Psychology: The science of behavior?

Sabrina F. Norwood, Malte Elson, & Ian Hussey
University of Bern

Stage 1 Registered Report (Submitted)

Psychology's self-image is that we are "the study of the mind and behavior" (APA, 2018). Indeed, across twenty commonly used undergraduate psychology textbooks and two dictionaries of psychology published between 2002 and 2024, the study of "behavior" is the only consistent element across all definitions (see Table 1), which speaks to our collective self-image and goals. Naturally, psychological science aims to explain behavior and its determinants in terms of psychological variables. Our field also continues to place high value on research that treats overt behavior as its explanandum, perhaps revealing our implicit scientific values. Our introductory textbooks are full of examples of studies with behavioral explananda, including classics such as Milgram's obedience studies (1974), Bandura et al.'s Bobo Doll experiment (1961), or Asch's conformity study (1951). These studies continue to be cited (collectively over 35,000 times), and used in psych 101 courses to excite students and inspire them to pursue the study of psychology.[1] Consider the counterfactual: if Milgram, instead of instructing participants to deliver what they believed to be actual shocks to others, had merely asked participants to rate how likely they would be to deliver a hypothetical shock on a scale of 1 (very unlikely) to 7 (very likely), would his study have the same impact or carry the same value in our field?

Table 1. Definitions of psychology in popular textbooks and dictionaries.

Source	Definition
APA dictionary of psychology (2018)	"the study of the mind and behavior."
Bernstein & Nash (2002) Essentials of psychology. 2nd edition.	"the science of behavior and mental processes" (p. 3)
Carlson et al. (2014). Psychology: The science of behavior. 7th edition.	"a science with a specific focus on behavior. The primary emphasis is on discovering and explaining the causes of behavior." (p. 19)
Ciccarelli & White (2017) Psychology. 5th edition.	"the scientific study of behavior and mental processes." (p. 44)
Coon & Mitterer (2016) Introduction to psychology: Gateways to mind and behavior. 14th edition.	"the scientific study of overt behavior and mental processes (i.e., covert behavior)" (p. 16)
Feldman (2017) Understanding psychology. 13th edition.	"the scientific study of behavior and mental processes" (p. 3)
Gazzaniga et al. (2016) Psychological science. 5th edition.	"the study of mental activity and behavior" (p. 4)
Hockenbury & Hockenbury (2014) Discovering psychology. 6th edition.	"the science of behavior and mental processes" (p. 2)
Kalat (2016). Introduction to psychology. 11th edition.	"the systematic study of behavior and experience." (p. 3)
King (2017) The science of psychology: An appreciative view. 4th edition.	"the scientific study of behavior and mental processes." (p. 4)
Lilienfeld et al. (2011) Psychology: From inquiry to understanding. 2nd edition.	"the scientific study of the mind, brain, and behavior." (p. 3)
Myers (2013) Psychology. 10th edition.	"the science of behavior and mental processes." (p. 6)
Nolen-Hoeksema et al. (2008) Atkinson & Hilgard's introduction to psychology. 15th edition.	"the scientific study of behavior and mental processes." (p. 5)
Oxford dictionary of psychology (2015) 4th edition.	"the study of the nature, functions, and phenomena of behaviour and mental experience."

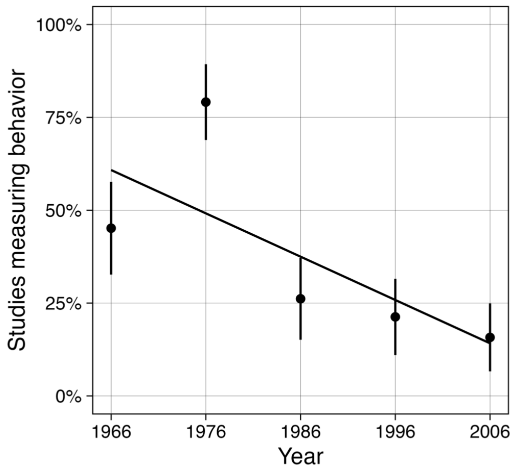
[1] Note that our point here is about the impact and fame of these studies, agnostic to their quality or replicability.

Passer & Smith (2011) Psychology: The science of mind and behavior. 5th edition	"the scientific study of behavior and the mind." (p. 2)
Pastorino & Doyle-Portillo (2016) What is psychology? Foundations, applications, and integration. 4th edition.	"the scientific study of behavior and mental processes" (p. 4)
Plotnik (2014) Introduction to psychology. 10th edition.	"the systematic, scientific study of behaviors and mental processes" (p. 4)
Santrock (2004) Psychology. 7th edition.	"the scientific study of behavior and mental processes" (p. 5)
Stangor & Walinga (2014) Introduction to psychology. 1st Canadian edition.	"the scientific study of mind and behaviour." (p. 2)
Wade et al. (2024). Psychology. 14th edition.	"the scientific discipline concerned with behavior and mental processes and how they are affected by an organism's physical state, mental state, and external environment" (p. 3)
Weiten (2015) Psychology: Themes and variations. 10th edition.	"the science that studies behavior and the physiological and cognitive processes that underlie behavior" (p. 15)
Zimbardo et al. (2017) Psychology: Core concepts. 8th edition.	"the science of behavior and mental processes" (p. 2)

For behavior to serve as our explanandum, it must be measured. Importantly, Baumeister and colleagues (2007) called this self-image of psychology as the science of behavior into question by showing that between 1966 and 2006 the proportion of psychology studies reporting direct observations of behavior dramatically decreased from 79% in its peak in 1976 to 16% in 2006 (see Figure 1[2]), suggesting that these have been supplanted by introspective self-reports, hypothetical scenarios, and reaction times. Provocatively, the authors argued that, even in 2007, psychology had transitioned from being a "science of behavior" to a "science of self-reports and finger movements." The explanandum has shifted from observable behaviors to purported mental states, processes, and other unobservable constructs. Baumeister et al. (2007) lament that "it cannot be blithely assumed that responding to questionnaires is enough to tell us all we need to know about actual life. It is necessary to study actual behavior sometimes." (p. 400).

Figure 1. Reanalysis of Baumeister et al.'s (2007) reported data. Intervals represent 95% Confidence Intervals.

[2] We extracted the percentages of articles employing direct behavioral measures from Baumeister et al. (2007) using WebPlotDigitizer (Rohatgi, 2017) and calculated 95% Confidence Intervals based on the number of articles they reported extracting data from. We then reanalyzed this data using meta-regression using a Bayesian model with normal(0,1) priors. Extracted data and the linear trend are plotted in Figure 1. Overall prevalence: 39.4%, 95% CR [34.6, 44.0], trend: -1.3%, 95% CR [-1.6, 1.0] points per year. [for peer review: analysis can be found in the github under code/baumeister et al. 2007 reanalysis/analysis.Rmd]



Of course, many would argue that psychological variables can also be legitimate explananda, and we would agree. To argue otherwise would be to say that it is not useful to study, for example, whether depression can be treated with psychotherapy or whether personality changes over the lifespan. However, at the same time, whereas Baumeister et al.'s (2007) results suggest that psychological science's explananda have changed over time, our self-image of our field does not appear to have been updated: even recently published dictionaries and introductory textbooks retain definitions of psychology as the science of behavior (see Table 1). This raises the question of misalignment between our self-image and our collective research activities. A science of behavior must routinely observe and measure actual overt behavior, just as cell biology as the science of cell structure and function must routinely observe actual cells for its explicit goals to be aligned with its collective research activity.

Technology has greatly expanded the opportunity to collect behavioral data

It is worth asking why psychology studies measured behavior at reduced rates from the 1960s through the 2000s. Was this due to a genuine shift in our collective explananda of interest? Or was it perhaps due to more mundane and pragmatic concerns? Baumeister et al. (2007) argued that the decrease was driven by the latter, and described self-report measures as a contrast to "difficult and expensive behavioral observation." This may have been true when direct behavioral observation was at its most prevalent in the 1960s and 70s. Indeed, even in the early 2000s it was time consuming to directly observe and record overt behavior. Potentially, the pressure to publish more frequently and employ larger sample sizes may have contributed to the apparent decline in the use of Direct Behavioral Measures. Recruiting participants for lab-based tasks involving clicking keys or completing self-report questionnaires was far quicker and cheaper than directly observing behavior in the real world or in artificial, constructed environments. However, in the nearly twenty years since Baumeister et al. (2007), our technological landscape, and the ubiquity of devices in everyday life, has transformed significantly. Coincidentally, the first iPhone was released in 2007 around the same time as Baumeister et al. (2007) was published. The subsequent mass adoption of smartphones has, in principle, greatly expanded our ability to measure and study behavior unobtrusively and lowered the pragmatic barriers to collecting direct behavioral data.

