*Journal of the International Neuropsychological Society* (2025), 1–8
doi:10.1017/S135561772510115X

## Research Article

# Recognizing improved Complex Figure memory assessment: The Emory 4-choice Complex Figure recognition task

David W. Loring[1,2], Felicia C. Goldstein[1], James J. Lah[1] and Daniel M. Bolt[3]

[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 Educational Psychology, University of Wisconsin, Madison, WI, USA

### Abstract

**Objective:** We compare the Emory 10-item, 4-choice Rey Complex Figure (CF) Recognition task with the Meyers and Lange (M&L) 24-item yes/no CF Recognition task in a large cohort of healthy research participants and in patients with heterogeneous movement disorder diagnoses. While both tasks assess CF recognition, they differ in key aspects including the saliency of target and distractor responses, self-selection versus forced-choice formats, and the length of the item sets. **Participants and Methods:** There were 1056 participants from the Emory Healthy Brain Study (EHBS; average MoCA = 26.8, SD = 2.4) and 223 movement disorder patients undergoing neuropsychological evaluation (average MoCA = 24.3, SD = 4.0). **Results:** Both recognition tasks differentiated between healthy and clinical groups; however, the Emory task demonstrated a larger effect size (Cohen's $d = 1.02$) compared to the M&L task (Cohen's $d = 0.79$). d-prime scoring of M&L recognition showed comparable group discrimination (Cohen's $d = 0.81$). Unidimensional two-parameter logistic item response theory analysis revealed that many M&L items had low discrimination values and extreme difficulty parameters, which contributed to the task's reduced sensitivity, particularly at lower cognitive proficiency levels relevant to clinical diagnosis. Dimensionality analyses indicated the influence of response sets as a potential contributor to poor item performance. **Conclusions:** Emory CF Recognition task demonstrates superior psychometric properties and greater sensitivity to cognitive impairment compared to the M&L task. Its ability to more precisely measure lower levels of cognitive functioning, along with its brevity, suggests it may be more effective for diagnostic use, especially in clinical populations with cognitive decline.

**Keywords:** Neuropsychological tests; recognition (psychology); psychometrics; diagnostic techniques; neurological; memory; short-term; neuropsychology / methods

(Received 20 February 2025; final revision 26 May 2025; accepted 17 June 2025)

### Statement of research significance

**Research question:** This study compared the psychometric performance of two Rey Complex Figure (CF) recognition tasks: the 4-choice Emory CF Recognition (based on 10 streamlined scoring elements) and the 24-item yes/no task developed by Meyers and Lange (1995; M&L). Key CF Recognition differences included item saliency, response format, and task length.

**Main findings:** Both tasks distinguished healthy controls from patients with movement disorders, but the Emory task showed larger effect sizes. Despite being shorter, Emory CF Recognition demonstrated better sensitivity at lower cognitive levels and more effective group discrimination. The M&L task, which combines target identification and foil rejection, was less precise, more multidimensional, and more prone to response biases. d-prime scoring did not improve its diagnostic value.

**Study contributions:** Results support the Emory 4-choice task as a more efficient and psychometrically sound measure of CF recognition memory, particularly in cognitively impaired populations, due to its unidimensionality and improved sensitivity.

### Introduction

The Rey CF is a widely utilized assessment tool for evaluating visual constructional skills and visual memory (Rabin et al., 2016). Introduced in 1941 (Corwin & Bylsma, 1993; Rey, 1941), the CF originally included only a copy and single free recall memory condition. Rey's scoring system assigned two points each to four core elements of the figure (diamond, circle, and two line groups) and one point to remaining segments, totaling 47 points. Osterrieth subsequently revised the scoring to reduce scoring burden by focusing on 18 larger CF components, each of which were scored for accuracy and placement resulting in a maximum of 36 points (Osterrieth, 1944). Osterrieth's scoring method is the most widely used approach for CF scoring for both copy and memory recall conditions (Zhang et al., 2021).

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

**Cite this article:** Loring D.W., Goldstein F.C., Lah J.J., & Bolt D.M. Recognizing improved Complex Figure memory assessment: The Emory 4-choice Complex Figure recognition task. *Journal of the International Neuropsychological Society*, 1–8, https://doi.org/10.1017/S135561772510115X

Although common in clinical assessments, the CF was not originally developed as a standardized test, with varying administration protocols yielding different performance patterns (Loring et al., 1990). To enhance its clinical utility, Meyers and Lange (M&L; Meyers & Lange, 1994; Meyers & Meyers, 1995) developed a standardized administration protocol that includes a copy trial, 3-minute free recall, 30-minute delayed recall, and normative data for each condition. They also introduced a recognition task to help determine whether poor free recall is due to retrieval deficits. This recognition task includes 12 Osterrieth scoring elements from the Rey Figure (targets) with 12 scoring elements from the Taylor Figure (a parallel CF; Taylor, 1969) resulting in a 24-item yes/no recognition test.

To both simplify scoring complexity and decrease the scoring time associated with Osterrieth's 18-element CF scoring system, we developed a streamlined approach for copy and memory conditions using only 10 CF components (Loring et al., 2024). Unlike the M&L recognition that combines components from two different CFs into a yes/no recognition task, Emory CF recognition includes the same 10 CF copy and memory scoring elements using a 4-choice recognition format. CF distractors were designed based on errors observed in patients with lateralized right temporal lobe epilepsy during free recall (Loring et al., 1988), which are also commonly encountered across a range of neurological diagnoses in clinical practice.

The current study contrasts M&L 24-item, yes/no CF Recognition performance with performance on the Emory 10-item, 4-choice Recognition task in a large cohort of cognitively healthy research participants and in patients with various movement disorders undergoing neuropsychological evaluation for deep brain stimulation(DBS) or focused ultrasound. Given their elevated risk for memory and other cognitive impairments, individuals with movement disorders offer an ecologically valid clinical context in which to assess the relative utility of the M&L and Emory recognition tasks (Loring et al., 2024).

