Applying Artificial Intelligence to Expand the Measurement Toolkit in Clinical Psychological Science: Moving Beyond Self-Reports

Catharine E. Fairbairn[1] and Nigel Bosch[2]

[1]Department of Psychology, University of Illinois Urbana-Champaign

[2]School of Information Sciences and Department of Educational Psychology, University of Illinois Urbana-Champaign

This research was supported by National Institutes of Health Grants R01AA025969 to Catharine E. Fairbairn and R01AA028488 to Catharine E. Fairbairn and Nigel Bosch. Our thanks to Alexa Boland, Scott Jung, Camille Lansang, Silvia Murgia, Jay Park, Yujung Son, and the students and staff of the Alcohol Research Laboratory for help in the conduct of this research. Our thanks also to Thomas Kwapil for thoughtful comments on an earlier draft of this manuscript.

Correspondence concerning this article should be addressed to Catharine Fairbairn, Ph.D., Department of Psychology, University of Illinois Urbana-Champaign, 603 East Daniel St., Champaign, Illinois, 61820, United States. Email: cfairbai@illinois.edu.

Abstract

Research exploring correlates of, precursors to, and consequences of psychological disorders has often relied on designs wherein both predictor and outcome are measured by self-reports. In this article, co-authored by a clinical psychologist (CEF) and a data scientist (NB), we offer information surrounding an evolving class of machine learning models as these inform an expanding measurement toolkit in clinical psychological science. Specifically, we note the development of deep learning applications for image analysis, language analysis, and the analysis of physiological time series data, reviewing implications of these advances for measurement in behavioral research. We weigh strengths and limitations of these automated methods in comparison to self-reports, including the specific form of error likely yielded via each (random vs. systematic), with the aim of fostering a replicable, sustainable, and reputationally strong field of clinical psychological science.

Keywords: artificial intelligence, machine learning, common methods bias, automated measurement, self-reports

Applying Artificial Intelligence to Expand the Measurement Toolkit in Clinical Psychological Science: Moving Beyond Self-Reports

The science of clinical psychology has long been a science of self-reports (Frances & Widiger, 2012; Garcia & Gustavson, 1997; Grove & Tellegen, 1991). From DSM-based diagnoses to depression inventories, basic research exploring psychological health and dysfunction has relied heavily on participant reporting of experience (American Psychiatric Association, 2013; Beck et al., 1988; Hathaway & McKinley, 1951; Sobell & Sobell, 1992). In a field uniquely interested in internal subjective state, an emphasis on reported experience is unsurprising, as self-reports can at times reflect the most appropriate and direct means of assessment (G. W. Allport, 1961; Garcia & Gustavson, 1997; Lilienfeld & Fowler, 2006). Yet reliance on survey-based assessment extends widely within clinical psychological science, encompassing constructs ranging from behavior, to physiology, to automatic cognitive processes, to events in the distant past—domains in which self-reports are susceptible to both random and systematic forms of error (Baldwin et al., 2019; Nisbett & Wilson, 1977; Schwarz, 1999; Widom & Shepard, 1996). In a field where such error can translate into the prolongation or even exacerbation of suffering among individuals in distress, matching construct to method of measurement in a manner that minimizes bias emerges as a central concern.

In the current article we bring into conversation the fields of computer science and clinical psychology with the aim of introducing an expanding measurement toolkit for behavioral researchers of psychopathology. We review not only the strengths and utilities of questionnaire and interview-based measures, but also the limitations of self-reports when employed for use as a primary method of assessment across construct domains. A class of representational machine learning models is introduced, and specific applications of this model class are illustrated and

reviewed as these might facilitate the measurement of key constructs of interest for psychosocial researchers. Specific concerns are noted relevant to the pervasiveness of self-report assessment, particularly the potential for false-positive effects from "common methods bias." Finally, we weigh strengths and limitations of automated,[1] self-report, and other measurement approaches as they relate to the broader goal of fostering a replicable and robust field of clinical psychological science.

Self-Reports in Clinical Psychological Science

Clinical psychology is a science that is uniquely concerned with the experience of the individual (Garcia & Gustavson, 1997). Participant self-reports offer notable advantages for studying psychological constructs, including their informational richness, causal force, and sheer practicality (Paulhus & Vazire, 2007). This method of measurement has the potential to tap into the expansive body of knowledge humans carry surrounding themselves and their own functioning, a knowledge base that can extend not only to external characteristics but also internal thoughts and feelings (Vazire, 2010). In the study of psychopathology more specifically, self-reports arguably hold particular value in that they have the potential to capture the form of suffering that lies at the core of what we seek to understand. As clinical researchers, we are interested not only in understanding reality, as it exists in truth, but also in understanding each participant's unique experience of that reality (Frances & Widiger, 2012; Garcia & Gustavson, 1997; Lilienfeld & Fowler, 2006). Is there a truly "objective" measure of whether life is worth living? Isn't it arguably the individual's own subjective sense of self-esteem, hopelessness, and/or wellbeing that concerns us singularly? An interview or questionnaire-based measure

[1] In the current article, we use the term "automated" as a shorthand to refer to measures developed on the basis of machine learning techniques.

might not capture such experiential processes in a manner free of error. Yet arguably, when it comes to the measurement of such intrinsically experiential constructs, no other measure is likely to do better.

Reflecting the field's commitment to subjective assessment, self-report measures have had a prominent place within clinical psychology. With a few exceptions (Hathaway & McKinley, 1951), such measures primarily take the form of direct, closed-ended questionnaire or interview-type items, requiring the explicit reporting of symptoms and experiences by patients and, further, the subsequent patient or interviewer-mediated translation of these experiences according to a pre-defined scale or fixed set of response options (American Psychiatric Association, 2013; Beck et al., 1988; Sobell & Sobell, 1992). For the purposes of this article, we conducted a review aimed at quantifying more precisely the prevalence of self-reports in clinical psychological science, systematically coding measure type employed in empirical research published in three highly ranked clinical psychology outlets during the years 2021-2024: *Clinical Psychological Science*, *Journal of Consulting and Clinical Psychology*, and *Journal of Psychopathology and Clinical Science* (preregistration: https://osf.io/cvnk8). Of 546 studies included in this review, 87% (k=477) tested primary study aims using at least one closed-ended questionnaire or interview-style self-report. At the level of the measure, 71% (n=969) of measure codes assigned in our review reflected closed-ended reports. Regarding the assessment target for these self-report measures, results indicated that self-reports were employed for assessing a wide variety of domains of human experience, from behavior to past events to physiological sensations (see Figure 1), reflecting a broad commitment to self-reports for measuring phenomena well beyond the inherently subjective.

Importantly, while acknowledging their strengths, behavioral researchers have long raised concerns surrounding the potential fallibility of self-reports (F. H. Allport, 1927; Baumeister et al., 2007; Nisbett & Wilson, 1977; Schwarz, 1999). Congruence between contemporaneous and retrospective measures of major life events is markedly low (Baldwin et al., 2019; Widom & Shepard, 1996), participants' self-assessments of their own skills/performance consistently yield inflated results (Dunning et al., 2004), and, regarding internal experience, the automaticity of many domains of human cognition can limit participant ability to accurately report (Nisbett & Wilson, 1977). Misreporting is especially widespread for items that touch on socially-sensitive topics: to offer just two examples, studies estimate that only 52% of abortions (Fu et al., 1998) and 30-70% of drug use (Hilario et al., 2015; see Tourangeau & Yan, 2007) are reported in survey measures.

Regarding the study of psychopathology more specifically, characteristics of clinical contexts, populations, and the assessments themselves have the potential to exacerbate self-report bias. Regarding clinical *assessments*, these measures are particularly likely to involve socially sensitive topics—including items assessing self-harm, trauma, and substance use—and so are especially vulnerable to misreporting (Kelly, 1998; Tourangeau & Yan, 2007; Yeater et al., 2012). Beyond the assessments themselves, clinical *contexts* can be linked with the perception of secondary gains surrounding measurement outcomes, leading to substantial concern surrounding under- or over-reporting on clinical measures (Berry et al., 1991; Rees et al., 1998; Schretlen, 1988). Finally, clinical *populations* comprise individuals exhibiting diverse levels of functioning, with specific forms of psychopathology being defined in part or in whole by skill deficits that would substantially impact self-reports. These include deficits along domains core to accurate reporting, including impaired verbal expression, impaired attention, lack of insight, and

propensity towards deceit (Ditzer et al., 2023; Lilienfeld & Fowler, 2006; Williams et al., 2008). In sum, while self-reports are widely employed in clinical psychology, clinical psychology is arguably a discipline that should be uniquely concerned with the biases associated therewith.

Given the well-documented forms of bias linked with subjective assessment, why has clinical psychology so often turned to self-reports, including to measure constructs well beyond those that might be considered inherently experiential? One possible answer to this question is the lack of perceived alternatives available within the field. While biological researchers have recourse to brain scans and biological assays, objective assessment options available to psychosocial researchers have historically been relatively sparse. Regarding the measurement of behavior, direct assessment has often relied on the application of human-based coding systems (Bakeman & Gottman, 1997), an endeavor that can consume hundreds and sometimes thousands of hours for a single participant sample (e.g., Ekman et al., 2002). Regarding the measurement of context, no established systems currently exist for the objective assessment of participants' everyday environments (Ariss et al., in press; Rauthmann et al., 2014). Studies of personality/traits have sometimes sought to address issues of subjective bias via informant reports, a practice that can yield valuable information yet is associated with its own form of subjective bias (McDonald, 2008). Finally, while physiological data can often be valuable, the output from physiological sensors is often highly complex and thus links between such sensor output and underlying psychological processes have been difficult to disentangle (e.g., the relationship between skin conductance and stress; Dawson et al., 2007). In sum, in many cases, researchers in clinical psychological science have turned to self-reports because it has been challenging to conduct adequately powered, affordable, and interpretable research using other methods. Yet, with the advent of new machine learning model types, automated measurement

options available to behavioral researchers are likely to rapidly expand, resulting in a substantially more diverse toolkit of methods for use in psychological science. In this article, co-authored by a behavioral researcher with a doctorate in clinical psychology (CEF) and a machine learning researcher with a doctorate in computer science (NB), we offer a cross-disciplinary perspective on these methodological developments.

