Psychological Measurement in the Information Age: Machine-Learned Computational Models

Sidney K. D'Mello[1]

Louis Tay[2]

Rosemary Southwell[1]

[1]University of Colorado Boulder

[2] Purdue University

[1]Address correspondence to:

Sidney D'Mello, 594 UCB, Boulder CO 80309, USA

sidney.dmello@colorado.edu

Abstract

Psychological science can benefit from and contribute to emerging approaches from the computing and information sciences driven by the availability of real-world data and advances in sensing and computing. We focus on one such approach, machine-learned computational models (MLCMs) – computer programs learned from data, typically with human supervision. The article introduces MLCMs and how they contrast with traditional computational modeling and assessment in the psychological sciences. Examples of MLCMs from cognitive and affective science, neuroscience, education, organizational psychology, and personality and social psychology are provided. We consider the accuracy and generalizability of MLCM-based measures, cautioning researchers to consider the underlying context and intended use when interpreting their performance of such measures. We conclude that in addition to known data privacy and security concerns, the use of MLCMs entails a reconceptualization of fairness, bias, interpretability, and responsible use.

Psychological Measurement in the Information Age: Machine-Learned Computational Models

If measurement is the cornerstone of science, psychological science has accomplished a lot. We have designed clever experiments to measure complex social phenomena, honed the measurement of ill-defined constructs to a precise science, made inferences about the mind through probing of behavior, begun to delve into the brain, and have applied our findings to improve the human condition. Meanwhile, the trifecta of the information age – new, improved, and cost-effective sensing, anywhere/anytime computing, and a new generation of digital natives – has led to a data and computing revolution which has enhanced multiple research areas and created new ones (e.g., computational social science, cyber-physical systems, quantitative biology). Can such advances similarly enhance psychological science? We think so and describe how the core of psychological science – psychological measurement – can benefit from an information-age update.

Consider one simplified view of psychological measurement: *measurement = data + inference*. The data typically comes from humans (e.g., posts on social media) and is converted to a structured format (e.g., human coders count the number of pronouns). Computers can automate and scale this task and discover complex associations in the data, revealing multivariate interactions and nonlinearities. However, they cannot make meaning of any patterns they discover, at least in any deep sense. We rely on human knowledge and expertise to make inferences from data. Even when measurement is automated, for example, computerized adaptive testing (Wainer et al., 2000), the items and inference are preprogrammed into the computer.

But what if we could design computers to *learn how to* make human-like inferences from data? The resultant measure would combine the pattern-finding prowess of computers with the inferencing abilities of humans, resulting in transformative impacts. For one, such a measure would enable the analysis of less-structured datasets with the scope and scale to address thorny issues of reproducibility and generalizability. By leveraging modern sensing/analysis capabilities, these measures can focus on real-world human behavior rather than curated responses. Measurement could also be done in real-time, opening the door for just-in-time interventions, individualized experimental manipulations, and discoveries currently precluded by measurement latencies. The measures would potentially be more objective provided that bias is mitigated in their design. Because the measures are learned, not preprogrammed, analysis of the measures themselves can deepen understanding of the underlying phenomena.

If this all seems too fanciful, rest assured there is a systematic approach to developing these measures. It is called *computational modeling,* a representation of a phenomenon *in silico* – i.e., performed or simulated by a computer. This is not an advance in itself – the novelty is that the computational models are directly learned from data rather than preprogrammed.

Machine Learned Computational Models (MLCMs)

A computational model is a computer program that produces a desired *output* given *input*. Applied to psychological measurement, this entails converting input data into a higher-level representation (*features*) usable by a computer, which are transformed into a measurement estimate (i.e., output) via various algorithmic *structures* (approaches). For example, a computational model of mind wandering during reading (Faber et al., 2018) based on eye tracking can map features, such as the number and duration of gaze fixations, onto estimates of mind wandering using one of the structures in Figure 1A.

$$MW = B_0 + B_1 \times NFix + B_2 \times FixDur$$

Regression structure

if [NFix < T1 and FixDur > T2];
then [MW = true]

Rule-based (or tree-based) structure

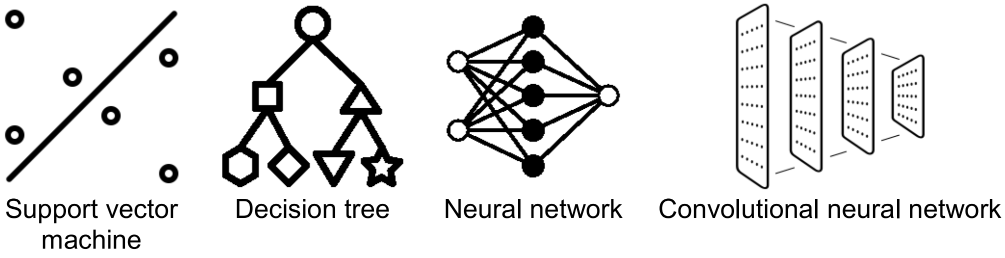
$$P(MW \mid NFix = x, FixDur = y) = [P(NFix = x \mid MW) \times P(FixDur = y \mid MW) \times P(MW)]/P(NFix = x, FixDur = y)$$

(Naïve) Bayesian structure (simplified)

$$\begin{bmatrix} w_{NFix\_h1} & w_{FixDur\_h1} \\ w_{NFix\_h2} & w_{FixDur\_h2} \end{bmatrix} \begin{matrix} w_{h1\_MW} \\ w_{h2\_MW} \end{matrix}$$

