AI and Cognitive Testing: A New Conceptual Framework and Roadmap

Maithilee Kunda (mkunda@vanderbilt.edu)
Department of Electrical Engineering and Computer Science, Vanderbilt University
Nashville, TN 37235 USA

Abstract

Understanding how a person thinks, i.e., measuring a single individual's cognitive characteristics, is challenging because cognition is not directly observable. Practically speaking, standardized cognitive tests (tests of IQ, memory, attention, etc.), with results interpreted by expert clinicians, represent the state of the art in measuring a person's cognition. Three areas of AI show particular promise for improving the effectiveness of this kind of cognitive testing: 1) behavioral sensing, to more robustly quantify individual test-taker behaviors, 2) data mining, to identify and extract meaningful patterns from behavioral datasets; and 3) cognitive modeling, to help map observed behaviors onto hypothesized cognitive strategies. We bring these three areas of AI research together in a unified conceptual framework and provide a sampling of recent work in each area. Continued research at the nexus of AI and cognitive testing has potentially far-reaching implications for society in virtually every context in which measuring cognition is important, including research across many disciplines of cognitive science as well as applications in clinical, educational, and workforce settings.

Keywords: artificial intelligence; behavioral sensing; cognitive modeling; computational psychiatry; neuropsychology.

Introduction

The meat of the matter is often *how* a patient solves a problem or approaches a task rather than what the score is.
(Lezak et al., 2012, *Neuropsychological Assessment*, p. 160)

Different people think in different ways. This seemingly obvious statement masks many deep scientific mysteries about the human mind and also has enormous implications for individual and societal well-being.

How a person thinks is central to everything that they do: it affects how they learn, work, communicate, set goals, make decisions, etc. Thus, the scientific study of individual cognitive variations is critical not just for (1) advancing our basic understanding of human cognition and development across the lifespan, including research on genes, brain, and behavior, but also for (2) improving evidence-based practices in education and special education, workforce training, clinical diagnosis and treatment, rehabilitation, and more.

However, measuring cognition is uniquely challenging, as cognitive entities and processes are not observable in the same way that genetic, physiological, behavioral, and even neural characteristics can be measured using physical sensing technologies. **We have no way (at least at present) of *directly* measuring a person's mental representations.**

Even with advances in neuroimaging technologies that can capture subtle characteristics of neural activity, measuring such activity is only a rough proxy for actual cognitive activity; the question remains of how to "allow the brain measurements to make contact with putative cognitive processes" (Forstmann & Wagenmakers, 2015, p.144).

Currently, the gold standard for individual cognitive evaluations are those carried out by expert clinicians, usually psychologists or neuropsychologists.[1] These evaluations might be done to diagnose learning or developmental disabilities in children, detect signs of cognitive decline in elderly patients, or identify cognitive deficits after stroke or other brain injury (Lezak et al., 2012). Such evaluations combine two types of information about an individual: (1) information about how that individual is functioning *outside* the clinic, through self-report measures, interviews or questionnaires given to parents or caregivers, etc.; and (2) information about how that individual is functioning *inside* the clinic, usually through the administration of standardized cognitive tests, e.g., tests of memory, IQ, visuospatial reasoning, language, etc.

It is the second item in this list—cognitive testing—that is the focus of this paper. Distinct research paradigms within artificial intelligence (AI) have the potential to advance cognitive testing in (at least) three key ways:

1. *Behavioral sensing*: to more robustly quantify individual test-taker behaviors.
2. *Data mining*: to identify and extract meaningful patterns from behavioral datasets.
3. *Cognitive modeling*: to help map observed behaviors onto hypothesized cognitive strategies.

Before getting into the details of these three areas, however, it is important to first understand how conventional cognitive testing works. **This paper presents a new conceptual framework that explains the strengths and limitations of current methods for cognitive testing and highlights specific ways in which AI can help.** We also provide a sampling of recent AI research in each area.

