**Codebook instructions**

*Before you begin trying to code measures, you will need to familiarize yourself with the modes of measurement. These are listed on the last page with some examples of common measures that would fit into each mode. Please read them thoroughly.*

1. For each journal, start with the first empty row in the codebook. The order of the rows has already been randomly shuffled for you. Your goal is to extract **15 articles from each journal, not all 45 that are included in the codebook**. Once you meet 15 articles per journal, assessed in the order the rows are presented, you should move on to the next journal.
2. Check this article against inclusion and exclusion criteria. If it meets criteria for inclusion, mark the “included” column as TRUE, otherwise mark as FALSE.
   1. Only articles reporting original research data are considered.
   2. Secondary data analyses, meta-analyses, reviews, opinion and theoretical articles, editorials, and commentaries are excluded.
   3. You may need to access the website for that article to see more details to determine whether to include it or not.
   4. If included == FALSE, go to the next article for this journal in the codebook.
   5. If included == TRUE, proceed to the next step.
3. Obtain pdf for first listed article.
   1. These should be kept and organized. I suspect a Zotero workflow would be fastest, but this might need describing for the RA.
4. Go to the measures/procedure subsection (usually under the method section).
5. Record the names of each measure listed or described as being part of the study in the relevant column in the codebook.
   1. The following are excluded from consideration / do not constitute a measure for our purposes
      1. Demographics data collection (e.g. age, gender, education level, etc.)
      2. Intervention, manipulation checks, attention checks or exclusion variables
   2. You may encounter measures that are not given a name but merely a description, e.g., “participants evaluated the stimuli on a 7-point Likert scale from very negative to very positive”. In such cases, enter the name using the format “unnamed [description]”, for example “unnamed rating scale”. Using the word “unnamed” is important for data processing later.
6. Code each measure as one of the 7 categories.
   1. Step 1: is the measure already in the codebook from a previously scored study, either under the same *or a similar name*, given that these names can vary between studies and authors? If so - briefly, but thoughtfully, review the mode it was previously assigned to, and if that coding seems correct and without errors then reuse it for this study. You will naturally become familiar with some popular measures with practice.
   2. Step 2: If the measure has not already been coded for a previous article, try to classify the measure based on the information available in the article using the decision heuristics and flowchart.
   3. Step 3: If you are unable to make a classification based on the information available in the article, look up more details about the scale (assuming it has a name or something that would allow for this). We have access to APA PsycTests which contains the details of thousands of psychological measures, although bear in mind that a) the details recorded in PsycTests are not perfect and it does have errors (e.g., similarly named measures being confused), and b) it contains mostly Self-Reports about Subjective States, Self-Reports about Behavior, and Behavioral Proxy Measures.
   4. Step 4: If no determination can be made within a reasonable amount of time (~5 minutes), code the measure as “indeterminable or other”.
      1. This could happen because a) the details of the measure are clear but the classification of it based on those details is unclear, or b) there are insufficient details to make a classification.
   5. Please do not use ChatGPT to find the items from a scale (as it is known for hallucinating responses).
   6. If the measure is self-report – read through all items listed. This is important because if even one item asks to report about behavior (e.g., sleep, physical activity, screen time, etc.) then the whole measure needs to be coded as Self-Report about Behavior

Decision making heuristic:

1. If the measure has words like “subjective” or “perceived” in its name, then you know it will be Self-Report about Subjective States
2. If the description mentions a Likert scale or a Likert-type scale, then you know it will be either Self-Report about Subjective States or Self-Report about Behavior
3. If the article states that it is a self-report scale and mentions a few items as examples:
   * if one or more of the examples is a Self-Report about Behavior (i.e., asks participants to report about their own overt behavior) such as “I slept poorly in the last month” or “I have left the house more than twice this week” then this will be coded as Self-Report about Behavior
   * if the only examples listed are Self-Report about Subjective States (i.e., asks participants to report about their mood, attitudes, beliefs, etc.) such as “In the past two weeks I have felt more sad than usual” or “I don’t feel valued in my romantic relationships” then you will still need to find the full list of items, if possible. This is because if one of the items does ask about behavior and is just not listed as an example in this article, it would need to be coded as Self-Report about Behavior. Please use PsycTests in situations like these.
4. If the procedure section talks about participants being seated in a cubicle or at a desk or at a computer, it is less likely that Direct Behavioral Measures were being utilized.
5. If the description of the measure mentions collecting behavior (frequency, latency, duration), read further to find what the construct of interest is. This may be made evident by phrases like “to collect attitudes about…” or “to determine XXX traits…”.
   * If the construct of interest is behavioral itself (e.g., weight loss) and the measurement is collected in behavior (e.g., kilograms of weight lost per month), then code it as a Direct Behavioral Measure.
   * If the construct of interest is internal, cognitive, or goes beyond the observed variable itself (e.g., implicit bias, moral judgement) and the measurement is collected in behavior (e.g., latency of keyboard presses, number of laws broken while driving a simulated car) then code it as Behavioral Proxy Measure.
6. If the construct of interest or the measurement type is neurological, physiological, or biological (e.g., neuron activation, eye movement, blood cells) then code it as Neuro/Bio/Psychophys Measure.
7. If the construct of interest requires a trained coder or device to score it, and it assesses individuals’ capacity to produce responses that are predefined as being correct in order to index an ability that varies between individuals (such as IQ or memory), code it as Measure of Cognitive Ability.
8. Repeat these steps until you have coded every measure from every study in the article. Your task is to extract all **distinct measures reported in each article, not each study**. If a measure is used in study 1 and then again in study 2, do not record it in the codebook a second time.

