Journal of the International Neuropsychological Society (2024), **30**, 992–997
doi:10.1017/S1355617724000584

Research Article

Simplifying Complex Figure scoring: Data from the Emory Healthy Brain Study and initial clinical validation

David W. Loring[1,2], Najé Simama[1], Katherine Sanders[1], Jessica R. Saurman[1], Liping Zhao[3], James J. Lah[1] and Felicia C. Goldstein[1]

[1]Department of Neurology, Emory University School of Medicine, Atlanta, GA, USA, [2]Department of Pediatrics, Emory University School of Medicine, Atlanta, GA, USA and [3]Department of Biostatistics and Bioinformatics, Rollins School of Public Health (Emory University), Atlanta, GA, USA

Abstract

Objective: To introduce the Emory 10-element Complex Figure (CF) scoring system and recognition task. We evaluated the relationship between Emory CF scoring and traditional Osterrieth CF scoring approach in cognitively healthy volunteers. Additionally, a cohort of patients undergoing deep brain stimulation (DBS) evaluation was assessed to compare the scoring methods in a clinical population. **Method:** The study included 315 volunteers from the Emory Healthy Brain Study (EHBS) with Montreal Cognitive Assessment (MoCA) scores of 24/30 or higher. The clinical group consisted of 84 DBS candidates. Scoring time differences were analyzed in a subset of 48 DBS candidates. **Results:** High correlations between scoring methods were present for non-recognition components in both cohorts (*EHBS*: Copy $r = 0.76$, Immediate $r = 0.86$, Delayed $r = 0.85$, Recognition $r = 47$; *DBS*: Copy $r = 0.80$, Immediate $r = 0.84$, Delayed Recall $r = 0.85$, Recognition $r = 0.37$). Emory CF scoring times were significantly shorter than Osterrieth times across non-recognition conditions (all $p < 0.00001$, individual Cohen's d: 1.4–2.4), resulting in an average time savings of 57%. DBS patients scored lower than EHBS participants across CF memory measures, with larger effect sizes for Emory CF scoring (Cohen's d range = 1.0–1.2). Emory CF scoring demonstrated better group classification in logistic regression models, improving DBS candidate classification from 16.7% to 32.1% compared to Osterrieth scoring. **Conclusions:** Emory CF scoring yields results that are highly correlated with traditional Osterrieth scoring, significantly reduces scoring time burden, and demonstrates greater sensitivity to memory decline in DBS candidates. Its efficiency and sensitivity make Emory CF scoring well-suited for broader implementation in clinical research.

Keywords: Assessment; Visual construction; Visual memory; Recognition memory; Deep brain stimulation; Inter-rater reliability

(Received 28 May 2024; final revision 7 October 2024; accepted 7 October 2024; First Published online 14 November 2024)

Introduction

The Rey Complex Figure (CF) is a widely used test of visual constructional ability and visual memory (Rabin et al., 2016). Successful CF performance relies on several cognitive domains, with visual–perceptual skills and executive functions being the most crucial for accurate copying. These abilities in addition to memory also play a key role in CF recall (Beebe et al., 2004; Temple et al., 2006).

The CF was developed by Swiss psychologist André Rey to detect cognitive impairments (Corwin & Bylsma, 1993; Rey, 1941). Rey identified four primary CF elements for scoring –the diamond, circle, and line groupings in the upper left and lower right quadrants – each worth 2 points. Additional line segments were given a value of 1 point each, with a total possible CF score of 47.

Paul-André Osterrieth, a student of Rey, revised CF scoring to simplify the process and reduce ambiguity (Osterrieth, 1944). While retaining several individual lines, Osterrieth prioritized scoring larger, more complex components. His system evaluates 18 CF elements based on accuracy and placement, with a maximum score of 36 points. Osterrieth's method remains the most commonly used for scoring the CF, although alternative quantitative (Breier et al., 1996; Denman, 1984; Fastenau, 1996; Waber & Holmes, 1985) and qualitative (Loring et al., 1988; Stern et al., 1999) scoring approaches have also been developed.

The CF has a rich neuropsychology history, although its use in clinical research protocols is limited due to the extensive training and time burden needed to ensure scoring accuracy (NINDS, 2022). Consequently, simpler figures with fewer elements have been developed (de Paula et al., 2016; Poreh et al., 2020; Possin et al., 2011; Randolph, 1998). While simplified designs shorten administration and scoring, they also reduce organizational complexity, potentially decreasing its sensitivity to executive function deficits. As Osterrieth (1944) noted, simple copy tasks "have the advantage of being simple, … (but) past the age of success, they no longer give much useful information as to the organization of perception" (*Ces tests de copie ont l'avantage d'être*

Corresponding author: David W. Loring; Email: dloring@emory.edu

Cite this article: Loring D.W., Simama N., Sanders K., Saurman J.R., Zhao L., Lah J.J., & Goldstein F.C. (2024) Simplifying Complex Figure scoring: Data from the Emory Healthy Brain Study and initial clinical validation. *Journal of the International Neuropsychological Society*, **30**: 992–997, https://doi.org/10.1017/S1355617724000584

simples, . . . et passé l'âge de réussite ils ne donnent plus guère de renseignements utiles quant à l'organisation de la perception, p. 209).