[1] Cognitive evaluations also often occur in education and workforce settings, though these are typically less detailed but more domain-specific than clinical evaluations. In many human research studies across all areas of science, cognitive evaluations are used for participant inclusion/exclusion, group matching, and/or covariate analyses. While this paper focuses primarily on the clinical setting, our observations pertain to these other settings as well.

How Cognitive Testing Works

The rationale behind cognitive tests is straightforward. A given test poses problems for a test-taker to solve. Problems are specifically designed to tap certain cognitive representations and processes, which we refer to as *cognitive strategies*. Test designs are often validated (i.e. to "prove" that a test indeed is tapping into the right cognitive strategies) through converging evidence from many different sources, including data from neuroimaging studies, patients with known cognitive or neurological issues, and/or other cognitive tests. A person's test score thus provides an indirect measure of these hypothesized cognitive strategies.

However, a well known issue with most cognitive tests is **ambiguity**: while test scores do indicate *how well* a person solves test problems, i.e., that person's level of ability, they do not indicate *how* a person solves test problems, i.e., their actual cognitive strategy. In other words, two people can get the same test score using very different cognitive strategies. Moreover, this ambiguity can occur with low or high scores:

> "'There are many reasons for failing and there are many ways you can go about it. And if you don't know in fact which way the patient was going about it, failure doesn't tell you very much' (Darby & Walsh, 2005). There can also be more than one way to pass a test." (Lezak et al., 2012, p. 160)

Because of this ambiguity, expert clinicians often combine scores with other observed behaviors, such as errors, eye gaze, emotions, general demeanor, etc., in order to better interpret a person's test performance. This supports the rationale for why only "trained" clinicians should administer cognitive tests, and also why clinicians develop such deep expertise with their particular population and goal (e.g. screening children for learning disabilities versus working with elderly patients to detect memory issues).

In reality, as mentioned in the introduction, clinicians likely never rely on results from a single cognitive test to make judgments about a person's cognition. They combine results from many tests with additional information about a person's performance outside the clinic (e.g. school performance, medical history, etc.). For the purposes of this paper, however, we focus on thinking about just a single cognitive test and what it can tell us.

Proposed framework

In this section, we propose a new formalism for describing what is happening during a conventional cognitive test. For added clarity, we also use the Raven's Progressive Matrices (RPM) intelligence test as a running example. The RPM is a well studied standardized test that poses problems similar to geometric analogies: a matrix of visual figures is presented with one missing, and the missing figure must be chosen from among a set of candidate answer figures (i.e., multiple choice). The RPM is one of the best single-format measures of intelligence among all cognitive tests (Snow, Kyllonen, & Marshalek, 1984) and thus is very widely used.

Definition 1. Let the set X_{human} represent all possible cognitive strategies that a person can use to attempt to perform a given cognitive test, successful or not.

Definition 2. Let the set Y represent all possible scores that can be earned on a given cognitive test.

Definition 3. Let the function F represent a mapping from a person's use of a particular cognitive strategy onto the resulting test score:

$$F(x_i \in X_{human}) = y_i \in Y$$

We do not concern ourselves with how X_{human} might be represented. The set is infinite, even if we exclude obviously irrelevant strategies.[2] An individual person probably can access at least a few strategies from X_{human}, and certainly they can also be taught to use particular strategies.

Though not, perhaps, designed this way on purpose, the RPM is amenable to multiple distinct strategies. For example, there is evidence that many neurotypical individuals often use verbal, inner-speech-like strategies, whereas many individuals on the autism spectrum use visually mediated, mental-imagery-like strategies (Soulières et al., 2009). In fact, some argue that the reason the RPM is such a good intelligence test may be because it is actually testing metacognitive flexibility, in terms of strategy selection/adaptation (Kirby & Lawson, 1983)...a point that we return to later on in this paper.

For simplicity, let us assume that a person uses a single strategy $x_i \in X_{human}$ to solve a given cognitive test. Using this strategy x_i, they receive a score y_i. In other words, the act of taking the test is what "computes" the function F.

In the case of the RPM, the test is scored as number of correct answers, and so possible scores (for the standard version of the test) range from 0 to 60. So, suppose someone uses a verbally mediated strategy, and they get a score of 50/60.