Machine Learning

Machine learning is a branch of data science in which algorithms learn to make predictions directly from patterns observed in data without requiring explicit programming. In contrast to conventional statistical approaches, wherein the relationship between variables is first theorized and then imposed on the data by the researcher (e.g., prespecified interactions between predictors, linear relationship between predictor and outcome), machine learning models learn relationships directly from observed patterns within the data itself. Machine learning models in fact undergo an explicit "training" stage, wherein a model is exposed to observations ("training data") and subsequently learns associations in the context of this data. Machine learning models tend to be exceptionally flexible and well-suited to capturing complex associations, including non-linear relationships and interactions among multiple predictors, especially when the shape of associations is unknown or too complex to be prespecified. As such, machine learning models can have particular utility when applied to data that might be characterized as unstructured or *organic* in nature (e.g., text from webpages, speech in audio recordings, or image data), a data-type that often yields highly complex datasets in which the number of features (i.e., predictors) is large and patterns are best recognized through an understanding of multiple variables and variable interactions (e.g., how pixels work together to represent an object in an image; Adjerid & Kelley, 2018; Tay et al., 2022).

Many in psychology are familiar with the enthusiasm (even "hype") that surrounds the field of machine learning. Yet less well understood is its larger history, including roots spanning back over eighty years (McCulloch & Pitts, 1943) and, particularly relevant, the rapid rate of progress that has distinguished the years since 2012 (Holzinger et al., 2018; Maclure, 2020). The field has in fact seen epochs characterized by highly divergent levels of success and public enthusiasm, encompassing periods of intense acceleration and also lackluster progress (Norvig & Russell, 2021). According to many, progress in machine learning application had reached a plateau in the years prior to 2012, with automated systems failing to reach human performance levels across a variety of key task domains. However, within the past decade a specific sub-class of representational machine learning models known as "deep learning" experienced a resurgence, ultimately heralding an era of unprecedented progress for artificial intelligence (AI)[2] application (LeCun et al., 2015).

Deep learning methods were proposed by LeCun and colleagues in the 1990s, although it took nearly two decades and the introduction of fast graphical processing units (GPUs) for the promise of these methods to be realized (LeCun & Bengio, 1995; see Norvig & Russell, 2021). Deep learning methods are characterized by multiple hierarchically organized structural levels, each of which constitutes a "layer" of representation for pattern decomposition (LeCun et al., 2015). Whereas traditional machine learning models required sophisticated "feature engineering" of raw data into high-level, semantically meaningful units by the programmer prior to analysis, deep learning methods were unique in that they could automatically detect patterns in low-level raw data itself, leading to programs capable of detecting patterns at a complexity level not

[2] Artificial intelligence (AI) is used as an umbrella term to refer to a broad field of study that aims to build machines capable of achieving cognitive abilities historically characteristic of humans. Machine learning reflects one dominant sub-domain within AI.

previously possible. In 2009 deep learning methods achieved high accuracy on a standard speech recognition task (Grove et al., 2000; Mohamed et al., 2011) and, several years later, in the context of the widely publicized ImageNet competition, a deep learning-based program nearly halved error rates for image recognition (Krizhevsky et al., 2012). This period also gave rise to developments in methods for controlling model complexity, so increasing the likelihood that results of complex models would generalize to new datasets (e.g., regularization; Tian & Zhang, 2022). Reflecting these developments, the years since 2012 have sometimes been referred to as the "AI Spring," marking a period of notable growth for the field of machine learning: AI publications increased 20-fold between 2010-2019, error rates for image-based object detection improved from 28% in 2010 to 2% in 2017, and language-based question answering accuracy had exceeded that of humans by 2019 (Norvig & Russell, 2021; see also aiindex.stanford.edu).

Clinical psychological researchers were quick to recognize the potential of machine learning methods for identifying those at risk for negative mental health outcomes (Dwyer et al., 2018; Walsh et al., 2017; Yarkoni & Westfall, 2017) and for matching patients to clinical interventions (Delgadillo & Gonzalez Salas Duhne, 2020; Webb et al., 2020). The question of machine-based prediction is not one that is new within clinical psychological research as, for decades, theorists and researchers have pondered the relative utility of "mechanical" (e.g., algorithmic, numbers-based/automatic) vs. clinical (e.g., human/subjective) prediction of mental health (Grove et al., 2000; Meehl, 1954). In the present day, machine learning methods have gained attention through their application in studies seeking to predict clinically relevant outcomes using standard survey data types (Belsher et al., 2019; Kuo et al., 2024; Passos et al., 2016; Vieira et al., 2022; Walsh et al., 2017; Yarkoni & Westfall, 2017), as well as, in some

instances, their lackluster performance and/or problematic application (Jacobucci, Littlefield, et al., 2021; Christodoulou, Ma et al 2019; Jacobucci & Grimm, 2020).

To this point, the most widely cited machine learning applications in clinical psychology have often centered around the use of machine learning as an analytic/predictive tool applied to a standard set of pre-defined measures—e.g., predicting suicide risk on the basis of survey responses. However, some of the most powerful machine learning model types currently available to researchers require massive training datasets rich in both instances (e.g., observations; Adjerid & Kelley, 2018; Bahri et al., 2024; Hestness et al., 2017) and also reliably-measured predictors (Jacobucci & Grimm, 2020), a group of attributes unlikely to characterize even the largest of survey studies. As such, the application domain for which machine learning is best suited arguably diverges from that for which it has to date received the most attention in clinical psychological science: not in transforming the manner in which we analyze constructs of interest, but rather by transforming the manner in which we measure these constructs to begin with. Put differently, in light of recent advances in deep learning applications, machine learning may prove most revolutionary for psychological science not as a novel analytic technique applied to existing datasets, but rather as a method for re-imagining these datasets from the ground up.

With the advent of deep learning and enhanced regularization, machine learning models have demonstrated strengths in three key application areas with implications for measurement in clinical psychological science: 1-*speech* transcription and *language* understanding/analysis; 2-analysis of *images*; 3-recognition of patterns within *complex time-series data*. Below, focusing on these three areas of application, we review impactful automated measure development relevant to behavioral scientists. In the context of this review, we focus on specific content

domains high in observability/externality (Vazire, 2010). Specifically, our review focuses on measures of social processes, behavior, and situation/context. However, note that this review is not intended to be comprehensive—in light of the pace of progress in the field, such a reference would be dated immediately upon publication—but is rather aimed at offering a broad overview of key advances relevant for psychosocial/behavioral researchers. A summary of core domains of progress, together with associated publications and tools, is provided in Table 1, and a narrative summary of selected research is provided below.

Regarding *language*, within the past several years, machine learning algorithms have emerged capable of understanding speech at high accuracy, with novel programs avoiding common transcription errors observed in earlier iterations and now also capable of identifying changes in speakers in social exchange (PoornaPushkala et al., 2022). Advances in speech recognition have come hand-in-hand with advances in language analysis, with recent automated tools capable of the rapid analysis and categorization of large quantities of unstructured written text, a data type that previously posed formidable analytic challenges for human coders. These developments have yielded a subfield of machine learning application often referred to as natural language processing (NLP). NLP has been employed for the analysis of various features of language, from broad emotional tone (e.g., positive, negative, neutral) to the content and structure of speech (e.g., pronoun usage; beliefs on a specific topic). These advances in both spoken-language transcription and language analysis have packed a powerful one-two punch for the examination of social processes in clinical research, having been employed successfully for the assessment of expressed empathy (Xiao et al., 2015), alliance (Goldberg et al., 2020), stereotyping (Nicolas et al., 2022), and therapist fidelity and effectiveness (Althoff et al., 2016; Atkins et al., 2014). In even more recent developments, large language models (e.g., ChatGPT)

have been leveraged to analyze the content and form of social communication. Trained on text corpora of unprecedented size, large language models have been successful in detecting characteristics of social media posts, such as sentiment, expressed emotion, and offensiveness with high accuracy levels (Rathje et al., 2024).

In the domain of *image analysis*, machine learning models have been developed with accuracy levels that exceed that of humans for the recognition of elements within videos and photographs (e.g., object and face recognition; LeCun et al., 2015; Norvig & Russell, 2021). Such algorithms can be used for the swift and accurate detection of a variety of elements of the physical environment, from the level of overall lighting (e.g., dim, bright), to the position of objects in the physical surroundings, to the nature of activities being performed, to the size and composition of social groups. Leveraging these advances, tools for image analysis have been applied to photographs and videos taken by participants in everyday life to identify context-level drivers of psychological disorder and disorder symptomatology, including setting-level predictors of heavy drinking (Ariss et al., in press; see also Redmon, 2016). Also in the realm of image analysis, machine learning models have been employed for the examination of a range of social processes, including socially expressed emotions (e.g., smiles) and affiliative behaviors (e.g., physical proximity). Regarding the latter of these, algorithms have been developed with high accuracy for recognizing human body position and posture from images (Cao et al., 2019). Such posture detection algorithms have been employed for a variety of purposes, including for the analysis of factors impacting changes in physical proximity (i.e., social distance) among participants in social exchange from video recordings (Gurrieri et al., 2021).