Neural network representation (weight matrix)

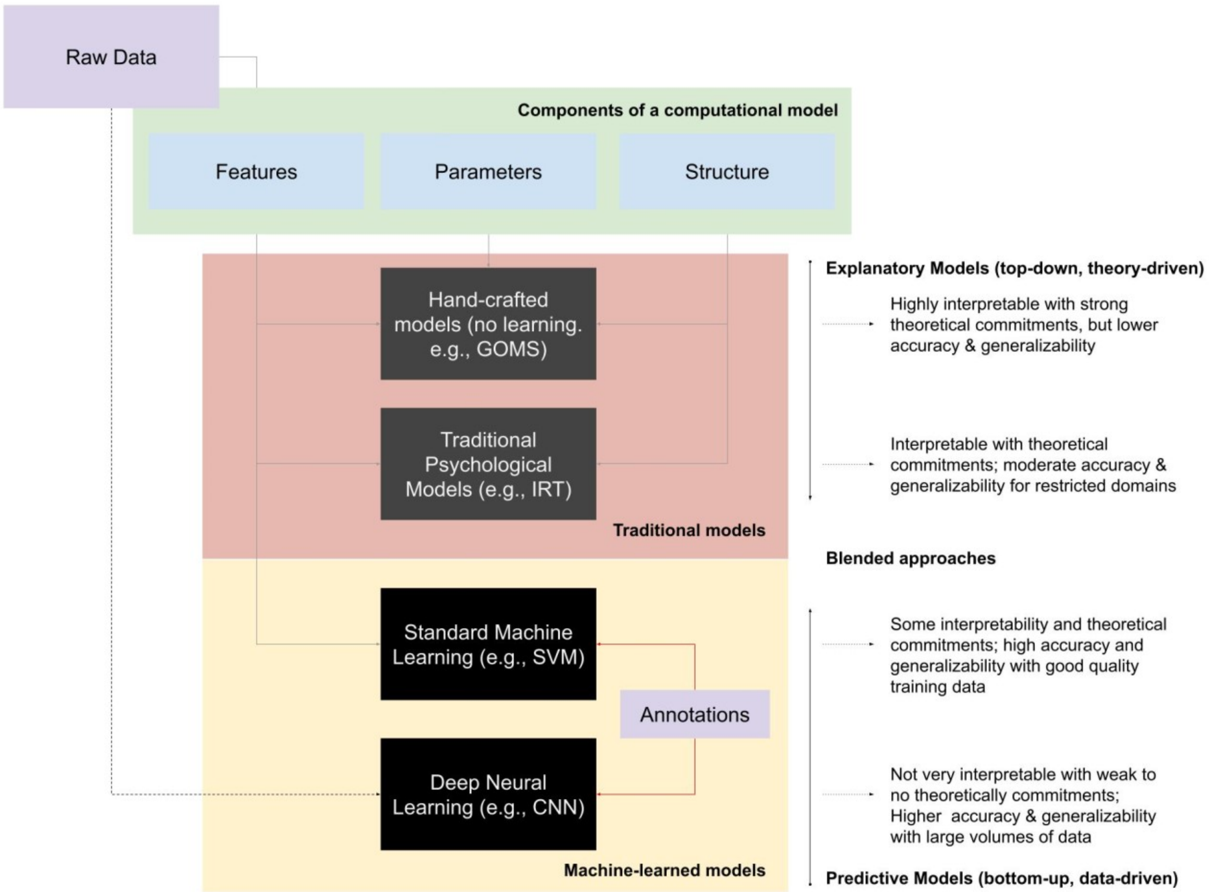
(A)



(B)

Figure 1. (A). Example structures for computational models of mind wandering (MW) based on two eye gaze features (number of fixations [NFix] and fixation duration [FixDur]). B = parameter; T = threshold; P = probability and W = weight (top). (B). Graphical representation of some common machine-learned computational models (bottom)

Computational models differ in how features, structure, and parameters (e.g., regression weights) are specified. Traditionally, human experts pre-programmed the models by specifying all components (Figure 2) as in the classic GOMS models in human factors (Card et al., 1983). *Hand-crafted* models are rare due to difficulties in specifying a generalizable set of parameters (amongst other factors). An intermediate step is to pre-specify the features and structure but learn the parameters from data as with traditional psychological models, such as item response theory used in assessment and classic Bayesian models of cognition.


Raw Data
Components of a computational model
Features
Parameters
Structure
Hand-crafted models (no learning. e.g., GOMS)
Traditional Psychological Models (e.g., IRT)
Traditional models
Standard Machine Learning (e.g., SVM)
Annotations
Deep Neural Learning (e.g., CNN)
Machine-learned models
Explanatory Models (top-down, theory-driven)
Highly interpretable with strong theoretical commitments, but lower accuracy & generalizability
Interpretable with theoretical commitments; moderate accuracy & generalizability for restricted domains
Blended approaches
Some interpretability and theoretical commitments; high accuracy and generalizability with good quality training data
Not very interpretable with weak to no theoretically commitments; Higher accuracy & generalizability with large volumes of data
Predictive Models (bottom-up, data-driven)

GOMS: Goals, Operators, Methods, and Selection; **IRT**: Item-response theory; **SVM**: Support vector machine; **CNN**: Convolutional neural network

Figure 2. The four main approaches to computational modeling. Note that the approaches are not mutually exclusive and can be combined in multiple ways. Annotations are only needed in the model training phase.