Using these definitions, ambiguity exists because F is a many-to-one function. There are many possible strategies in X_{human} that may lead to a score of 50. As a result, the inverse function $F^{-1}(y_i) = x_i \in X_{human}$ is ill-defined.

To help with this problem, we expand our definitions to include additional test-taker behaviors, beyond just test score:

Definition 4. Let the vector B_{human} represent a sequence of observable behaviors generated by a person taking a cognitive test, including test score y as well as response times, types of errors made, patterns of eye gaze, etc.

Definition 5. Let the function G represent a mapping from a person's use of a particular cognitive strategy onto the sequence of resulting behaviors:

$$G(x_i \in X_{human}) \rightarrow B_{human}$$

For example, for a person taking the RPM, one might include in B_{human} the time taken to complete each problem, the

[2] Making a peanut butter and jelly sandwich is one possible strategy for solving RPM problems. It is, however, an exceedingly poor strategy, and so let's exclude it from X_{human}.

answer choice that is selected, the pattern of eye gaze between different visual elements, etc.

Now, while the function G is still a many-to-one function (i.e., multiple strategies might still map onto the same sequence of behaviors), it is "less" many-to-one than our earlier function F that mapped strategies onto scores. Each behavioral observation that is made places an additional constraint on the subset of strategies in X_{human} that could have produced the full sequence of behaviors. Therefore, given a sequence of observable behaviors B_{human}, the inverse function $G^{-1}(B_{human}) = x_i \in X_{human}$ provides a better estimate of a person's cognitive strategy than does the inverse function F^{-1} that relies on test score alone.

For example, research on geometric analogies has shown that different patterns of eye gaze seem to be indicative of different high-level problem-solving strategies (Bethell-Fox, Lohman, & Snow, 1984). Some people look at the "problem" part and come up with their own answer before looking at the answer choices, while others look at the answer choices early and use more of a trial-and-error approach, mentally plugging in each answer choice to see which one looks best.

One problem remains: where does the sequence of behaviors B_{human} come from? For traditional cognitive tests, usually administered in a pencil-and-paper or objects-on-a-table format, there is no perfect record of B_{human}. Clinicians observing a person taking a test use their own, expertly trained powers of perception, memory, and note-taking to process B_{human} in real time in order to extract meaningful patterns:

Definition 6. Let the function P represent a mapping from a sequence of low-level behaviors B_{human} to a selected set of patterns (i.e., a subset and/or transformed view of individual observations in B_{human}).

We use P_{expert} to denote the function that a clinician applies to extract meaningful patterns from the raw behavioral sequence B_{human}. **Thus, when a clinician observes a person's test performance to infer information about that person's cognition, they are implicitly computing the function:**

$$G^{-1}_{expert}(P_{expert}(B_{human})) = x_i \in X_{human} \quad (1)$$

Where do the functions G^{-1}_{expert} and P_{expert} come from? In general, they are learned over years or decades of administering cognitive tests to certain segments of the population. For example, a clinician with expertise in learning disabilities likely uses G^{-1} and P functions that are tuned to patterns of behavior most relevant for diagnosing these conditions in children. Another clinician who works mostly with brain injury patients would likely use different G^{-1} and P functions, even when administering similar tests.

The problem with implicit functions, and current, non-AI-based solutions

The main problem with these learned G^{-1}_{expert} and P_{expert} functions is that they are implicit in a clinician's expertise. Not only are they implicit, but they are also very difficult to make explicit, even if a clinician tries to do so. This difficulty in turn complicates efforts to measure the validity or reliability of these functions, both for individual clinicians and for the field of cognitive assessment as a whole.

The Boston Process Approach to neuropsychology was essentially an attempt to "write down" these functions using a combination of expert judgment and carefully designed research studies, so that the resulting functions could be more rigorously evaluated for validity and reliability, and also so these functions could be explicitly taught as part of professional neuropsychology training. However, while the ideas of the Boston Process Approach have been influential, the complexity of its methods and the challenges of real-time data collection during testing sessions limited its widespread adoption (Milberg, Hebben, Kaplan, Grant, & Adams, 2009).