In combination with the more recent proliferation of other technologies such as smart watches and smart home devices, there is now a vast array of sensors that routinely record overt behavior, including location data, microphones, cameras, pedometers, and more, as well as to access user behavior within apps running on these devices, such as social media screen time, content, communications, physical activity, and sleep (Harari et al., 2016). Moreover, with more than 80% of the global population owning smartphones (Statista, 2023), individuals are already generating significant behavioral observation data through their devices at no cost or effort to them, often without them even noticing. This technological shift has therefore significantly expanded our data collection options, especially for measures of direct behavior - at least in principle.

This raises an important question: has psychological science capitalized on these technological opportunities for greater collection of behavioral data and realigned collective research activity with its goals over the past two decades? And, more broadly, what types or modes of measurement does psychological science employ, and how have they changed in recent history since the proliferation of smartphones and smart devices? Our own reanalysis of Baumeister et al.'s (2007) data suggested that over their observation period of 1966 to 2006 there was a linear decrease in the use of behavioral measures by -1.3 (95% CR [-1.6, 1.0]) percentage points per year (see footnote 2 and Supplementary Materials for full details [for peer review: analysis can be found in the github under code/baumeister et al. 2007 reanalysis/analysis.Rmd]). Nearly two decades later, either that downward trend has continued and measures of behaviors, our self-declared primary explanandum, are now extremely rare, or perhaps the disruptive effects of technology have greatly changed our data collection practices. These questions provided a central motivation for this work.

Modes of measurement

In order to answer these questions, it is useful to define general classes or modes of measurement. It is worth noting that Baumeister et al. (2007) only quantified the prevalence of "behavioral measures", and even then did not provide a definition or report the reliability of these ratings, instead adopting a 'we know it when we see it' approach. One key implication of this is the conclusion in the title, that psychology risks becoming a "science of self-reports and finger movements" was not directly evidenced by the data, as a low rate of behavioral measures does not necessarily imply that all remaining measurement is done via "reaction times and questionnaire responses" (p. 396).

In this study, we therefore go beyond Baumeister et al.'s (2007) approach by creating a coding scheme for multiple mutually exclusive modes of measurement that we think can provide good coverage of the types of measurement employed in psychological science. Direct Behavioral Measures involve the frequency, latency, or duration of overt behaviors. Behavioral Proxy Measures use observable behavior to infer internal or mental states, such as tasks that measure reaction time to draw conclusions about implicit associations or attentional bias. Despite reaction times being a form of overt behavior, like Baumeister et al. (2007), we make this distinction on the basis that for Behavioral Proxy Measures, the construct of interest is not the behavior itself, unlike Direct Behavioral Measures. We discuss this further below. Neuro/Bio/Psychophys Measures capture internal physiological and/or biological behavior (e.g., brain behavior, cell behavior, Galvanic skin response, or eye movements). Measures of Cognitive Ability assess 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. Lastly, we make a distinction between two types of self-report measures: Self-Reports about Subjective States (e.g., their mood, attitudes, beliefs) versus Self-Reports about Behavior (i.e., their own overt behavior). We return to the classification of these in the methods section.

Alignment between research question and mode of measurement

When should we conclude that a given mode of measurement is being used enough? It is important to recognize that this study is neither concerned with superiority or differences of validity between modes, nor with validity of measures within each mode. Rather, the relevance and utility of a given measurement method depends on its alignment with the researchers' explanandum (thing to be explained) and their measurement choice. For example, clinical psychologists aiming to reduce patient attrition during treatment may find Direct Behavioral Measures of patient attendance (e.g., observational data from the session therapist) more closely aligned with their explanandum than Self-Reports about Subjective States, such as intentions to attend future sessions, as attendance (behavior) is more aligned with concept they seek to explain. The same principle applies across various research domains. Emotion researchers studying subjective experiences may find better alignment by using Self-Reports about Subjective States, as these measures capture the internal experiences they aim to explain. Similarly, cognitive neuroscientists investigating brain function are likely to achieve better alignment by using Neuro/Bio/Psychophys Measures (e.g., fMRI, EEG), as these methods reflect the internal physiological processes that constitute their explanandum.

At first glance, these examples may seem obvious, but misalignment between modes of measurement and research questions may be more prevalent than expected and can serve to undermine the credibility of claims. For example, in the field of media effects research, studies examining the impact of social media use on adolescent well-being should prioritize Direct Behavioral Measures of social media usage, such as smartphone logs of app usage duration over self-reports about usage (Parry et al., 2022). This is because Direct Behavioral Measures more accurately reflect the behavior in question. Indeed, this concern is well-documented in the literature, where multiple authors have pointed out that Direct Behavioral Measures of duration of social media use and self-reported social media use are very poorly correlated (Boyle et al., 2022; Ernala et al., 2020; Verbeij et al., 2021), and yet many claims in the literature are based on things other than Direct Behavioral Measures of social media use (Kaye et al., 2020), potentially undermining the credibility of those claims.

A second example can be found in depression research. Among seven of the most commonly used self-report depression scales, all seven included items related to sleep disturbances (Fried, 2017). However, research elsewhere suggests that self-reported sleep duration (Self-Report about Behavior) is only weakly correlated with sleep duration measured via wrist-mounted accelerometer (Direct Behavioral Measure; $r = .28$, $N = 821$: Jackson et al., 2020). This could have theoretical implications, as it raises the question of whether depression is defined by actual sleep disturbances or perceived sleep disturbances.

Other examples of poor correspondence between self-reported behaviors and directly observed behaviors have been reported in many other domains, including smartphone use (Ellis et al., 2019; Ohme et al., 2021; Parry et al., 2021), physical activity (Affuso et al., 2011; Marasso et al., 2021; Prince et al., 2008), and medication adherence (Garfield et al., 2011; Rickles et al., 2023; Wells et al., 2022). This is not to say that self-report data about behavior in these domains are inherently flawed, or that it is not interesting to study how people self-report these things. Rather, they serve to highlight the potential for misalignment between modes of measurement and specific research questions or claims.

When research strays too far from using psychological variables to explain behavior, and instead uses psychological variables to explain other psychological variables (e.g., perceived sleep to explain depression), there is a risk that resulting explanations become arbitrary, disconnected from observable realities, or even tautological. While research that employs explananda other than behavior can be easier to conduct, it introduces significant challenges. These explanations are harder to disprove because they lack clear, observable benchmarks; without a behavioral anchor, distinguishing between a useful explanation and a flawed one becomes more difficult, because it can become less clear what the explanation is of. Further, this type of research may face challenges in producing findings consumers of research might find useful or effective in guiding their behaviors, as the research was not designed to inform behaviors in the first place. When the behavior is removed as the explanandum, the line between a sound explanation and a speculative one may blur, leaving the field vulnerable to producing theories with internal coherence but little external relevance.

The current research

While Baumeister et al. (2007) raised important questions about the trends in measurement practices in psychology, particularly the shift away from direct behavioral observations as its primary explanandum, it's important to note that their conclusions were based on an analysis with limited robustness, depth and scope. First, their study used just one rater per article, did not provide a definition of behavioral measures, and did not report inter-rater reliability, making it unclear how accurate their extractions were. Second, their study extracted data from just a single journal (Journal of Personality and Social Psychology), leaving open the possibility that their results may not generalize to other subfields or the general field of psychology. Third, by quantifying only one mode of measurement (Direct Behavioral Measures), Baumeister et al. (2007) leaves the reader to infer that studies not collecting behavioral data are likely collecting other modes such as "self-reports and finger movements", without actually providing direct evidence of this. Fourth and finally, results were presented simply as percentages of articles per year rather than employing formal modeling to estimate prevalence and trends.