But what about complex, poorly understood phenomena, where neither the model structure nor the parameters can be pre-specified? Using *supervised machine learning*, it is possible to learn both from data (Jordan & Mitchell, 2015). Starting with a set of *training examples* which link features with corresponding *annotations* (e.g., human ratings), these methods *learn* the model by identifying patterns in the training data. After training is complete, the resultant *machine-learned computational model* (MLCM; Figure 2) produces computer estimates (i.e., measurements) on new input data (without annotations).

Turning to our example, training data are collected by tracking eye gaze (to compute features) and self-reports of mind wandering (annotations) as participants read. Training examples are created by aligning the gaze features with the mind wandering reports over a temporal window (e.g. a page), upon which supervised learning methods are applied to learn an MLCM, which produces estimates of mind wandering from gaze features.

What are these supervised learning techniques? Linear regression is one example. However, in the psychological sciences, where the goal is *explanation,* the models are fit on the entire data, and the emphasis is on statistical significance of the coefficients (Yarkoni & Westfall, 2017). For machine learning, where the goal is *prediction,* the focus is on the extent to which MLCM outputs align with some measure of "ground truth" when applied to *holdout* (i.e., different from training) data*,* including data from different people, paradigms, populations, and contexts (*generalizability*); e.g., alignment of the MLCM's predictions of mind wandering to self-reports (accuracy) from a *different* set of readers on a *new* text (generalizability).

A highly accurate model might overfit to the training data and perform poorly on holdout data (lower generalizability), whereas a highly generalizable model might underfit to the data (lower accuracy). Because regression and its variants (e.g., generalized linear models) are limited in both respects, researchers have developed numerous approaches to improving accuracy (e.g., modeling nonlinearity and feature interactivity) and generalizability (e.g., penalizing models with more parameters, using an ensemble of multiple models). The resultant models (Figure 1) have different representations (e.g., probabilities, parameter weights), structures (e.g., equations, rules, networks of artificial neurons), and assumptions (e.g., some assume feature independence whereas feature interdependence is critical in others). But they are all computer programs.

With the resurgence of *deep neural learning* (Jordan & Mitchell, 2015)*,* which combines massive data (e.g., the entirety of English Wikipedia), computing (e.g., thousands of parallel processors), and advanced algorithms, MLCM complexity (up to billions of parameters) has increased, yielding greater performance. Some innovations include *representational learning,* where the features themselves are learned from raw data rather than being pre-specified. An extension is *end-to-end learning,* where everything (features, structure and parameters) are simultaneously learned from raw data (Figure 2). For example, forgoing human-engineered features, the model automatically extracts internal representations most useful for predicting mind wandering when presented with raw gaze data. Another is *fine tuning,* when a model is first *pre-trained* on massive data in a domain-agnostic fashion and then adapted for a given domain using a small amount of annotated data.

As Figure 2 indicates, computational models can be broadly divided into *explanatory* – where the primary aim is understanding the underlying mechanisms – and *predictive* – where accurate and generalizable predictions are the main goal. MLCMs fall into the predictive family in that they have fewer theoretical commitments and are more bottom-up and data-driven. As a result, MLCMs with very different structures can yield similar predictions, which limits their ability to provide causal or mechanistic explanations. However, because they are powerful, fine-grained predictive machines, MLCMs can be useful tools for scientific inquiry (in addition to applications in assessment and intervention; see Introduction). For example, they can be designed to compare the diagnosticity of various input modalities, investigate whether combining modalities results in superadditive, additive, or redundant effects, understand the time course of phenomena, model nonlinearity and interactivity among inputs, contrast model predictions with human judgments, and investigate generalizability across people, domains, contexts, etc. Thus, MLCMs can complement explanation-based approaches, especially for complex, ill-defined phenomena, and are valuable tools in the arsenal of a pluralistic scientist.

It should also be noted that distinctions among the four main modeling approaches (Figure 2) are not crisp. For example, when theoretical commitments are important, it is possible to pre-specify a subset of the structure and parameters based on theory and/or plausibility while allowing others to be learned (e.g., Hinaut & Dominey, 2013). Some deep learning architectures, for example, convolutional neural networks, which have revolutionized image processing, are inspired by the neural pathways in the visual cortex (Le Cun et al., 2015). When data is abundant but annotations are sparse, a useful approach is to use deep representational learning to automatically learn the features in an unsupervised (i.e., without annotations) fashion, but then revert to standard supervised learning (i.e., with annotations) to learn the MLCM. The main message is that MLCM development should not be dogmatic – the goals of the enterprise, availability of data, and expertise of those involved should determine the approach.

Illustrative Example

To illustrate, Jensen et al. (2021) automatically analyzed audio recordings of teachers' classroom discourse to estimate the prevalence of seven discourse categories (e.g., questions, elaborated evaluations) linked to student achievement growth. The main steps to construct the MLCM – which are common to multiple MLCMs – are shown in Figure 3A. First, the researchers recorded teacher audio from 127 authentic class sessions of 16 English Language Arts (ELA), which were segmented and transcribed into 35k utterances via an automatic speech recognizer. Trained coders annotated 16k utterances for the presence of each discourse category.