Finally, specific types of psychophysiological data can manifest as *dense time series*, featuring individual series spanning thousands or even millions of observations and characterized

by complex temporal fluctuations (i.e., Big Data in the "temporal" sense; Adjerid & Kelley, 2018). Within such data, parsing signal from noise—recognizing meaningful temporal fluctuations—has represented a challenge for conventional statistical methods relying on theory-derived models (Dawson et al., 2007; Fairbairn & Bosch, 2021). Automated methods have been developed that can quickly extract hundreds of over-time features from such time-series data (e.g., slope, number of peaks, cyclical elements; Christ et al., 2018). These extracted features can be entered into machine learning algorithms to yield models capable of accounting for complex interactions between over-time processes in a manner not possible using conventional analytic approaches (Fairbairn & Bosch, 2021). Where larger datasets are available for model training, deep learning methods reduce this two-step process into a single stage, directly learning features from over-time patterns in the dataset itself and so being capable of more flexibly extracting patterns uniquely predictive within the dataset vs. constrained by a pre-defined set of features. Machine learning has been employed to analyze and extract a range of behavioral indicators from such time-series data, including everyday activities and eating behaviors from accelerometer data (Ordóñez & Roggen, 2016; Thomaz et al., 2015), sleep duration and sleep stage from multi-sensor data (Boe et al., 2019), and alcohol use episodes from sweat-based sensors (Didier et al., 2023; Fairbairn et al., 2020, 2025; Fairbairn & Bosch, 2021).

In sum, progress in the field of machine learning over the past decade has been rapid, giving rise to new methods well suited to the recognition of complex patterns within unstructured data. These novel methods have limitations, including the potential for bias and lack of transparency, with accuracy levels varying across applications (see Weighing Risks and Benefits section below). Machine learning has nonetheless yielded automated tools for assessing a range of constructs in clinical psychological science, including through the analysis of speech,

photographs, video, and physiological data. However, researchers in clinical psychology have tended so far to approach cautiously, with uptake rates for many of these methods not keeping pace with progress even as these tools exceed the accuracy of human coders. We next consider these novel tools in the context of the current measurement landscape of clinical research, with specific consideration given to the type of error linked with various measurement options and the potential for false positive effects associated therewith.

Selecting a Method of Measurement: Common Methods Bias and False Positive Effects

As the field of machine learning was transformed by applications of deep learning, psychological science meanwhile underwent a transformation of its own. Often referred to as the replicability movement, the past several decades have witnessed a sharp rise in concern surrounding the prevalence of false-positive findings in psychological science (Open Science Collaboration, 2015) paired with awareness of the prevalence of research practices within the discipline that invite such outcomes (John et al., 2012; Simmons et al., 2011). Engagement with the replicability movement within clinical psychological science was somewhat delayed from its initial inception in the field at large (Tackett et al., 2017). Nonetheless, and driven in part by concern surrounding low reliability of clinical diagnoses (Frances & Widiger, 2012), awareness surrounding harm linked with false positive effects has increased among researchers of psychopathology (Miller et al., 2025; Tackett et al., 2017, 2019). Highlighting the importance of both *direct* and also *conceptual* replications, the movement brought to the fore potential reputational damage and also lost time and resources linked with false-positive findings (Francis, 2014; Gelman & Loken, 2013; Simmons et al., 2011). In a field where such wasted resources might translate to the prolongation of suffering among individuals in need, the stakes of such false positive effects in clinical psychology might be considered especially high. As such, clinical

scientists have advocated for vigilance surrounding research methods that might systematically inflate associations, including methods employed for the initial measurement of constructs.

While concern has risen among clinical psychological researchers surrounding false-positive findings, discussion has been ongoing within various psychological sub-disciplines concerning potential inflationary effects associated with self-reports (Bagozzi & Yi, 1991; J. R. Edwards, 2008; Francis, 2014; Podsakoff et al., 2012). The process of responding to a self-report measure is a complex social-communicative act that relies on the functioning of multiple discrete stages of cognitive processing, including question interpretation, memory retrieval, judgment formation, and response selection (Nisbett & Wilson, 1977; Tourangeau, 1984). In moving through these stages, participants engage with self-reports not simply as a means of information transfer, but rather as a process for social communication, influenced by all of the interpretative challenges involved therein (Schwarz, 2007). Such a process requires participants to not only interpret the literal meaning of the question but to make inferences surrounding the questioner's intent, including integration of a given question with other questions and with additional contextual information offered within the assessment context (Schwarz, 1999). As such, self-reports can be influenced by a range of factors beyond the content targeted by the question itself, from participants' state or trait-based self-reflective capacity (Andrews & Herzog, 1986; Knäuper et al., 1997), to their response style (Couch & Keniston, 1960), to the order and phrasing of questions (Schwarz et al., 1991), to the letterhead a given survey was printed on (Norenzayan & Schwarz, 1999), to architectural and structural features of the setting in which reporting takes place (Chaikin et al., 1976). Given cognitive and communicative challenges evident within many clinical populations, paired with the sensitive and (sometimes) cognitively taxing nature of some clinical assessment contexts, the potential for influence by such extraneous

factors can run particularly high in the study of psychopathology. Importantly, the type of error linked with such measures might be *systematic* in nature, leading to significant associations linked not only with underlying constructs but also with the manner in which these constructs were measured (Campbell & Fiske, 1959).

Variability in measured constructs can be divided into three discrete elements: variability in the true underlying construct, random error variance, and systematic error variance. In the examination of relationships between multiple variables, significant correlations can emerge attributable to not only shared variation in the underlying constructs, but also to shared systematic forms of error linked with the method of measurement (see Figure 2). This type of error co-variance, often referred to as *shared methods bias* or *common methods bias*, has long been documented across various psychological sub-disciplines as a major factor influencing results of behavioral research (Campbell & Fiske, 1959; J. R. Edwards, 2008; Podsakoff et al., 2003). As observed by Campbell and Fiske (1959), measures that use the same method to measure different constructs will often produce more similar results than do measures that assess the same construct using different techniques. In light of the ubiquity of questionnaire and interview methods, and the specific systematic forms of error associated therewith, concern has been particularly high surrounding potential effects of common methods bias as these relate to self-reports (Bodner, 2006; Podsakoff et al., 2012). In hypothetical terms, common methods bias might either magnify (inflate) or rather diminish observed effects, depending on the size of the true underlying correlation in relation to the size of shared systematic forms of error. However, when examined with respect to studies employing self-report methods within psychology, observed effects of common methods bias have been overwhelmingly inflationary. A mega-analysis of 233 bivariate correlations from 59 meta-analyses estimated common methods

inflation rates ranging from 120-160%, depending on the specific domain of methods overlap (shared reporter vs. shared time point; Podsakoff et al., 2024).

Clinical psychology is notable as a sub-discipline in which discussion of common methods effects has been relatively scarce, yet the potential for bias linked with these effects runs high. Several factors point to a potentially outsized influence of common methods bias in clinical psychology, including the ubiquity of self-report assessments as well as the amplification of specific drivers of common methods effects within clinical assessment contexts. Regarding the latter of these, factors that have been established or theorized as potential contributors to common methods bias include participants' affective states (Watson et al., 1987), social desirability concerns (A. L. Edwards, 1957), implicit theories (Eden & Leviatan, 1975; Nisbett & Wilson, 1977), consistency motifs (Podsakoff & Organ, 1986; Schmitt, 1994), and response style (Baumgartner & Steenkamp, 2001). In light of the (often) sensitive nature of clinical assessment and the characteristics and deficits observed in clinical populations, each of these elements has the potential for amplification in research on psychopathology. Specifically, pronounced state- or trait-level affect (e.g., depression) has long been proposed as a source for common methods effects, often leading individuals to respond to all self-report items in an affect-consistent manner, including items extending well beyond mood to include current events, behaviors, and attitudes (Thoresen et al., 2003; Watson et al., 1987). Further, self-report items in clinical investigations are especially likely to trigger social desirability concern (e.g., sensitive items; Tourangeau & Yan, 2007), potentially leading individuals to respond to items in a manner most likely to cast themselves in a positive light, irrespective of item content (A. L. Edwards, 1957). Third, in light of the perceived urgency and centrality of mental health problems among many clinical samples, studies of psychopathology may be especially likely to concern topics

surrounding which participants have their own fully crystallized lay theories (i.e., my drinking and my marital stress are related), leading to self-report studies that reflect primarily the prevalence of these implicit theories rather than true underlying relationships (Schleider et al., 2015). In addition, in clinical assessment contexts, cognitive resources available for processing individual items within self-reports may at times be diminished (Bruijnen et al., 2019; Robinson et al., 2006), leading participants to rely more heavily on consistency motifs and/or a specific response style (Podsakoff & Organ, 1986; Schmitt, 1994). Finally, beyond these established contributors to common methods bias, the (real or perceived) instrumental consequences associated with clinical assessment contexts might yield drivers of common methods bias unique to the field, including *malingering* as well as (related but less widely discussed) *distress broadcasting* (Berry et al., 1991; Rees et al., 1998; Schretlen, 1988), each of which might lead participants to report symptoms broadly across domains in a manner that artificially inflates associations yet fails to capture true underlying links.

