Data Quality Check - FAQ

**What this FAQ is about:**

This FAQ is related to the data quality check task for Gates Meta-analysis project we started in early November 2024. The team working on this consists of undergraduate RAs (Kevin, Arthi, Samir, Ben) and Research Analysts (Romil, Leo).

*Part 1* of this FAQ first provides a quick review of the motivation of the task, such as why this data quality check is necessary and how the results will be used subsequently. Then, *Part 2* will take you through each component of the check sheet. It will also document some key decisions and rules of thumb we applied throughout the checking process, as well as their respective justifications.

# **1. Why are we doing this?[[1]](#footnote-0)**

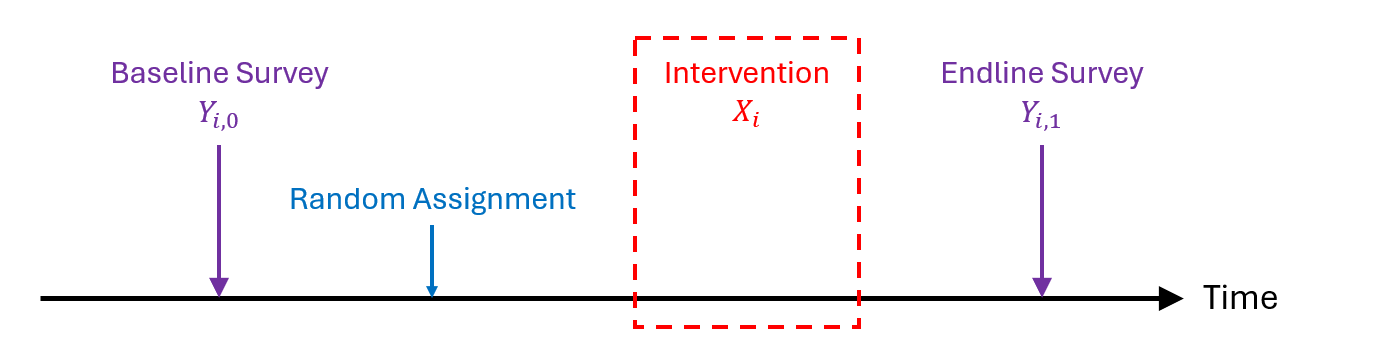
## **1.1. The Project**

In order to understand this task we’ve gotten ourselves into, we will first revisit the entire meta-analysis project, but in its simplest possible form. Recall what meta-analysis is:

* ***Meta-analysis*** is a method of synthesis of quantitative data from multiple independent studies addressing a common research question. (Thanks, Wikipedia!)

In other words, we *systematically[[2]](#footnote-1)* select and harmonize a list of relevant studies, and then use appropriate statistical methods to estimate the treatment effect of interest, i.e., the common research question, based on these studies as a whole.

Let us now make the wildest assumption: there only exist 3 papers studying the intervention of our interest and all of them used the same, simplest possible research design:



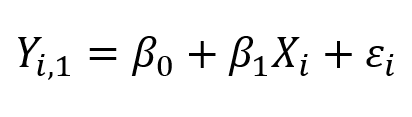
This graph should be fairly self-explanatory. Authors of these 3 papers are curious about the same thing:

* ***The common research question***: How much change in outcome is ***caused*** by the intervention for **any** individual in our sample?

Some side notes to help you understand this research question:

* ***“Caused”***: economists (or social scientists, in general) can’t stop thinking about the “*correlation vs. causality*” distinction even when they are taking a shower. In the current context, the power of drawing causal interpretation comes from the *random assignment* of treatment, which is a luxury you normally do NOT have in your later research career.
* “***Any***”: It is always important to acknowledge the subject upon which we are drawing conclusions. In this case, we are trying to say something about our *entire* sample, which we label as average treatment effect (ATE). Depending on the desired implication and feasibility of doing experiments, researchers sometimes care more about average treatment effect on the treated (ATT) or average treatment effect on the untreated (ATU).