Recognition memory performance can vary substantially depending on factors such as distractor saliency, response format (yes/no vs. forced-choice), and the number of test items. Tasks with distractors closely resembling target items demand greater cognitive discrimination and are typically more difficult, while less salient distractors make recognition easier. Similarly, the format of the recognition task influences underlying cognitive processes. Yes/no recognition requires individuals to make an independent judgment about each item, a process susceptible to response bias and confidence effects. In contrast, forced-choice formats mitigate these influences by requiring direct comparisons among options (Kroll et al., 2002). Additionally, the number of items affects the psychometric robustness of the task, with larger item sets generally offering greater reliability (Downing, 2004). To investigate performance differences, we apply logistic regression(LR), item factor analysis, and item response theory (IRT). Building on prior findings, we also employ d-prime scoring, which we have previously shown to effectively distinguish individuals with amnestic mild cognitive impairment (aMCI) who are positive for Alzheimer's disease (AD) biomarkers (amyloid-β, tau) from those who are biomarker-negative (Goldstein et al., 2019). Accordingly, we examine group-level differences in M&L CF Recognition using d-prime, based on hit and false alarm rates.

## Method

### Healthy volunteers

There were 1056 participants from the Emory Healthy Brain Study (EHBS). EHBS is a longitudinal AD' biomarker discovery project to identify predictors of cognitive trajectories of normal and pathological aging, with EHBS study visits scheduled every 2 years after enrollment (Goetz et al., 2019), with cognitive testing conducted at each study time point. This project was approved by the Emory University Institutional Review Board in accordance with the Declaration of Helsinki, and all participants provided written informed consent.

### Movement disorder patients

Movement disorder patients were 223 referrals for neuropsychological testing as part of their preoperative evaluation for DBS or for independent diagnostic characterization as part of a specialty Comprehensive Care Clinic. Diagnoses included 51 (60.7%) patients with Parkinson's disease (PD), 25 (29.8%) patients with Essential Tremor (ET), 1 (1.1%) mixed PD/ET patient, 4 (4.8%) patients with cervical dystonia, and one patient (1%) each with blepharospasm, tremor associated with normal pressure hydrocephalus, or tardive dyskinesia.

### Cognitive testing

Cognitive testing was conducted in person or via telehealth (Hewitt & Loring, 2020) Cognitive testing was conducted via telehealth for the majority of EHBS participants (telehealth: $n = 898$; in-person: $n = 158$), whereas all but one Movement Disorder patient had face-to-face evaluations. Although different assessment protocols were employed, both included the Montreal Cognitive Assessment (MoCA) (Nasreddine et al., 2012) and the Rey CF (Lezak et al., 2004; Loring et al., 2024). Equivalence of telehealth evaluation has been demonstrated for MoCA testing (Loring et al., 2023). Emory CF Recognition was obtained after M&L Recognition to prevent any potential unknown performance influences on the latter since M&L CF performance is included as part of the formal EHBS research protocol. There were no tests of visual memory or visual perceptual function administered between CF copy and the delayed memory conditions.

### Meyers and lange (M&L) Complex Figure recognition

The M&L recognition task is a 24-item yes/no recognition test that incorporates 12 of the 18 Osterrieth CF scoring elements with 12 (of 18) scoring elements from the Taylor CF. CF Recognition is assessed after the copy, immediate recall, and delayed CF recall trials. Between 4 and 9 CF elements are presented on a single page and participants indicate which elements are recalled from the Rey CF. The primary Recognition memory score is calculated as the sum of correctly identified Rey CF target items and correctly rejected Taylor CF foil items, yielding a maximum possible score of 24.

### Emory Complex Figure recognition

The Emory CF Recognition task employs a 4-choice recognition paradigm to evaluate the 10 scoring elements defined by the Emory CF system to characterize CF copy and memory performances (see Supplementary File for Emory CF Recognition Stimuli), Target items and distractors are presented in distinct spatial positions using the "Union Jack" as a frame of reference to serve as a visual anchor for all stimuli (see Supplementary File for Recognition Stimuli and scoring form). Emory Recognition prioritizes spatial configural memory, a domain considered sensitive to right hippocampal dysfunction (Breier et al., 1996).

https://doi.org/10.1017/S135561772510115X Published online by Cambridge University Press

**Table 1.** EHBS versus Movement disorder group differences

| | EHBS | Movement | *p*-value | Cohen's *d* | *d* 95% CI |
|---|---|---|---|---|---|
| MoCA | 26.8 (2.4) | 24.3 (4.0) | <.001 | 0.92 | 0.76, 1.06 |
| Emory delay | 13.5 (4.5) | 7.7 (4.4) | <.001 | 1.30 | 1.14 – 1.45 |
| Emory recognition | 7.6 (2.0) | 5.5 (2.1) | <.001 | 1.02 | 0.91, 1.21 |
| Osterrieth delay | 18.8 (7.1) | 12.3 (7.0) | <.001 | 0.92 | 0.78, 1.08 |
| M&L recognition | 20.7 (2.0) | 19.0 (2.3) | <.001 | 0.79 | 0.70, 1.00 |
| M&L hits | 10.2 (1.6) | 8.8 (1.8) | <.001 | 0.82 | 0.69, 0.98 |
| M&L false positives | 1.5 (1.6) | 1.9 (1.8) | <.001 | −0.23 | −0.40, −0.11 |
| M&L true negatives | 10.5 (1.6) | 10.1 (1.8) | <.001 | 0.23 | 0.11, 0.40 |
| M&L false negatives | 1.8 (1.6) | 3.2 (1.9) | <.001 | −0.80 | −0.98, −0.69 |
| M&L d-prime | 2.3 (0.6) | 1.8 (0.7) | <.001 | 0.81 | 0.64 0.97 |

