# Computer Vision Scoring of Figure Copy and Recall

**David L. Woods[1], Kathleen Hall[1], Isabella Jaramillo[1], Mike Blank[1], Kristin Geraci[1], Andrew Boghossian[1], and Peter Pebler[1]**

*[1]Neurobehavioral Systems, Inc., Berkeley, CA*

## Introduction

**Figure-drawing tests are among the most sensitive tools in neuropsychology, but labor-intensive and subjective scoring limits their clinical use and scalability for clinical trials. We developed a scalable, objective, and fully reproducible computer-vision scoring system that scored 9,117 drawings from 2,011 adults with accuracy indistinguishable from expert human raters.**

## Methods

**Participants.** A diverse normative sample of 2,011 community-dwelling adults ages 18–90 (mean 52.7 years; 37% White) completed the Figure Drawing and Recall task, a subtest of the California Cognitive Assessment Battery (CCAB).

**Apparatus.** Participants copied a 12-element figure with their index fingers on a touchscreen tablet computer, with the figure displayed on half of the screen and a blank canvas on the other half. Finger position was continuously recorded. In delayed recall conditions, participants were instructed to reproduce the drawing from memory.

**Vertex AI training.** Rendered drawings were homographically warped onto a canonical 512 x 512 pixel template to optimize alignment with the reference figure. One author (DLW) trained a convolutional object-detection model on the Vertex AI AutoML Vision platform in iterative training cycles by labeling bounding boxes surrounding approximately 20,000 elements in 1,774 randomly selected drawings.

**Vertex AI scoring.** Once trained, Vertex AI returned a confidence measure for each element present along with its bounding box coordinates. Element presence was scored as 1 or 0 based on element-specific Vertex AI confidence thresholds. Element location was scored as 1 if the center of mass of the detected bounding box fell within the corresponding template bounding box, 0.5 if the two boxes overlapped, and 0 otherwise. Element presence and location were summed to yield a total score. Vertex AI scores were obtained from the full set of 9,117 drawings.

**Human Scoring.** Three expert human raters were trained to use identical scoring rules. Each rater independently scored 500 randomly selected drawings for comparison with Vertex AI, including a common subset of 100 drawings scored by Vertex AI and all human raters.

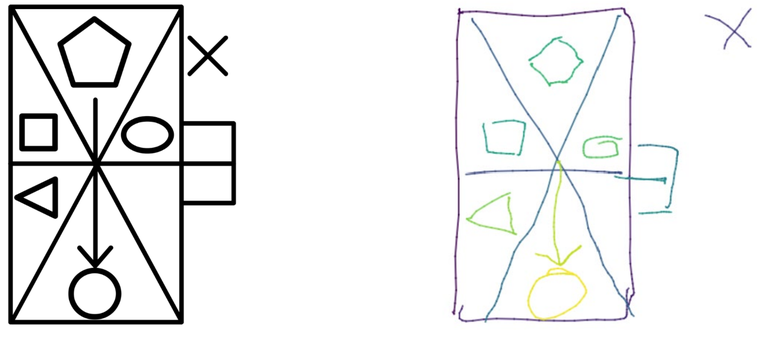

**Figure 1. LEFT. The CCAB figure drawing template.** The figure consists of twelve canonical elements arranged within a bounding rectangle: the outer large rectangle (LR); two crossing diagonals forming a large X (LX); a pentagon (PN); a small square (SS); an oval to the right of center (OV); a horizontal midline bisecting the large rectangle (HL); a triangle (TR); a vertical shaft (VS) terminating in an arrowhead (AH); a circle (CO); an attached external box (EB); and a small X outside the bounding rectangle (SX). Each element occupies a fixed canonical position, size, and shape and was annotated as a discrete object with a bounding box for use in training the Vertex AI object-detection model. **RIGHT.** Drawing from one subject in the Copy condition. Line color denotes drawing latency.

## Results

**Agreement of Vertex AI with human raters.** The mean Pearson correlation among the three human raters on Total scores across all trials (computed on the 100-drawing overlap) was r = 0.972, while the mean Vertex–human vs. human correlation across the three rater corpora was r = 0.966, corresponding to a Vertex–human gap of approximately 0.006. For individual element scores, Vertex element presence scores concorded with modal human scores (range 90%–100%), while Location scores showed lower concordance (65%–100%) because Vertex AI made objective pixel-level decisions that human raters could not match.