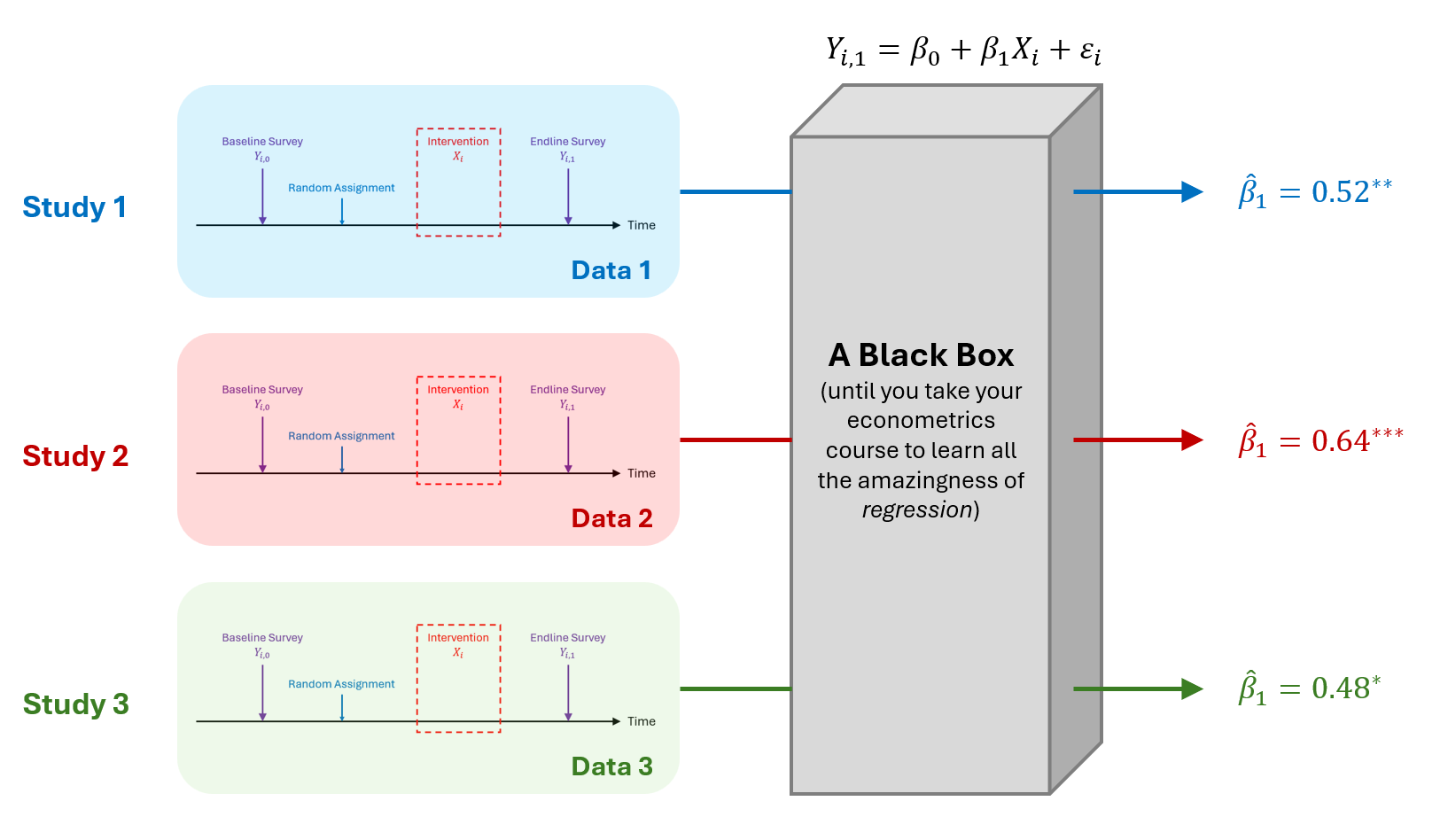
Now to answer this common question, the authors must obtain an unbiased estimate of the average treatment effect of the intervention. Econometricians told us that running the following regression should help[[3]](#footnote-2):



Note that is a indicator variable that equals 1 if individual received the intervention (“treated”) and 0 if otherwise (control). The left-hand-side of the equation is the endline outcome of interest (Spoiler alert: one of the s we care about the most is **consumption!**)[[4]](#footnote-3).