*Note: By convention, d = 0.2 is considered a small effect, d = 0.5 a moderate effect, d = 0.8 a large effect, d ≥ 1.0 a very large effect.*

### *Analysis*

Primary analyses consist of 2 (group) × 2 (CF method) mixed-design ANOVAs, followed by pairwise comparisons to examine differences across CF scoring approaches. To further assess group discrimination, we also independently applied d-prime scoring to the M&L Recognition task using hit and false alarm rates, and applied an independent *t*-test to evaluate mean d-prime differences across groups. d-prime, derived from signal detection theory, provides an unbiased estimate of recognition performance by separating true memory sensitivity from response bias (Yonelinas, 1994). While comparing clinical group performance to that of healthy controls is an indirect method of evaluating construct validity, the extent to which a task differentiates between groups, particularly with varying effect sizes, serves as an indicator of its sensitivity to disease-related cognitive changes. Patients with movement disorders represent a relevant clinical population in which the CF is included as a standard component of neuropsychological assessment protocols.

To better understand differences between measures, we will then explore individual item contributions to group discrimination using LR, and through IRT-based analyses. In these analyses, we use IRT as a strictly descriptive tool, recognizing that either (or both) of the measures could possess some multidimensionality, thus rendering the unidimensional IRT trait a weighted composite of the multiple dimensions (see e.g., Reckase, 2009, pp.126 – 127). One useful outcome from IRT analysis is the test information function, which is derived from the slopes of the estimated individual item response curves and ultimately quantifies the precision of the latent trait estimates at different levels of the latent trait (test information is inversely related to measurement error). Test information functions provide a useful way of comparing measures, not only with respect to their absolute levels of measurement precision, but also according to the relative measurement precision they provide at different locations along the latent trait continuum.

## Results

The average EHBS MoCA score was 26.8/30 (SD = 2.4). Age ranged from 50.1 years to 81.9 years ($M$ = 66.0, SD = 6.7). Education varied between 11 – 22 years ($M$ = 16.8, SD = 2.0). There were 670 (63.4%) women and 386 (36.6%) men, and included 833 White participants (78.9%), 188 Black participants (17.8%), 15 Asian participants (1.4%), 3 American Indian or Alaska Native (0.3%), 4 Native Hawaiian or Other Pacific Islander (0.4%) with 13 (1.2%) not further characterized.

The average Movement Disorder MoCA score was 24.3/30 (4.0), which is significantly lower than EHBS participants ($p$ < .0001), Cohen's $d$ = 0.92). The average age was 66.8 years (SD = 10.5), ranging from 20.9 to 88.5, which did not differ significantly from EHBS participants ($p$ = .180, Cohen's $d$ = 0.09). Movement Disorder education ranged from 8-20 years ($M$ = 15.4, SD = 2.5) which was significantly lower than healthy volunteers ($p$ < .0001, Cohen's $d$ = 0.62). There were 73 women (32.7%) and 150 men (67.30%), which included 203 White (91.0%) patients, 13 Black (5.8%) patients, 1 Asian (0.4%), and 8 Asian Indian (2.7%) patients.

### *Complex Figure performance*

Table 1 presents the mean performance scores for the EHBS and Movement Disorder groups, along with d-prime values for M&L Recognition. CF performance for M&L and Emory scoring approaches was analyzed using a 2 (group) × 2 (scoring approach) mixed-design ANOVA. The d-prime index is calculated as Z_HIT-Z_FA, where Z_HIT and Z_FA denote the inverse cumulative normal (i.e., z) values associated with the proportion correct on target items (i.e., hits) and the proportion incorrect on foil items (i.e., false alarms), respectively. As these Z values are infinitely high when the proportion of hits is 1 and infinitely low when the proportion of false alarms is 0, we followed convention in replacing such proportions by $(n_{TARGET}-.5)/n_{TARGET}$ and $.5/n_{FOIL}$ respectively (Macmillan & Kaplan, 1985), prior to calculating the Z values, where $n_{TARGET} = n_{FOIL} = 12$ for the M&L assessment.

When contrasting M&L and Emory CF performances, a statistically significant interaction effect was found for Recognition ($p$ = .045), indicating differential group sensitivity, with an even stronger interaction effect observed for delayed CF free recall ($p$ = .018). These interactions were further examined using simple main effects analyses contrasting group differences for each CF measure individually. As shown in Table 1, significant group differences were observed for both CF scoring methods, each associated with large effect sizes (Cohen's $d$). The largest effect sizes were for Emory CF delayed free recall and Emory Recognition, both with Cohen's $d$ values greater than 1.0, indicative of very large group differences. Significant differences were also present for Osterrieth delayed free recall and M&L Recognition also showed significant group differences, though with smaller, yet still large, effect sizes. Using confidence intervals to compare effect sizes, the d-prime scoring of the M&L task also demonstrated statistically significant group discrimination, though the effect size was comparable to that obtained with traditional M&L scoring. Emory Delayed Recall produced a larger effect size than the MoCA, while no other statistically significant differences in group-level effect sizes were

https://doi.org/10.1017/S135561772510115X Published online by Cambridge University Press