Osterrieth did not describe how CF scoring elements were identified, and the clinical significance and independence of each element remain unclear. Redundancy in CF performance likely allows for a more straightforward scoring approach that still maintains similar test sensitivity. Therefore, we developed a streamlined CF scoring system that includes only 10 key figure components. Additionally, we created a new 4-choice CF Recognition task based pm the same 10 CF elements.

The primary aim of this report is to introduce the Emory CF scoring system and evaluate its relationship with the traditional Osterrieth scoring method in individuals who self-identify as cognitively normal. Additionally, a cohort of patients undergoing deep brain stimulation (DBS) evaluation was assessed to compare the scoring methods in a clinical population. DBS candidates serve as a clinical contrast group due to their relatively homogeneous referral characteristics and comparable age to the EHBS cohort. While we present group differences using both CF scoring methods to highlight their neuropsychological sensitivity in patients with a high prevalence of spatial, executive function, and memory impairments, this study does not explore the specific neuropsychological mechanisms driving group differences in CF performance. We also evaluate the time differences between scoring methods and assess inter-rater reliability within a subgroup of DBS candidates.

Method

Healthy volunteers

A total of 315 participants were included in the Emory Healthy Brain Study (EHBS), an Alzheimer's disease biomarker discovery initiative aimed at identifying predictors of cognitive trajectories in both normal and pathological aging (Goetz et al., 2019). All EHBS participants completed the Montreal Cognitive Assessment (MoCA) (Nasreddine et al., 2012), and those with MoCA scores of 23/30 or lower were excluded. This project received approval from the Emory University Institutional Review Board in compliance with the Declaration of Helsinki, and all participants provided written informed consent.

The CF was administered during the baseline EHBS study visit via telehealth (Hewitt & Loring, 2020; Loring et al., 2023). With telehealth, participants are instructed to fold the page in half, with the drawing facing inward after completing the copy task, and to place the page on the floor. After the study visit, participants are asked to tear the figure sheets into pieces and dispose of them.

Movement disorder patients

The movement disorder cohort consisted of 84 programmatic referrals for neuropsychological evaluation as part of the preoperative evaluation for DBS. Diagnoses included 51 patients (60.7%) with Parkinson's disease, 25 patients (29.8%) with essential tremor, 1 patient (1.1%) with mixed PD/ET, 4 patients (4.8%) with cervical dystonia, 1 patient (1.1%) with blepharospasm, 1 patient (1.1%) with tremor related to normal pressure hydrocephalus, and 1 patient (1.1%) with tardive dyskinesia.

Complex Figure administration

CF administration involved three standard conditions: copy, immediate recall, and 30-minute delayed recall (Loring et al., 1990). Participants are not forewarned about subsequent memory testing after completing the CF copy. Osterrieth's 18 elements are scored based on their accuracy and placement, with scores ranging from 0.5 points for recognizable but inaccurately placed reproductions to 2.0 points for accurate and correctly located reproductions. Recognition testing (Meyers & Lange, 1994; Meyers & Meyers, 1995) was conducted following delayed free recall. Correct target identification and correct foil rejection are both scored, resulting in a maximum recognition score of 24 points.

Emory Complex Figure elements

Emory CF scoring was modeled after the Benson Figure in which single points are awarded independently for element accuracy and element location (NACC, 2015; Possin et al., 2011; Weintraub et al., 2018). The primary CF rectangle and inner CF lines forming the "Union Jack" are used as a spatial reference frame for recognition testing, precluding the use of Osterrieth elements 2, 3, 4, and 5. In addition, single lines are excluded since they lack spatial or figural complexity (i.e., Osterrieth elements 7, 10, 15, 16). There are no ½ point scores, and the maximum score for each figure element is 2 based on accurate and correctly placed components. CF elements are scored leniently with instructions to give credit (i.e., round up) if scoring uncertainty exists. Location points are not awarded if item presence is not scored. There are 10 elements yielding a possible total score range from 0 to 20. Scoring criteria are presented in Appendix A.