In order to address the prevalence of common methods effects in published research in clinical psychology, we return to the results of our journal review. As noted previously, self-reports were the most common measure type in reviewed studies, outpacing the highest-frequency alternative measure type by a factor of nearly six (see Figure 1). Of note, and most pertinent to the question of common methods effects, of 386 non-experimental studies coded, 209 (54.1%) tested primary hypotheses via an analysis that examined relationships between two closed-ended self-report assessments. To assess the extent to which this potential source of bias had received attention, we conducted a search for common methods effects and related terms[3]

[3] To capture potential variability in terminology employed we searched for the following terms: "common method* bias" OR "common method* variance" OR "common method* effects" OR "shared method* bias" OR "shared method* variance" OR "shared method* effects"

within these same three journals for the same span of years. This search indicated that in a field where over half of articles rely on a single method type for primary hypothesis tests, none mentioned shared methods bias in the article abstract, and less than 5% (N=20) mentioned these terms anywhere in the full text. In sum, attention to common methods effects in clinical psychological science appears to be low, whereas the potential for bias due to these effects is high. Such a measurement landscape points to the utility of alternative measurement types beyond self-reports. While such options in the past might have been sparse, automated tools may ultimately represent a game changer for expanding measurement options in clinical psychological science.

Recommendations for Researchers: Weighing Risks and Benefits

In this paper, we urge researchers to re-imagine the principal place of machine learning methods in clinical psychological research: from a tool useful primarily at the stage of data analysis to one useful also for the initial measurement of constructs. At the same time, it is important to consider these automated measurement tools in the context of their limitations. First, machine learning models often trade increased accuracy for reduced transparency. Although it is possible to understand machine learning output in a manner that breaks down which variables in the model were most impactful in determining predictions, such explanatory models typically leave many questions unaddressed, including the exact means by which a variable influenced the outcome and the nature of the relationship (Rudin, 2019). As such, machine learning models are poorly suited for use cases where knowledge of the processes through which a measurement was made is essential (e.g., insurance coverage eligibility, legal proceedings; Vollmer, 2020). Further, accuracy rates for machine learning models vary widely

across areas of application, often dependent on the size, representativeness, and measurement type reflected in the dataset available for model training. These models are typically trained and curated on datasets created by humans, and as such often contain biases and representational imbalances, including imbalances along contextual and also person-level factors (Belitz et al., 2021; Tay et al., 2022). Compared with conventional ("top-down") statistical and measurement approaches, data-driven models such as machine learning are particularly sensitive to characteristics of training data and, as such, final models will reflect biases and errors contained in datasets used in model training. Finally, machine learning models have the potential to yield systematic forms of error in cases where they are misapplied (Jacobucci & Grimm, 2020). Many psychologists lack training in the application of machine learning methods, and thus collaboration with data scientists or psychologists who have specifically developed skills in applying these methods may be required to protect against such errors.

Regarding the vulnerability of machine learning to specific inflationary effects described above, a range of measures have been employed as the label (i.e., target or ground truth) used for training and validating machine learning models, including closed ended self-reports (see Table 1). Each of these measure types is linked with its own potential measurement errors. To some extent, it seems apparent that machine learning models may recreate the same measurement errors present in the labels used during training, since models are indeed supposed to learn to replicate those labels. However, note that "leakage" of self-report bias into the newly trained automated measure requires that a given source of bias is discernable not only in the self-report measure, but also within the predictors (i.e., features) in the machine learning model. In practice, measurement biases in the labels are due to numerous factors (e.g., various cognitive biases in self-reports), including some that are not present in the features (e.g., unmeasured contextual

factors). As such, effects of common methods bias are likely to be substantially mitigated if not wholly erased within the context of such automated assessment tools, even those trained on a self-report as ground truth.[4]

The question of whether machine learning methods are relevant to any specific research project is a complex one, requiring consideration of details of both the study itself as well as the specific automated tool in question. A framework for approaching this process is provided in Table 2. In the event a study involves examining relationships between two or more measured variables, and one of these variables is already assessed using closed ended self-reports, then our recommendation is that researchers consider expanding beyond survey methods for measuring the remaining variables. The need for measurement diversification becomes particularly urgent where the experimental context features elements likely to exacerbate shared methods effects (e.g., sensitive questions, evaluation/stigma concern, limited cognitive capacity, positive or negative affect) and/or features impacting the validity of self-reports (e.g., perception of secondary gains or punishments, studies targeting primarily automatic processes, populations with low verbal/expressive ability).

Regarding measures that expand beyond self-reports, a broad range of behavioral, psychophysiological, lexical, and other methods have long been established capable of assessing constructs of interest in psychological science (Bakeman & Gottman, 1997; Dawson et al., 2007;

[4] A machine-learning based measure integrated into a study that otherwise employs survey-based assessments may mitigate common methods bias, even in cases where the automated measure is validated on the basis of a closed-ended self-report. Common methods bias can be driven by a variety of factors, including shared *reporter/person-level* effects as well as shared *context* for assessment. In most cases, automated measures of the type reviewed here (e.g., Table 1) will eliminate the latter context-driven sources of these biases, while substantially reducing the former. Regarding the former of these effects, the potential influence of shared reporter effects would depend on the medium for automated assessment and the extent to which potential person-level influences on this medium are likely to covary with influences on survey self-reports. However, assuming the medium of automated assessment is other than that of self-report, the effect of person-level influences on common methods bias for automated measures is likely to be substantially reduced compared to the direct integration of a survey measure.

Ekman et al., 2002; Holmqvist et al., 2011). The validity of many such measures was established well before the advent of automated tools for assessment. Regarding the specific choice to apply machine learning methods for measurement above-and-beyond these alternative tools, we recommend researchers consider two essential questions in approaching this decision (Table 2): 1) Is **transparency** surrounding the process of variable creation critical for the specific application in question? 2) How **accurate** are current machine learning models in measuring the identified constructs in representative data? Regarding the former of these questions, while machine learning-based measures can be replicated via the re-running of code, more complex model types cannot easily be explained or decomposed into component parts. In basic clinical research the issue of transparency in measure creation is unlikely to represent a major concern. However, in specific domains of applied clinical science, transparency issues may come to the fore—e.g., researchers seeking an indicator of recidivism that might be applied directly within the courts (Rudin, 2019). Regarding the latter of these questions, it is important to consider accuracy metrics in the context of the data employed in algorithm training. While some machine learning models can be trained on smaller datasets, more complex tasks require more powerful model types (e.g., deep learning) and thus exceptionally large datasets for model training. As such, large datasets typically yield higher accuracy metrics. Of note, while the size of the dataset is often linked with the generalizability of the sample to novel measurement scenarios, this is not always the case. For example, for facial recognition algorithms, while larger datasets are more likely to feature more faces of different races, many larger early training datasets still involved a remarkably low proportion of non-White individuals (Tay et al., 2022). Therefore, not only the size of the dataset but also its diversity and generalizability to the specific use case is relevant for interpreting accuracy metrics. Finally, many automated tools are developing at a rapid rate,

particularly in the broad area of deep learning. Therefore, an accuracy metric at a given moment in time is exactly that—accuracy at that moment in time. In the event accuracy metrics fail to meet expectations for a given measure type, it may be worth revisiting the literature in the (near) future to establish the extent to which such performance metrics improve.

Other considerations may be relevant to the question of whether machine learning measurement tools warrant consideration. Depending on the research environment and domain of measurement application, machine learning methods might offer specific advantages beyond other available assessment tools. In some cases, machine learning tools might offer enhanced accuracy. For example, accuracy for automated object and image-analysis now exceeds that of the average human coder (Norvig & Russell, 2021). Advantages might also include that of consistency, as automated methods are not reliant on the strength or training of a specific collection of individual human coders on a research team, and thus have the potential to "smooth over" rater idiosyncrasies and yield measures with superior psychometric properties (Hickman et al., 2025). Finally, machine learning can often offer pragmatic advantages, facilitating free or (comparatively) low-cost application, completing tasks in a matter of seconds that might take a team of human coders months or even years to complete.

Finally, note that, while interpretation of automated models may in some cases be facilitated when these algorithms are applied to the measurement of characteristics considered "observable" in nature (e.g., Table 1), these methods have nonetheless been applied to the analysis of a wide range of features beyond the inherently extrinsic. A variety of data sources from behavior to language to psychophysiology have long been employed as external markers of internal psychological characteristics, including as measures of personality, cognition, emotion and psychiatric symptomatology. These have included, among many others, eye tracking

measures for the assessment of attentional focus (Ariss et al., 2023; Holmqvist et al., 2011), observer ratings of personality (McDonald, 2008), and facial and prosodic measures of emotion (Bakeman & Gottman, 1997; Caumiant et al., 2025; Ekman et al., 2002; Fairbairn et al., 2015). In such investigations, wherein inherently internal characteristics are inferred via extrinsic/observable cues, the machine learning model is considered as acting in a role analogous to the informant in studies employing observer-reports of personality, inhabiting a valuable yet singular (often context-dependent) lens for viewing the subjects' characteristics that reflects one among several potential perspectives (Brunswik, 1956; Tay et al., 2020). Of note, where constructs of interest might be considered inherently experiential, and thus reports of internal experience are of central importance, natural language processing approaches can be applied to unstructured text derived from participants' written entries or transcribed interview responses, so circumventing issues of classification linked with traditional forced-choice self-report items. As such, automated analysis of both written and spoken text have been used to derive indicators of a range of internally felt subjective experiences. A selective review of machine learning applications for the measurement of personality, emotion, cognition, and psychiatric symptoms/diagnoses is provided in Supplemental Table S1. Finally, a compilation of useful resources for psychologists starting out in the domain of machine learning measurement, including reading recommendations and software packages/tools, is presented in Table S2.