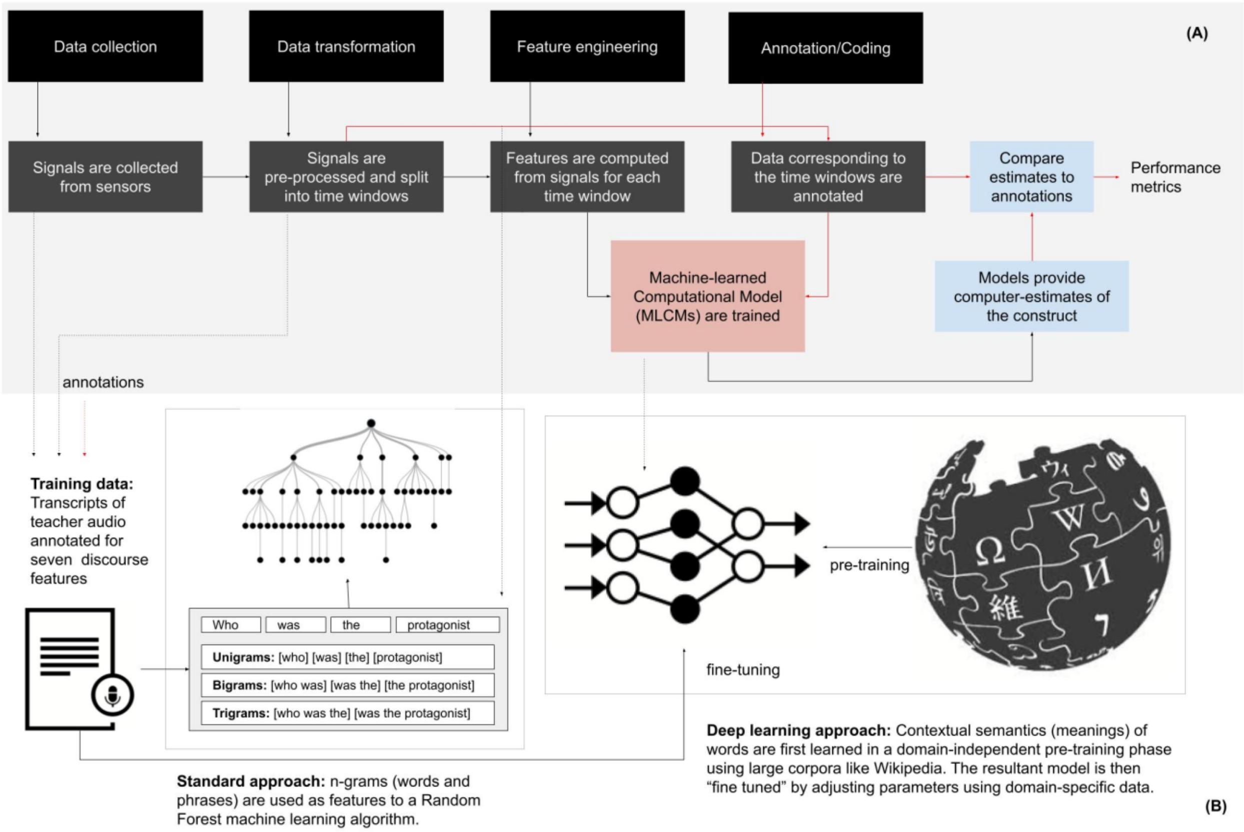

Data collection
Data transformation
Feature engineering
Annotation/Coding
(A)
Signals are collected from sensors
Signals are pre-processed and split into time windows
Features are computed from signals for each time window
Data corresponding to the time windows are annotated
Compare estimates to annotations
Performance metrics
Machine-learned Computational Model (MLCMs) are trained
Models provide computer-estimates of the construct
annotations
Training data: Transcripts of teacher audio annotated for seven discourse features
Who
was
the
protagonist
Unigrams: [who] [was] [the] [protagonist]
Bigrams: [who was] [was the] [the protagonist]
Trigrams: [who was the] [was the protagonist]
pre-training
fine-tuning
Standard approach: n-grams (words and phrases) are used as features to a Random Forest machine learning algorithm.
Deep learning approach: Contextual semantics (meanings) of words are first learned in a domain-independent pre-training phase using large corpora like Wikipedia. The resultant model is then "fine tuned" by adjusting parameters using domain-specific data.
(B)

Figure 3. Illustrative example of training an MLCM to identify spoken discourse features from audio. Panel (A) lists the basic pipeline for training an MLCM. Lines in red denote components where human input might be needed. Panel (B) contrast a standard and a deep machine learning approach.

The researchers contrasted two modeling approaches (Figure 3B). The standard approach used utterance-level counts of individual words and two- and three-word phrases (called n-grams) as features. Then, binary Random Forest classifiers (a supervised learning method) were individually trained to identify the presence/absence of each discourse category from the features. An examination of the n-grams most predictive of each discourse category provided an intuitive understanding of teacher talk. For the second approach, the researchers started with a deep neural network that was pre-trained on large text corpora containing over 3 billion words to learn the contextual semantics of words (e.g., distinguishing between "bank" in the context of a river vs. financial institution) and then fine-tuned (i.e., adapted parameters) it to identify each discourse category using the 16k annotated utterances.

Both approaches used cross-validation where the utterances were divided into eight partitions; MCLMs were trained on seven partitions (training set), and evaluated on the held-out partition (test set). The process was repeated until all partitions were included as the test set exactly once. To ensure generalization across teachers, utterances of a given teacher were only included in a training *or* testing partition in a given iteration.