Emory CF recognition

Emory Recognition task utilizes a 4-choice format. Targets and distractors are positioned differently relative to the Union Jack, which serves as the primary frame of reference (see Supplementary File for Recognition Stimuli). Spatial CF memory is reportedly more sensitive to right hippocampal dysfunction than figural CF features (Breier et al., 1996) and is emphasized in Emory CF Recognition. Distractors include errors that are commonly observed in patients with lateralized right temporal lobe epilepsy during free recall (Loring et al., 1988). Recognition scores range from 0 to 10.

Emory CF Recognition was obtained after completion of the Meyers and Meyers recognition to prevent any potential unknown performance influences on the latter. This sequence ensures that Meyers and Meyers CF performance, an integral part of the formal EHBS research protocol, remains unaffected.

Scoring duration and inter-rater reliability

A subset of DBS patients ($n = 48$) was scored independently by two experienced EHBS research assistants to assess inter-rater reliability and scoring times for both scoring methods (excluding recognition). The scoring sequence order was randomized across both patients and scoring methods for each research assistant independently to minimize carry-over effects when calculating the scoring time burden.

Statistical analysis

Means, standard deviations, and ranges for sample demographics, as well as Osterrieth and Emory scores, were calculated using SPSS 29.0. Group differences and effect sizes were established with ANOVAs for parametric data and chi-square for frequency data. Group classification was analyzed using logistic regression to predict group membership based on Copy, Immediate Recall,

https://doi.org/10.1017/S1355617724000584

Table 1. Complex Figure performances (means and standard deviations) for EHBS participants ($n = 315$) and DBS patients ($n = 84$). The lower portion of the table reports group difference effect sizes (Cohen's d) for each CF condition using both scoring methods

	CF Copy	CF Immediate	CF Delay	CF Recognition
EHBS Emory scoring	19.2/20 (1.9)	13.1/20 (4.3)	12.8/20 (4.4)	7.4/10 (1.8)
EHBS Osterrieth scoring	31.7/36 (4.8)	18.6/36 (6.5)	17.9/36 (6.7)	20.6/24 (2.0)
EHBS T-score	–	55.6 (13.8)	54.1 (14.6)	51.9 (11.9)
DBS Emory scoring	18.1 (2.9)	8.3 (3.9)	8.5 (3.9)	5.4 (2.0)
DBS Osterrieth scoring	30.0 (6.7)	13.7 (6.6)	13.1 (6.5)	19.0 (2.2)
DBS T-score	–	46.0 (15.4)	44.4 (15.1)	42.9 (12.3)
Emory Scoring Cohen's d	0.45	1.17	1.03	1.05
Osterrieth scoring Cohen's d	0.29	0.75	0.73	0.76
Meyers and Meyers T-score Cohen's d	–	0.66	0.66	0.74

Note: Cohen's $d = 0.2$ is considered a small effect, Cohen's $d = 0.5$ is considered a medium effect, Cohen's $d = 0.8$ is considered a large effect.

Delayed Recall, and Recognition scores from each scoring method. Univariate demographic influences on CF scores were assessed using Pearson correlations, while multivariate regression models were developed to control for demographics in predicting Osterrieth CF memory values from Emory CF scores (see Supplementary File).

Results

The average MoCA score for EHBS participants was 27.1/30 (SD = 1.8). Ages ranged from 50.1 to 79.6 years, with a mean age of 63.9 years (SD = 6.6). Education levels varied from 11 to 20 years, averaging 16.8 years (SD = 2.1). The sample comprised 201 women (63.8%) and 114 men (36.2%), including 243 White participants (77.1%), 67 Black participants (21.3%), and 2 Asian participants (0.6%), with 3 participants (1.0%) not further characterized.

The average MoCA score for DBS participants was 24.7/30 (SD = 3.5), which is significantly lower than that of EHBS participants ($p < 0.0001$ (exact value = 3.4E–17), Cohen's $d = 0.9$). Ages in the DBS group ranged from 42.1 to 83.4 years, with a mean age of 64.8 years (SD = 9.8), which did not differ significantly from EHBS participants ($p = .357$, Cohen's $d = 0.1$). Education levels in the DBS group ranged from 9 to 20 years, averaging 14.7 years (SD = 2.5), which was significantly lower than that of EHBS volunteers ($p < 0.0001$ (1.6E–14), Cohen's $d = 0.9$). The cohort included 26 women (31.0%) and 58 men (69.0%), comprising 73 White patients (86.9%), 8 Black patients (9.5%), 1 Asian patient (1.1%), and 2 Asian Indian patients (2.4%).

Concordance

Copy, Immediate Recall, Delayed Recall, and Recognition scores for both CF scoring approaches, including T-scores for Osterrieth scoring, are presented in Table 1 for both EHBS participants and DBS subjects. Univariate correlations reflecting demographic influences are detailed in the Supplemental File, where the strongest influences ranged from 5% to 6% of the shared variance. The Supplemental File also includes multivariate regression models that predict Osterrieth scores from Emory scores while controlling for demographic factors.