And the estimated coefficient provides those researchers the answers they need. In other words, all other things being equal[[5]](#footnote-4), receiving the intervention will, on average, increase the outcome by units for any individual in our sample. Question answered! Hoo-ray!

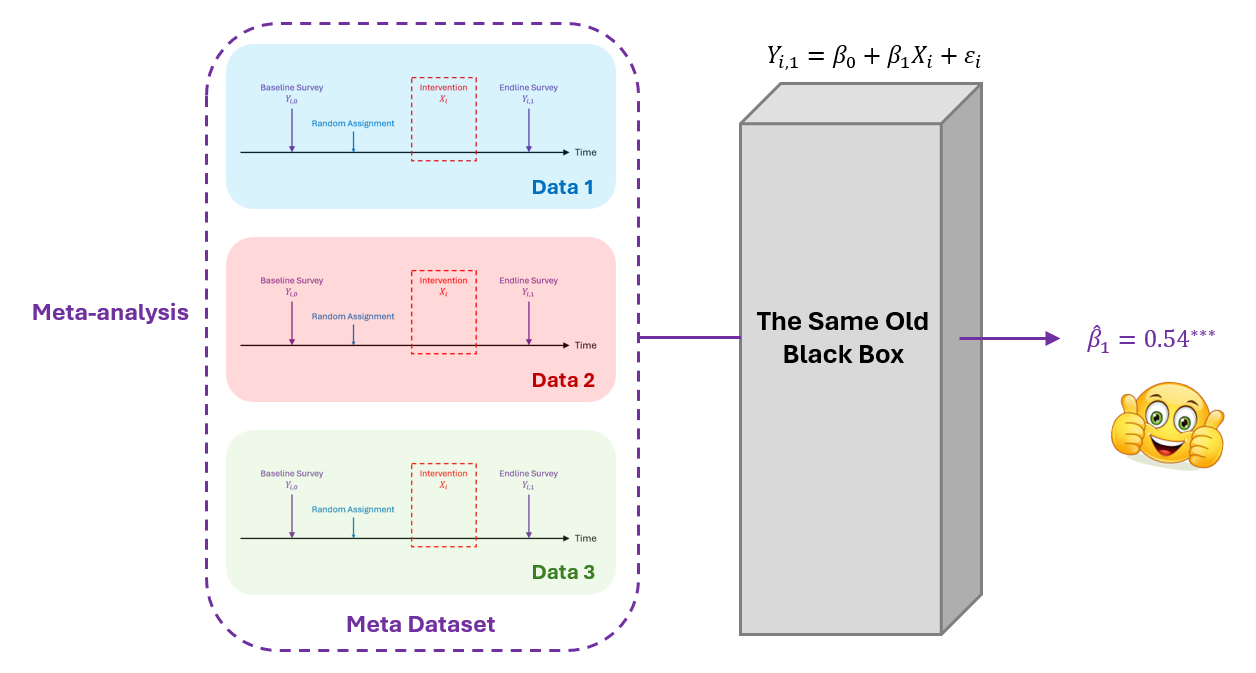
Although the authors of all 3 papers used the same research design as such, their studies were most likely implemented at different places/time and on different sets of people. It should come as no surprise that their estimation would be different:



This is where our meta-analysis comes in. Based on what we see in the existing (3) studies, we want to say something conclusive about the common research question:

* **Effect dispersion**: If we find the estimated ATE to differ considerably across different studies, e.g., with different signs, we will be interested in the direction of the effect when we aggregate all studies systematically.
* **Effect size**: If we find little dispersion among estimates from individual papers, we will be interested in the size of the effect instead. This will also be useful for conducting a cost-benefit analysis of the intervention.
* **Heterogeneity analysis[[6]](#footnote-5)**: estimates of heterogeneous treatment effects might not have sufficient statistical power within individual papers (because you are essentially cutting the sample in half, sometimes multiple times). When aggregating more data into the sample, you will have more power to detect smaller effects.