From Stage 1 Report (submitted)

## Coding scheme

Measures were coded as one of 6 mutually exclusive categories or as “indeterminable or other”. Brief descriptions of each category and their coding rules are provided here. Thorough explanations of the steps and decisions to be taken to classify each measure are provided in the codebook.

[Please note that the codebook will go through a piloting phase, using a separate set of articles not used in the analyses, before being finalized. The preregistration will be updated with the final version of the codebook. During the piloting phase, we will iteratively update, where necessary, definitions of the modes of measurement, both here in the manuscript text and also in the codebook and codebook instructions and decision flow chart. If reviewers would prefer to review the finalized codebook and materials prior to (updating) the preregistration, we are happy to supply this.]

### Direct Behavioral Measures

We take our definition of Direct Behavioral Measures from the behavioral psychology literature, which specifies that these require direct observation of the behavior of interest, that it must be quantified using metrics of frequency, latency, or duration (Cooper et al., 2020, p. 76). These observations and recordings must be made by a third party (e.g., a human observer in real time or reviewing recordings) or device rather than the individual themselves. For example: number of times per day an individual opens a social media app as recorded by the smartphone’s logs, or number of seconds that a participant keeps their hand submerged in a cold pressor task as recorded by the researcher observing the participant. These are also defined by what they exclude, specifically, they do not involve self-reports. As such, any mention of procedural features such as Likert scales excludes a measure from this category.

The [measure name] was [a commonly used / the most frequently used measure] in this category.

### Behavioral Proxy Measures

We define this mode as measures that collect measurements of overt behavior but use these to create metrics of psychological variables. For example, while the Implicit Association Test (IAT; Greenwald et al., 1998) does measure button pressing behavior and record reaction times, these primary data are used to create a metric of implicit associations. That is, the construct of interest is a psychological that goes beyond the observed behavior itself.

This distinction between Direct Behavioral Measures and Behavioral Proxy Measures is important as otherwise behaviors such as eating, sleeping, or suicide attempts would be treated as modally equivalent with measures such as the Iowa Gambling Task (IGT; Bechara et al., 1994), the Dot-Probe task (MacLeod et al., 1986), or the Balloon-Analogue Risk Task (BART; Lejuez et al., 2002), each of which involve counts of responses but are ultimately used to measure psychological variables (i.e., decision making, attentional bias to threat, and risk taking, respectively).

The [measure name] was [a commonly used / the most frequently used measure] in this category.

### Neuro/Bio/Psychophys Measures

This mode captures neurological, biological, psychophysiological and other internal variables, whether psychological or behavioral. Examples include EEG, fMRI, salivary cortisol swabs, eye-blinks, and Galvanic skin response.

The [measure name] was [a commonly used / the most frequently used measure] in this category.

### Self-Reports about Subjective States

This mode involves one or more items that are either open-ended (e.g., text responses) or closed-ended (e.g., Likert scales) and involve the respondent introspecting upon their internal subjective states. The object of reference being reported on can include things such as attitudes, beliefs, perceptions, feelings, thoughts, emotions, or preferences, but also intentions and planned future behavior. Examples include the Subjective Happiness Scale (Lyubomirsky & Lepper, 1999) and the Multidimensional Scale of Perceived Social Support (MSPSS; Zimet et al., 1988).

The [measure name] was [a commonly used / the most frequently used measure] in this category.

### Self-Reports about Behavior

This mode collects responses that are derived from introspection but whose object of reference is the individual’s own past behavior. For example, a self-report measure asking participants how often they open Instagram (regardless of whether response options are counts of the behavior or Likert scales of relative frequency).

Many multi-item self-report measures contain a combination of items that refer to subjective states and behaviors. For example, the Beck Depression Inventory II (BDI-II: Beck et al., 1996) contains both items such as “I sleep somewhat more than usual” (Self-Report of Behavior) as well as “I feel quite guilty most of the time” (Self-Report of Subjective State). For pragmatic reasons, our codebook classified self-reports containing mixed items such as this as Self-Reports of Behavior. That is, measures were rated as Self-Report of Behavior if they contained any such items.

The [measure name] was [a commonly used / the most frequently used measure] in this category.

### Measures of Cognitive Ability

This mode involves the assessment of individuals’ capacity to produce responses that are predefined as being correct in order to index an ability that varies between individuals. Examples include measures of IQ, memory, creativity, and moral reasoning.

The [measure name] was [a commonly used / the most frequently used measure] in this category.

### Indeterminable or other

Raters are instructed to label a measure as ‘indeterminable or other’ if they could not determine its mode within a reasonable amount of time or if it does not appear to meet any of the modes defined by the codebook.

## Instructions to coders

The coders extracted each measure’s name, or assigned it a generic one if no name was reported. In the first instance, the coders used the details of the measure provided in the article to try to classify the measure’s mode, including its name, description, its procedural features: descriptions of the variable, construct, or target object it was used to measure; the responses captured, including response options and method of responding; the metrics to quantify those responses (e.g., whether they used metrics frequency, latency, or duration); and the stimuli presented to participants, including stimuli, text, instructions, items, and feedback.

These decisions were aided via a decision flow chart and a set of heuristics. For example, use of Likert response options implies that the measure is a Self-Report about Subjective Experiences, Self-Report about Behavior, or in principle a Measure of Cognitive Ability. Comparably, measures referring to “subjective” or “perceived” are typically Self-Report about Subjective Experiences, but occasionally they are Self-Report about Behavior if they also contain items about self-reported behavior.

In the second instance, if the details of the measure reported in the article were insufficient to code its mode, the coder consulted the APA PsycTests database of psychological measures or the reference for the measure if one was cited in the article.