/time
* Session length
* Passage title (this is relevant because students get to choose from 5 texts tailored to their reading level based on the title)
* Passage text
* How many, and which kind of, micro-interventions Amira delivered - and exactly where in the story each micro-intervention was delivered
* Whether the student completed the session or abandoned it (could be due to disinterest or an external factor such as class period ending)

For any session read to completion by the student, we also have:

* Whether the student read each word correctly or not (scored by AI models)
* For correctly read words, timing data on when the student successfully read each word

We also have the audio of every session but sharing will be limited to the special districts as part of the project. Since Amira is primarily a reading (not clicking) product, most of the within-session signs of engagement or lack thereof will be found in the audio. We have all of the data longitudinally for each student.

We have dozens of different types of micro-interventions spanning phonological awareness, decoding, vocabulary, comprehension, and background knowledge. One question we have been highly interested in is the impact of the quantity and types of micro-interventions delivered—for different student ability bands and in different contexts—on learning outcomes. We could look at similar variables with respect to engagement if that’s a direction of interest.

## 

## March 23 Dataset: [Amira Interaction Data](https://docs.google.com/spreadsheets/d/16XuiW5VVKKsRmNBYqDwb9x6RaB4xKxl3ryGCJYpF9Bc/edit?gid=1871864116#gid=1871864116)

Data dictionary

**micro\_intervention\_data** tab: this contains information on what micro-interventions were delivered and at which moments in the session

**activity\_id:** UUID to denote a unique session. Students generally either read a text they pick out of five recommended texts for their reading ability level (most sessions are like this) or engage in foundational reading skill building activities.

**student\_id:** UUID to denote a unique student.  
  
**phrase\_index:** Each session/text is split up into “phrases” and phrases are a key data unit because our ASRs process one phrase at a time. Typically a phrase is 1-2 sentences long. The numbering is always zero-based.

**intervention\_scope:** Some micro-interventions happen right after a given word in a phrase; these will be labeled **“word”**. Others happen at the end of a phrase and will target a specific word from the phrase that the student had a miscue on; these will be labeled **“phrase”**.  
  
**word\_index:** If the **intervention\_scope** is **“word”** then this will provide the index of the word within the phrase (the first word in the phrase is 0, the next is 1, etc.)  
  
**intervention\_type:** The code name for the type of micro-intervention delivered at that moment. Information on each micro-intervention can be found here (indexed by code name): [Intervention Wiki.pdf](https://drive.google.com/file/d/11KSLqoEtL5OpOSX4IVzFuwz37adbVhn-/view?usp=drive_link)  
  
**intervention\_word:** The specific word within the phrase the micro-intervention was given for.

**reading\_and\_timing\_feats** tab: this contains detailed story text broken up such that each word in the text is a row, with student speech data & timings aligned to each story text word

**activity\_id:** UUID to denote a unique session. Students generally either read a text they pick out of five recommended texts for their reading ability level (most sessions are like this) or engage in foundational reading skill building activities. NOTE - these sessions should perfectly overlap with those covered in the **micro\_intervention\_data** tab if you join by **activity\_id** and **phrase\_index**.

**student\_id:** UUID to denote a unique student.  
  
**phrase\_index:** Each session/text is split up into “phrases” and phrases are a key data unit because our ASRs process one phrase at a time. Typically a phrase is 1-2 sentences long. The numbering is always zero-based. NOTE - these sessions should perfectly overlap with those covered in the **micro\_intervention\_data** tab if you join by **activity\_id** and **phrase\_index**.

**word\_index:** The index of the word within the phrase (the first word in the phrase is 0, the next is 1, etc.)

**Story Word**: The actual word in the phrase at the specific word index.

**AmiraW2V\_Rec\_Word :** The best AmiraW2V transcript segment of IPA that aligns to this specific Story Word based on Amira’s alignment algorithm.

**AmiraKaldi\_Rec\_Word:** The AmiraKaldi transcript word that best aligns to this specific Story Word based on Amira’s alignment algorithm.

**AmiraKaldi\_Start\_Time:** The start time (measured from the beginning of the phrase) of the transcribed aligned word. This is your main source of timing data for utterances which carries a lot of richness.

**AmiraKaldi\_End\_Time:** The end time (measured from the beginning of the phrase) of the transcribed aligned word. This is your main source of timing data for utterances which carries a lot of richness.

**annotator label:** Whether the student read the story word correctly or not, based on human annotation.

**prealignment\_asr\_file:** The file in <https://drive.google.com/file/d/1oXJCxJXGnOZJiSjOQfHRRwnKX85Myac_/view?usp=drive_link> that contains three different ASRs’ transcriptions (and timings for two of the ASRs) for the speech present in the entire phrase with no filtering or alignment to story text. This is going to be your main source of off-task speech signals.

**longitudinal stud data** tab: all recorded sessions for each student with sampled sessions in the other tabs. Provides longer-term, cross-session macro engagement data. It is possible to pull the micro-intervention data and the detailed session feature for every single session of a given student if you need this. The micro-intervention data is easy to pull. The detailed session/ASR feature data is more involved because we have to reconstruct some of the features, but it’s also doable.

If a session is “completed” displaystatus, they went through the entire session. **Timeread** is the total session time in milliseconds. If it is “started” that means they abandoned the session partway through. If this happens, the timeread is not recorded, so it will always be 0 (even if they read a lot of the text and then exited out before finishing).