Exactly how the synthesis is done is beyond the scope of this document. It could be as simple as a weighted average of the point estimates from 3 individual papers[[7]](#footnote-6). A more sophisticated approach would be aggregating 3 individual datasets and then perform the analysis of interest on such “*meta dataset*”:



Let’s assume that we only care about the size of the main effect in this meta-analysis. Using the appended meta dataset, we draw the conclusion that, *ceteris paribus*, receiving the intervention will, on average, increase the outcome by units for any individual in our sample. Meta-analysis done! Life is beautiful!

Or is it?

## **1.2. The Task**

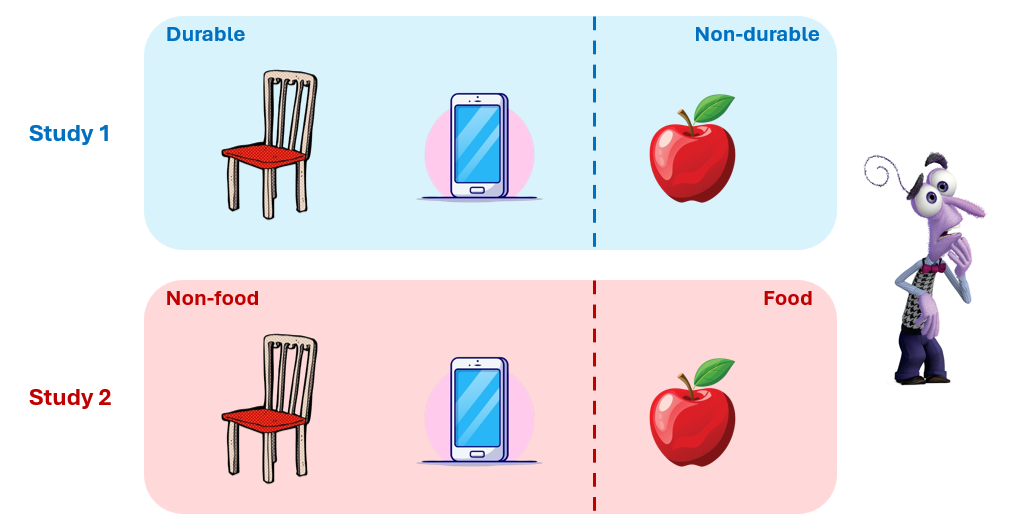
*“Ernest Hemingway once wrote, 'The world is a fine place and worth fighting for.' I agree with the second part.”*

*—* Detective Lt. William Somerset, *Se7en*

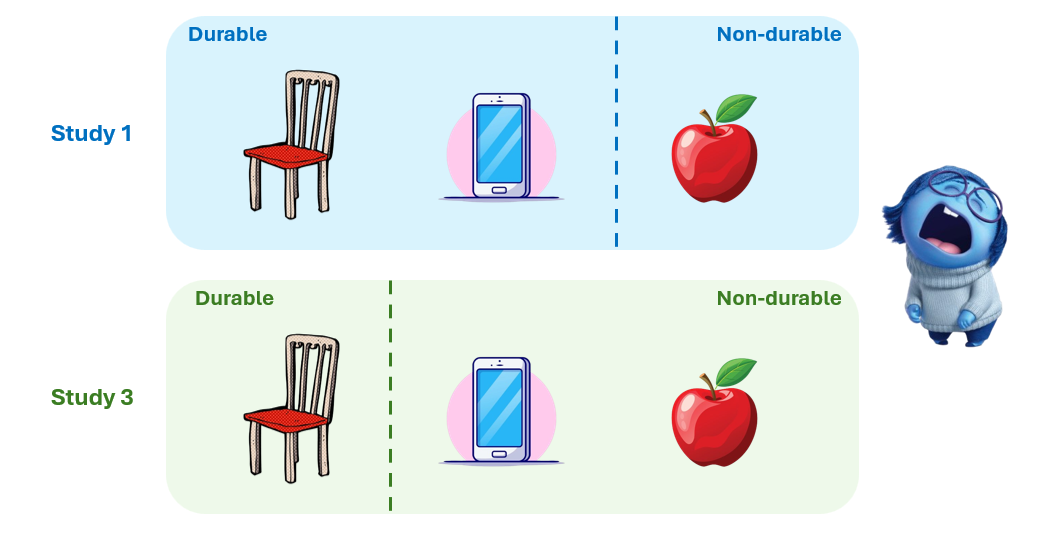
Throughout the previous section, we are assuming all three authors are doing the same things in their research design. That’s overly optimistic in practice: what if we allow for the slightest bit of deviation from this ideal assumption?