Accuracy was assessed by comparing each MLCM's utterance-level estimate with corresponding human annotations using the area under the receiver operating characteristic curve (AUC). Scores ranged from 0.73-0.90 for the deep learning approach compared to 0.71-0.85 for the standard approach; both substantially outperformed chance guessing (AUC of 0.5). The researchers have embedded the models into a software application that provides teachers with automated feedback on their own classroom discourse to enable reflection and improvement.

The example highlights some newsworthy points. First, developing MLCMs for complex phenomena such as spoken discourse classification often entails leveraging MLCMs developed for more primitive tasks (e.g., speech recognition, representing word semantics). Second, the example used minimal human knowledge engineering in that features were automatically computed (standard approach) or bypassed altogether (deep learning approach). An alternate approach would be to use hand-crafted features such as parts of speech (e.g., nouns, pronouns) that may have theoretical significance. Third, there is an accuracy-interpretability tradeoff, which favors the deep learning and standard approaches, respectively.

Selective Examples of MLCMs from the Psychological Sciences

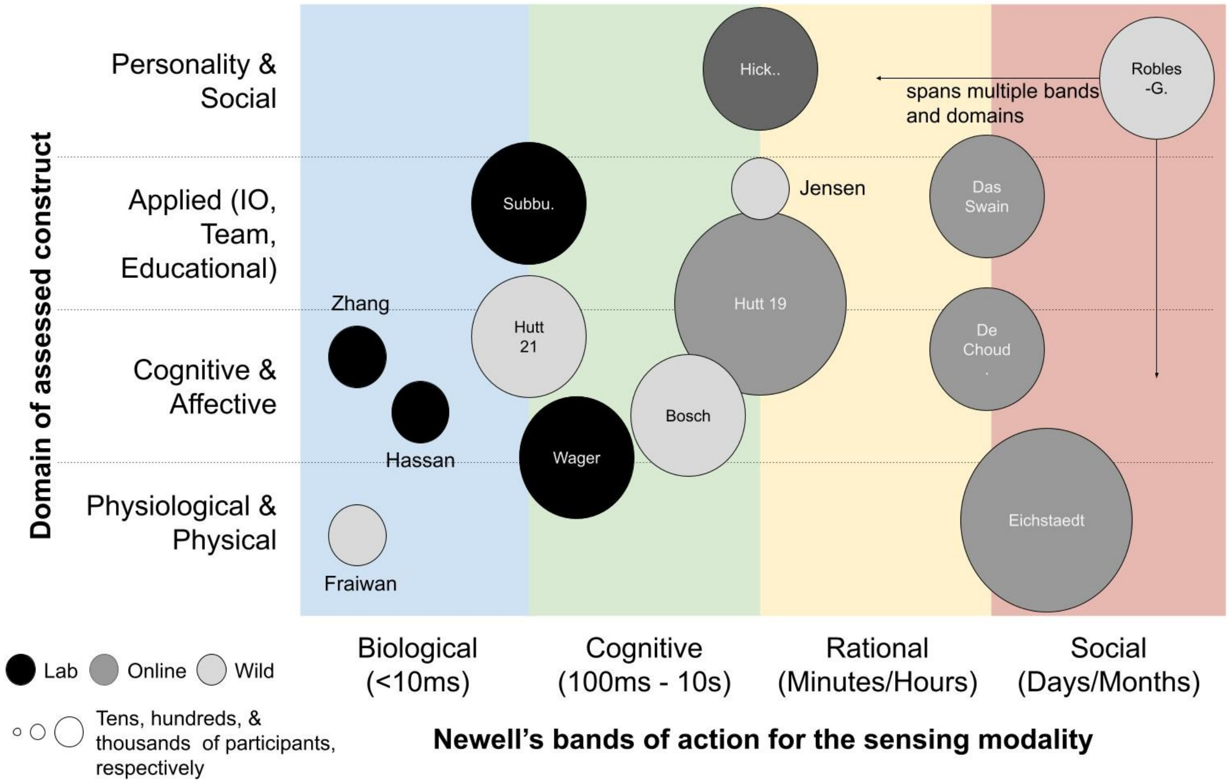
We now present further examples of MLCMs for measurement, which we have roughly organized across four levels of *sensing* timescale inspired by Newell (1990) bands of action (biological, cognitive, rational, and social –see Figure 4). We start with the biological band (<10ms), such as some measures of neuronal activity. In one example, Fraiwan et al. (2012) developed an EEG-based MLCM to accurately discriminate among the five main sleep stages (a time-consuming task for trained clinicians) in a thoracic clinic. While this study used predefined EEG features, Zhang et al. (2018) developed an end-to-end deep approach to learn spatio-temporal patterns directly from EEG data to distinguish between high and low workloads. As an example of integrating multiple modalities, Hassan et al. (2019) combined electrodermal activity (EDA), photoplethysmography (PPG), and electromyography (EMG) to discriminate among experimentally-elicited emotions in the lab. Turning to the cognitive band (100ms-10s), Wager et al. (2013) developed an MLCM that discriminates heat-induced pain from warmth, anticipation, recall of pain, and social pain based on whole-brain fMRI activity.

Whereas these examples used research-grade sensing and experimentally-induced responses in controlled settings, MLCMs can measure spontaneous responses with cost-effective sensing in the wild with many studies blending the biological/cognitive sensing bands. For example, Hutt et al. (2021) used $100 eye trackers to develop an MLCM of mind wandering while high-school students interacted with educational technology in classrooms. They used the MLCM's estimates to trigger dynamic interventions to reengage attention and improve learning. Similarly, Bosch et al. (2015) combined facial expressions from video with interaction patterns (clicks and click timings) to measure affect as students played an educational videogame, finding that a multimodal approach improved robustness to missing data but negligibly impacted accuracy. At the team level, Subburaj et al. (2020) used a multimodal (facial expressions, acoustics, eye gaze, and interaction patterns) and multiparty approach (integrating signals from three individuals) to predict collaborative problem solving outcomes in remote teams.