Our primary analyses involved Pearson correlations of CF scores between scoring approaches in healthy EHBS participants. High correlations were observed between the traditional Osterrieth criteria and the Emory CF scoring criteria across all non-recognition CF conditions for EHBS participants: Copy ($r = 0.76$), Immediate Recall ($r = .86$), and Delayed Recall ($r = 0.85$). Although all correlations were statistically significant, the magnitude was lower for Recognition ($r = 0.47$). Similarly, the DBS group exhibited strong correlations between the two scoring approaches: Copy ($r = 0.80$), Immediate Recall ($r = 0.84$), Delayed Recall ($r = 0.85$), and Recognition ($r = 0.37$).

Group differences

Except for CF copy scores using Osterrieth criteria, both scoring approaches effectively discriminated between groups at the $p < 0.0001$ level or better. Across all CF conditions, Emory criteria were associated with larger effect sizes (Cohen's d) compared to traditional scoring, with Emory effect sizes ranging from 0.4 to 1.2 and Osterrieth effect sizes ranging from 0.3 to 0.8 (see Table 1).

We evaluated group classification for both scoring methods using logistic regression incorporating all four CF scores (Copy, Immediate Recall, Delayed Recall, and Multiple Choice). The full logistic regression model utilizing Osterrieth/traditional scores was significant ($\chi^2 = 49.7$, $p < 0.0001$ [4.1E–10]), correctly classifying 80.5% of cases. Similarly, the full model based on Emory scoring was also statistically significant ($\chi^2 = 87.4$, $p < 0.0001$ [4.6E–18]) and resulted in a slight increase in overall correct classification to 82.5%. Classification tables are provided in the supplement; while there was a slight decrease in classification accuracy for EHBS control participants from 97.5% to 95.9%, there was a notable increase in classification accuracy for DBS participants from 16.7% to 32.1%.

Emory CF recognition

Tables 2 and 3 present Emory recognition item level frequencies for EHBS and DBS groups. Correct recognition for EHBS participants ranged from a high of 97.5% for the Left Exterior Cross (Emory element 10) to a low of 49.2% for the Exterior Box (Emory element 5), with an average correct total score of 7.4 (SD = 1.8). In the DBS cohort, correct recognition for Emory CF elements ranged from a high of 89.6% for the Left Interior Cross (Emory element 10) to a low of 33.3% for the Small Inner Rectangle (Emory element 9), with an average total score of 5.4 (SD = 2.0).

Distinct recognition patterns were observed between EHBS volunteers and DBS patients. Elements that differed by more than 25% included the upper triangle (element 3), railroad tracks (element 6), bowling ball (element 7), and small inner triangle (element 9), all of which showed significant differences at the $p < 0.0001$ level. The most frequently chosen distractor in the DBS group was response C for the small inner triangle (31.0%, element 9), followed by response D for the railroad tracks (29.8%, element 6), suggesting that these distractors may share similar visual features. Additionally, responses A (20.2%) and B (21.4%) for the upper triangle (element 3) were also common among the DBS

https://doi.org/10.1017/S1355617724000584

Table 2. Emory CF Recognition item level performance for EHBS volunteers ($n = 315$). Correct responses are italicized and bolded

	Response options			
	A	B	C	D
Diamond	1 (0.3%)	***244 (77.5%)***	18 (5.7%)	52 (16.5%)
Parallel lines	44 (14.0)	32 (10.2%)	24 (7.6%)	***215 (68.3%)***
Upper triangle	32 (10.2%)	25 (7.9%)	***232 (73.7%)***	26 (8.3)
Lower horizontal cross	32 (10.2%)	38 (12.1%)	***229 (72.7%)***	16 (5.1%)
Exterior box	98 (31.1%)	31 (9.8%)	31 (9.8%)	***155 (49.2%)***
Railroad tracks	***242 (76.8%)***	15 (4.8%)	7 (2.2%)	51 (16.2%)
Bowling ball	21 (6.7%)	***284 (90.2)***	1 (0.3%)	9 (2.9%)
Nose triangle	29 (9.2%)	19 (6.0%)	***217 (68.9%)***	50 (15.9%)
Small inner rectangle	22 (7.0%)	11 (3.5%)	76 (24.1%)	***206 (65.4%)***
Left exterior cross	***307 (97.5%)***	2 (0.6%)	6 (1.9%)	0 (0.0%)

Table 3. Emory CF Recognition item level performance for DBS patients ($n = 84$). Correct responses are italicized and bolded