In sum, the question of whether to integrate novel automated methods for measurement will always be a complex one. Consideration should be given to the specific construct under investigation, potential bias (if any) linked with self-reports, availability of alternative, non-automated measurement types, and the phase of development of the automated tool. In such circumstances, the draw to favor a flawed yet familiar survey measure can be strong. Yet the

specific form of error linked with assessment type is a factor worthy of consideration—whereas error variance linked with automated assessment may in certain cases be sizable, *systematic* forms of error can result from over-reliance on self-reports. Thus, in a field in which, among top journals, 87% of tests of primary study aims rely on closed-ended self-reports, and concern surrounding research practices likely to yield false-positive findings runs high, measurement diversity should rank as a priority for the field.

Conclusion

With each major advent in scientific measurement technology comes the siren song of hope (Borup et al., 2006; Maclure, 2020). From fMRI to EMA to eye-tracking, the arrival of new methods has consistently inspired exaggerated arguments surrounding their utility for psychological research—their ability to offer the long-coveted "window into the soul." The draw of such inflated visions can be strong, especially within behavioral research, where "soft" constructs are often of critical interest yet have long posed challenges for objective assessment (Meehl, 1978). Here, in considering the potential for automated methods of measurement, we offer no rose-colored view of the future, but rather a vision of measures now. We argue that error from machine learning models is likely to be substantial in some cases, and the exact size and nature of this error for all applications is currently unknown. Yet, in a research landscape dominated overwhelmingly by survey methods vulnerable to systematic forms of bias, these new tools may offer clinical scientists a relatively unglamorous, but nonetheless precious, "lessor of two evils." By offering a source of error more likely to serve as *noise* as compared with *confound*, we posit that the past decade of development in automated assessment techniques could ultimately prove a point of inflection for measurement in clinical psychological research.

In sum, clinical psychology has long been characterized by a focus on self-reports. This emphasis may spring in part from the nature of clinical constructs, which can manifest as inherently subjective in nature. In other cases, reliance on self-reports springs from a lack of viable alternatives for behavioral researchers. Here, we review advances at the intersection of computer science and clinical psychology, where recent developments bid fair to expand measurement tools for researchers of psychopathology, offering paths towards a methodologically varied, replicable, and reputationally sound field of clinical psychological science.

References

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist*, *73*(7), 899–917.

Allport, F. H. (1927). Self-evaluation: A problem in personal development. *Mental Hygiene*, *11*, 570–583.

Allport, G. W. (1961). *Pattern and growth in personality.* Holt, Reinhart & Winston. https://psycnet.apa.org/record/1962-04728-000

Althoff, T., Clark, K., & Leskovec, J. (2016). Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, *4*, 463–476.

American Psychiatric Association (Ed.). (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed). American Psychiatric Association.

Andrews, F. M., & Herzog, A. R. (1986). The quality of survey data as related to age of respondent. *Journal of the American Statistical Association*, *81*(394), 403–410. https://doi.org/10.1080/01621459.1986.10478284

Ariss, T., Caumiant, E. P., Fairbairn, C. E., Kang, D., Bosch, N., & Morris, J. K. (in press). Exploring associations between drinking contexts and alcohol consumption: An analysis of photographs. *Journal of Psychopathology and Clinical Science*.

Ariss, T., Fairbairn, C. E., Sayette, M. A., Velia, B. A., Berenbaum, H., & Brown-Schmidt, S. (2023). Where to look? Alcohol, affect, and gaze behavior during a virtual social interaction. *Clinical Psychological Science*.

Atkins, D. C., Steyvers, M., Imel, Z. E., & Smyth, P. (2014). Scaling up the evaluation of psychotherapy: Evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, *9*(1), 49. https://doi.org/10.1186/1748-5908-9-49

Bae, S., Chung, T., Ferreira, D., Dey, A. K., & Suffoletto, B. (2018). Mobile phone sensors and supervised machine learning to identify alcohol use events in young adults: Implications for just-in-time adaptive interventions. *Addictive Behaviors*, *83*, 42–47.

Bagozzi, R. P., & Yi, Y. (1991). Multitrait-multimethod matrices in consumer research. *Journal of Consumer Research*, *17*(4), 426–439.

Bahri, Y., Dyer, E., Kaplan, J., Lee, J., & Sharma, U. (2024). Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, *121*(27), e2311878121. https://doi.org/10.1073/pnas.2311878121

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). Cambridge University Press.

Baldwin, J. R., Reuben, A., Newbury, J. B., & Danese, A. (2019). Agreement between prospective and retrospective measures of childhood maltreatment: A systematic review and meta-analysis. *JAMA Psychiatry*, *76*(6), 584–593.

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*, 396–403.

Baumgartner, H., & Steenkamp, J. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Beck, A. T., Steer, R. A., & Carbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, *8*, 77–100.

Belitz, C., Jiang, L., & Bosch, N. (2021). Automating procedurally fair feature selection in machine learning. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 379–389. https://doi.org/10.1145/3461702.3462585

Belsher, B. E., Smolenski, D. J., Pruitt, L. D., Bush, N. E., Beech, E. H., Workman, D. E., Morgan, R. L., Evatt, D. P., Tucker, J., & Skopp, N. A. (2019). Prediction models for suicide attempts and deaths: A systematic review and simulation. *JAMA Psychiatry*, *76*(6), 642–651.

Berry, D. T. R., Baer, R. A., & Harris, M. J. (1991). Detection of malingering on the MMPI: A meta-analysis. *Clinical Psychology Review*, *11*(5), 585–598.

Bodner, T. E. (2006). Designs, participants, and measurement methods in psychological research. *Canadian Psychology*, *47*(4), 263–272.

Boe, A. J., McGee Koch, L. L., O'Brien, M. K., Shawen, N., Rogers, J. A., Lieber, R. L., Reid, K. J., Zee, P. C., & Jayaraman, A. (2019). Automating sleep stage classification using wireless, wearable sensors. *NPJ Digital Medicine*, *2*(1), 131.

Borup, M., Brown, N., Konrad, K., & Van Lente, H. (2006). The sociology of expectations in science and technology. *Technology Analysis & Strategic Management*, *18*(3–4), 285–298. https://doi.org/10.1080/09537320600777002

Bruijnen, C. J. W. H., Dijkstra, B. A. G., Walvoort, S. J. W., Markus, W., VanDerNagel, J. E. L., Kessels, R. P. C., & De Jong, C. A. J. (2019). Prevalence of cognitive impairment in patients with substance use disorder. *Drug and Alcohol Review*, *38*(4), 435–442. https://doi.org/10.1111/dar.12922

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. University of California Press. https://books.google.com/books?hl=en&lr=&id=xTwwQvk6XCUC&oi=fnd&pg=PA1&dq=Perception+and+the+representative+design+of+psychological+experiments&ots=tgj-Ck4f91&sig=NNWQCVveyVjx5yMMgGtzk7BTuiY

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105.

Cao, Z., Hidalgo, G., Simon, T., Wei, S., & Sheikh, Y. (2019). OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(1), 172–186.

Caumiant, E. P., Kang, D., Girard, J. M., & Fairbairn, C. E. (2025). Alcohol and emotion: Analyzing convergence between facially expressed and self-reported indices of emotion under alcohol intoxication. *Psychology of Addictive Behaviors*. https://psycnet.apa.org/record/2025-64188-001

Chaikin, A. L., Derlega, V. J., & Miller, S. J. (1976). Effects of room environment on self-disclosure in a counseling analogue. *Journal of Counseling Psychology*, *23*(5), 479–481.

Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018). Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing*, *307*, 72–77.

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *The Journal of Abnormal and Social Psychology*, *60*(2), 151.

Dawson, M. E., Schell, A. M., & Filion, D. L. (2007). The electrodermal system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of Psychophysiology* (Vol. 2, pp. 200–223). http://apsychoserver.psych.arizona.edu/JJBAReprints/PSYC501A/Readings/Chapter%208.pdf

Delgadillo, J., & Gonzalez Salas Duhne, P. (2020). Targeted prescription of cognitive–behavioral therapy versus person-centered counseling for depression using a machine learning approach. *Journal of Consulting and Clinical Psychology*, *88*(1), 14–24.

Didier, N. A., King, A. C., Polley, E. C., & Fridberg, D. J. (2023). Signal processing and machine learning with transdermal alcohol concentration to predict natural environment alcohol consumption. *Experimental and Clinical Psychopharmacology*, *32*(2), 245–254. https://doi.org/10.1037/pha0000683

Ditzer, J., Wong, E. Y., Modi, R. N., Behnke, M., Gross, J. J., & Talmon, A. (2023). Child maltreatment and alexithymia: A meta-analytic review. *Psychological Bulletin*, *149*(5–6), 311–329.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, *5*(3), 69–106. https://doi.org/10.1111/j.1529-1006.2004.00018.x

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*, 91–118.

