# Creation of bespoke metadata (`__metadata-xvars-60.xlsx`)

We started with the publicly released metadata file that can be accessed in DEAP.

> Went to "My datasets" tab

> Clicked on "Pre-assembled datasets"

> Downloaded file called "data_dictionary_levels.xlsx" underneath the heading "Metadata files (complete dataset):"

We then added the following columns:

| Column | Meaning | Possible values | Values should be present for which rows? | How did we create it? |
|---|---|---|---|---|
| has_branching_logic | Was this variable behind any skip logic? | yes/no | All rows should have a value | Programatically, off the metadata already released in DEAP:<br>ifelse(is.na(branching_logic), "no", "yes") |
| could_be_a_survey_item | Could this variable be a question on a survey? | 0/1 | All rows should have a value | We took a very inclusive approach here by including any and all variables that could be asked on a survey – thus, only excluding neuroimaging, linked external data, biophysiological measures, actigraphy, neurocognition tasks, etc. |
| flag_is_substance_use | Is this variable a measure of substance use (amount, frequency, etc.)?<br>Do not include things that can't be items (e.g., hair tox) in this column | 0/1/NA | Only rows where could_be_a_survey_item == 1 should have a value | We flag any data that indicates youth substance use because the initial proof-of-concept for TRS-ABCD used early, high-risk substance use as its outcome. As such, we flag any potential item encompassing youth substance use outcomes to leave out of our predictors.<br>Other people might want to include these on their predictive screeners, so we are going to process them like other items later. |
| excluded_because_from_substudy | Was this variable not administered in the core study (e.g., COVID substudy) and, as such,was there no possibility for it to be administered to most participants | 0/1/NA | Only rows where could_be_a_survey_item == 1 should have a value | We exclude data from substudies (e.g., COVID, ABCD-SD, IRMA) because that data comes from a subsample and is administered at different timepoints from the rest of the data. But the end-user could theoretically use these data if they put in more work (e.g., build a whole screener using just questions from the ABCD-SD study). |
| excluded_because_from_screener | Is this variable administered during the ABCD Screener? | 0/1/NA | Only rows where could_be_a_survey_item == 1 should have a value | We exclude data from the ABCD Study screener because that data was obtained solely for study eligibility and enrollment logistics |

| Column | Meaning | Possible values | Values should be present for which rows? | How did we create it? |
|---|---|---|---|---|
| excluded_because_of_skip_logic | If an item dependent on skip logic cannot be combined with the gating question in any straightforward way to produce a single question, with a single set of response options, that would be sensible/practical to include on a screener, then exclude it. The resulting recoded variable must be expected to have a monotonic (e.g., linear) relationship with outcomes. If the recode felt chaotic/awkward, we erred towards exclusion. | 0/1/NA | Only rows where could_be_a_survey_item == 1 should have a value | We exclude heavily gated variables that cannot have the majority of their non-responses recoded as the median response or as an obvious result of previous gating-questions because these heavily gated variables would heavily bias the model. Additionally, we want items that are not just predictive of high-risk outcomes, but are generalizable / applicable to the general public, such that they can be presented on a predictive screener on its own / without heavy-handed explanation |
| excluded_bc_other_reason | There is some other reason aside from the other excluded_columns as to why this potential item was excluded from *our* screener build | 0/1/NA | Only rows where could_be_a_survey_item == 1 should have a value | There are several other reasons for which an item is not suitable for inclusion in TRS-ABCD. These reasons are further explained in excluded_bc_other_reason_note |
| excluded_bc_other_reason_note | Transcribed reason for excluded_bc_other_reason | Text string | Only rows where excluded_for_other_reason_0/1 == 1 should have a value | A note explaining in further detail why a variable was in excluded_bc_other_reason |
| requires_recoding | The variable requires some form of recoding of current response options in order to run analyses | 0/1 | Only rows where could_be_a_survey_item == 1 and all excluded_bc cols == 0 | We used this column to indicate broadly if a variable needed any kind of recoding at all. This allowed us to filter on variables needing recoding in the metadata. |
| coalesce_recode | The relevant variable values are split up across multiple variables (e.g., variables ending in "_l"). These variables essentially have the same content, but are just split into different variables due to slight changes in wording. | 0/1 | Only rows where requires_recoding == 1 | We used this column to keep track of any variables that needed to be combined in R through the function "coalesce". This was reserved for variable values that had data split up between multiple variables (e.g., ending in "_l" or "_v01") |
| other_recode | The relevant variable requires some form of complex recoding (usually involving case_when statements) | 0/1 | Only rows where requires_recoding == 1 | We used this column to keep track of variables that required more in-depth recodes – typically for variables affected by skip logic that we could impute values for. (e.g., we can impute 0s for a question about the child's biological father having problems with alcohol if, in the gating question asking about any family member having problems with alcohol, the parent indicated that no family member had these problems) |
| delete_after_recoding | Should the original variable be deleted after recoding?<br>Sometimes we recode in place, in which case it shouldn't be deleted, other times we | 0/1 | Only rows where requires_recoding == 1 | Typically, variables were determined as requiring deletion after recoding if they were recoded by coalescing into another variable OR if they were split up into several categorical variables (with the variable itself not being monotonic). This column serves as a quick |