	Response options			
	A	B	C	D
Diamond	6 (7.1%)	**47 (56.0%)**	10 (11.9%)	21 (25.0%)
Parallel lines	15 (17.9%)	11 (13.1%)	9 (10.7%)	**49 (58.3%)**
Upper triangle	17 (20.2%)	18 (21.4%)	**40 (47.6%)**	8 (9.5%)
Lower horizontal cross	17 (20.2%)	12 (14.3%)	**47 (56.0%)**	8 (9.5%)
Exterior box	22 (26.2%)	14 (16.7%)	14 (16.7%)	**34 (40.5%)**
Railroad tracks	**37 (44.0%)**	14 (16.7%)	8 (9.5%)	25 (29.8%)
Bowling ball	14 (16.7%)	**53 (63.1%)**	4 (4.8%)	13 (15.5%)
Nose triangle	11 (13.1%)	17 (20.2%)	**42 (50.0%)**	14 (16.7%)
Small inner rectangle	23 (27.4%)	7 (8.3%)	26 (31.0%)	**28 (33.3%)**
Left exterior cross	**75 (89.3%)**	1 (1.2%)	7 (8.3%)	1 (1.2%)

participants; both responses are positioned correctly above the Union Jack, but one is oriented correctly above the left quadrant, while the other is a mirrored arrow pointing upward, resembling the nose triangle to the right of the figure.

Inter-rater reliability

Single-rater intraclass correlations (ICCs) for both absolute agreement and consistency were calculated for 48 DBS patients scored by two EHBS research assistants using a two-way random effects model (Koo & Li, 2016). Both scoring approaches demonstrated consistency and absolute agreement across CF conditions (Osterrieth ICC: Copy = 0.89/0.66, Immediate Recall = 0.83/0.71, Delayed Recall = 0.92/0.56; Emory ICC: Copy = 0.67/0.75, Immediate Recall = 0.87/0.81, Delayed Recall = 0.84/0.75), with ICC correlations between 0.75 and 0.90 indicating good inter-rater reliability. Because recognition items are selected by the individual rather than explicitly scored by the tester, the reliabilities of recognition scores were not analyzed.

Scoring time

CF scoring times were analyzed in the same DBS cohort using a series of mixed-design ANOVAs, with rater as the between-subject factor and CF scoring approach as the within-subject factor. A significant Rater × Scoring method interaction was found for CF copy ($p < 0.0005$), but not for either recall condition. Scoring times were significantly shorter when using Emory criteria for each non-recognition CF condition: Copy (37.5 s *vs.* 97.5 s, $p < 0.0001$ [1.4E–23]), Immediate Recall (37.1 s *vs.* 78.1 s, $p < 0.0001$ [1.9E18]), and Delayed Recall (35.7 s *vs.* 75.9 s, $p < 0.0001$ [1.5E–17]), reflecting an average time savings of 57% across conditions. Individual rater effect sizes (Cohen's *d*) were large: Rater 1 had effect sizes of Copy ($d = 2.1$), Immediate Recall ($d = 2.4$), and Delayed Recall ($d = 2.0$), while Rater 2 had effect sizes of Copy ($d = 1.9$), Immediate Recall ($d = 1.4$), and Delayed Recall ($d = 1.4$).

Discussion

Emory CF scoring yields results that are highly correlated with traditional Osterrieth scoring, with correlations >0.75 observed for immediate and delayed CF recall in both healthy volunteers and DBS participants. The lower correlations for CF copy are likely influenced by narrow performance variability, while the differences in recognition highlight meaningful variations in task demands.

The high correlations between scoring approaches reflect performance redundancy. For example, visual construction impairment should manifest across multiple CF components rather than being restricted to specific elements, especially in the absence of hemispatial attentional deficits. Osterrieth (1944) also reported high correlations between scoring methods. When comparing Rey's original 47-point scoring system to his 36-point system, he found strong rho correlations ($\rho = 0.95$ in 50 adults and $\rho = 0.92$ in 20 six-year-olds). These robust correlations further demonstrate the effectiveness of various scoring approaches in characterizing CF performance.

DBS patients were included not to investigate the mechanisms of CF impairment, but rather to compare and contrast CF scoring approaches within a clinical context. DBS patients represent a significant portion of referrals to our neuropsychology service and have an age range comparable to that of EHBS participants. Other patient groups, such as those with multiple sclerosis or traumatic brain injury, could also be used to assess the relative sensitivity of both CF scoring approaches in clinical applications. However, the

https://doi.org/10.1017/S1355617724000584

larger effect sizes and greater classification accuracy observed with Emory CF scoring compared to traditional Osterrieth scoring suggest increased neuropsychological sensitivity.