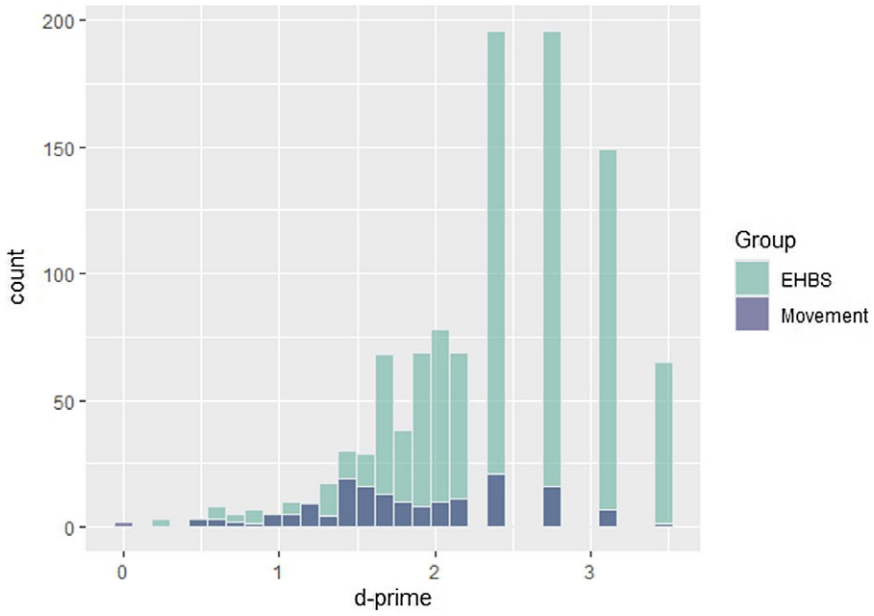


**Figure 1.** Histogram representing d-prime distribution of EHBS and movement disorder participants.

observed. Figure 1 displays empirical histograms of d-prime for the M&L task across the EHBS and Movement Disorder groups.

*Group Prediction.* Each recognition item's contribution to group identification was analyzed independently for both recognition approaches using LR. Item discriminability was evaluated both through univariate analyses, which predict the significance of each item if entered as a sole predictor of group membership, and through multivariate logistic regression (MLR), where the functioning of individual items toward group prediction are evaluated in the simultaneous presence of all other items as predictors. Despite anticipated multicollinearity, the MLR analysis was viewed as a useful way of further contrasting measures according to the extent to which individual items provided incremental prediction in distinguishing groups. As shown in Table 2, all 10 Emory Recognition Elements are predicted to successfully distinguish group membership, with the Bowling Ball item exhibiting the highest anticipated individual discriminability. In contrast (Table 3), individual M&L CF Recognition score coefficients are more variable and often small, with many failing to reach statistical significance. Distractor elements from the Taylor figure exhibit the lowest values, and compared to the Emory Recognition Elements, M&L Target elements frequently have coefficients of smaller magnitude.

A similar conclusion is reached from the MLR analysis. Although the ability to meaningfully isolate individual item predictive effects is likely affected by intercorrelations among items and thus multicollinearity, we use the MLR analysis as a means of evaluating the extent to which the entire collection of items is contributing to group differentiation. In this regard, nearly all Emory items remain significant contributors to predicting clinical status, whereas more elements for M&L Recognition are identified as nonsignificant (Tables 2 & 3). The effect sizes for the LR coefficients (exp(B)), which represent the change in relative odds of group membership given a correct versus incorrect response (i.e., a B value of 1 represents no effect), tend to be lower for Emory compared to M&L items, likewise suggesting greater discrimination from Emory items. Individuals who score correct on the items consistently have a lower odds of being in the movement disorder group relative to the EHBS control group. Collectively, the M&L elements provide less predictive accuracy (i.e, lower $R^2$s) than the Emory items, despite there being more M&L items. These findings indicate that the Emory Recognition offers a more efficient approach to evaluating clinical status.

### *Measurement precision*

To further evaluate the psychometric performance of CF items, we applied unidimensional two-parameter logistic (2PL) models using the multidimensional item response theory (mirt) package in R (Chalmers, 2012). The 2PL provides descriptive information about item discrimination and difficulty against a latent unidimensional recognition proficiency as defined collectively by the items from each measure. The analyses also yield test information functions, providing quantifications of measurement precision (test information is inversely related to measurement error) in relation to the underlying construct (recognition proficiency, denoted as Theta; see Figure 2). For each recognition method, a two-group (EHBS Healthy Volunteer vs. Movement Disorder) model was specified such that the latent proficiency Theta is assigned a mean 0 and variance of 1 in the EHBS Healthy Volunteer group. This information function indicates how measurement precision (recognition proficiency/ Theta) varies across the proficiency continuum.

Item parameter estimates for the Emory and M&L items are shown in Tables 4 and 5, respectively. The 2PL model includes a *Discrimination Parameter* (*a*), which indicates how well an item discriminates different levels of recognition proficiency (greater more positive *a* indicates a greater sensitivity of the item to the latent proficiency, implying better measurement) and a *Difficulty Parameter (b),* the latent proficiency (Theta) level associated with a 50% probability of correctly answering the item (higher *b* values reflecting greater item difficulty). Due to the scaling of Theta mentioned above, items with larger more positive *a*'s and *b*'s close to −1.5 can be viewed as ideal given the goal of precise measurement at the cut point of classification (1.5 standard deviations below the mean).

https://doi.org/10.1017/S135561772510115X Published online by Cambridge University Press

**Table 2.** Emory recognition item group discrimination including logistic regression analyses predicting group membership (Healthy volunteer vs. movement disorder)

| | *Univariate logistic regression* | | | *Multivariate logistic regression* | | |
|---|---|---|---|---|---|---|
| Emory element | *ULR B* | *ULR p*-value | *Exp(B)* | *MLR B* | *MLR p*-value | *Exp(B)* |
| 1. Diamond | −1.12 | <.001 | .33 | −0.63 | <.001 | .53 |
| 2. Parallel lines | −0.69 | <.001 | .50 | −0.02 | .913 | .98 |
| 3. Upper triangle | −1.00 | <.001 | .37 | −0.42 | .019 | .66 |
| 4. Lower cross | −0.80 | <.001 | .45 | −0.09 | .607 | .91 |
| 5. Extra box | −0.60 | <.001 | .55 | −0.30 | .071 | .74 |
| 6. RR tracks | −1.20 | <.001 | .30 | −0.40 | .028 | .67 |
| 7. Bowling ball | −1.93 | <.001 | .15 | −1.48 | <.001 | .23 |
| 8. Nose triangle | −0.83 | <.001 | .44 | −0.04 | .834 | .96 |
| 9. Inner box | −1.49 | <.001 | .23 | −0.71 | <.001 | .49 |
| 10. Left cross | −1.51 | <.001 | .22 | −1.02 | <.001 | .36 |