This study therefore sought to provide robust and general evidence of prevalences and trends in different modes of measurement in psychology research during the period between 2009 and 2023. In particular, it sought to understand how widespread Direct Behavioral Measures were during this period and whether this was changing, perhaps due to the technological changes that have come about during this period. Lastly, we were interested in the use of other modes of measurement during this period and the associations between the rate of change in the use among the different modes. We did this by going beyond Baumeister et al.'s (2007) data extraction and analyses in multiple ways: by quantifying multiple modes of measurement; providing definitions, scoring guidelines, and estimates of inter-rater reliability of our extracted data; by extracting data from multiple subfields and multiple journals within each subfield; and by formally modeling data in a multilevel model that produces estimates of prevalence and trend at different levels (field, subfield, and journal).

Method

[Please note that placeholder text is highlighted in yellow. This text will be finalized after Stage 1 Acceptance and data collection]

Transparency statement

Availability of data and code

All data and code are available at osf.io/ykw4s and github.com/sabrinanorwood/measurement-trends-in-psychology.

Preregistration and deviations from preregistration

The preregistration for this study is available at [OSF prereg URL]. The preregistration includes the sample size, the analyses, the code implementing them, hypotheses, and decision making rules.

[There were no deviations from preregistration / We deviated from the preregistration in the following ways: …]

Subfields and journals covered

The generality of Baumeister and colleagues' (2007) results were limited by the fact that they extracted data from just a single journal (Journal of Personality and Social Psychology). In order to increase generality and afford comparisons, we extracted data from five journals in each of six subfields of psychology (general, clinical, cognitive, developmental, industrial and organizational, and social and personality).

Journals were selected based on a combination of their citation metrics, their representativeness for that field, and the type of work they tend to publish (e.g., we tried to include both experimental and non-experimental journals from fields that publish both, and cover the range of populations considered by a subfield, such as child and aging in developmental). We excluded review and theoretical journals, journals that publish a low volume of articles per year (<100), and highly interdisciplinary journals (e.g., Trends in Cognitive Science, Developmental Science) on the basis that we are interested in measurement trends in psychology specifically and our coding scheme was designed for psychological modes of measurement. We consulted with colleagues in each subfield about our candidate lists of journals and made modifications based on their recommendations. The journals sampled are depicted in Table 2.

Table 2. Included journals.

Subfield	Journal
General Psychology	Collabra: Psychology
	Journal of Experimental Psychology: General
	Nature Human Behavior
	PLOS ONE psychology
	Psychological Science
Clinical	Behaviour Research and Therapy
	Clinical Psychological Science
	Journal of Clinical Child and Adolescent Psychology
	Journal of Consulting and Clinical Psychology
	Journal of Psychopathology and Clinical Science
Cognitive	Cognition
	Journal of Experimental Psychology: Human Perception and Performance
	Journal of Experimental Psychology: Learning Memory and Cognition
	Journal of Memory and Language
	Memory and Cognition
Developmental	Child Development
	Developmental Psychology
	Journal of Experimental Child Psychology
	Journal of Research on Adolescence
	Psychology of Aging
Industrial & Organizational	Journal of Applied Psychology
	Journal of Experimental Psychology: Applied
	Journal of Organizational Behavior
	Organizational Behavior and Human Decision Processes
	Personnel Psychology
Social & Personality	Emotion
	Journal of Experimental Social Psychology
	Journal of Personality and Social Psychology
	Journal of Research in Personality
	Personality and Social Psychology Bulletin

Inclusion criteria

Inclusion criteria were publications in the specified journal whose publication dates were between 2009 and 2023. Articles were included if they reported the results of quantitative research involving original data collection. Meta-analyses, review articles, and theoretical papers were therefore not included.

Sample size determination

Based on piloting, we estimated that extraction takes 20 minutes per article. Based on the availability of resources, we therefore elected to extract 15 articles per journal for a total of 450 articles. To estimate inter-rater reliability, a randomly sampled 25% of articles were scored by a second rater. This gave us an estimated total of around 190 hours of work between the raters, which we determined to be feasible within the project time. While the proportion of articles sampled per journal across the 15-year period is small, partial pooling within the multilevel model we use to analyze the data is nonetheless capable of providing meaningfully precise estimates of prevalence and trend at the field and subfield level. We illustrate this by fitting the model to plausible simulated datasets prior to data collection, although no full-blown simulation study was conducted.

Data extraction

We first obtained a complete list of publications in each journal between these time periods. We then used a reproducible R script to randomly sample and randomize the order of 45 articles from each journal (i.e., three times more articles than would be extracted, on the basis that some would not meet inclusion criteria). We employed random sampling across years (e.g., rather than stratifying by year) to provide less biased estimates of prevalence and slope.

Inclusion criteria are applied item by item during extraction: raters started with the first publication by order and assessed whether it met inclusion criteria. If it did, they extracted the relevant data. If it did not, it was discarded. This process continued until data were extracted for 15 articles in each journal. Data extraction was performed by two coders. Of the articles that were determined to meet the inclusion criteria by the first coder, a randomly chosen 25% were double scored by the second coder.

In order to streamline the extraction process, once a given mode of measurement was extracted from an article, no further instances of that same mode are collected. This also fits within our analysis plan as each article is analyzed as either having an instance of each mode or not, but not the number of instances per mode of measurement. For example: if the first measure listed in Study One of an article utilized a Self-Report about Subjective States measure, then no further instances of Self-Report about Subjective States were noted for that article.

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

Coders were trained to use the codebook (for peer review: https://docs.google.com/spreadsheets/d/10M1C6YqZDElwCEIQq51Q43nplzg93ELXGjujYnJ6t4/edit?usp=sharing) using a set of instructions (for peer review: https://osf.io/jd2cr) and a decision flow chart (for peer review: https://osf.io/8pf9y). 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.

Data processing

The output of the coding scheme is (a) a list of all measures used in each article from which data were extracted, and (b) a classification of each of these measures into a single category or "indeterminable or other". From this, binary classifications were calculated at the article level, which represent whether each article did (1) or did not (0) employ one or more measures in each category. For example, if a study had participants complete the Beck Depression Inventory-II and the Beck Hopelessness Scale, the Self-Report about Behaviors and Self-Report about Subjective States were each be scored as 1 and all other categories were scored as 0.

Inter-rater reliability

The inter-rater reliability of the article-level category scorings for each measurement mode are presented in Table 3. Cohen's κ was used to determine the reliability of the codings. Modes with $\kappa > .20$ are often regarded as fair, $\kappa > .40$ as moderate, $\kappa > .60$ as substantial, and $\kappa > .80$ as almost perfect. We preregistered that the results of analysis for modes whose codings were not at least moderate are interpreted with additional caution on the basis that the reliability of the data they are based on is lower. This could indicate either issues with the codebook or with the underlying conceptualization of the modes themselves. [Results suggest that the reliability of the codings is at least [fair] for [all] modes. / …]

Table 3. Inter-rater reliability for each measurement mode.

Measurement mode	Agreement	Cohen's κ
Behavioral Proxy	80.0%	0.70
Cognitive Ability	80.0%	0.70
Direct Behavioral Measure	80.0%	0.70
Neurophysiological	80.0%	0.70
Self-Report about Behavior	80.0%	0.70
Self-Report about Subjective Experience	80.0%	0.70

Resolution of disagreements

Disagreements between coders were resolved by discussion and mutual agreement, with the input of an additional coder where needed. For efficiency, disagreements are assessed at the level of the binary category scores rather than the individual measures' scores, on the basis that it is not necessary to resolve all disagreements about individual measures when the analyses are conducted on the article-level category data.