The advent of computer-based testing has provided new methods for recording sequences of test-taker behaviors, such as detailed reaction times, errors, etc. Some , like the California Verbal Learning Test (Delis, Freeland, Kramer, & Kaplan, 1988), have been designed specifically to enable the use of these additional behaviors to infer more and better information about a person's cognitive strategy than would be obtainable from their score alone.

These and similar efforts from the neuropsychology research community have been analyzed more recently under the heading of the Quantified Process Approach (Poreh, 2012), which emphasizes the critical need to understand cognitive strategies, i.e. "process," using quantifiable measures, in addition to the subjective and often qualitative judgments of individual clinicians (what we describe here as the implicit G^{-1}_{expert} and P_{expert} functions). The Quantified Process Approach outlines three categories of potential solutions: 1) using additional tests to essentially triangulate a person's strategy using multiple points of measurement; 2) using additional measures of behavior from a single test to develop new indices of interest; and 3) decompose scores into subscores that might reflect different underlying factors. Of these three categories, the latter two would fall into our proposed framework as efforts to come up with explicit G^{-1} and P functions, depending on whether the behaviors B_{human} considered are taken from behavioral dimensions above and beyond scores (category 2) or from behavioral dimensions within scores that pinpoint more detailed subscores (category 3).

However, these various pockets of research have yet to transform the daily practice of cognitive testing. Problems remain in how to quantify G^{-1} and P functions in a scalable way that can be applied across many different cognitive tests and many populations, while also ensuring that methods are readily usable by practicing clinicians.

AI to the Rescue[3]

Using this framework, we now describe ways in which AI can help solve some of these problems through 1) behavioral sensing, 2) data mining, and 3) cognitive modeling.

[3]Possibly.... https://xkcd.com/1831/

Behavioral sensing

The first, and perhaps most obvious, role for AI in cognitive testing is in recording behavioral observations, i.e., in obtaining the sequence of behaviors B_{human} from a test session.

Part of behavioral sensing involves advances in hardware, such as the development of more advanced (and more affordable) eye trackers. Computer-based testing platforms can easily log many kinds of behaviors, including mouse movements, key presses, etc. Tablet-based tests are being used to capture more detailed manual behaviors such as velocity of pen strokes (Davis, Libon, Au, Pitman, & Penney, 2014).

While behavioral sensing in computer-based environments is currently more common, one of the most exciting new areas for behavioral sensing involves sensing in real, 3D environments, which often calls for a combination of advances in hardware and in signal processing algorithms. Eye tracking technology is now getting to the point where head-mounted eye trackers are relatively lightweight and affordable (Kassner, Patera, & Bulling, 2014), and computer vision algorithms can be used to help analyze the video stream coming from such eye trackers. These advances enable scalable eye-tracking in 3D environments, which, in previous years, would have been virtually unthinkable in the context of cognitive testing from usability or scalability perspectives. Physiological sensors are also now often incorporated into cognitive assessments, e.g., using skin conductance sensors to obtain measurements of heart rate, etc. as a proxy for measuring cognitive stress or other affective variables during a testing session (Fletcher et al., 2010).

In addition, even data recorded from regular sensors (cameras, microphones, etc.) can now be analyzed automatically using AI algorithms coming from computer vision, natural language processing, etc. The term *behavioral imaging* has been coined to describe this new subfield of AI directed at producing robust and reliable measurements of human behavior in 3D assessment settings (Rehg et al., 2014).

Behavioral sensing can thus be understood in terms of its two components: sensors to record raw signals coming from a testing session(e.g., pixels from a video camera), plus algorithms to process those raw signals into measurements of behavior (e.g., computer vision algorithm to detect, from a raw video stream, when a person moves an object on a table).