*Note: The Univariate Logistic Regression coefficients correspond to a model with only the studied item as predictor of group (Healthy Volunteer = 0 vs. Movement Disorder = 1). The Multivariate Logistic Regression coefficients correspond to a model with all items as predictors (Cox & Snell $R^2$ = .161; Nagelkirke $R^2$ = .267). The regression coefficients (B) are evaluated for significance using z-tests.*

**Table 3.** M&L recognition item discrimination using logistic regression analyses predicting group membership (Healthy volunteer versus movement disorder)

| | *Univariate logistic regression* | | | *Multivariate logistic regression* | | |
|---|---|---|---|---|---|---|
| M&L Element | ULR *B* | *ULR p*-value | *Exp(B)* | *MLR B* | *MLR p*-value | *Exp(B)* |
| *1. Foil* | *−.89* | *.067* | .41 | −0.52 | .291 | .59 |
| 2. Target | −.42 | .018 | .66 | −0.35 | .062 | .71 |
| *3. Foil* | *−.08* | *.646* | .93 | −0.05 | .803 | .96 |
| *4. Foil* | *.01* | *.988* | 1.01 | 0.48 | .326 | 1.61 |
| 5. Target | −.89 | .003 | .41 | −0.52 | .083 | .59 |
| *6. Foil* | *.10* | *.899* | 1.10 | 0.17 | .787 | 1.19 |
| 7. Target | −.52 | .028 | .60 | −0.19 | .418 | .82 |
| 8. Target | −.53 | .009 | .59 | −0.41 | .042 | .67 |
| 9. Target | −1.09 | <.001 | .34 | −0.68 | .002 | .51 |
| *10. Foil* | *−.68* | *.237* | .51 | −0.50 | .385 | .61 |
| *11. Foil* | *−.62* | *.028* | .54 | −0.33 | .290 | .72 |
| 12. Target | −.52 | .037 | .59 | 0.02 | .935 | 1.02 |
| 13. Target | −.97 | <.001 | .38 | −0.51 | .009 | .60 |
| *14. Foil* | *−.44* | .021 | .65 | −0.48 | .021 | .62 |
| 15. Target | −1.44 | <.001 | .24 | −1.06 | <.001 | .35 |
| *16. Foil* | *−.22* | *.374* | .81 | 0.09 | .739 | 1.09 |
| *17. Foil* | *−.05* | *.784* | .95 | −0.11 | .608 | .90 |
| *18. Foil* | *−.99* | .030 | .37 | −0.91 | .040 | .40 |
| 19. Target | −.68 | <.001 | .51 | −0.59 | .001 | .56 |
| 20. Target | −.78 | <.001 | .46 | −0.34 | .064 | .71 |
| *21. Foil* | *−.38* | *.452* | .69 | 0.78 | .158 | 2.19 |
| *22. Target* | *−1.21* | <.001 | .30 | −0.77 | <.001 | .47 |
| *23. Foil* | *−.40* | *.033* | .67 | −0.37 | .059 | .69 |
| 24. Target | −.60 | <.001 | .55 | −0.27 | .114 | .77 |

*Note:* The Univariate Logistic Regression coefficients correspond to a model with only the studied item as predictor of group (Healthy Volunteer = 0 vs. Movement Disorder = 1). The Multivariate Logistic Regression coefficients correspond to a model with all items as predictors (Cox & Snell $R^2$ = .122; Nagelkirke $R^2$ = .202). The regression coefficients (B) are evaluated for significance using *z*-tests.

The 2PL item parameter estimates for the Emory Recognition items, presented in Table 4, indicate consistently positive discrimination parameters (*a*), generally large in magnitude, and difficulty parameters (*b*) clustered near −1.5. One slight exception is Emory Item 5 (Extra Box), although its *a* is still positive. By contrast, the M&L items frequently show items with *a's* close to, and in many instances below, 0, suggesting a number of the items are not functioning well in measuring a common underlying latent proficiency. More commonly it is the target items that appear to be

**Table 4.** 2PL item parameter estimates for the emory items

| Item | *a* (Discrimination) | *b* (Difficulty) |
|---|---|---|
| 1. Diamond | 1.10 | −1.54 |
| 2. Parallel lines | 0.99 | −1.27 |
| 3. Upper triangle | 1.32 | −1.03 |
| 4. Lower cross | 0.96 | −1.23 |
| 5. Extra box | 0.40 | −0.47 |
| 6. RR tracks | 1.72 | −1.19 |
| 7. Bowling ball | 1.37 | −1.91 |
| 8. Nose triangle | 1.17 | −1.13 |
| 9. Inner box | 1.33 | −0.86 |
| 10. Left cross | 1.66 | −2.53 |

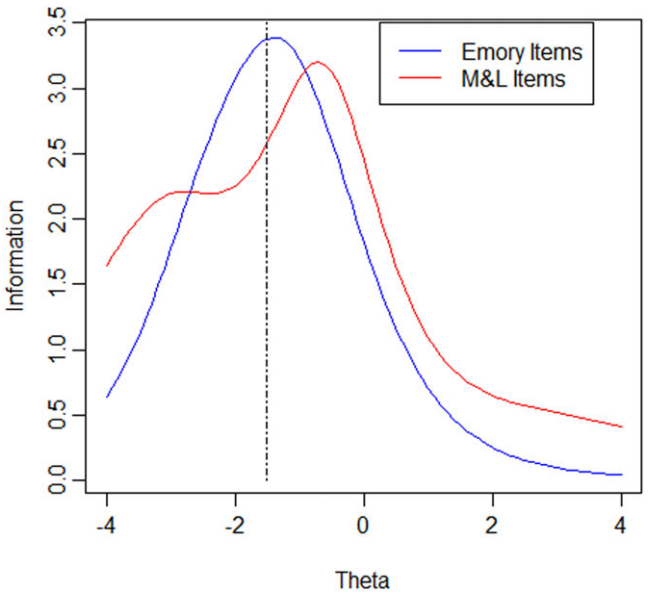


**Figure 2.** Comparison of test information functions. *Note:* test information functions showing measurement precision across levels of recognition proficiency (Theta). Theta reflects underlying recognition ability, scaled to have a mean of 0 and variance of 1 in the EHBS group. Higher curves indicate greater precision (lower measurement error) at that level of proficiency.

functioning poorly, suggesting a latent dimension more closely associated with successful performance on the foil items. While the *b*'s are also quite variable, their interpretation is complicated by the low item *a*'s.