Based on what we’ve seen in the paper, data, and instrument, we found that different authors might be constructing the same outcome variables differently. Take ***consumption*** variables as an example: Suppose all 3 papers in our sample asked about consumption value at the *itemized* level, which include mobile phones, chairs, and apples.

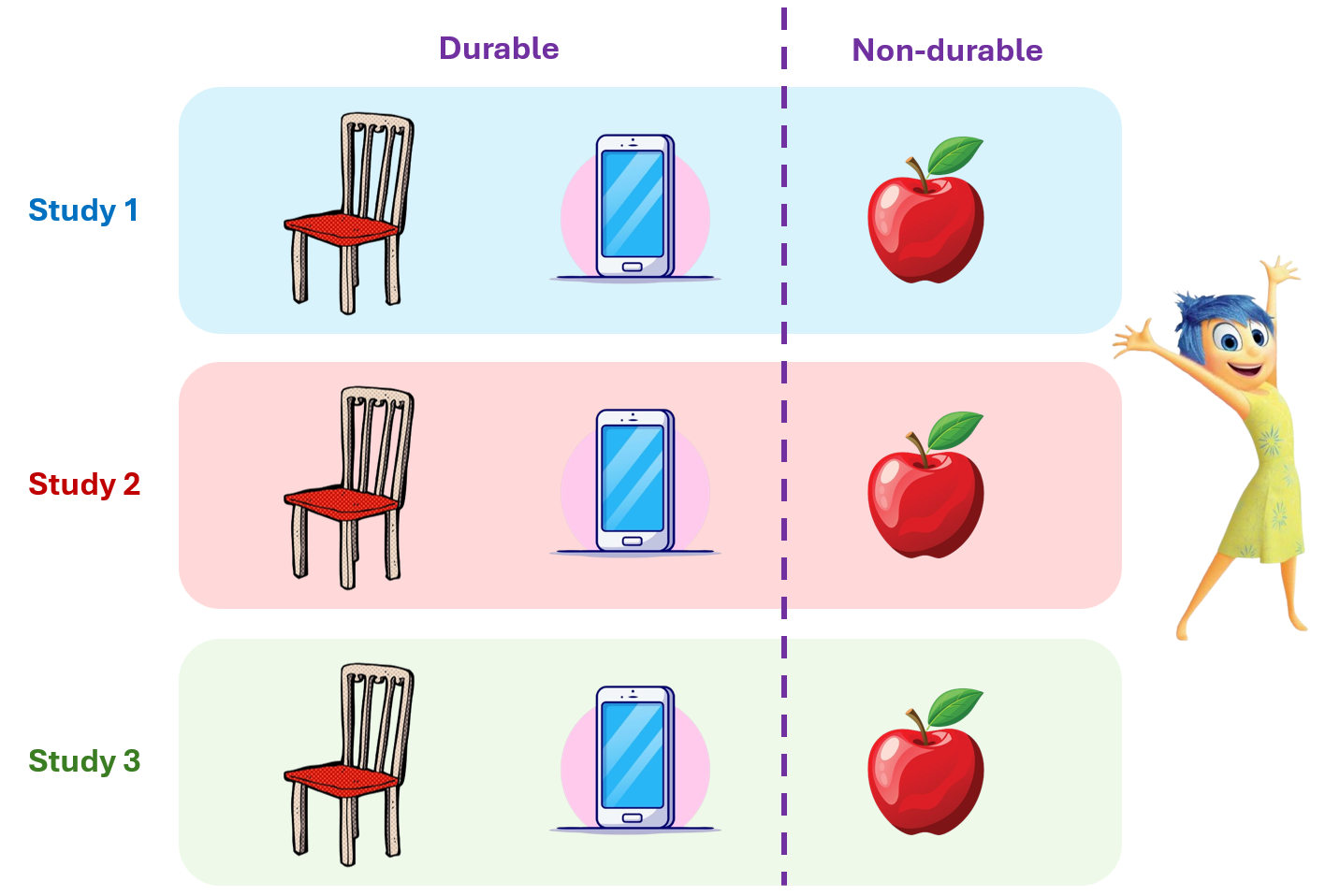
* **Creating different categories**: two papers might aggregate these consumption values by different categorizations, e.g., durable/non-durable and food/non-food (presumably for different purposes of their analysis). The problem is, if we only take the aggregate level amounts, the data does not speak to each other - There is not a single consumption variable that exists in all papers.
  + In this case, we *cannot* run the regression that includes study 1 and 2 since we cannot find such to put on the LHS of the model.