Eden, D., & Leviatan, U. (1975). Implicit leadership theory as a determinant of the factor structure underlying supervisory behavior scales. *Journal of Applied Psychology*, *60*(6), 736–741.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research.* Dryden. https://psycnet.apa.org/record/1958-00464-000

Edwards, J. R. (2008). To prosper, organizational psychology should … overcome methodological barriers to progress. *Journal of Organizational Behavior*, *29*(4), 469–491. https://doi.org/10.1002/job.529

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *Facial Action Coding System*. Network Information Research.

Fairbairn, C. E., & Bosch, N. (2021). A new generation of transdermal alcohol biosensing technology: Practical applications, machine learning analytics, and questions for future research. *Addiction*, *116*(10), 2912–2920.

Fairbairn, C. E., Han, J., Caumiant, E. P., Benjamin, A. S., & Bosch, N. (2025). A wearable alcohol biosensor: Exploring the accuracy of transdermal drinking detection. *Drug and Alcohol Dependence*, *266*, 112519.

Fairbairn, C. E., Kang, D., & Bosch, N. (2020). Using machine learning for real-time BAC estimation from a new-generation transdermal biosensor in the laboratory. *Drug and Alcohol Dependence*, *216*, 108205. https://doi.org/10.1016/j.drugalcdep.2020.108205

Fairbairn, C. E., Sayette, M. A., Amole, M. C., Dimoff, J. D., Cohn, J. F., & Girard, J. M. (2015). Speech volume indexes sex differences in the social-emotional effects of alcohol. *Experimental and Clinical Psychopharmacology*, *23*(4), 255–264. https://doi.org/10.1037/pha0000021

Frances, A. J., & Widiger, T. (2012). Psychiatric Diagnosis: Lessons from the DSM-IV Past and Cautions for the DSM-5 Future. *Annual Review of Clinical Psychology*, *8*(1), 109–130. https://doi.org/10.1146/annurev-clinpsy-032511-143102

Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review*, *21*(5), 1180–1187. https://doi.org/10.3758/s13423-014-0601-x

Fu, H., Darroch, J. E., Henshaw, S. K., & Kolb, E. (1998). Measuring the extent of abortion underreporting in the 1995 National Survey of Family Growth. *Family Planning Perspectives*, 128–138.

Garcia, J., & Gustavson, A. R. (1997). The science of self-report. *APS Observer*, *10*. https://www.psychologicalscience.org/observer/the%C3%A2%E2%82%AC%20science%C3%A2%E2%82%AC%20of%C3%A2%E2%82%AC%20self%C3%A2%E2%82%AC%20report

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, *348*(1–17), 3.

Goldberg, S. B., Flemotomos, N., Martinez, V. R., Tanana, M. J., Kuo, P. B., Pace, B. T., Villatte, J. L., Georgiou, P. G., Van Epps, J., & Imel, Z. E. (2020). Machine learning and natural language processing in psychotherapy research: Alliance as example use case. *Journal of Counseling Psychology*, *67*(4), 438–448.

Grove, W. M., & Tellegen, A. (1991). Problems in the classification of personality disorders. *Journal of Personality Disorders*, *5*(1), 31–41. https://doi.org/10.1521/pedi.1991.5.1.31

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30.

Gurrieri, L., Fairbairn, C. E., Sayette, M. A., & Bosch, N. (2021). Alcohol narrows physical distance between strangers. *Proceedings of the National Academy of Sciences*, *118*(20), e2101937118. https://doi.org/10.1073/pnas.2101937118

Hathaway, S. R., & McKinley, J. C. (1951). *Minnesota multiphasic personality inventory; manual, revised*. Psychological Corporation. https://psycnet.apa.org/record/1952-03407-000

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., & Zhou, Y. (2017). *Deep learning scaling is predictable, empirically* (arXiv:1712.00409). arXiv. http://arxiv.org/abs/1712.00409

Hickman, L., Tay, L., & Woo, S. E. (2025). Are automated video interviews smart enough? Behavioral modes, reliability, validity, and bias of machine learning cognitive ability assessments. *Journal of Applied Psychology*, *110*(3), 314–335.

Hilario, E. Y., Griffin, M. L., McHugh, R. K., McDermott, K. A., Connery, H. S., Fitzmaurice, G. M., & Weiss, R. D. (2015). Denial of urinalysis-confirmed opioid use in prescription opioid dependence. *Journal of Substance Abuse Treatment*, *48*(1), 85–90.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press. https://books.google.com/books?hl=en&lr=&id=5rIDPV1EoLUC&oi=fnd&pg=IA2&dq=eye+tracking+measurement&ots=_y2COYwInO&sig=Zbq2r5SCZw4ISmK4oY3QAOCgbeM

Holzinger, A., Kieseberg, P., Weippl, E., & Tjoa, A. M. (2018). Current advances, trends and challenges of machine learning and knowledge extraction: From machine learning to explainable AI. *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 1–8.

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, *15*(3), 809–816. https://doi.org/10.1177/1745691620902467

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kelly, A. E. (1998). Clients' secret keeping in outpatient therapy. *Journal of Counseling Psychology*, *45*(1), 50–57. PsycARTICLES. https://doi.org/10.1037/0022-0167.45.1.50

Knäuper, B., Belli, R. F., Hill, D. H., & Herzog, A. R. (1997). Question difficulty and respondents' cognitive ability: The effect on data quality. *Journal of Official Statistics*, *13*, 181–199.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 84–90.

Kuo, P. B., Tanana, M. J., Goldberg, S. B., Caperton, D. D., Narayanan, S., Atkins, D. C., & Imel, Z. E. (2024). Machine-learning-based prediction of client distress from session recordings. *Clinical Psychological Science*, *12*(3), 435–446. https://doi.org/10.1177/21677026231172694

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. In *The Handbook of Brain Theory and Neural Networks* (Vol. 3361). MIT Press.

https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=e26cc4a1c717653f323715d751c8dea7461aa105

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lilienfeld, S. O., & Fowler, K. A. (2006). The self-report assessment of psychopathy. In C. Patrick (Ed.), *Handbook of Psychopathy* (pp. 107–132). Guilford Press. https://books.google.com/books?hl=en&lr=&id=c8JWDwAAQBAJ&oi=fnd&pg=PA211&dq=use+of+self-report+assessments+in+clinical+psychology&ots=qscvMNyE6l&sig=0GVPp4mc4Hqnxr6ARM9yQvYutYU

Maclure, J. (2020). The new AI spring: A deflationary view. *AI & SOCIETY*, *35*, 747–750. https://doi.org/10.1007/s00146-019-00912-z

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*, 115–133.

McDonald, J. D. (2008). Measuring personality constructs: The advantages and disadvantages of self-reports, informant reports and behavioural assessments. *Enquire*, *1*(1), 1–19.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* University of Minnesota Press. https://psycnet.apa.org/record/2006-21565-000

Miller, J. D., Phillips, N. L., & Lynam, D. R. (2025). Questionable research practices violate the American Psychological Association's Code of Ethics. *Journal of Psychopathology and Clinical Science*, *134*, 113–114.

Mohamed, A., Dahl, G. E., & Hinton, G. (2011). Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 14–22.

Nicolas, G., Bai, X., & Fiske, S. T. (2022). A spontaneous stereotype content model: Taxonomy, properties, and prediction. *Journal of Personality and Social Psychology*, *123*(6), 1243–1263.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259.

Norenzayan, A., & Schwarz, N. (1999). Telling what they want to know: Participants tailor causal attributions to researchers' interests. *European Journal of Social Psychology*, *29*(8), 1011–1020. https://doi.org/10.1002/(SICI)1099-0992(199912)29:8<1011::AID-EJSP974>3.0.CO;2-A

Norvig, P., & Russell, S. J. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson. https://thuvienso.hoasen.edu.vn/handle/123456789/8967

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Ordóñez, F. J., & Roggen, D. (2016). Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, *16*(1), 115.

Parrigon, S., Woo, S. E., Tay, L., & Wang, T. (2017). CAPTION-ing the situation: A lexically-derived taxonomy of psychological situation characteristics. *Journal of Personality and Social Psychology*, *112*(4), 642–681.

Passos, I. C., Mwangi, B., & Kapczinski, F. (2016). Big data analytics and machine learning: 2015 and beyond. *The Lancet. Psychiatry*, *3*(1), 13–15.

Paulhus, D. L., & Vazire, S. (2007). The self-report method. *Handbook of Research Methods in Personality Psychology*, *1*(2007), 224–239.

Podsakoff, P. M., MacKenzie, S. B., Lee, J., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, *63*(1), 539–569. https://doi.org/10.1146/annurev-psych-120710-100452

Podsakoff, P. M., & Organ, D. W. (1986). Self-reports in organizational research: Problems and prospects. *Journal of Management*, *12*(4), 531–544. https://doi.org/10.1177/014920638601200408

Podsakoff, P. M., Podsakoff, N. P., Williams, L. J., Huang, C., & Yang, J. (2024). Common method bias: It's bad, it's complex, it's widespread, and it's not easy to fix. *Annual Review of Organizational Psychology and Organizational Behavior*, *11*(1), 17–61. https://doi.org/10.1146/annurev-orgpsych-110721-040030

PoornaPushkala, K., Samundeswari, S., & Megala, J. (2022). Real time system for handling customer queries using Twilio, Assembly Ai and NLP. *2022 1st International Conference on Computational Science and Technology (ICCST)*, 111–115. https://ieeexplore.ieee.org/abstract/document/10040469/?casa_token=n2UYUHHse1UAAAAA:OAbyQ823yfEh8jXWwhIcdh16yEZQBGvnFx6AOVkB0uLnVzc8DoQi3QKwwZ54mJLLxAQ47Mmeah7l

Rathje, S., Mirea, D., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34), e2308950121. https://doi.org/10.1073/pnas.2308950121

Rauthmann, J. F., Gallardo-Pujol, D., Guillaume, E. M., Todd, E., Nave, C. S., Sherman, R. A., Ziegler, M., Jones, A. B., & Funder, D. C. (2014). The Situational Eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology*, *107*(4), 677–718.