Figure 1 displays the test information functions associated with the IRT analyses above. In addition to the two analyses above, the figure also shows the estimated test information when an IRT analysis is applied to only the Target M&L items. While the underlying proficiencies associated with each IRT analysis are likely not the same, we draw each function against a common Theta metric to illustrate differences in the psychometric functioning of the instruments.

Emory Recognition demonstrates greater information than M&L Recognition at lower proficiency levels and, importantly, maximizes its information near proficiency levels critical for diagnosis (e.g., −1.5 SD, equivalent to Theta = −1.5). In contrast, the M&L Recognition scale provides considerably less information at the lower end of the proficiency spectrum, with its peak information occurring at higher proficiency levels. This distinction underscores the Emory Recognition's potential for more effectively identifying respondents above or below the diagnostic threshold. Furthermore, despite its significantly shorter length (10 items compared to 24), the Emory scale delivers greater information at its point of maximum precision than M&L Recognition.

https://doi.org/10.1017/S135561772510115X Published online by Cambridge University Press

**Table 5.** 2PL item parameter estimates for the M&L elements

| Item | *a* (Discrimination) | *b* (Difficulty) | *Item* | *a* (Discrimination) | *b* (Difficulty) |
|---|---|---|---|---|---|
| 1. Foil | 1.68 | −3.25 | *13. Target* | −0.12 | 14.07 |
| 2. Target | −0.81 | 1.41 | *14. Foil* | 1.63 | −1.13 |
| 3. Foil | 1.72 | −0.16 | *15. Target* | 0.26 | 7.66 |
| 4. Foil | 0.75 | −5.00 | *16. Foil* | 0.74 | −3.01 |
| 5. Target | 0.20 | −14.43 | *17. Foil* | 2.19 | −0.64 |
| 6. Foil | 0.73 | −6.20 | *18. Foil* | 1.25 | −3.66 |
| 7. Target | −0.68 | 3.33 | *19. Target* | 0.50 | 2.12 |
| 8. Target | −0.14 | 11.27 | *20. Target* | 0.64 | 1.76 |
| 9. Target | 0.24 | −8.50 | *21. Foil* | 0.66 | −6.21 |
| 10. Foil | 0.80 | −5.74 | *22. Target* | 0.02 | −110.37 |
| 11. Foil | 1.34 | −2.49 | *23. Foil* | 1.04 | −1.48 |
| 12. Target | −0.60 | 4.08 | *24. Target* | 0.06 | 12.22 |

**Table 6.** Item parameter estimates from confirmatory two-dimensional IRT model of M&L elements

| Recognition Item | $a_1$ | $a_2$ | *d* | $b = \frac{-d}{\sqrt{a_1^2+a_2^2}}$ |
|---|---|---|---|---|
| 1. Foil | 0 | 1.84 | 5.36 | −2.91 |
| 2. Target | 0.79 | 0 | 1.13 | −1.43 |
| 3. Foil | 0 | 1.97 | 0.28 | −0.14 |
| 4. Foil | 0 | 0.84 | 3.78 | −4.50 |
| 5. Target | 0.42 | 0 | 2.94 | −7.00 |
| 6. Foil | 0 | 0.86 | 4.60 | −5.35 |
| 7. Target | 1.13 | 0 | 2.52 | −2.23 |
| 8. Target | 0.53 | 0 | 1.65 | −3.11 |
| 9. Target | 0.57 | 0 | 2.11 | −3.70 |
| 10. Foil | 0 | 0.89 | 4.63 | −5.20 |
| 11. Foil | 0 | 1.46 | 3.36 | −2.30 |
| 12. Target | 0.86 | 0 | 2.57 | −2.99 |
| 13. Target | 0.84 | 0 | 1.88 | −2.24 |
| 14. Foil | 0 | 1.84 | 1.91 | −1.04 |
| 15. Target | 1.30 | 0 | 2.56 | −1.97 |
| 16. Foil | 0 | 0.86 | 2.28 | −2.65 |
| 17. Foil | 0 | 2.20 | 1.34 | −0.61 |
| 18. Foil | 0 | 1.34 | 4.58 | −3.42 |
| 19. Target | 0.58 | 0 | 1.08 | −1.86 |
| 20. Target | 1.26 | 0 | 1.34 | −1.06 |
| 21. Foil | 0 | 0.83 | 4.16 | −5.01 |
| 22. Target | 0.81 | 0 | 2.33 | −2.88 |
| 23. Foil | 0 | 1.13 | 1.55 | −1.37 |
| 24. Target | 0.49 | 0 | 0.81 | −1.65 |

*Note:* Item parameters from the 2PL-2D model: $a_1$ indicates discrimination along the target proficiency dimension (higher values = better differentiation by proficiency); $a_1$ reflects discrimination on the foil proficiency dimension; d denotes multidimensional item difficulty, with larger positive values indicating easier items. Because each item measures only one latent proficiency, item difficulty can also be reported using the *b* parameter of traditional unidimensional IRT.