The rational band consists of measurement in the range of minutes to hours and studies often aggregate more fine-grained sensing (cognitive band) over longer time frames (rational band). The Jensen et al. (2021) example discussed above is one example. Another is Hickman et al. (2021), who automated scoring of personality based on language, facial expressions, and prosody in mock video interviews for personnel selection. In a large-scale study, Hutt et al. (2019) developed an MLCM to infer engagement from interaction patterns as approximately 70,000 students interacted with an online learning platform.

Studies at the social band have largely relied on social media posts (individual posts are on the rational band) using timeframes from days to months. De Choudhury et al. (2013) developed an MLCM to identify individuals diagnosed with depression based on their Twitter usage. Eichstaedt et al. (2015) also used Twitter data, but at the societal level – their MLCM was a better predictor of county-level atherosclerotic heart disease mortality rates than established demographic and health indicators. At the organization level, Das Swain et al. (2020) analyzed language used in over 600,000 Glassdoor reviews from 92 Fortune 500 companies to infer 41 dimensions of organizational culture, which then were used to predict job performance.

MLCMs can span all four bands. In a year-long study of 757 information workers, Robles-Granda et al. (2021) measured physical and physiological signals from wearable sensors, communications from a smartphone app, relative location using Bluetooth beacons, contextual cues (e.g., weather), and social media data to develop MLCMs of personality, cognitive ability, health, well-being, and job performance using a robust (to missing/noisy data) approach.



Study	Sensing Band	Signals	Context	N	Construct	Level	Machine Learning Model
Fraiwan 2012	Bio	EEG	Clinic	16	Sleep stage	Within Indv.	Random forest classifier
Zhang 2019	Bio	EEG	Lab	20	Mental workload	Within Indv.	Deep neural network
Hassan 2019	Bio	Physiology (EDA, PPG and EMG)	Lab	32	Affect	Within Indv.	Deep belief network, Support vector machine
Hutt 2021	Bio/Cog	Eye gaze	Classroom	287	Mind wandering	Within Indv.	Bayesian network
Subburaj 2020	Bio/Cog	Task context, eye gaze, text (speech), facial expressions	Lab	303	Team performance	Within Grp.	Random forest
Wager 2013	Cog	fMRI	Scanner	114	Pain	Within Indv.	Regularized regression
Bosch 2015	Cog	Clicks, task context, Facial expressions	Classroom	133	Affect	Within Indv.	Standard ML (various)
Jensen 2021	Cog/Ratnl.	Text (speech)	Classroom	16	Discourse	Within, Between Indv.	Transformer (Deep neural network)
Hickman 2021	Cog/Ratnl.	Text (speech), acoustics, facial expressions	Lab, Online	1082	Personality	Between Indv.	Regularized regression
Hutt 2019	Cog/Ratnl.	Actions (clicks)	Online	69174	Engagement	Within Indv.	Standard ML (various)
De Choudhury 2013	Ratnl./Soc	Text (Twitter)	Online	476	Depression	Between Indv.	Support vector machine
Eichstaedt 2015	Ratnl./Soc	Text (Twitter)	Online	1347 (counties)	Heart disease	Between Grps.	Regularized regression
Das Swain 2020	Ratnl./Soc	Text (Glassdoor reviews)	Online	341	Organizational culture	Between Indv.	Linear regression
Robles-Granda 2021	Multiple	Text (Facebook), physiology, activity, location, comms, context	Home, office	757	Well-being, health, cog. ability and job performance	Between Indv.	Standard ML (various)

Bio. Biological; Cog. Cognitive; Ratnl.; Rational; Soc. Social; Indv. Individuals. Grps. Groups; ML. Machine learning.

Figure 4. Selective example MLCMs aligned with respect to Newell's four bands of action for the temporal granularity of the assessment and psychology domain for the assessed construct (top). Additional details for the examples (bottom).

Accuracy and Generalizability of MLCMs

MLCMs are typically evaluated across dimensions of accuracy and generalizability (defined above). Accuracy is higher with well-engineered features and sophisticated algorithms that can infer complex patterns without overfitting than simpler approaches which risk underfitting to the data. It is often assumed that "big data" is better, but this is an oversimplification; it is not the volume but what matters is the quality of the data and how well it represents the phenomenon to be measured. Another assumption is that multimodality improves accuracy, but this is not always the case (e.g., the Bosch study); often its main advantage is increasing robustness (D'Mello & Kory, 2015). All things equal, the quality of the annotations matters most because it provides the "supervisory" signal (Figure 2) for learning and performance evaluation. High-quality annotations should reach the same standards of construct validation as any psychological measure (e.g., reliability, convergent validity).

In terms of generalizability, MLCMs developed in very specific contexts are unlikely to generalize beyond the specific paradigm (e.g., Hassan affect-induction study), which can be somewhat alleviated by training on multiple stimuli/tasks (e.g., Zhang used both spatial and arithmetic tasks). The Hutt engagement study made domain generalizability a design principle in selecting features, and their MLCM trained on Math data generalized to Geometry data without retraining. Temporal generalizability is of concern for language models as new terms enter the lexicon (e.g., "COVID-19" for Jensen et al. (2021) which used 2018 data). The gold standard is to collect broad, diverse, and voluminous training data, such as the social media examples, but this is challenging for sensor-based models (e.g., Robles-Granda study) without mass surveillance. When applicable, as in the Jensen et al. (2021), example, starting with models pre-trained on large datasets across multiple domains and customizing them using limited data in a target domain is a promising approach.