Analytic strategy

Model specification and interpretation

Data were analyzed using a Bayesian multilevel multivariate logistic regression fit using the R package brms (v2.14.4; Bürkner, 2017). The six binary article-level usage data for each category of measure were entered as the DVs. Correlations between the dependent variables were estimated by the model rather than assuming these to be zero. Publication year was centered relative to the middle year in the dataset (i.e., 2016) to make the centered variable years before or after 2016. Subfield and journal were entered in the model as nested random intercepts, to acknowledge the dependencies in the data (i.e., articles are nested in journals which are nested in subfields), and in order to estimate prevalence and slopes for journals and subfields. It is more appropriate to enter these as random effects rather than fixed on the basis that subfields and journals are non-exhaustive categories (i.e., we wish to make results more generalizable to other unobserved levels of these variables). The centered year variable was entered in the model as both a fixed effect and a random slope in order to acknowledge that trends over time can vary between subfields and journals.

Lastly, we sought to make generalizations about the prevalences and trends in different modes of measurement in the field of psychology as a whole, on the basis of the subfields and journals that we extracted data from. However, we recognize that some subfields and journals produce a larger number of articles and studies than others. In order to take this into account, we therefore included as weightings the total number of empirical articles published per journal during the time period as a percentage of the total number of empirical articles published in all the journals. Note that this is not an attempt to weight the model by the relative size or influence of the subfields, which would be much harder to quantify and not directly related to our research question. Rather, this was a way to acknowledge that subfields and journals publish different volumes of research when making generalizations about the field as a whole, as represented by our sample of journals.

The Wilkinson notation for the each DV in the model takes the following form (see the Supplementary Materials for the full model specification): [for peer review: analysis can be found in the github under code/registered report/analysis.Rmd]

```
dv | weights(percent_of_all_articles) ~ 1 + year_centered + (1 +
year_centered | p | subfield) + (1 + year_centered | q | subfield:journal
```

The centering of year makes the model results more interpretable: the estimated marginal mean for the intercept more closely corresponds with the prevalence of a given measure category across the years; and the estimated marginal slope more closely corresponds with the linear trend across years. However, no attempt was made to interpret the model parameters themselves, which are more difficult to interpret due to the relative complexity of the model and as they are on the log-odds scale.

Instead, all results were interpreted via the estimates of the posterior prevalences and trends and their 95% percentile-based Credible Intervals. These were computed using the R package marginaleffects (Arel-Bundock, 2023; Arel-Bundock et al., 2023). In order to be precise about the type of estimates we extracted, we adopted the nomenclature employed by Heiss (2021, 2022). In order to make inferences about prevalences and trends in psychology as a field, we computed Marginal Global Means (aka Average Marginal Effects). These do not take observed variation between the levels of the random effect into account. Additionally, in order to make inferences about typical psychological articles, we extracted Conditional Global Means (aka Group Average Marginal Effects), which do take observed variation between the levels of the random effect into account. As such, the 95% CR of a Conditional Global Mean will tend to be wider than those for the Conditional Global Means. In simpler terms, whereas the Marginal Global Mean allows us to make inferences about the abstract field of psychology as a whole (e.g., "typically, psychology articles …"), whereas the Conditional Global Mean allows us make inferences about articles in the field of psychology while recognizing variations between subfields and journals (e.g., "a typical psychology article …").

Additionally, we extracted Conditional Means at two different levels: by subfield and by journal. These allow inferences to be made about subfields and journals respectively, not taking heterogeneity at lower levels of the random effect into account. Results for the individual journals are reported in the Supplementary Materials [for peer review: analysis can be found in the github under code/registered report/plots/ and data/registered report/results/posterior_prevalence_and_trends_journals.csv] due space constraints and the relative imprecision of estimates at that level given that each is based on 15 data points.

Model priors

We provide a brief description and justification of our priors here (as well as an in-depth justification and visualization of all priors in the Supplementary Materials [for peer review: analysis can be found in the github under code/registered report/analysis.Rmd]. We employed weakly informative priors on the Intercepts which differ between the DVs, in order to aid model convergence based on our prior beliefs (informed by Baumeister et al., 2007) about the prevalence of these measures. These priors are sufficiently weak that it is more useful to describe what values they characterized as less probable to be observed (negative beliefs) than which beliefs they asserted (positive beliefs). For example, for the intercept for Self-Reports about Subjective States, we employ a normal prior with $M = 0.85$ (on the log-odds scale) and $SD = 1$, which represents a weak prior belief that they are unlikely to be rare or uncommon (<10% prevalence). In contrast, the normal prior on the intercept for Direct Behavioral Measures uses $M = -1.10$ and $SD = 1$, which represents a weak prior belief that they are unlikely to be extremely common (>80% prevalence). The intercept of all other measurement modes employs normal priors with $M = 0.50$ and $SD = 1.4$, which represents an even weaker belief that only very low or very high prevalences are less probable (e.g., prevalences < 5% or > 95%).

We employed weakly regularizing priors on all other parameters, which serve to aid model convergence by excluding impossible or extremely implausible values without imposing any strong prior beliefs. We employed a normal prior with $M = 0$ and $SD = 0.20$ on all slopes for (centered) year on the basis these represented a weak belief that only very large changes in prevalence between the start and the end of our extraction period were improbable (e.g., from 5% to 95% or vice-versa). Following the recommendations of McElreath (2020), we placed exponential priors with $\lambda = 1$ on the *SD* of all random intercepts (i.e., representing the variability between subfields and between journals nested within subfields). This provides coverage for a wide range of plausible *SD* values on the log-odds scale while making implausibly large ones unlikely. Lastly, because our model modeled correlations among random effects, the distribution of correlations also requires a prior. We retained brms's default weakly regularizing Lewandowski-Kurowicka-Joe (LKJ) prior with $\eta = 1$, which is merely skeptical of large correlations, in line with our goal of employing weakly regularizing priors.

Model convergence

We assessed model convergence using multiple metrics. First, we examined the Gelman-Rubin diagnostic ($\hat{R}$) to ensure all values are close to 1.0, indicating satisfactory convergence, and that effect sample sizes are sufficiently large. We visually inspected trace plots for each parameter to confirm that the Markov chains exhibited good mixing and had fully explored the posterior distribution. [No issues with model convergence were detected. / Issues with model convergence were detected and resolved by [e.g. increasing the number of warmup samples, increasing the adapt-delta parameter to increase how cautiously the NUTS sampler explores the probability space] so that the final model had adequate convergence.]

Research questions and decision making

For decision making purposes regarding prevalences, we created category labels: Rare < 5% ≤ Uncommon < 10% ≤ Occasional < 25% ≤ Frequent < 50% ≤ Common. These labels are descriptive rather than evaluative (e.g., ‘rare’ does not inherently imply ‘problematically low’). To make decisions about trends, we used the 95% Credible Intervals to determine whether there was a detectably positive trend or negative trend (i.e., whether the intervals excluded zero) or whether no trend was detectable.

Our first research question is related to the prevalence of Direct Behavioral Measures. If psychology is the science of behavior - a science that regularly treats behavior as its explanandum - then psychology research must measure overt behavior regularly. However, based on Baumeister et al.’s (2007) results and our own subjective reading of the literature, we have the impression that psychology’s research activities are not aligned with this. If the prevalence of Direct Behavioral Measures is Rare or Uncommon (≤ 10%) for the field of psychology as a whole (Figures 2 and 4, and Table 4, Marginal Global Mean), we would conclude that psychology did not, within the observed time period, regularly collect the necessary measurements to treat behavior as its explanandum, and the field is therefore not accurately described as the ‘science of human behavior’.

Our second research question is related to changes in the use of Direct Behavioral Measures over time. If psychology has capitalized on the technological changes of the last two decades to collect greatly more data on overt behavior (e.g., via smartphones and smart devices), we would see an increase in the use of this mode of measurement over time. Unfortunately, we have the subjective impression that it has not done so. If the trend for Direct Behavioral Measures is detectably positive for the field of psychology as a whole (Figures 2 and 5, and Table 4, Marginal Global Mean), we would conclude that the use of this mode increased during the time period, providing supportive evidence for the idea that psychology has embraced the availability of behavioral data collection afforded by technological changes.