### *Factor analysis*

To better understand the poorer discrimination of M&L items, a multidimensional IRT analysis, also referred to as an item factor analysis, was conducted. Standardized Root Mean Square Residual (SRMR) reflects the average difference between the observed and model-implied correlation matrices, indicating how well the latent factor structure explains the relationships among items. SRMR values range from 0 to 1, with lower values indicating better fit ( $\leq 0.05$ very good model fit, SRMR 0.05 – 0.08 suggests acceptable fit, and SRMR >0.08 indicates a poor model fit). A single unidimensional factor model resulted in a poor fit for M&L Recognition items (SRMR = .14), with Emory Recognition items demonstrating a good model fit (SRMR = .04). The primary cause of the multidimensionality in M&L Recognition appears due to dimensional distinctions between correct target recognition and correct foil rejections.

We next performed a two-factor IRT model assigning targets and foils items to separate factors. This model demonstrated improved fit for the M&L items (SRMR = .09). Despite improved model fit, the model continues to be a poor fitting model (i.e., SRMR >0.08) but helps confirm the target/foil distinction as a primary cause of the multidimensionality. Table 6 displays the resulting coefficients for this model, where *a1* represents item discrimination in relation to the first latent trait (i.e., target proficiency), with higher values indicating that the item is more effective at differentiating between individuals at different target proficiency levels. Similarly, *a2* represents the item discrimination in relation to the second latent trait (i.e., foil proficiency), while *d* reflects item difficulty, scaled such that larger, more positive values indicate easier items. When the M&L items are separated this way, all items exhibit positive discrimination on their respective proficiencies. However, the correlation between dimensions (r = −0.31) is negative, suggesting that individuals who perform well on target item identification tend to perform worse on rejecting foil items, and vise versa. In other words, those with high proficiency in correctly identifying target items are also more likely to mistakenly classify a foil item as part of the CF. This multidimensionality undoubtedly played a primary role in the poorer item performance seen for the M&L items in the unidimensional IRT analysis.

This pattern suggests the presence of a response set (Cronbach, 1950) to the M&L items in which respondents appear disproportionately prone toward either yes or no responses regardless of the target/foil distinction. The effect appears so strong that better performance on the target proficiency dimension is actually associated with poorer performance on the foil proficiency dimension. Such a response set interpretation would also explain the poor item performance seen for many of the M&L items in both the earlier LR and 2PL IRT analyses.

## Discussion

The Emory CF recognition task demonstrates better group discrimination between cognitively healthy volunteers and patients with various movement disorders compared to the popular M&L Recognition task. Applying d-prime analysis of hits and false-positive recognition to M&L recognition did not meaningfully improve group discrimination. Importantly, when examined across all study participants, Emory CF Recognition demonstrates better psychometric properties, particularly in measuring lower cognitive proficiency levels. Emory CF Recognition's superior group discrimination is attributed to its improved item discrimination and unidimensionality compared to the M&L test, which

https://doi.org/10.1017/S135561772510115X Published online by Cambridge University Press

suffers from multidimensionality and less sensitivity at lower proficiency levels.

Larger group effects sizes were present for delayed CF recall using each scoring approach compared to their associated CF recognition performances. This pattern reflects a common memory retrieval deficit seen in movement disorders, which is characteristic of a "subcortical" cognitive profile. Individuals with this profile typically show impaired free recall but demonstrate significant improvement when recognition-based memory tasks are used. Such a discrepancy suggests that the primary issue lies in memory retrieval, rather than in encoding or storage, consistent with the known cognitive effects of subcortical dysfunction.

There is a clear empirical statistical distinction between performance on targets and foils in M&L Recognition. The negative relationship observed between the underlying dimensions of these item sets strongly suggests response set heterogeneity. This may arise from differences in prior beliefs about the proportion of targets versus foils in the assessment or from variations in response thresholds based on confidence in identifying an item as a target. Alternatively, it may reflect response bias related to confidence when identifying whether an element was present in the figure or not. In yes-no recognition memory testing, response set bias can influence the likelihood of identifying an element as being a target. This bias may stem from individual differences in response tendencies, such as a general inclination to endorse items as previously seen or, conversely, a more conservative approach that limits affirmative responses. Such biases can distort recognition performance by affecting hit and false alarm rates in parallel, making it difficult to distinguish genuine memory ability from response tendencies. Thus, response set bias can impact the overall reliability of recognition assessments by introducing systematic variability unrelated to true memory performance. Regardless of the underlying cause, this distinction appears to hinder M&L Recognition to provide a singular measure of recognition proficiency. We consider it likely that the 4-choice recognition helps address this concern regarding false positives in recognition memory and provides a more homogeneous measure of recognition memory.

Performance on recognition memory tasks can be influenced by factors unrelated to the construct being assessed. Cronbach (Cronbach, 1950) highlighted how individual response sets and response biases can affect tests, potentially decreasing the validity of the assessment. In recognition memory tasks, factors such as confidence in response accuracy, expectations about the proportion of correct versus incorrect responses, and the saliency of correct versus incorrect elements may shape performance. Additionally, clinical factors such as impaired executive function can increase the likelihood of false-positive responses, particularly in yes/no recognition designs, further influencing likelihood of item selection. Response set bias may influence differences in recognition test formats, particularly regarding confidence thresholds when determining whether an item has been previously encountered in yes-no or true-false formats. Cronbach suggests that multiple-choice formats are less prone to response set bias, as they require a response for every item. This approach also helps minimize intrusion errors.

The M&L approach in combining hits and true negatives into a single memory score assumes that false positives reflect impaired memory function, although false-positive errors in recognition memory testing are also often linked to executive function difficulties. As a result, combining both scores may underestimate memory in individuals whose actual recognition memory is intact but whose executive impairments distort their performance, and decreased executive function is associated across a variety of movement disorder diagnoses. This conflation of executive and memory deficits complicates clinical interpretation, highlighting the need for independent response characterization.