* **Using different definitions**: two papers might use the same categorization yet apply different definitions. Directly using these aggregate variables constructed by the authors themselves leads to ambiguity in the interpretation of our results. In other words, garbage in, garbage out.
  + In this case, we *can* run the regression that includes study 1 and 3, yet the interpretation of the estimate cannot be definitively made.



To resolve this issue, our PI made the suggestion that we dig into the itemized consumption values. Imagine: ideally, if we can obtain the consumed amounts for ALL items for ALL papers, we can use whichever categorization we want (thus solved the 1st issue) based on whatever definition we want (thus solved the 2nd issue). That will solve this issue, namely different authors constructed consumption variables differently, once and for all!



There. So this is a lengthy but detailed discussion of why we are doing these itemized checks.

To accomplish this task, our team was divided into two routes:

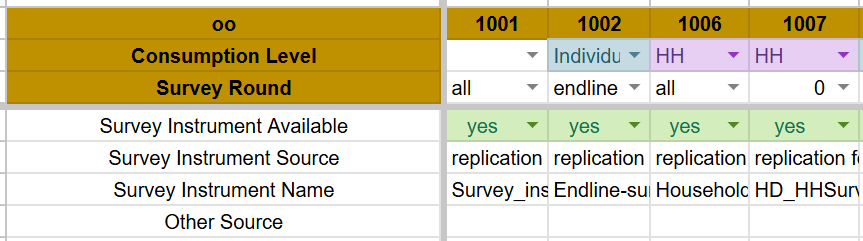
* **Survey instruments**: Kevin and Arthi will dig into the survey instruments to see which items had their consumption values surveyed in which paper. This gives us an idea which itemized data *should* exist in our datasets.
* **Datasets**: Samir and Ben will dive into the current dataset to see which items had their consumption values included in our data. This tells us which itemized data *actually* exist in our datasets.

For those papers where itemized data should exist (included in survey instrument) but does not currently exist (not in data), we will get in touch with the authors to request the raw-raw data that contains this information.

On the other hand, for those papers where itemized data should not exist (either survey instrument unavailable or does not contain these questions) but currently exist (in data), we should check the correctness of the data at first instance. If it is indeed sensible, well, it must be our birthday.

# **2. How are we doing this? (Survey instruments)**

## **2.1. Section 1 - Basic Information**



* **Paper ID:** the IDs of individual paper that are consistent with the harmonization tracker.
  + **When do we create more than one column for a single paper?**
    - Items surveyed differ across different *consumption levels*.
    - Items surveyed differ across different *survey rounds*.

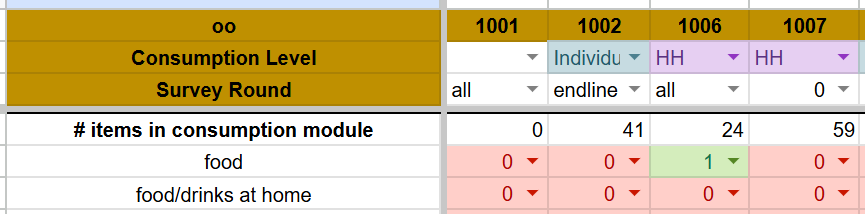
(If you find other cases beyond these two, please inform us.)