Despite the assumption that more detailed scoring will enhance test sensitivity, increased CF scoring complexity does not outperform the traditional Osterrieth approach. Attempts to incorporate improvements, such as equal component weighting and criteria for acceptable angle variance, proved no more effective than traditional scoring and were, importantly, less efficient (Fastenau et al., 1996). Therefore, while additional scoring features may seem beneficial, it does not necessarily lead to improved performance characterization in clinical practice. For instance, when comparing patients with temporal lobe epilepsy to healthy controls, formal scoring methods were more likely to misclassify a control's performance as impaired than the clinical assessments made by trained neuropsychologists (LeMonda et al., 2022). Consequently, detailed scoring of multiple CF features appears unnecessary for identifying abnormal production.

Good inter-rater reliability was established for both single and average measure ICC calculations, with no clear advantage observed between the scoring methods. Across both approaches, CF copy exhibited the lowest correlations, likely due to ceiling effects in copy performance. ICC values ranging from 0.75 to 0.90 are generally considered indicative of good inter-rater reliability, and all memory conditions except for the Osterrieth absolute score (ICC = 0.71) fell within this range. Although absolute ICCs were lower than consistency ICCs, the differences were minimal (average difference: Osterrieth memory = 0.115; Emory memory = 0.075), indicating no significant bias. Both scoring approaches exhibited strong overall inter-rater reliability.

The average time savings for Emory CF was 57%, associated with extremely large individual effect sizes ranging from $d = 1.4$ to $d = 2.4$. Both raters were experienced research assistants resulting in high overall scoring efficiency, although the influence of scoring the same figure with repeated methods also lowers mean scoring times for both conditions. However, both the scoring method and subject order were randomized to minimize the introduction of systematic differences.

Although this report was not specifically designed to examine CF performance in movement disorder patients, comparisons between the EHBS and DBS groups provide valuable insights into the clinical applicability of the Emory CF scoring system. Although the high correlations between scoring systems suggest comparability, they should not be considered strictly equivalent. MoCA scores, which serve as a general estimate of overall cognitive performance, were associated with a group effect size of $d = 0.9$. The average effect size for the three Osterrieth CF memory conditions was $d = 0.7$, lower than MoCA, while the average effect size for the three Emory CF memory conditions was $d = 1.1$, exceeding the MoCA effect size. Although both the MoCA and the two CF scoring methods show large effect sizes, the larger effect sizes associated with the Emory CF scoring suggest a potential for greater neuropsychological sensitivity compared to the Osterrieth criteria, as further supported by logistic regression analyses. Both the effect size and the differences in logistic regression classification between groups highlight the lack of strict equivalence in clinical application.

A study limitation is the lack of counterbalancing between the traditional and Emory recognition conditions. This was necessary because Emory CF scoring was not originally included in the National Institute of Aging (NIA) approved study protocol. Therefore, administering the Emory CF Recognition task after the completion of the Meyers and Meyers Recognition task avoids potential exposure effects that could not otherwise be accurately characterized. While this is necessary for research protocol adherence, this introduces potential order effects.

The administration of the EHBS CF was conducted via telehealth due to the COVID-19 pandemic, which necessitated a transition to remote methods for non-critical research activities to prioritize safety. This shift required modifications to clinical research protocols, including the National Alzheimer's Coordinating Center Uniform Data Set (Weintraub et al., 2018) and our EHBS study (Goetz et al., 2019). Although video telehealth cognitive assessments are reliable and valid (Bilder et al., 2020; Geddes et al., 2020; Marra et al., 2020), there is a lack of formal studies specifically examining the effects of video telehealth assessment on CF performance. This gap underscores the need for additional research investigating how telehealth delivery may influence CF performance and other cognitive assessments.

The Emory CF scoring system presents several advantages. It provides a streamlined approach that saves approximately 57% in time compared to the traditional 18-point Osterrieth system, all while maintaining diagnostic sensitivity as reflected by contrasts between EHBS and DBS groups. Like the Benson Figure, the Emory scoring system, including its Recognition test stimuli, is freely accessible, which should facilitate its adoption in clinical research settings.

In conclusion, this report demonstrates that the Emory CF scoring system is significantly correlated with the traditional Osterrieth scoring for both immediate and delayed recall conditions. Moreover, Emory scoring may enhance neuropsychological test sensitivity, as indicated by larger effect sizes and greater classification differences between groups. Error patterns in Emory 4-choice CF Recognition present an opportunity to explore altered spatial memory. The impaired spatial recall is similar to false positives or semantic intrusion errors in verbal memory, which are diagnostically relevant in mild cognitive impairment (Thomas et al., 2018). In contrast to a yes/no recognition task that can be partially influenced by the verbal encoding of individual items (e.g., "diamond," "bowling ball"), Emory recognition can provide insight into spatial distortions of CF memory. We anticipate that these findings will bolster the use of the CF as a Common Data Element for evaluating visual constructional ability and visual memory.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S1355617724000584.