Behavioral sensing can help in measuring many types of behaviors. Some behaviors are already easily measured by humans, but automated approaches may increase the scalability or accuracy of such measurements (e.g., counting how many errors a person makes while solving a table-top block copying task). Other behaviors might be currently detectable by human clinicians but only in qualitative ways. For example, many social assessments for the diagnosis of autism use "quality of eye contact" as a measurement of interest, which is often recorded as a subjective overall impression by a human clinician, but could be broken down into quantified components by an algorithm (Ye et al., 2015). Still other behaviors might not be detectable by human clinicians at all; for example, being able to capture the exact velocities and pressures manually applied by a person performing a tablet-based drawing test (Davis et al., 2014).

Data mining

The next role for AI is in quantifying the function P that takes in a sequence of behaviors B_{human} and extracts meaningful patterns. Meaningful patterns can be created in many different ways, including by identifying subsets of behaviors that are particularly relevant, or by producing transformations of low-level behaviors into higher-level constructs.

For example, there has been work that first uses a tablet-based version of the clock drawing test to record low-level manual drawing behaviors, and then applies machine learning classification algorithms to these data to help diagnose Alzheimer's, Parkinson's, and other cognitive conditions (Souillard-Mandar et al., 2016).

In another effort, eye tracking data from a visual recognition test (the Visual Paired Comparison test) have been used to train classifiers to detect early signs of mild cognitive impairment, which is often a precursor to Alzheimer's (Lagun, Manzanares, Zola, Buffalo, & Agichtein, 2011). A clever extension of this work aims to see if mouse movement data from a non-eye-tracking variant of the task can support comparable classification performance, which would greatly increase the scalability of the test by removing the need for an eye tracker (Agichtein et al., 2017).

In general, the broad umbrella of data mining approaches for cognitive testing can include the use of: 1) new algorithms applied to existing behavioral datasets; 2) conventional statistical analyses applied to new behavioral datasets; and 3) new algorithms applied to new datasets. All of these approaches represent important routes for improving our understanding of the low-level behaviors that come out of cognitive tests, i.e., to identify which behaviors or combinations of behaviors are most important for a given clinical goal.

Cognitive modeling

The third important role that AI can play in cognitive testing is through cognitive modeling. What does a computational cognitive model actually accomplish? To answer this question, we begin by supposing that we have created a particular type of AI system—a computational cognitive architecture—that can employ different problem-solving strategies to solve problems from a given cognitive test.

Critically, such an AI system is not just a mathematical model of relationships between hypothesized cognitive entities involved in solving the test. It is a computational model of the hypothesized entities themselves; it provides a mechanism-level view of what might be going on. The key difference between a mathematical model and a computational model is that a computational model bears an analogical relationship with what it is trying to model; there is some structural correspondence between the model and what it represents (Hunt, Ropella, Park, & Engelberg, 2008).

Definitions 7 through 12 (below) refer to concepts related to this kind of computational model, which are also analogous (but not identical) to the concepts given in Definitions 1 through 6 (above) for human test-takers.

Definition 7. Let X_{AI} represent the set of problem-solving strategies that an AI system can use to solve a given cognitive test, including successful and unsuccessful strategies.

Definition 8. Let y_{AI} represent the score the AI system receives on a given cognitive test.

Definition 9. Let the function F_* represent a mapping from an AI system's use of a particular strategy onto the resulting test score, i.e., $F_*(x_i \in X_{AI}) \rightarrow y_{AI}$.

Definition 10. Let B_{AI} represent the sequence of *simulated* observable behaviors b_i generated by an AI system taking a cognitive test. These behaviors can include test scores y_{AI} as well as response times, types of errors made, patterns of eye gaze, etc.

Definition 11. Let the function G_* represent a mapping from an AI system's use of a particular strategy onto the resulting test score plus simulated behaviors, i.e., $G_*(x_i \in X_{AI}) \rightarrow B_{AI}$.

Definition 12. Let the function P_* represent a mapping from a sequence of low-level behaviors B_{AI} to higher-level features.