Expectations of accuracy and generalizability must be calibrated with respect to the complexity of the construct and availability of quality training data (especially annotations). Accuracy will be higher for well-defined, experimentally-induced phenomena in the lab (Wager pain example) compared to spontaneously occurring, ill-defined phenomena in the wild (the Hutt engagement example). Similarly, generalizability is difficult when the phenomenon is highly context-specific, such as emotion (D'Mello et al., 2018). Here, it is prudent to learn context-specific models, live with modest accuracy, and temper performance claims rather than completely write off the approach. We suggest channeling Tukey when interpreting the value of such models: "far better an approximate answer to the right question, which is often vague, than the exact answer to the wrong question, which can always be made precise." (1962, p. 13).

Bias, Fairness, and Explainability of MLCMs

There was a recent media frenzy when it was revealed that commercial face recognition technology routinely underperforms for dark- compared to lighter-skinned individuals with shocking disparities (error rates up to 34.7% compared to 0.8%) (Buolamwini & Gebru, 2018). Whereas the idea of biased algorithms dates back to the 1970s, similar high-profile revelations have renewed interest in raising awareness of algorithmic bias and approaches to mitigate it.

Tay et al. (2021) proposed a theoretical framework that addresses bias in MLCMs for psychological assessment. They consider evidence for *bias* in a measurement context to arise when, for some subgroups, MLCM scores systematically depart from the actual scores, but there are *no* actual subgroup differences. Though often used interchangeably, *fairness* is distinct from bias. It is a subjective perspective based on the values and beliefs of individuals and societies.

As an example, consider a MLCM that assesses personality from automated video interviews, which are increasingly used in real-world hiring (e.g., Hickman et al., 2021). If the MLCM yields higher scores for men compared to women and nonbinary individuals on the personality dimension of conscientiousness when there are *no* gender differences in the annotations used to train the model (e.g., expert-rated conscientiousness), this would be *prima facie* evidence of bias. On the other hand, fairness is a broader subjective evaluation of a MLCM's predictions and its outcomes. If there are higher scores for men on the personality dimension of agreeableness in the actual annotations, and the MLCM reproduces this (i.e., it is not biased), some would view the MLCM as fair because its measurements reflect actual scores. Others would view it as unfair because it gives unequal group outcomes.

It is sometimes assumed that bias is purely a factor of the representativeness of data used to develop the models, but it arises from decisions made throughout the modeling process. The framework identifies and contextualizes potential sources of bias at both the data and algorithm level while also recommending tests and mitigation strategies.

A related concept is *explainability*, where the inner-workings of the model are interpretable by humans, a critical concern for both scientific inquiry and real-world use. Explainability can pertain to the structure of an MCLM itself (e.g., how do the features combine? what are the representations?) and/or the MLCM outputs (e.g., why did the model predict X for data point Y). The four modeling approaches in Figure 2 align along an explainability-performance tradeoff, with the hand-crafted models and deep learning approaches on either extreme. Whereas methods from the nascent field of explainable AI (XAI) can help improve the interpretability of MLCMs (e.g., Lundberg et al., 2020), it is unlikely that the tradeoff will be entirely eliminated akin to the *no-free-lunch-theorem* of mathematical folklore.

MLCMs in a Well Measured Life

What role do MLCMs play in an information age obsessed with measurement? As the examples illustrate, MLCM-based measures have been developed across multiple areas of psychological sciences ranging from neuroscience, cognitive/affective science, education, organizational personality, and personality/social psychology (Figure 4). They reflect measurements in the scanner, the lab, online, workplaces, homes, schools, and the community. Whereas most MLCMs focus on within- and between-individual differences, some produce measurements at the level of the team, organization, or society. MLCMs have been used for scientific inquiry, automated scoring, assessment, and intervention. They extend our capacity to harness natural data sources, in each case drastically increasing the speed, scale, and convenience of psychological measurement. Psychological scientists have a vital role to play in the future of MLCMs by providing guidance on human behavior, construct validity, statistical rigor, theoretical grounding, and evaluations of bias and fairness.

At the same time, a proliferation of such measures increases privacy, security, and ethical concerns over what and how data is collected, processed, and stored. It also raises long-established concerns of bias and fairness. Whereas researchers have historically emphasized accuracy and generalizability, achieving unbiased, fair, and interpretable models has garnered considerable interest over the past decade. As research and recommendations emerge, one immediate step is to adopt a culture where ethical design is a core goal. For example, the NSF National AI Institute on Student-AI Teaming[3] has adopted a Responsible Innovation Framework (Stilgoe et al., 2013) that guides its vision, values, methods, and success criteria. Of course, words must be followed with action in terms of how, when, why, from whom, and for what purpose are data collected and analyzed so that research artifacts (MLCMs here) are instruments that reflect and promote justice rather than perpetuate inequality.

