Combining Data Across Studies: Merge Early or Harmonize First?

A conceptual guide for deciding when to pool datasets in multi-study trauma research

The challenge

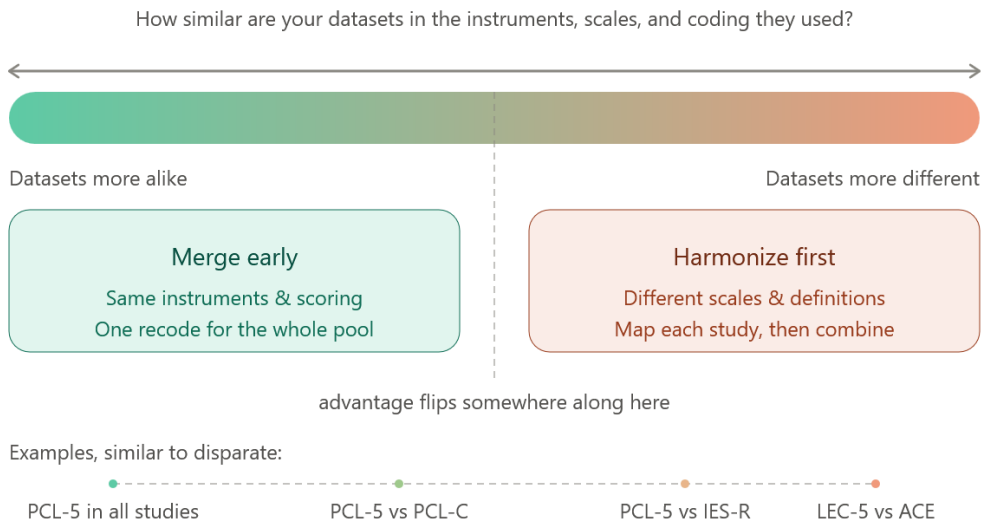
When working to combine datasets, you have to harmonize the target variables you are interested in to ensure they can be meaningfully synthesized and/or compared. For example, suppose you want to combine several data sets that all measured PTSD and trauma exposure. One study assess PTSD symptoms with the PCL-5, another with the IES-R, and a third with a clinician interview. Trauma exposure might be a count of event types in one dataset and a simple presence or absence flag in another.

Harmonization involves taking actions like renaming, recoding, rescaling, and aligning your target variables so that all share a common operational definition. When doing this work, trauma researchers often encounter the following choice:

Do you perform these operations separately within each dataset and then combine, or do you first stack everything into one data frame and do the work there?

Start by placing your data on the continuum

There is no single right answer, but a major factor influencing your decision is how similar your datasets already are in their measures and coding. It can be helpful to think of this as a continuum. At one end are studies using the same instruments with the same response options and scoring, and the datasets differ mostly in surface formatting. At the other end are studies using different instruments, scales, and definitions to capture the "same" construct. The closer your datasets sit to the similar end, the more merging early might be the better choice. The closer they sit to the disparate end, the more it might pay off to harmonize each dataset first (see Figure below).



When merging early helps

If your datasets are largely alike, combining first lets you write each transformation once rather than repeating near-identical code across studies. You compute a PCL-5 total, apply a provisional PTSD cutoff, or reverse-score an item a single time, for the whole pooled sample. This reduces redundancy in your work and, importantly, guarantees consistency: when the recode lives in one place, you have less to manage. Separate scripts cannot quietly drift apart. And if you need to make a change, a single fix does not need to be copied into five files. The cost is up-front formatting. You still have to align variable names and tag each row in each dataset with a variable that identifies the dataset so that you can trace and filter by source later when all the datasets are combined.

When harmonizing first helps

As datasets diverge, the recodes you would write in a pooled frame have to become conditional on the source dataset anyway. At that point, keeping each dataset's cleaning self-contained is often more efficient. Harmonizing study by study also forces you to state what each variable actually measures before anything is combined as you think through what each variable shares in common with others across datasets. Per-study processing keeps errors localized and easy to trace to their source, scales as your project grows (a new study means one new mapping, not a rewrite of every recode), and keeps "not measured in this study" distinct from "missing." It also leaves you with self-contained, reusable processing scripts, which supports the reuse that FAIR practices are meant to enable.

In thinking about whether to harmonize first, it is important to also weigh what you and others will want to do with these data later. If the vision is to maintain harmonized versions of each study so they can be recombined in flexible ways for questions you cannot yet foresee, then keeping the datasets as self-contained, well-documented units serves that goal. A researcher who later wants only the studies that used a CAPS interview, or only community samples, can draw on the pieces they need without having to untangle them from one large file. If instead the integrated multi-study dataset is itself the end product, and you do not expect to pull the studies apart again, that goal points toward building and maintaining the single combined dataset. Clarifying which of these is your true aim helps you decide where to invest your effort.

A human-factors lens

It can be helpful to think through approach to think about which workflow leaves the clearest documentation (ideally , a digital/reproducible trail). A reviewer, a collaborator, or you in two years should be able to follow how each original variable became a harmonized one. Sometimes the tidier-looking pooled script is harder to audit than a set of transparent per-study scripts, and sometimes it is the reverse. Whichever path you choose, you will still

need a study or site indicator in the final data, and you will still validate the recoding within each study.

Questions to guide your decision

- How similar are your datasets in the instruments, scales, and coding they used? The more similar, the stronger the case for merging early.
- For each construct you plan to pool, are you confident the variables measure the same thing, or only share a name? Where you are unsure, harmonize first.
- How many of your transformations would a merged script have to make conditional on the source dataset? Many conditionals is a signal to harmonize first.
- Do you expect the harmonized study datasets to be re-used later in configurations you cannot yet anticipate, or is a single integrated dataset the end product?
- How many datasets do you expect to add later? If the project will grow, per-study harmonization often scales more easily.
- Which workflow will be easiest for someone else, or future you, to follow and audit?
- Have you planned a study or site indicator and a per-study validation step, regardless of which path you take?

Part of the Toolkit for FAIR Traumatic Stress Data. See also the "Checklist for Reducing Re-Identification Risk in Traumatic Stress Research" and the "Information on Choosing a Data Repository – A Comparison of Repositories."