Acknowledgments. This research was supported by funding from the NIA (Emory Healthy Brain Study: R01-AG070937, J.J. Lah, M.D., Ph.D. Principal Investigator).

Competing interests. The authors have no competing interests or conflicts of interest to report.

A preliminary version of the report was presented at the 2023 Annual Meeting of the *International Neuropsychological Society* held in San Diego, CA, USA.

References

Beebe, D. W., Ris, M. D., Brown, T. M., & Dietrich, K. N. (2004). Executive functioning and memory for the Rey-Osterreith complex figure task among community adolescents. *Applied Neuropsychology*, *11*(2), 91–98. doi: 10.1207/s15324826an1102_4.

Bilder, R. M., Postal, K. S., Barisa, M., Aase, D. M., Cullum, C. M., Gillaspy, S. R., Harder, L., Kanter, G.,Lanca, M.,Lechuga, D. M.,Morgan, J. M.,Most, R.,

Puente, A. E.,Salinas, C.M., & Woodhouse, J. (2020). Inter Organizational Practice Committee recommendations/guidance for teleneuropsychology in response to the COVID-19 pandemic. *Archives of Clinical Neuropsychology*, *35*(6), 647–659. doi: 10.1093/arclin/acaa046.

Breier, J. I., Plenger, P. M., Castillo, R., Fuchs, K., Wheless, J. W., Brookshire, B. L.,Willmore, L. J.,Thomas, A. B., & Papanicolaou, A. (1996). Effects of temporal lobe epilepsy on spatial and figural aspects of memory for a complex geometric figure. *Journal of the International Neuropsychological Society*, *2*(6), 535–540.

Corwin, J., & Bylsma, F. W. (1993). Translations of excerpts from André Rey's psychological examination of traumatic encephalopathy and P.A. Osterrieth's The Complex Figure Copy Test. *The Clinical Neuropsychologist*, *7*(1), 3–15.

Paula, J. J. de, Costa, M. V., Andrade, G. F., Ávila, R. T., & Malloy-Diniz, L. F. (2016). Validity and reliability of a "simplified" version of the Taylor Complex Figure Test for the assessment of older adults with low formal education. *Dementia & Neuropsychologia*, *10*(1), 52–57. doi: 10.1590/s1980-57642016dn10100010.

Denman, S. (1984). *Denman Neuropsychology Memory Scale*. S.B. Denman.

Fastenau, P. S. (1996). Development and preliminary standardization of the "Extended Complex Figure Test" (ECFT). *Journal of Clinical and Experimental Neuropsychology*, *18*(1), 63–76.

Fastenau, P. S., Bennett, J. M., & Denburg, N. L. (1996). Application of psychometric standards to scoring system evaluation: Is "new necessarily "improved"? *Journal of Clinical & Experimental Neuropsychology*, *18*(3), 462–472.

Geddes, M. R., O'Connell, M. E., Fisk, J. D., Gauthier, S., R., Camicioli, & Ismail, Z. (2020). Remote cognitive and behavioral assessment: Report of the Alzheimer Society of Canada Task Force on dementia care best practices for COVID-19. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *12-1*(1), e12111. doi: 10.1002/dad2.12111.

Goetz, M. E. Hanfelt, John, J. J., Bergquist, S. E., Loring, S. H., Quyyumi, D. W., Lah, A., & J., J. (2019). Rationale and design of the Emory Healthy Aging and Emory Healthy Brain Studies. *Neuroepidemiology*, *53*(3-4), 187–200. doi: 10.1159/000501856.

Hewitt, K. C., & Loring, D. W. (2020). Emory University telehealth neuropsychology development and implementation in response to the COVID-19 pandemic. *The Clinical Neuropsychologist*, *34*(7-8), 1352–1366. doi: 10.1080/13854046.2020.1791960.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. doi: 10.1016/j.jcm.2016.02.012,

LeMonda, B. C., MacAllister, W.,Morrison, C., L., Vaurio,Blackmon, K., Maiman, M., Liu, A.,Liberta, T., &Bar, W.B. (2022). Is formal scoring better than just looking? A comparison of subjective and objective scoring methods of the Rey Complex Figure Test for lateralizing temporal lobe epilepsy. *The Clinical Neuropsychologist*, *36*(7), 1637–1652. doi: 10.1080/13854046.2020.1865461.

Loring, D. W., Lah, J. J., & Goldstein, F. C. (2023). Telehealth equivalence of the Montreal cognitive assessment (MoCA): Results from the Emory Healthy Brain Study (EHBS). *Journal of the American Geriatrics Society*, *71*(6), 1931–1936. doi: 10.1111/jgs.18271.