Notes

[1] Address correspondence to Sidney D'Mello, 594 UCB, Boulder CO 80309, USA; sidney.dmello@colorado.edu

[2] This research was supported by the National Science Foundation (DRL 2019805; DRL 1920510; SES 2030599; SES 1928612; SES 1921111; IIS 1921087; IIS 1523091/ 1748739; DUE 1745442/1660877) and Institute of Education Sciences (R305A170432 and R305C160004). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

[3] The NSF National AI Institute for Student-AI Teaming (iSAT) is one of the seven inaugural AI Institutes. www.isat.ai

References

Bosch, N., Chen, H., Baker, R., Shute, V., & D'Mello, S. K. (2015). Accuracy vs. Availability Heuristic in Multimodal Affect Detection in the Wild. In *Proceedings of the 17th ACM International Conference on Multimodal Interaction (ICMI 2015)*. ACM.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Conference on fairness, accountability and transparency,

Card, S. K., Moran, T., & Newell, A. (1983). *The psychology of human-computer interaction*. Lawrence Earlbaum Associates.

D'Mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys*, *47*(3), 43:41-43:46.

D’Mello, S. K., Kappas, A., & Gratch, J. (2018). The Affective Computing Approach to Affect Measurement. *Emotion Review*, *10*(2), 174-183.

Das Swain, V., Saha, K., Reddy, M. D., Rajvanshy, H., Abowd, G. D., & De Choudhury, M. (2020). Modeling organizational culture with workplace experiences shared on glassdoor. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-15). ACM.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting Depression via Social Media. In *Proceedings of the International Conference on Web and Social Media (ICWSM-13)* (pp. 2).

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., . . . Sap, M. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, *26*(2), 159-169.

Faber, M., Bixler, R., & D'Mello, S. K. (2018). An automated behavioral measure of mind wandering during computerized reading. *Behavior Research Methods*, *50*(1), 134-150.

Fraiwan, L., Lweesy, K., Khasawneh, N., Wenz, H., & Dickhaus, H. (2012). Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier. *Computer Methods and Programs in Biomedicine*, *108*(1), 10-19.

Hassan, M. M., Alam, M. G. R., Uddin, M. Z., Huda, S., Almogren, A., & Fortino, G. (2019). Human emotion recognition using deep belief network architecture. *Information Fusion*, *51*, 10-18.

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2021). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*.

Hinaut, X., & Dominey, P. F. (2013). Real-time parallel processing of grammatical structure in the fronto-striatal system: A recurrent network simulation study using reservoir computing. *Plos One*, *8*(2), e52946.

Hutt, S., Grafsgaard, J., & D'Mello, S. K. (2019). Time to Scale: Generalizable Affect Detection for Tens of Thousands of Students across An Entire Schoolyear. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI 2019)*. ACM.

Hutt, S., Krasich, K., Brockmole, J., & D'Mello, S. K. (2021). Breaking out of the Lab: Mitigating Mind Wandering with Gaze-Based Attention-Aware Technology in Classrooms. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI 2021)*. ACM.

Jensen, E., Pugh, S., & D'Mello, S. K. (2021). A Deep Transfer Learning Approach to Automated Teacher Discourse Feedback In *Proceedings of the 11th Learning Analytics & Knowledge Conference (LAK 2021)*. ACM.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Le Cun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. *Nature*, *521*, 436-444.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., . . . Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, *2*(1), 56-67.

Newell, A. (1990). *Unified theories of cognition*. Harvard Univ Press.

Robles-Granda, P., Lin, S., Wu, X., D'Mello, S., K., Martinez, G. J., Saha, K., . . . De Choudhury, M. (2021). Jointly predicting job performance, personality, cognitive ability, affect, and well-being. *IEEE Computational Intelligence Magazine*, *16*(2), 46-61.

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research policy*, *42*(9), 1568-1580.

Subburaj, S. K., Stewart, A. E., Ramesh Rao, A., & D'Mello, S. K. (2020). Multimodal, Multiparty Modeling of Collaborative Problem Solving Performance. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (pp. 423-432). ACM.

Tay, L., Woo, S. E., Hickman, L., Booth, B., & D'Mello, S. (2021). A Conceptual Framework for Investigating and Mitigating Machine Learning Bias for Psychological Assessment (in review).

Tukey, J. W. (1962). The future of data analysis. *The annals of mathematical statistics*, *33*(1), 1-67.

Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C.-W., & Kross, E. (2013). An fMRI-based neurologic signature of physical pain. *New England Journal of Medicine*, *368*(15), 1388-1397.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer* (2nd, Ed.). Routledge.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100-1122.

Zhang, P., Wang, X., Zhang, W., & Chen, J. (2018). Learning spatial–spectral–temporal EEG features with recurrent 3D convolutional neural networks for cross-task mental workload assessment. *IEEE Transactions on neural systems and rehabilitation engineering*, *27*(1), 31-42.

Recommended Readings

1. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255-260. *Provides an accessible tutorial of machine learning and review of recent advances.*
2. Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. *Influential paper on building predictive (vs. explanatory) models in psychology.*
3. D'Mello, S. K., Kappas, A., & Gratch, J. (2018). The Affective Computing Approach to Affect Measurement. *Emotion Review*, 10(2), 174-183. *Provides a general tutorial on how to construct machine-learned computational models in the domain of emotion.*
4. Tay, L., Woo, S. E., Hickman, L., Booth, B., & D'Mello, S. (2021). A Conceptual Framework for Investigating and Mitigating Machine Learning Bias for Psychological Assessment (in review). *Provides a framework integrating psychometric concepts of bias with machine learning for psychological assessment.*
5. Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annual Review of Clinical Psychology*, 14(1), 91–118. *Review paper on using machine learning for assessment in clinical psychology.*