**Demographic effects.** Performance declined with age, with a substantially larger influence on Recall (r = −0.32) than on Copy (r = −0.16 Higher education and vocabulary scores were associated with better performance with larger correlations on Copy than Recall for both education (r = +0.19 Copy vs. +0.13 Recall) and vocabulary ( r = +0.25 Copy vs. +0.13 Recall).

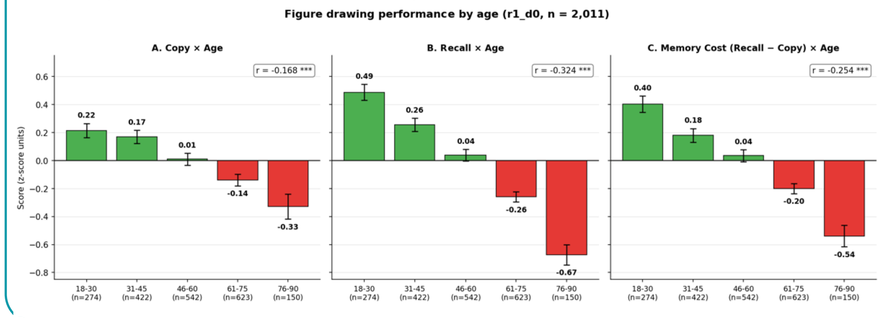


**Figure 2. Vertex AI scores of figure drawing performance by age quintile (n = 2011).** **Left:** Total (location +element) z-scores for the Copy condition. **Center:** Total (location + element) z-scores for the Recall condition. **Right:** A Memory Cost z-score (Recall − Copy) that isolates the memory component independently of copy performance.

## Key Findings

- **Unlike other AI scoring approaches that output a single black-box score, computer-vision methods replicate manual protocols, providing drawing quality and location scores for each picture element as well as summary scores.**
- **Automated scores agreed with those from expert human raters as closely as human experts agree with each other.**
- **Automated scoring captured the figure-drawing memory signal: Delayed recall performance showed approximately twice the age-related decline observed for the Copy trial, revealing an age-dependent "memory cost" score.**
- **Objective and fully reproducible computer vision scoring makes Figure Drawing tests scalable for clinical trials and eliminates barriers to use in clinical assessment.**

**Additional information is available including methodological details, test-retest reliability, duplicate element scoring, drawing organization scoring, retest learning effects, and kinetic analyses of drawing motion.**

Woods, D. L., Hall, K., Jaramillo, I., Blank, M., Geraci, K., Boghossian, A., & Pebler, P. (2026). Computer vision scoring of figure copy and recall [Preprint]. *medRxiv*. *https://doi.org/10.64898/2026.06.10.26355298*

## Discussion

Automated scoring of the CCAB figure drawing task offers four concrete advantages over manual scoring.

1. The pipeline requires no additional time from the examiner: drawings are scored automatically without inter-rater variability or drift.
2. Scores are fully replicable: the same drawing processed twice by the same pipeline produces identical scores, without intra-rater drift or disagreement among reviewers.
3. The pipeline applies the location-scoring rubric with single-pixel precision that expert human raters cannot match.
4. The pipeline also provides motor and timing measures unavailable through manual scoring.

- **Summary.** Other convolutional neural networks approaches (Andersen et al., 2026) report agreement against expert raters similar to that of the Vertex AI pipeline described here, but provided only total scores. Similarly, hybrid feature-engineering systems (e/g/. DCTclock, Souillard-Mandar et al., 2021) report multiple hybrid scores, , leaving element-level scores unreviewable. In contrast, the Vertex AI pipeline mirrors manual scoring decisions and preserves what end-to-end and hybrid systems discard: every scoring decision is exposed at the element level, visualized as a labeled bounding box on the drawing, and can be directly compared to manually scored data.

## References

Andersen, S. L., Lundervold, A. J., & Ronold, E. H. (2026). Automated versus human scoring of the Rey-Osterrieth Complex Figure Test: A rapid review. *Frontiers in Psychiatry, 16*, 1746720.

Souillard-Mandar, W., Penney, D., Schaible, B., Pascual-Leone, A., Au, R., & Davis, R. (2021). DCTclock: Clinically-Interpretable and Automated Artificial Intelligence Analysis of Drawing Behavior for Capturing Cognition. *Frontiers in Digital Health, Volume 3 - 2021.*

## Contact us

drdlwoods@neurobs.com
ccabresearch.com
neurobs.com

## Visit us at booth 1417!

**Supported by NIA R44AG062076, NIA R44AG080951**