Our third research question is related to the more general patterns in the different modes of measurement. Researchers have finite resources for data collection and analysis and measurement choices necessarily involve opportunity costs. Patterns in the prevalences and trends in the field of psychology of all six modes of measurement are discussed, in a more exploratory way, based on their prevalence category determinations and trend decisions (Figures 2, 4, 5, and Table 4; again using the Marginal Global Means). This is done in combination with the (a) correlations among the prevalences of the different modes of measurement (Table 5) and (b) correlations among their trends (Table 6). That is, we extracted the posterior correlations among the random intercepts and slopes, respectively, at the level of journals. These provide information about whether, respectively, the use of one mode of measurement was associated with the use of other modes of measurement, and whether changes in the use of one mode during the time period was associated with changes in the use of other modes.

Our fourth research question is related to recognizing the potential for heterogeneity between subfields and journals. For example, perhaps psychology as a field rarely employs a given mode of measurement, but it could nonetheless be the case that a given subfield or specific journal employs it very regularly. This nuance would be important to recognize. As such, while our research questions 1-3 sought to make generalizations about the field as a whole, our fourth research question was about heterogeneity in typical articles. This was done by changing the metric used to obtain the posterior estimates. Whereas research questions 1-3 were addressed through the use of Marginal Global Means, which allow us to make inferences about psychology articles on average, for this research question we instead rely on the Conditional Global Means, which allow us to make inferences about typical psychology articles. The relative width of the 95% Credible Intervals for the Conditional Global Means for the field of psychology (Figure 3, 4, and 5, and Table 4) will be considered descriptively and contrasted with the width of the intervals of the Marginal Global Means[3]. Wide conditional intervals would imply the presence of substantial variation between typical psychology articles due to the presence of heterogeneity. This could suggest, for example, that the probability of a given typical article employing a given mode of measurement may diverge substantially from the estimated average observed at the level of the field. In contrast, comparable interval widths between the Marginal and Conditional Global Means would imply there is little heterogeneity between subfields and journals, and that the patterns observed in the field as a whole are generally reflected in each subfield and journal.

Results are also presented at the level of the subfields (i.e., Conditional Means; see Table 4) on the basis that they are likely to be of interest to readers, and they may aid in describing notable specific sources of heterogeneity. However, no specific research questions or comparisons were preregistered at the level of the subfields. Additionally, results at the level of the individual journals can be found in the Supplementary Materials [for peer review: analysis can be found in the github under code/registered report/plots/ and data/registered report/results/posterior_prevalence_and_trends_journals.csv]. This is for reasons of conciseness and on the basis that estimation at the journal level is relatively poorer.

Results

Prevalence of Direct Behavioral Measures

[describe results for prevalences of Direct Behavioral Measures from Figure 2, 4 and 5, and Table 4 Field of Psychology (marginal), following the research question and decision rules outlined above.]

Trends in the use of Direct Behavioral Measures

[describe results for trends for Direct Behavioral Measures from Figure 2, 4 and 5, and Table 4 Field of Psychology (marginal), following the research question and decision rules outlined above.]

Prevalence and trends among modes of measurement

[describe results for prevalences and trends for other modes of measurement from Figure 2, 4 and 5, and Table 4 Field of Psychology (marginal), following the research question and descriptive approach outlined above.]

Heterogeneity between subfields

[describe observed heterogeneity using Figure 3 and Table 4 Field of Psychology (conditional), following research question and descriptive approach outlined above.]

[3] Note that this distinction has a conceptual parallel with Confidence Intervals vs. Prediction Intervals in frequentist random-effects meta-analyses, where the Confidence Interval estimates represent the uncertainty around the estimate of the mean of the distribution of the random effect (akin to the Marginal Global Mean), whereas the Prediction Interval takes observed heterogeneity in the random effect into account (akin to the Conditional Global Mean).

Figure 2. Prevalence and trends in the modes of measurement in the field of psychology (Marginal Global Means).

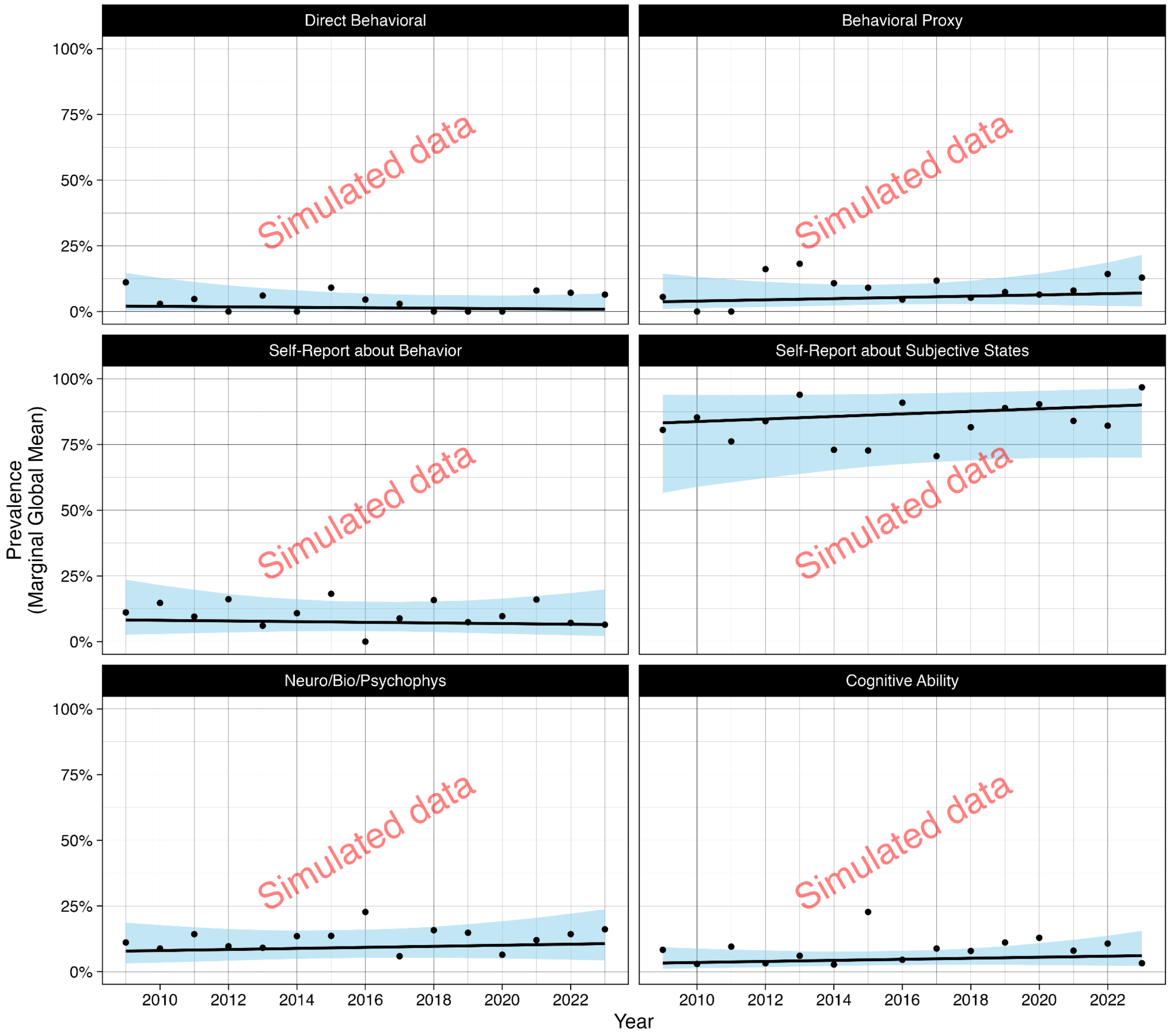


Figure 3. Prevalence and trends in the modes of measurement in typical psychology articles (Conditional Global Means).