Redmon, J. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. http://175.27.250.89:10000/media/attachment/2024/08/YOLOV1.pdf

Rees, L. M., Tombaugh, T. N., Gansler, D. A., & Moczynski, N. P. (1998). Five validation experiments of the Test of Memory Malingering (TOMM). *Psychological Assessment*, *10*(1), 10–20.

Robinson, L. J., Thompson, J. M., Gallagher, P., Goswami, U., Young, A. H., Ferrier, I. N., & Moore, P. B. (2006). A meta-analysis of cognitive deficits in euthymic patients with bipolar disorder. *Journal of Affective Disorders*, *93*(1–3), 105–115.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Schleider, J. L., Abel, M. R., & Weisz, J. R. (2015). Implicit theories and youth mental health problems: A random-effects meta-analysis. *Clinical Psychology Review*, *35*, 1–9. https://doi.org/10.1016/j.cpr.2014.11.001

Schmitt, N. (1994). Method bias: The importance of theory and measurement. *Journal of Organizational Behavior*, 393–398.

Schoedel, R., Kunz, F., Bergmann, M., Bemmann, F., Bühner, M., & Sust, L. (2023). Snapshots of daily life: Situations investigated through the lens of smartphone sensing. *Journal of Personality and Social Psychology*. https://psycnet.apa.org/record/2023-87986-001

Schretlen, D. J. (1988). The use of psychological tests to identify malingered symptoms of mental disorder. *Clinical Psychology Review*, *8*(5), 451–476.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*(2), 93–105. https://doi.org/10.1037/0003-066X.54.2.93

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology*, *21*(2), 277–287. https://doi.org/10.1002/acp.1340

Schwarz, N., Strack, F., & Mai, H. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly*, *55*(1), 3–23.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Sobell, L. C., & Sobell, M. B. (1992). Timeline follow-back. In R. Z. Litten, J. P. Allen, & N. J. Totowa (Eds.), *Measuring alcohol consumption* (pp. 41–72). Humana Press.

Tackett, J. L., Brandes, C. M., King, K. M., & Markon, K. E. (2019). Psychology's replication crisis and clinical psychological science. *Annual Review of Clinical Psychology*, *15*(1), 579–604. https://doi.org/10.1146/annurev-clinpsy-050718-095710

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., Oltmanns, T. F., & Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science*, *12*(5), 742–756. https://doi.org/10.1177/1745691617690042

Tay, L., Woo, S. E., Hickman, L., Booth, B. M., & D'Mello, S. (2022). A conceptual framework for investigating and mitigating machine-learning measurement bias (MLMB) in psychological assessment. *Advances in Methods and Practices in Psychological Science*, *5*(1), 25152459211061337. https://doi.org/10.1177/25152459211061337

Tay, L., Woo, S. E., Hickman, L., & Saef, R. M. (2020). Psychometric and validity issues in machine learning approaches to personality assessment: A focus on social media text mining. *European Journal of Personality*, *34*(5), 826–844. https://doi.org/10.1002/per.2290

Thomaz, E., Essa, I., & Abowd, G. D. (2015). A practical approach for recognizing eating moments with wrist-mounted inertial sensing. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 1029–1040. https://doi.org/10.1145/2750858.2807545

Thoresen, C. J., Kaplan, S. A., Barsky, A. P., Warren, C. R., & De Chermont, K. (2003). The affective underpinnings of job perceptions and attitudes: A meta-analytic review and integration. *Psychological Bulletin*, *129*, 914–945.

Tian, Y., & Zhang, Y. (2022). A comprehensive survey on regularization strategies in machine learning. *Information Fusion*, *80*, 146–166.

Tourangeau, R. (1984). Cognitive sciences and survey methods. In M. Jabine, J. Straf, & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (Vol. 15, pp. 73–100). National Academy Press.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883.

Vazire, S. (2010). Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*(2), 281–300.

Vieira, S., Liang, X., Guiomar, R., & Mechelli, A. (2022). Can we predict who will benefit from cognitive-behavioural therapy? A systematic review and meta-analysis of machine learning studies. *Clinical Psychology Review*, *97*, 102193.

Vollmer, N. (2020, May 22). *Recital 71 EU General Data Protection Regulation (EU-GDPR)* [Text]. SecureDataService. http://www.privacy-regulation.eu/en/recital-71-GDPR.htm

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science*, *5*(3), 457–469. https://doi.org/10.1177/2167702617691560

Watson, D., Pennebaker, J. W., & Folger, R. (1987). Beyond negative affectivity: Measuring stress and satisfaction in the workplace. *Journal of Organizational Behavior Management*, *8*(2), 141–158. https://doi.org/10.1300/J075v08n02_09

Webb, C. A., Cohen, Z. D., Beard, C., Forgeard, M., Peckham, A. D., & Björgvinsson, T. (2020). Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: A comparison of machine learning approaches. *Journal of Consulting and Clinical Psychology*, *88*(1), 25–38.

Widom, C. S., & Shepard, R. L. (1996). Accuracy of adult recollections of childhood victimization: Part 1. Childhood physical abuse. *Psychological Assessment*, *8*(4), 412–421.

Williams, D., Botting, N., & Boucher, J. (2008). Language in autism and specific language impairment: Where are the links? *Psychological Bulletin*, *134*(6), 944–963.

Xiao, B., Imel, Z. E., Georgiou, P. G., Atkins, D. C., & Narayanan, S. S. (2015). " Rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing. *PloS One*, *10*(12), e0143055.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yeater, E., Miller, G., Rinehart, J., & Nason, E. (2012). Trauma and sex surveys meet minimal risk standards: Implications for Institutional Review Boards. *Psychological Science*, *23*(7), 780–787. https://doi.org/10.1177/0956797611435131

Table 1. A Non-Comprehensive Review of Machine Learning Applications Relevant for Behavioral Researchers in Clinical Psychological Science

	Data Type	Ground Truth	Study
Behavior			
Detecting **human motion** (e.g., gestures) from multi-input **wearable sensors**	Sensor/Time-series	Researcher annotations	(Ordóñez & Roggen, 2016)
Recognizing **eating behaviors** through **wearable accelerometer**	Sensor/Time-series	Researcher annotations	(Thomaz et al., 2015)
Predicting **drinking episodes** through from multi-input **smartphone data**	Sensor/Time-series	Participant self-reports	(Bae et al., 2018)
Estimating **alcohol consumption levels** from **wearable alcohol sensors**	Sensor/Time-series	Physiological data	(Fairbairn et al., 2020, 2025)
Differentiate **drinking episodes** via **wearable alcohol sensors**	Sensor/Time-series	Participant self-reports	(Didier et al., 2023)
Detecting **sleep behavior** and stage from multi-input **wearable sensors**	Sensor/Time-series	Physiological data	(Boe et al., 2019)
Social Processes			
Evaluating **therapist effectiveness** through analysis of **text-messages**	Language	Participant self-reports	(Althoff et al., 2016)
Evaluating **therapist fidelity** through analysis of **transcribed sessions**	Language	Researcher-annotations	(Atkins et al., 2014)
Predicting **therapeutic alliance** from **transcribed sessions**	Language	Participant self-reports	(Goldberg et al., 2020)
Analyzing therapist **empathy** from **transcribed sessions**	Language	Researcher-annotations	(Xiao et al., 2015)
Analyzing **social-emotional expression** from **social media posts**	Language	Other/Mixed	(Rathje et al., 2024)
Recognizing **social distance** from social interaction **videos**	Images	Researcher-annotations	(Gurrieri et al., 2021)
Context/Environment			
Detecting **contextual features** (e.g., lighting, group size) from **photographs**	Images	Other/Mixed	(Ariss et al., in press)
Establishing psychological **contexts** (e.g., adversity) from **unstructured text**	Language	N/A	(Parrigon et al., 2017)
Examining psychological **context** characteristics from **smartphone data**	Sensor/Time-series	Participant self-reports	(Schoedel et al., 2023)
Real-time **object detection** (e.g., person, dog, car) from **photographs**	Images	Researcher annotations	(Redmon, 2016)

Studies reflect a mixture of application (in which a pre-existing model/tool is applied), validation (in which a model is trained and/or its accuracy is assessed) and studies integrating elements of both. Ground truth reflects the metric used to validate and train automated measures. In cases where models were unsupervised, no ground truth is listed.

The above review focuses on automated models that might offer alternatives to closed-ended reports and therefore does not cover applications of machine learning to strictly biological constructs (e.g., genetic or neuroscientific data). We focus here on the measurement of constructs high in externality/observability. Machine learning has also been applied to data derived from external indicators of internally experienced state- and trait-level constructs (e.g., emotion, cognition, personality, symptomatology). A non-comprehensive review of these applications is supplied in supplemental material.