To take our previous example of the Raven's Progressive Matrices test, many computational cognitive models of this kind have been developed over the years (Carpenter, Just, & Shell, 1990; Lovett, Tomai, Forbus, & Usher, 2009; Kunda, McGreggor, & Goel, 2013; Strannegård, Cirillo, & Ström, 2013). There has also been much work in the cognitive architectures community (e.g. using SOAR, ACT-R, etc.) to develop richly detailed models of many different tasks.

Given such a computational cognitive model, we can run experiments that have the model use a variety of different strategies to solve a given cognitive test. We can measure data from these experiments to obtain test scores and behaviors, just as we do for human test takers. **The key difference here is that cognitive strategies in a cognitive model are directly observable!** We have the "ground truth" for our model in a way that is (at least currently) impossible to obtain for human test takers.

At minimum, we can study the function F_* to understand more about potential ambiguities on a particular cognitive test, which would itself a valuable contribution to the field of cognitive testing.

Also, such a cognitive model provides a systematic way to obtain quantified functions for mapping from the space of observed behaviors back onto cognitive strategies, i.e., the function G^{-1}. This is still not easy (though it is much easier when we have the ground truth for X!). There are probably many possible approaches for obtaining the G^{-1} function. One might be to run a large set of computational experiments to get two linked datasets X_{AI} and B_{AI}, and then use machine learning and data mining algorithms to find relevant patterns and predictors within these.

One important area for research using computational cognitive models is to more effectively capture individual differences. Much of the research on cognitive architectures, for example, focuses on modeling generalized human performance or broad group differences. As the quantity and quality of behavioral measurements increase, through behavioral sensing and data mining, cognitive models should also be able to take advantage of these datasets to create more precise explanations of individual variations.

Another extremely interesting open question is: where do the strategies in X_{AI} come from? For now, X_{AI} is defined by the AI system's designers, informed by research on human cognition. An important AI frontier is to develop AI systems that *learn* strategies through instruction, observation, and experience, as people do (Laird et al., 2017). This research would not only expand the capabilities of our cognitive models, but results would also help us better understand human cognitive strategies at the metacognitive level. As mentioned earlier, for example, work on the Raven's Progressive Matrices test suggests that a person's methods for strategy selection are just as important for test performance as are the strategies themselves (Kirby & Lawson, 1983).

A Call to Action

Similar observations have been compiled under the heading of computational psychiatry (Montague, Dolan, Friston, & Dayan, 2012; Huys, Maia, & Frank, 2016), though the specific formalism given here is (to our knowledge) new.

What our analysis suggests is that interdisciplinary collaboration is critical for advancing the science of cognitive testing, not just between clinicians and AI researchers in general, but between clinicians and AI researchers coming from the distinct subfields of behavioral sensing, data mining, and cognitive modeling.

In addition, one extremely promising horizon is to think about the development of new cognitive tests that are enabled by the types of technological advances described above. For example, now that we can measure and understand very rich sets of behavior, and also map these onto detailed hypotheses about cognitive strategies, can we begin to measure complex forms of cognition in more naturalistic tasks? So much of current test design has been shaped by the limitations in the scalability of these elements in previous decades. Previously, cognitive test designers had to construct very constrained tasks, that would only measure one or two cognitive constructs at a time, and that would produce easily measurable scores. Now, for example, could we give people a realistic search task in a complex, 3D environment to test their attention and/or memory? There is a great opportunity here to begin coming up with much more creative and naturalistic ways to tap into a person's realistic cognitive processes.

Acknowledgment

This work was supported in part by NSF Award #1730044.

References

Agichtein, Y. E., Buffalo, E. A., Lagun, D., Manzanares, C., & Zola, S. (2017, April 25). *Internet-based cognitive diagnostics using visual paired comparison task.* Google Patents. (US Patent 9,629,543)

Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence*, *8*(3), 205–238.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test. *Psychological review*, *97*(3), 404.

Darby, D., & Walsh, K. W. (2005). *Walsh's neuropsychology: A clinical approach*. Churchill Livingstone.

Davis, R., Libon, D. J., Au, R., Pitman, D., & Penney, D. L. (2014). Think: Inferring cognitive status from subtle behaviors. In *Aaai* (pp. 2898–2905).