Loring, D. W., Lee, G. P., & Meador, K. J. (1988). Revising the Rey-Osterrieth: Rating right hemisphere recall. *Archives of Clinical Neuropsychology*, *3*(3), 239–247.

Loring, D. W., Martin, R. C., Meador, K. J., & Lee, G. P. (1990). Psychometric construction of the Rey-Osterrieth Complex Figure: Methodological considerations and interrater reliability. *Archives of Clinical Neuropsychology*, *5*(1), 1–14.

Marra, D. E., Hamlet, K. M., Bauer, R. M., & Bowers, D. (2020). Validity of teleneuropsychology for older adults in response to COVID-19: A systematic and critical review. *The Clinical Neuropsychologist*, *34*(7-8), 1411–1452. doi: 10.1080/13854046.2020.1769192.

Meyers, J. E., & Lange, D. (1994). Recognition subtest for the Complex Figure. *The Clinical Neuropsychologist*, *8*(2), 153–186

Meyers, J. E., & Meyers, K. R. (1995). *Rey Complex Figure Test and Recognition Trial*. Psychological Assessment Resources.

NACC. (2015). NACC Uniform Data Set: Instructions for the Neuropsychological Battery (Form C2). https://files.alz.washington.edu/documentation/uds3-np-c2-instructions.pdf.

NINDS. (2022). NINDS Common Data Elements. https://www.commondataelements.ninds.nih.gov/report-viewer/23967/Rey-Osterrieth%20Complex%20Figure%20Test.

Nasreddine, Z. S., Phillips, N., Chertkow, H., Rossetti, H., Lacritz, L., M., Collum, &Weiner, M. (2012). Normative data for the Montreal Cognitive Assessment (MoCA) in a population-based sample. *Neurology*, *78*(10), 765-78 10 765–766.

Osterrieth, P. A. (1944). Le test de copie d'une figure complexe. *Archives de Psychologie*, *30*, 206–356

Poreh, A., Levin, J. B., & Teaford, M. (2020). Geriatric Complex Figure Test: A test for the assessment of planning, visual spatial ability, and memory in older adults. *Applied Neuropsychology: Adult*, *27*(2), 101–107.

Possin, K. L., Laluz, V. R., Alcantar, O. Z., Miller, B. L., & Kramer, J. H. (2011). Distinct neuroanatomical substrates and cognitive mechanisms of figure copy performance in Alzheimer's disease and behavioral variant frontotemporal dementia. *Neuropsychologia*, *49*(1), 43–48. doi: 10.1016/j.neuropsychologia.2010.10.026.

Rabin, L. A., Paolillo, E., &Barr, W. B. (2016). Stability in test-usage practices of clinical neuropsychologists in the United States and Canada over a 10-year period: A follow-up survey of INS and NAN members. *Archives of Clinical Neuropsychology*, *31*(3), 206–230. doi: 10.1093/arclin/acw007.

Randolph, C. (1998). *RBANS manual: Repeatable Battery for the Assessment of Neuropsychological Status*. The Psychological Corporation.

Rey, A. (1941). L'examen psychologique dans les cas d'encéphalopathie traumatique. *Archives de Psychologie*, *28*, 286–340

Stern, R. A., Jovorksy, D. J., Singer, E. A., Singer Harris, N. G., Somerville, J. A., Duke, L. M., & Kaplan, E. 1999). *The Boston Qualitative Scoring System for the Rey-Osterrieth Figure*. Psychological Assessment Resources.

Temple, R. O., Davis, J. D., Silverman, I., & Tremont, G. (2006). Differential impact of executive function on visual memory tasks. *The Clinical Neuropsychologist*, *20*(3), 480–490. doi: 10.1080/13854040590967540.

Thomas, K. R., Eppig, J., Edmonds, E. C., Jacobs, D. M., Libon, D. J., Au, R., & Bondi, M. W. (2018). Word-list intrusion errors predict progression to mild cognitive impairment. *Neuropsychology*, *32*(2), 235–245. doi: 10.1037/neu0000413.

Waber, D. P., & Holmes, J. M. (1985). Assessing children's copy production of the Rey-Osterrieth Complex Figure. *Journal of Clinical and Experimental Neuropsychology*, *7*(3), 264–280.

Weintraub, S. Besser, Dodge, L., Teylan, H. H., Ferris, M., Goldstein, S., Morris, F. C., & J., C. (2018). Version 3 of the Alzheimer disease centers' Neuropsychological Test battery in the Uniform Data Set (UDS). *Alzheimer Disease & Associated Disorders*, *32*(1), 10–17. doi: 10.1097/wad.0000000000000223.