* **Consumption level**: the questions in the survey instrument should always explicitly specify at which level it is asking (otherwise it would be a terrible question). In our sheet we provided three options:
  + **Individual**: If the question focuses on individual level consumption, regardless of the identity of this individual (respondent, participant, beneficiary, HH head, etc.), we select “individual” option.
  + **HH**: If the question focuses on HH level consumption, we select the “HH” option.
  + **Both**: If the survey asks the consumption of the *SAME* list of items for both individual and HH, we select “both”. If you spot the slightest discrepancy between questions asked about individuals and HH, we should create separate columns for each.
  + **Sidenote: Why do we care?**

When we are later aggregating consumption data and doing analysis, individual and HH consumption values are treated as completely separate variables. We do not want to mix consumption data at different levels: for instance, we might want to put “HH apple consumption” and “HH banana consumption” into the same bucket, but never with “individual banana consumption”.

* **Survey rounds**: the survey instrument itself should clarify which round of survey it is associated with. Depending on the research purpose, changes in research budget, or refinement of instruments over time, you might find different sets of consumption items were being asked at different rounds. Two types of options are available for this row:
  + **All**: This option applies if you find surveys in all rounds asking about the same set of consumption items. (The survey might still be different in other sections.) Again, if you find the slightest discrepancy across rounds, create a separate row for each.
  + **Baseline, 1, 2, …**: This option applies if this column is only about one or more specific survey rounds.
* **Survey instrument availability (source, name, etc.)**: This section should be self-explanatory.
  + **One little thing on “*Source*”**: If you find the instrument not in the replication folder but from an online source, please put the link into the “Survey Instrument Source” row. Please do NOT download the instrument, paste into the replication folder, then write “replication folder” in the “Survey Instrument Source” row. (This way we can check the online source if we later need to).

## **2.2. Section 2 - Itemized record**



* **Number of items in consumption module**: numerically, this number should equal the number of “1”s in subsequent rows. If a separate item can be identified for each survey question in the consumption module, this number also equals the number of questions in that module.
  + **If you combine several “items” to be one item**: you should update this number to still be the sum of all “1”s. Under these circumstances, this number no longer equals the number of questions in the consumption module. But that’s OK. Please also leave a comment to record and explain this decision you made.
* **Item list (left 1st column)**: This list should be ongoingly updated as you check more and more papers. When you come across an item that has never been surveyed in previous papers, add a row here.
  + **Ordering of rows**: It would be most ideal if you could order these rows such that items that are similar *in category* to each other are gathered in one place. For example, such categories could be fruits, food, furniture, clothing, etc.
  + **IMPORTANT: What do we see as the same or separate items?**

There isn’t a precise, comprehensive definition I can think of to explain what we mean by “items”. Instead, I develop a list of cases where the answer to this question is not obvious. This list will also be ongoingly updated.

* + - **Case 1**: “Apple” and “Apple / banana / peach” - **Separate**

-> When one item is a subset to another “item”, we list them as separate items. It’s not obvious how we should use the data on the item “Apple / banana / peach” since the authors kind of already formed their own bucket when designing questions. Let’s keep them separate to keep track of such cases, but organize these two rows to be close to each other.

* + - **Case 2**: “Apples purchased outside” and “Apples self-grown” - **Same**

-> They are both about apples! I don’t see any reason why we should treat them separately. You can put “1” for the item “Apple” in both cases. Later when we work on the data, we will only use the item “apples” and then pool these two into the same variable.

* + - **Case 3**: “Food eaten outside HH” and “Food eaten within HH” - **Separate**

-> When the question is already a bucket of, rather than an individual, item, let’s separate them into two rows. Otherwise, we will have a single row for “food” and it’s by itself already a category, not an item. (I suspect that surveys that asked about “food” in general in single questions will also have further questions asking about specific items. Well, whatever.)