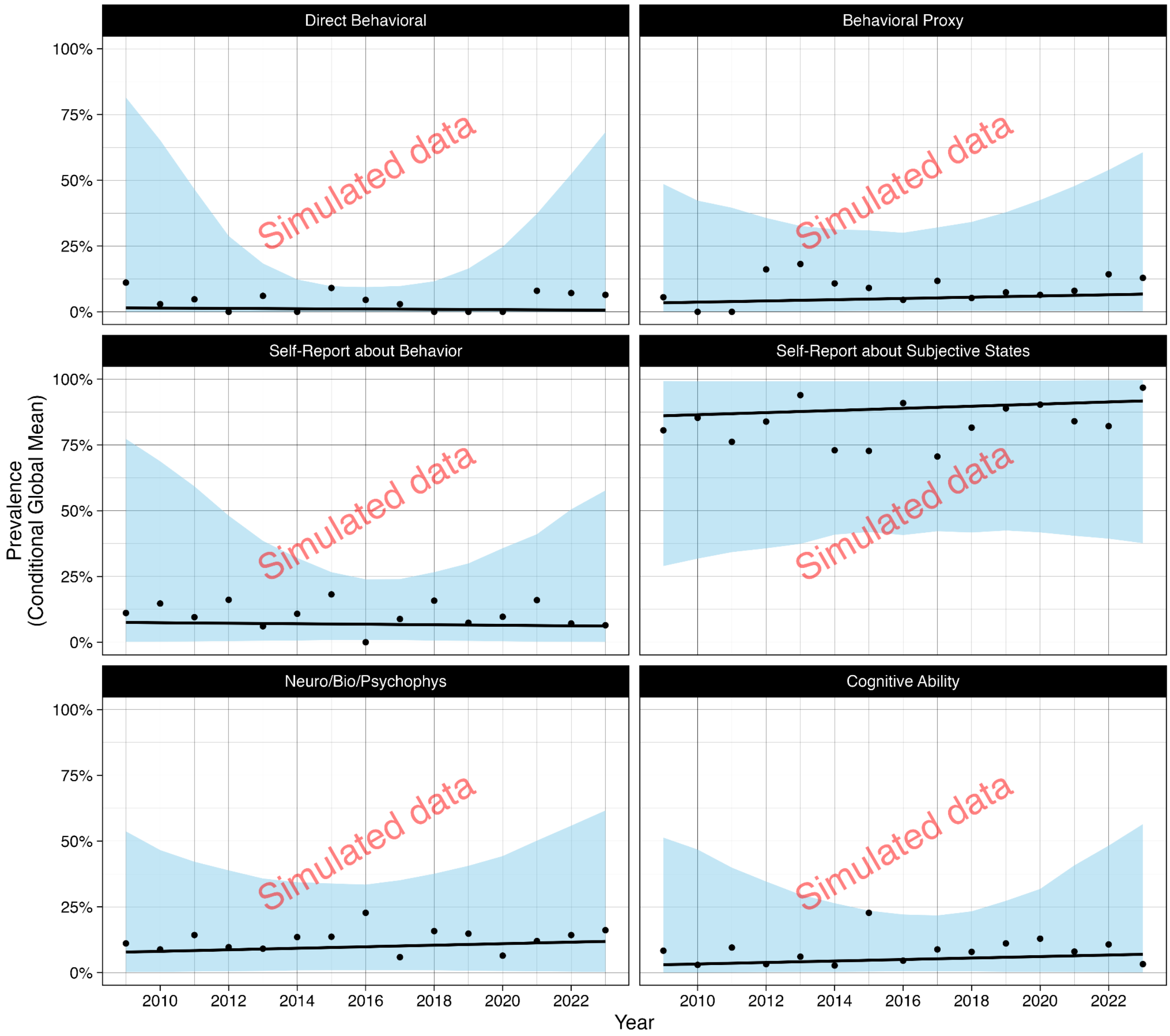


Figure 4. Prevalences of the different modes of measurement at the field and subfield levels.

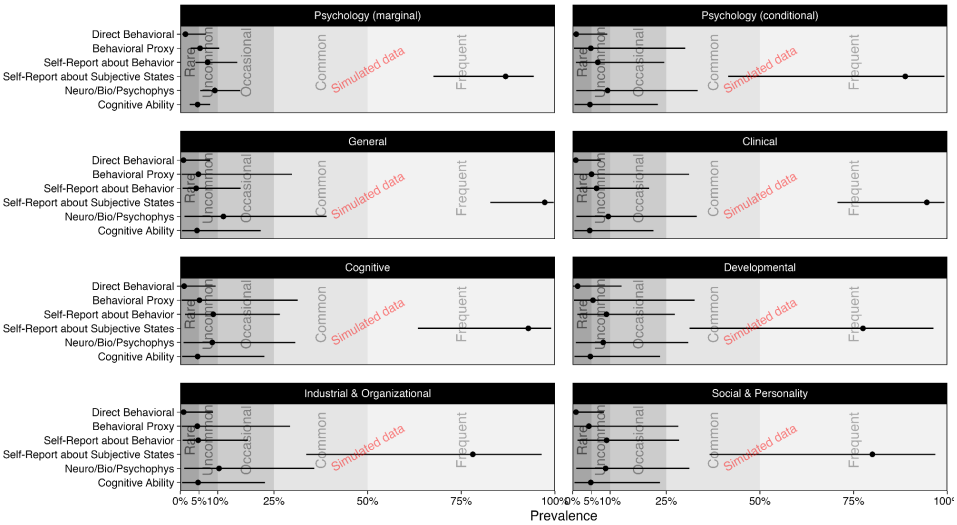


Figure 5. Trends of the different modes of measurement at the field and subfield levels.

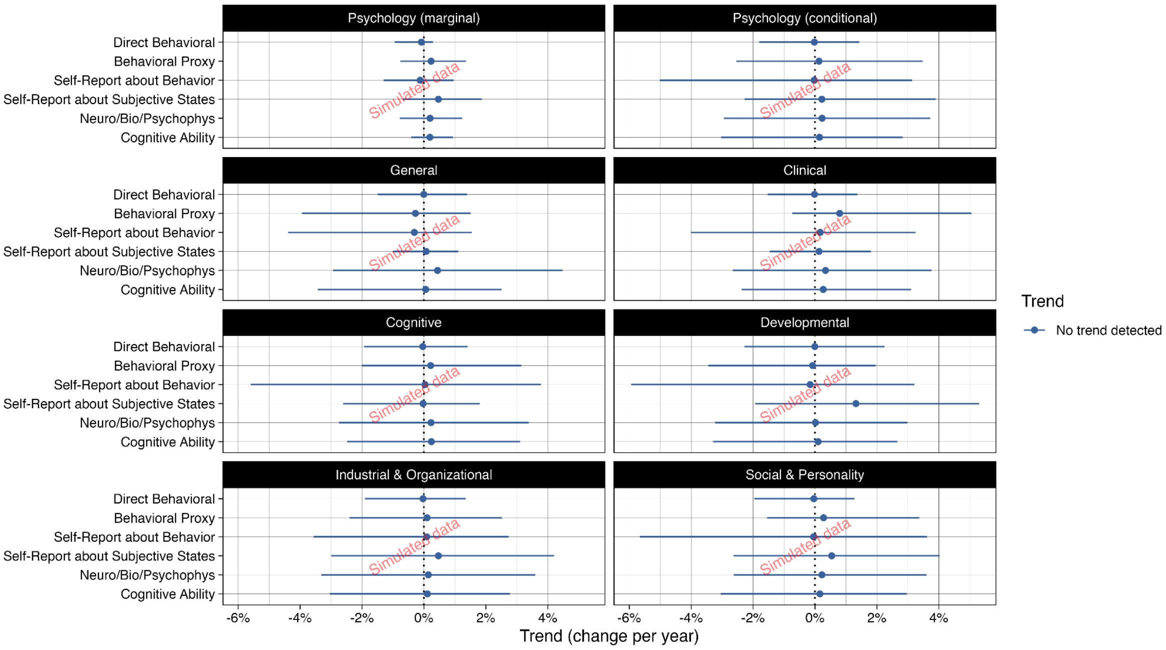


Table 4. Prevalences and trends for the different modes of measurement at the field and subfield levels.