Signal detection theory, and specifically the use of d-prime, has been proposed as a valuable method for characterizing recognition memory performance. A meta-analysis of recognition memory in schizophrenia found that d-prime produced more informative effect sizes than traditional accuracy metrics alone (Pelletier et al., 2005). In the present study, d-prime analysis of M&L Recognition data effectively differentiated group membership; however, the observed effect size (Cohen's $d = 0.81$) was not larger than that obtained using the traditional M&L scoring method, which combines correct target identifications and correct foil rejections (Cohen's $d = 0.82$). These findings suggest that, at least for this AVLT-based recognition measure, d-prime analysis does not appear to provide incremental benefit over conventional scoring.

It is important to recognize that the type of bias revealed in our MIRT analysis, and that we frame in relation to response sets, is commonly observed in recognition task measures. MIRT models can also be formulated to explicitly capture multidimensionality in terms of latent d-prime and bias dimensions, effectively representing a rotation of the MIRT solution we examined. The interested reader is referred to Thomas et al. (2018) and DeCarlo (2011) for illustration. One advantage of such models is their potential to clarify how items may differentially reflect sensitivity to *d*-prime bias. As our goal was primarily one of understanding the poorer performance of M&L in relation to Emory Recognition under traditional forms of scoring, we did not pursue such an analysis here, but recognize its value, especially if d-prime were applied in routine scoring of the M&L clinical assessment.

M&L item discrimination estimates (*a's*) are frequently near 0, indicating an item is not discriminating with respect to a unidimensional latent proficiency. Such items contribute little to IRT information, thus explaining why overall test information is not greater for M&L. Also apparent is the tendency for a large number of items with greater *a*'s to have *b*'s above 0 (implying more difficult items). This indicates that many of the M&L items are of high difficulty even for a normal proficiency population, making the scale less useful in measuring individuals with low levels of proficiency.

In conclusion, both CF recognition measures effectively distinguished between groups, but the Emory CF Recognition demonstrated a larger effect size than the M&L Recognition, even when the latter was scored using signal detection theory to account for hits and false positives, surpassing the MoCA as well. Despite its shorter length, the Emory Recognition measure showed superior psychometric properties, particularly in its precision and unidimensionality at lower levels of cognitive functioning. These qualities make it a more effective tool for diagnostic use, especially in populations with cognitive impairment.

**Supplementary material.** For supplementary materials referred to in this article, please visit https://doi.org/10.1017/S135561772510115X

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

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

A preliminary version of the report was presented at the 2025 Meeting of the International Neuropsychological Society, New Orleans, Louisiana, February 13, 2025.

Norms for Emory Complex Figure System and Recognition Task are available at https://med.emory.edu/departments/neurology/_documents/emory_cf_scaled_score_norms.pdf.

## References


Breier, J. I., Plenger, P. M., Castillo, R., Fuchs, K., Wheless, J. W., Thomas, A. B., Brookshire, B. L., Willmore, L. J., & 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.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1–29.

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*, 3–15.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, *10*, 3–31.

DeCarlo, L. T. (2011). Signal detection theory with item effects. *Journal of Mathematical Psychology*, *55*, 229–239.

Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education*, *38*, 1006–1012.

Goetz, M E., Hanfelt, J J., John, S E., Bergquist, S H., Loring, D W., Quyyumi, A., Clifford, G D., Vaccarino, V., Goldstein, F., Johnson 2nd, T M., Kuerston, R., Marcus, M., Levey, A I., & Lah, J J. (2019). Rationale and design of the emory healthy aging and emory healthy brain studies. *Neuroepidemiology*, *53*, 187–200.

Goldstein, F. C., Loring, D. W., Thomas, T., Saleh, S., & Hajjar, I. (2019). Recognition memory performance as a cognitive marker of prodromal Alzheimer's disease. *Journal of Alzheimer's Disease*, *72*, 507–514.

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*, 1352–1366.

Kroll, N. E., Yonelinas, A. P., Dobbins, I. G., & Frederick, C. M. (2002). Separating sensitivity from response bias: Implications of comparisons of yes-no and forced-choice tests for models and measures of recognition memory. *Journal of Experimental Psychology: General*, *131*, 241–254.

Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological Assessment* (4th edn). Oxford University Press.

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*, 1931–1936.

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*, 1–14.

Loring, D. W., Simama, N., Sanders, K., Saurman, J. L., 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.

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin*, *98*, 185–199.

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

Meyers, J. E., & Meyers, K. R. (1995). *Rey complex figure test and recognition trial*. Psychological Assessment Resources.

Nasreddine, Z. S., Phillips, N., Chertkow, H., Rossetti, H., Lacritz, L., Cullum, M., & Weiner, M. (2012). Normative data for the montreal cognitive assessment (MoCA) in a population-based sample. *Neurology*, *78*, 765–766, doi: *78*/10/765-a [pii].

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

Pelletier, M., Achim, A. M., Montoya, A., Lal, S., & Lepage, M. (2005). Cognitive and clinical moderators of recognition memory in schizophrenia: A meta-analysis. *Schizophrenia Research*, *74*, 233–252.

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*, 206–230.

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer.

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

Taylor, L. B. (1969). Localisation of cerebral lesions by psychological testing. *Clinical Neurosurgurgery*, *16*, 269–287.

Thomas, M. L., Brown, G. G., Gur, R. C., Moore, T. M., Patt, V. M., Risbrough, V. B., & Baker, D. G. (2018). A signal detection-item response theory model for evaluating neuropsychological measures. *Journal of Clinical and Experimental Neuropsychology*, *40*(8), 745–760.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1341–1354.

Zhang, X., Lv, L., Min, G., Wang, Q., Zhao, Y., & Li, Y. (2021). Overview of the complex figure test and its clinical application in neuropsychiatric disorders, including copying and recall. *Frontiers in Neurology*, *12*, 680474.


https://doi.org/10.1017/S135561772510115X Published online by Cambridge University Press