* + - Other cases to be updated…
* **0 or 1**: For each item, “1” means that this item is asked 1) in terms of consumption 2) in monetary amounts.
  + **In terms of consumption?** The confusion would mostly likely take place in the distinction between *consumption* and *assets*. See the following example:
    - “*Asset*” question: How many mobile phones do you currently own?
    - “*Consumption*” question: How much did you spend on buying mobile phones in the past month?

A simple rule is asking yourself: whether the question is about stock (asset) or flow (consumption)? More specifically, an asset question is about a state, a snapshot in time, whereas an consumption question is about money spent during a specific period (the recall period).

* **Small tip**: if you find your question to be in the “Consumption” module, you can be less worried about this distinction. Otherwise, you should make some effort to distinguish.
  + **In Monetary amounts?** Depending on the research purpose or data collection process, the authors might not care about monetary amounts, or they used a more convoluted way to get to this amount. For example, the authors may ask: 1) Did you consume any XXX 2) if yes, how many units of XXX did you consume and 3) what was the market price of XXX per unit. Answers to these questions enable the authors to calculate the monetary amounts of consumption, yet you would not find a single survey question asking about monetary amounts. In light of the situation, let’s do this:
    - **If you find a survey question about monetary amounts**: Great! Put “**1**” for this item.
    - If you find a survey question about itemized:
      * *Indicator*: consumed or not (Yes/No question)
      * *Quantity*: how many units consumed
      * *Price*: what was the market price per unit

Put “**0**” for this item, but leave a comment to record what type of question you found it to be.

* Other issues to be updated…

## **2.3. Some remaining loose ends**

***– Recall period***

One thing we chose to ignore for now is the “recall period” of the itemized consumption questions. We made this choice to not overwhelm ourselves at a preliminary stage. For now, we are more interested in what the bigger picture looks like at the itemized level, in order to assess the feasibility of certain things we’d like to do.

However, it might be reassuring for you to know that the recall period is something we can deal with. It is true that different surveys were asking these consumption questions at different recall periods (yesterday, last week, last month, etc.). To reconcile such differences, we developed an intuitive formula to harmonize all these questions to be under an unified recall period:

**Rule**: daily (\*\_1d) \* 30

= weekly (\*\_1wk) \* 4.3

= bi-weekly (\*\_2wk) \* 2.15

= monthly (\*\_1m)

= yearly (\*\_1yr) / 12

(from the “Harmonization FAQ” document)

I hope you are less worried after you know this!

# **3. How are we doing this? (Datasets)**

(Currently only Samir is working on this task in full. Before I write up this section, I need to talk to him to familiarize myself with the process and hear any question or difficulty he had)

1. During the first meeting between Kevin and I, we talked about what now has been written in this part. However, that was my first time telling this from my head to another person, so my verbalization is far from perfect. I felt guilty. Hence, I tried to organize my thoughts better in writing this document and hopefully resolve some loose ends. [↑](#footnote-ref-0)
2. “Systematic” is a word you see everywhere in economics. It means largely the similar thing, but still different in precise meanings in different contexts. In the current context, inquire a source on the distinction between *narrative review* and *systematic review*. [↑](#footnote-ref-1)
3. “Unbiased”, “average treatment effect”, “regression”... These are all technical jargons used in the field of economics and statistics. If you have no clue or only a vague idea of what they mean, please feel free to refer to an econometrics/statistics textbook, lecture notes, or simply ask any RA in the lab for help! [↑](#footnote-ref-2)
4. You might wonder how the baseline survey is useful, since it does not appear in our regression model. Good question! We frequently use the baseline survey to perform ***balance checks*** between treatment and control groups to ensure that they are not systematically different, at least in those observable characteristics that we have data on. It gives us an idea how well the random assignment is performed, while also validating our *causal* interpretation of the results later on. [↑](#footnote-ref-3)
5. Researchers sometimes say “*ceteris paribus*”, which is this phrase in latin. No idea why they do that. [↑](#footnote-ref-4)
6. Depending on your context and research question, this could be of main interest in your study. [↑](#footnote-ref-5)
7. The weight is determined by the *precision* of estimate from each paper, which is highly associated with, as you probably already guessed, sample sizes. [↑](#footnote-ref-6)