Field/subfield	Measurement mode	Prevalence	95% CR	Decision	Trend	95% CR	Decision
Field of Psychology (marginal)	Behavioral Proxy	0.08	0.04, 0.12	Uncommon	0.00	-0.01, 0.01	No trend detected
	Cognitive Ability	0.12	0.08, 0.17	Occasional	0.00	-0.01, 0.01	No trend detected
	Direct Behavioral Measure	0.03	0.01, 0.05	Rare	0.00	-0.01, 0.00	No trend detected
	Neurophysiological	0.07	0.04, 0.10	Uncommon	0.00	-0.01, 0.01	No trend detected
	Self-Report about Behavior	0.1	0.06, 0.16	Occasional	0.00	-0.01, 0.01	No trend detected
	Self-Report about Subjective Experience	0.86	0.74, 0.95	Frequent	0.00	-0.01, 0.01	No trend detected
Typical articles in Psychology (conditional)	Behavioral Proxy	0.08	0.04, 0.12	Uncommon	0.00	-0.01, 0.01	No trend detected
	Cognitive Ability	0.12	0.08, 0.17	Occasional	0.00	-0.01, 0.01	No trend detected
	Direct Behavioral Measure	0.03	0.01, 0.05	Rare	0.00	-0.01, 0.00	No trend detected
	Neurophysiological	0.07	0.04, 0.10	Uncommon	0.00	-0.01, 0.01	No trend detected
	Self-Report about Behavior	0.1	0.06, 0.16	Occasional	0.00	-0.01, 0.01	No trend detected
	Self-Report about Subjective Experience	0.86	0.74, 0.95	Frequent	0.00	-0.01, 0.01	No trend detected
General Psychology	Behavioral Proxy	0.07	0.01, 0.16	Uncommon	0.00	-0.01, 0.01	No trend detected
	Cognitive Ability	0.11	0.03, 0.22	Occasional	0.00	-0.01, 0.01	No trend detected
	Direct Behavioral Measure	0.03	0.00, 0.06	Rare	0.00	-0.01, 0.00	No trend detected
	Neurophysiological	0.07	0.02, 0.12	Uncommon	0.00	-0.01, 0.01	No trend detected
	Self-Report about Behavior	0.09	0.03, 0.15	Uncommon	0.00	-0.01, 0.01	No trend detected
	Self-Report about Subjective Experience	0.86	0.73, 0.95	Frequent	0.00	-0.01, 0.01	No trend detected
Clinical	Behavioral Proxy	0.08	0.01, 0.19	Uncommon	0.00	-0.01, 0.01	No trend detected
	Cognitive Ability	0.14	0.04, 0.26	Occasional	0.00	-0.01, 0.01	No trend detected
	Direct Behavioral Measure	0.03	0.00, 0.06	Rare	0.00	-0.01, 0.00	No trend detected
	Neurophysiological	0.07	0.02, 0.13	Uncommon	0.00	-0.01, 0.01	No trend detected
	Self-Report about Behavior	0.09	0.03, 0.15	Uncommon	0.00	-0.01, 0.01	No trend detected
	Self-Report about Subjective Experience	0.94	0.87, 0.99	Frequent	0.00	-0.01, 0.01	No trend detected
Cognitive	Behavioral Proxy	0.08	0.01, 0.17	Uncommon	0.00	-0.01, 0.01	No trend detected
	Cognitive Ability	0.12	0.03, 0.21	Occasional	0.00	-0.01, 0.01	No trend detected
	Direct Behavioral Measure	0.03	0.00, 0.06	Rare	0.00	-0.01, 0.00	No trend detected

	Neurophysiological	0.07	0.02, 0.12	Uncommon	0.00	-0.01, 0.01	No trend detected
	Self-Report about Behavior	0.1	0.04, 0.18	Occasional	0.00	-0.01, 0.01	No trend detected
	Self-Report about Subjective Experience	0.94	0.86, 0.99	Frequent	0.00	-0.01, 0.01	No trend detected
Developmental	Behavioral Proxy	0.08	0.00, 0.19	Uncommon	0.00	-0.01, 0.01	No trend detected
	Cognitive Ability	0.14	0.05, 0.27	Occasional	0.00	-0.01, 0.01	No trend detected
	Direct Behavioral Measure	0.03	0.00, 0.06	Rare	0.00	-0.01, 0.00	No trend detected
	Neurophysiological	0.07	0.02, 0.12	Uncommon	0.00	-0.01, 0.01	No trend detected
	Self-Report about Behavior	0.11	0.04, 0.21	Occasional	0.00	-0.01, 0.01	No trend detected
	Self-Report about Subjective Experience	0.81	0.65, 0.93	Frequent	0.00	-0.01, 0.01	No trend detected
Industrial & Organizational	Behavioral Proxy	0.07	0.00, 0.15	Uncommon	0.00	-0.01, 0.01	No trend detected
	Direct Behavioral Measure	0.02	0.00, 0.06	Rare	0.00	-0.01, 0.00	No trend detected
	Neurophysiological	0.07	0.02, 0.13	Uncommon	0.00	-0.01, 0.01	No trend detected
	Self-Report about Behavior	0.07	0.02, 0.14	Uncommon	0.00	-0.01, 0.01	No trend detected
	Self-Report about Subjective Experience	0.68	0.51, 0.87	Frequent	0.00	-0.01, 0.01	No trend detected
	Cognitive Ability	0.13	0.04, 0.23	Occasional	0.00	-0.01, 0.01	No trend detected
Social & Personality	Behavioral Proxy	0.07	0.01, 0.15	Uncommon	0.00	-0.01, 0.01	No trend detected
	Cognitive Ability	0.13	0.04, 0.23	Occasional	0.00	-0.01, 0.01	No trend detected
	Direct Behavioral Measure	0.02	0.00, 0.05	Rare	0.00	-0.01, 0.00	No trend detected
	Neurophysiological	0.07	0.03, 0.14	Uncommon	0.00	-0.01, 0.01	No trend detected
	Self-Report about Behavior	0.11	0.05, 0.20	Occasional	0.00	-0.01, 0.01	No trend detected
	Self-Report about Subjective Experience	0.71	0.53, 0.87	Frequent	0.00	-0.01, 0.01	No trend detected

Table 5. Correlations between the prevalences of the modes of measurement (random intercepts at the level of journals).

	Direct Behavioral	Behavioral Proxy	Self-Report about Behavior	Self-Report about Subjective States	Neuro/Bio/ Psychophys	Cognitive Ability
Direct Behavioral	1.00	0.16	0.29	0.21	-0.02	-0.14
Behavioral Proxy	0.16	1.00	0.19	0.05	0.15	-0.31
Self-Report about Behavior	0.29	0.19	1.00	0.29	0.03	-0.06
Self-Report about Subjective States	0.21	0.05	0.29	1.00	0.13	0.03
Neuro/Bio/Psychophys	-0.02	0.15	0.03	0.13	1.00	0.07
Cognitive Ability	-0.14	-0.31	-0.06	0.03	0.07	1.00

Table 6. Correlations between the trends of the modes of measurement (random slopes at the level of journals).

	Direct Behavioral	Behavioral Proxy	Self-Report about Behavior	Self-Report about Subjective States	Neuro/Bio/ Psychophys	Cognitive Ability
Direct Behavioral	1.00	0.12	-0.09	0.11	0.05	0.48
Behavioral Proxy	0.12	1.00	-0.01	0.00	-0.10	0.13
Self-Report about Behavior	-0.09	-0.01	1.00	-0.20	0.14	-0.21
Self-Report about Subjective States	0.11	0.00	-0.20	1.00	0.13	0.11
Neuro/Bio/Psychophys	0.05	-0.10	0.14	0.13	1.00	0.02
Cognitive Ability	0.48	0.13	-0.21	0.11	0.02	1.00

Discussion

[results will be summarized and interpreted following the research questions and decision making rules or descriptive guidelines laid out for each above.]

[results will be related back to Baumeister et al.'s (2007) results in terms of possible changes over time, however the differences in our methods will be highlighted as an important caveat on over-interpreting differences as evidence of substantive changes in the prevalence of measures of overt behavior over time, given that such changes could instead be driven by a change in measurement.]

Limitations

Generalizability of the results

It is worth considering the generalizability of the results. On the one hand, the use of random effects modeling and the inclusion of a range of subfields and journals within each subfield increase the model's ability to speak to other unobserved levels of the random effects (i.e., other subfields and journals). At the same time, the subfields and journals were not randomly sampled from the population but were preselected for their size and influence. As such, the degree to which our results would generalize to less prominent journals or other subfields which were not represented in our analysis (such as health psychology, forensic psychology, or environmental psychology) is unclear, and this must be acknowledged as an important limitation.