Table 2. Key Questions for Selecting Method of Measurement in Psychological Research

Question 1. Should I consider methods of measurement for my construct of interest beyond closed-ended self-reports?

a. Do I plan to explore the relationship between the identified construct of interest and **at least one other construct** that will be **assessed by survey/self-report**?

b. Will the study have context, population, or other **design elements** that might **exaggerate biases linked with survey methods**?
   - ☐ Feature socially sensitive or potentially stigmatizing questions
   - ☐ Require evaluation of self along socially desirable/undesirable dimensions
   - ☐ Target automatic/unconscious processes
   - ☐ Trigger positive/negative affective states and/or target populations characterized by high state-level affect
   - ☐ Involve the potential for secondary gains/punishments (real or perceived)
   - ☐ Examine topics surrounding which participants are likely to have lay theories
   - ☐ Involve a cognitively demanding task environment and/or a population with restricted cognitive resources

Answers and Implications: If the answer to at least one of the above questions is "yes," you should consider integrating non-survey methods into your research. If the answer to both of the above questions is "yes," relying on surveys is likely to carry high risk of common methods effects, including false positive findings. Non self-report measurement options here are varied, but common approaches include psychophysiological assessment, human coding/behavioral observation, informant report, and the analysis of language.

Question 2. Should I choose an automated measurement tool for my construct of interest?

a. Is there an automated tool available for the measurement of my construct of interest and, if so, what is its current **accuracy level**?
   - How large was the dataset used to train this measure and assess its accuracy?
   - How diverse were instances/individuals included in this dataset and are they representative of the instances/individuals I aim to study?
   - Is the context for model training similar to the context where it will be applied in my study (e.g., in the lab, in the wild)?

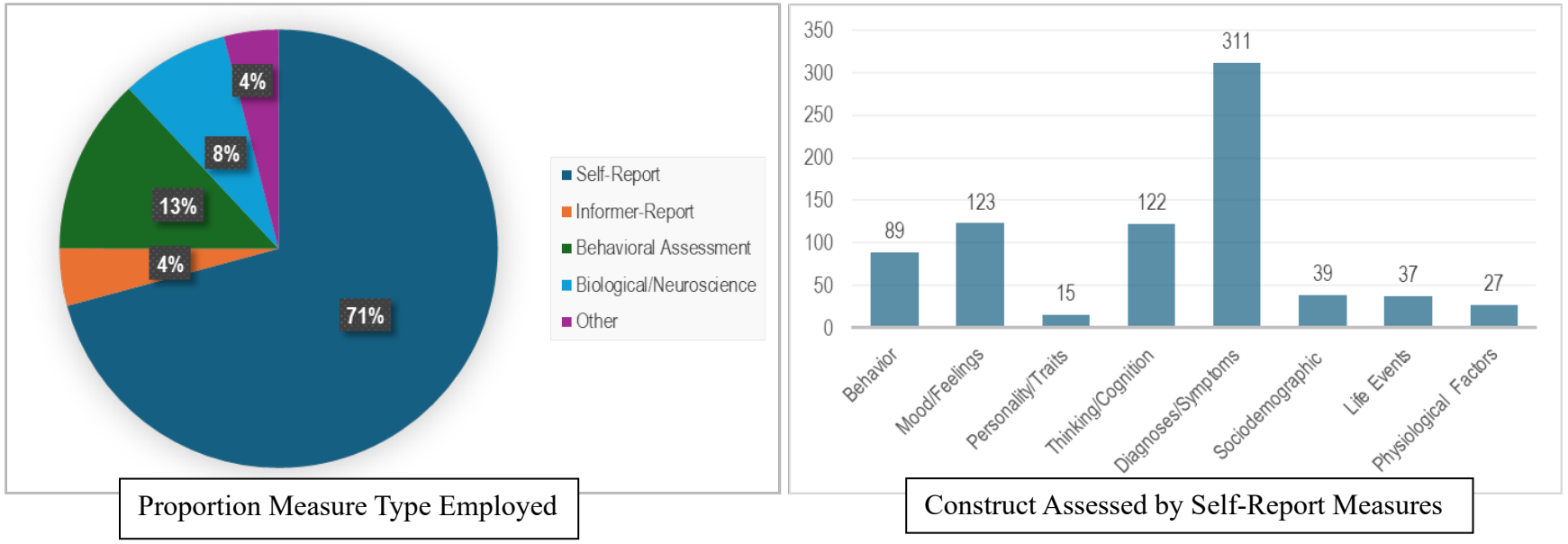
b. Are there **specific advantages** linked with automated measurement compared with other available measurement methods?
   - Financial and temporal resources (is the automated tool faster or cheaper?)
   - Rater consistency/reliability (is consistency higher for the automated measure vs other methods available to me?)

c. Is **transparency** in the measurement process critical for my specific application?

Answers and Implications: If a study already features at least one key variable assessed via traditional survey methods, and accuracy levels of automated tools are comparable or higher to other methods, automated tools can represent a useful measurement alternative. Automated tools may be particularly useful in cases where other alternatives (e.g., human coding of behavior) are time/resource intensive and/or yield inconsistent results. In

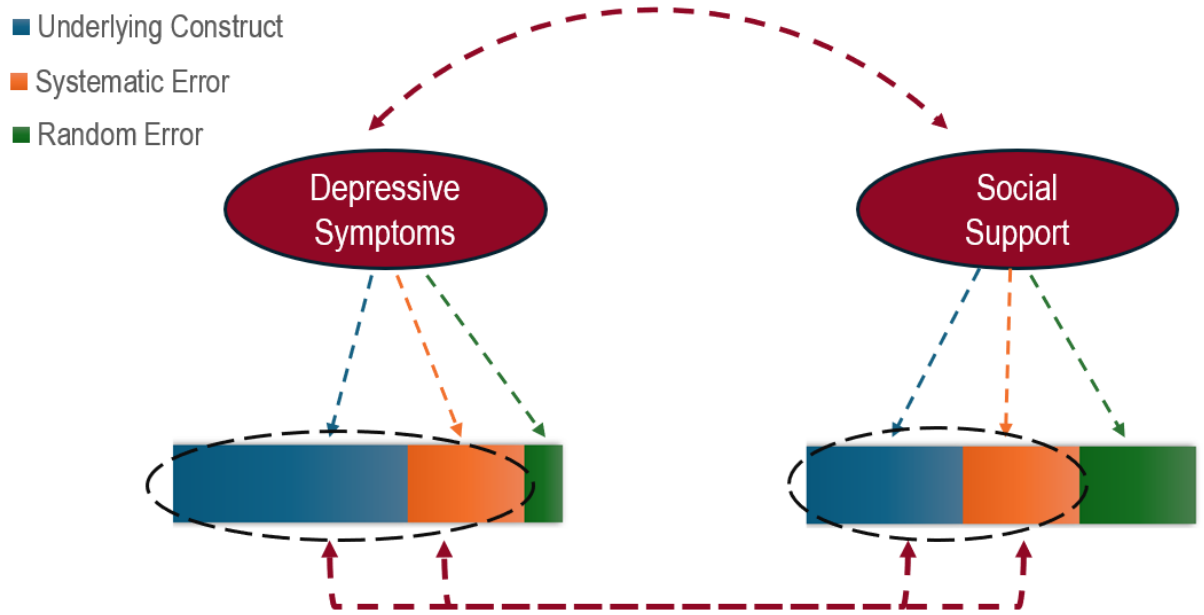
interpreting accuracy levels of automated tools, researchers should consider carefully the dataset employed to train and test the tool, including its size, the representativeness of the population, and similarity of the task environment to the researcher's current application. Automated tools are not appropriate in cases where transparency/explainability in measurement method is critical (e.g., research with direct legal application; specific diagnostic applications), except where interpretable algorithms can be used.

Figure 1. Review of Measure Types Employed in Highly Ranked Clinical Psychology Outlets 2021-2024.



Note: A literature review was conducted for all empirical articles published in the journals *Clinical Psychological Science*, *Journal of Consulting and Clinical Psychology*, and *Journal of Psychopathology and Clinical Science* from July 1, 2021-July 1, 2024. All articles were assessed for study type (experimental vs. correlational), measure type, and construct coded in tests of primary study aims. Primary aims were determined according to language employed by articles' authors or, where the authors did not indicate the primacy of individual aims over others, according to the order in which aims/hypotheses were presented within the article. A random sample of 10% of studies were double coded by a second-rater blind to the original rater's codes (κ =.80). A total of 386 correlational and 160 randomized studies were identified and coded for this search. The lefthand panel above represents proportion measure type at the level of the assessment. The righthand panel above reflects constructs assessed specifically within the measure type category of self-report. Overlapping codes were permitted—where a given assessment met criteria for more than one measure type or construct, it is represented within each. For the purposes of this review, "self-report" assessments are defined as closed questionnaire or interview-style assessments in which participants self-report according to a pre-defined set of categorical or numerical responses. Questionnaire or interview assessments involving open-ended response options that required no immediate categorization are listed under "other" measure type. Review procedures and hypotheses were pre-registered at Open Science Framework (see https://osf.io/cvnk8).

Figure 2. Example of Common Methods Effects on Correlations between Measured Constructs in Clinical Psychological Science



Note. Variability in measured constructs can be divided into three elements: variability in the true underlying construct, random error variance, and systematic error variance. In the examination of relationships between multiple variables, significant correlations can emerge attributable to not only shared variation in the underlying constructs, but also to shared systematic forms of error linked with the method of measurement (see also Podsakoff et al., 2024). Negative affect represents one contributor to common methods effects in self-report studies, influencing participants' reports of not only their internal subjective state, but also others' behaviors, past events, and attitudes.