Delis, D. C., Freeland, J., Kramer, J. H., & Kaplan, E. (1988). Integrating clinical assessment with cognitive neuroscience: construct validation of the california verbal learning test. *Journal of consulting and clinical psychology*, *56*(1), 123.

Fletcher, R. R., Dobson, K., Goodwin, M. S., Eydgahi, H., Wilder-Smith, O., Fernholz, D., ... Picard, R. W. (2010). icalm: Wearable sensor and network architecture for wirelessly communicating and logging autonomic activity. *IEEE Transactions on Information Technology in Biomedicine*, *14*(2), 215–223.

Forstmann, B. U., & Wagenmakers, E.-J. (2015). Model-based cognitive neuroscience: A conceptual introduction. In *An introduction to model-based cognitive neuroscience* (pp. 139–156). Springer.

Hunt, C. A., Ropella, G. E., Park, S., & Engelberg, J. (2008). Dichotomies between computational and mathematical models. *Nature biotechnology*, *26*(7), 737.

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience*, *19*(3), 404.

Kassner, M., Patera, W., & Bulling, A. (2014). Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 acm international joint conference on pervasive and ubiquitous computing: Adjunct publication* (pp. 1151–1160).

Kirby, J. R., & Lawson, M. J. (1983). Effects of strategy training on progressive matrices performance. *Contemporary Educational Psychology*, *8*(2), 127–140.

Kunda, M., McGreggor, K., & Goel, A. K. (2013). A computational model for solving problems from the ravens progressive matrices intelligence test using iconic visual representations. *Cognitive Systems Research*, *22*, 47–66.

Lagun, D., Manzanares, C., Zola, S. M., Buffalo, E. A., & Agichtein, E. (2011). Detecting cognitive impairment by eye movement analysis using automatic classification algorithms. *Journal of neuroscience methods*, *201*(1), 196–203.

Laird, J. E., Gluck, K., Anderson, J., Forbus, K. D., Jenkins, O. C., Lebiere, C., ... others (2017). Interactive task learning. *IEEE Intelligent Systems*, *32*(4), 6–21.

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment, fifth edition*. Oxford University Press, USA.

Lovett, A., Tomai, E., Forbus, K., & Usher, J. (2009). Solving geometric analogy problems through two-stage analogical mapping. *Cognitive science*, *33*(7), 1192–1231.

Milberg, W. P., Hebben, N., Kaplan, E., Grant, I., & Adams, K. (2009). The boston process approach to neuropsychological assessment. *Neuropsychological assessment of neuropsychiatric and neuromedical disorders*, 42–65.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in cognitive sciences*, *16*(1), 72–80.

Poreh, A. M. (2012). *The quantified process approach to neuropsychological assessment*. Psychology Press.

Rehg, J. M., Rozga, A., Abowd, G. D., & Goodwin, M. S. (2014). Behavioral imaging and autism. *IEEE Pervasive Computing*, *13*(2), 84–87.

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. *Advances in the psychology of human intelligence*, *2*, 47–103.

Souillard-Mandar, W., Davis, R., Rudin, C., Au, R., Libon, D. J., Swenson, R., ... Penney, D. L. (2016). Learning classification models of cognitive conditions from subtle behaviors in the digital clock drawing test. *Machine learning*, *102*(3), 393–441.

Soulières, I., Dawson, M., Samson, F., Barbeau, E. B., Sahyoun, C. P., Strangman, G. E., ... Mottron, L. (2009). Enhanced visual processing contributes to matrix reasoning in autism. *Human brain mapping*, *30*(12), 4082–4107.

Strannegård, C., Cirillo, S., & Ström, V. (2013). An anthropomorphic method for progressive matrix problems. *Cognitive Systems Research*, *22*, 35–46.

Ye, Z., Li, Y., Liu, Y., Bridges, C., Rozga, A., & Rehg, J. M. (2015). Detecting bids for eye contact using a wearable camera. In *Automatic face and gesture recognition (fg), 2015 11th ieee international conference and workshops on* (Vol. 1, pp. 1–8).