Relatedly, our model weights only took into account the number of papers published by the journals in the total time period. These may not be fully representative of the size of the subfields, which may influence the generalizability of the results at the field or subfield levels.

It is also worth considering the generalizability of the results to other time periods such as the future. Given the potential for system changes to our field's goals and incentive structures over time, we recommend caution in seeking to generalize our results to future time periods. Future prevalences and trends should be reassessed with contemporaneous data.

[possibly other limitations that emerge after Stage 1 review.]

References

American Psychological Association. (2018). APA Dictionary of Psychology. American Psychological Association. https://dictionary.apa.org/psychology

Arel-Bundock, V., Greifer, N., & Heiss, A. (2023). How to Interpret Statistical Models Using marginaleffects for R and Python. *Journal of Statistical Software.* https://marginaleffects.com.

Arel-Bundock, V. (2023). marginaleffects: Predictions, Comparisons, Slopes, Marginal Means, and Hypothesis Tests. R package version 0.14.0, https://CRAN.R-project.org/package=marginaleffects

Affuso, O., Stevens, J., Catellier, D., McMurray, R. G., Ward, D. S., Lytle, L., ... & Young, D. R. (2011). Validity of self-reported leisure-time sedentary behavior in adolescents. *Journal of negative results in biomedicine, 10*, 1-9.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (Ed.), *Groups, leadership and men; research in human relations* (pp. 177–190). Carnegie Press.

Bandura, A., Ross, D., & Ross, S. A. (1961). Transmission of aggression through imitation of aggressive models. *The Journal of Abnormal and Social Psychology, 63*(3), 575–582. https://doi.org/10.1037/h0045925

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the Science of Self-Reports and Finger Movements: Whatever Happened to Actual Behavior? *Perspectives on Psychological Science, 2*(4), 396–403. https://doi.org/10.1111/j.1745-6916.2007.00051.x

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition, 50,* 7-15.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). Manual for the Beck Depression Inventory-II. San Antonio, TX: Psychological Corporation.

Boyle, S. C., LaBrie, J., Trager, B. M., & Baez, S. (2022). Discrepancies between Objectively Assessed and Self-Reported Daily Social Media Time in the Age of Platform Swinging. *International Journal of Environmental Research and Public Health.* https://doi.org/10.3390/ijerph19169847

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software, 80(1). https://doi.org/10.18637/jss.v080.i01

Cooper, J. O., Heron, T. E., & Heward, W. L. (2020). *Applied Behavior Analysis*, Third Edition. Pearson Education Inc.

Ellis, D. A., Davidson, B. I., Shaw, H., & Geyer, K. (2019). Do smartphone usage scales predict behavior? *International Journal of Human-Computer Studies*, *130*, 86–92. https://doi.org/10.1016/j.ijhcs.2019.05.004

Ernala, S. K., Burke, M., Leavitt, A., & Ellison, N. B. (2020, April). How well do people report time spent on Facebook? An evaluation of established survey questions with recommendations. *In Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-14).

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders, 208,* 191–197. https://doi.org/10.1016/j.jad.2016.10.019

Garfield, S., Clifford, S., Eliasson, L., Barber, N., & Willson, A. (2011). Suitability of measures of self-reported medication adherence for routine clinical use: a systematic review. *BMC medical research methodology, 11*, 149. https://doi.org/10.1186/1471-2288-11-149

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology, 74*(6), 1464.

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using Smartphones to Collect Behavioral Data in Psychological Science: Opportunities, Practical Considerations, and Challenges. *Perspectives on Psychological Science, 11*(6), 838–854. https://doi.org/10.1177/1745691616650285

Heiss, A. (2021). A guide to correctly calculating posterior predictions and average marginal effects with multilievel Bayesian models. https://www.andrewheiss.com/blog/2021/11/10/ame-bayes-re-guide/#overall-summary-of-different-approaches

Heiss, A. (2022). Marginalia: A guide to figuring out what the heck marginal effects, marginal slopes, average marginal effects, marginal effects at the mean, and all these other marginal things are. https://www.andrewheiss.com/blog/2022/05/20/marginalia/#tldr-overall-summary-of-all-these-marginal-effects-approaches

Jackson, C. L., Ward, J. B., Johnson, D. A., Sims, M., Wilson, J., & Redline, S. (2020). Concordance between self-reported and actigraphy-assessed sleep duration among African-American adults: Findings from the Jackson Heart Sleep Study. *Sleep, 43*(3), zsz246. https://doi.org/10.1093/sleep/zsz246

Kaye, L. K., Orben, A., Ellis, D. A., Hunter, S. C., & Houghton, S. (2020). The conceptual and methodological mayhem of "screen time". *International Journal of Environmental Research and Public Health, 17*(10), 3661. https://doi.org/10.3390/ijerph17103661

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., ... & Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: the Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied, 8*(2), 75.

Lyubomirsky, S., & Lepper, H. S. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. *Social indicators research, 46,* 137-155.

MacLeod, C., Mathews, A., & Tata, P. (1986). Attentional bias in emotional disorders. *Journal of abnormal psychology, 95*(1), 15.

Marasso, D., Lupo, C., Collura, S., Rainoldi, A., & Brustio, P. R. (2021). Subjective versus Objective Measure of Physical Activity: A Systematic Review and Meta-Analysis of the Convergent Validity of the Physical Activity Questionnaire for Children (PAQ-C). *International journal of environmental research and public health*, *18*(7), 3413. https://doi.org/10.3390/ijerph18073413

McElreath, R. (2020). Statistical Rethinking: A Bayesian Course with Examples in R and STAN (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429029608

Milgram, S. (1974). *Obedience to authority*.

Ohme, J., Araujo, T., de Vreese, C. H., & Piotrowski, J. T. (2021). Mobile data donations: Assessing self-report accuracy and sample biases with the iOS Screen Time function. *Mobile Media & Communication, 9*(2), 293-313. https://doi.org/10.1177/2050157920959106

Parry, D. A., Davidson, B. I., Sewall, C. J. R., Fisher, J. T., Mieczkowski, H., & Quintana, D. S. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature human behaviour*, *5*(11), 1535–1547. https://doi.org/10.1038/s41562-021-01117-5

Parry, D. A., Fisher, J. T., Mieczkowski, H., Sewall, C. J., & Davidson, B. I. (2022). Social media and well-being: A methodological perspective. *Current Opinion in Psychology, 45*, 101285. https://doi.org/10.1016/j.copsyc.2021.11.005

Prince, S. A., Adamo, K. B., Hamel, M. E., Hardt, J., Connor Gorber, S., & Tremblay, M. (2008). A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *The international journal of behavioral nutrition and physical activity*, *5*, 56. https://doi.org/10.1186/1479-5868-5-56

Rickles, N. M., Mulrooney, M., Sobieraj, D., Hernandez, A. V., Manzey, L. L., Gouveia-Pisano, J. A., Townsend, K. A., Luder, H., Cappelleri, J. C., & Possidente, C. J. (2023). A systematic review of primary care-focused, self-reported medication adherence tools. *Journal of the American Pharmacists Association : JAPhA, 63*(2), 477–490.e1. https://doi.org/10.1016/j.japh.2022.09.007

Rohatgi, A. (2017). Web Plot Digitizer. https://automeris.io/WebPlotDigitizer

Statista. (2023). Number of smartphone users worldwide. https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/

Verbeij, T., Pouwels, J. L., Beyens, I., & Valkenburg, P. M. (2021). The accuracy and validity of self-reported social media use measures among adolescents. *Computers in Human Behavior Reports, 3,* 100090. https://doi.org/10.1016/j.chbr.2021.100090

Wells, J., Crilly, P., & Kayyali, R. (2022). A Systematic Analysis of Reviews Exploring the Scope, Validity, and Reporting of Patient-Reported Outcomes Measures of Medication Adherence in Type 2 Diabetes. *Patient preference and adherence, 16*, 1941–1954. https://doi.org/10.2147/PPA.S375745

Zimet, G. D., Dahlem, N. W., Zimet, S. G., & Farley, G. K. (1988). The multidimensional scale of perceived social support. Journal of personality assessment, 52(1), 30-41.