Title Page

Title: The validation of a mobile sensor-based neurobehavioral assessment with machine learning analytics

Authors:

Kelly L. Sloane[1], MD, Joel A. Mefford[2*], PhD, Shenly Glenn[3], BS, Zilong Zhao[3], PhD, Man Xu[3], MA, Guifeng Zhou[3], BS, Rachel Mace[4], MS, Amy E Wright[4], BS, Argye E. Hillis[4], MD

[1] Department of Neurology, University of Pennsylvania, Philadelphia, PA, USA
[2] Department of Neurology, University of California, Los Angeles, CA, USA
[3] Miro Health, Inc., San Francisco, CA, USA
[4] Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA
*Corresponding Author

Abstract:

Background: A mobile platform for the self-administration of sensor-based cognitive and behavioral assessments was developed. In addition to measurements typical of legacy neuropsychological tests, Miro also quantifies functions that are currently left to subjective clinical impressions such as motor function, language, and speech.

Objective: Studies were conducted to measure Miro's concurrent validity, test-retest reliability, and amnestic MCI classification performance.

Method: Spearman correlations were calculated to estimate the concurrent validity of Miro and legacy variables using data from 207 study participants. Twenty-six healthy controls were assessed at three time points to evaluate the test-retest reliability of Miro test scores. Reliability was quantified with the scores' intraclass correlations. Learning effects were measured as trends over three assessments. A machine learning algorithm combined Miro scores into a Risk Score to distinguish 65 healthy controls, 21 amnestic MCI (aMCI) participants, and 17 non-amnestic MCI (naMCI) participants.

Results: Significant correlations of Miro variables with legacy neuropsychological test variables were observed. Longitudinal studies show agreement of subsequent measurements and minimal learning effects. The Risk Score distinguished aMCI from healthy controls with an Area Under the Receiver Operator Curve (AUROC) of 0.97; the naMCI participants and controls were separated with an AUROC of 0.80, and the combined (aMCI + naMCI) group was separated from healthy controls with an AUROC of 0.89.

Conclusion: Miro includes valid and reliable versions of scores from legacy neuropsychological test scores and a machine-learning derived risk score that effectively distinguishes healthy controls and individuals with MCI.

Keywords:
neuropsychological assessment, MCI (Mild Cognitive Impairment), amnestic MCI, non-amnestic MCI, neurocognitive, Alzheimer's disease, ADRD, brief screen, cognition, cognitive, neurobehavioral, behavioral, speech, language, acoustic, neurodegenerative

Introduction

Mild Cognitive Impairment (MCI) is a complex clinical disorder that may be caused by many conditions that present with different cognitive profiles. MCI can be an intermediate state between normal aging and dementia. Though individuals with MCI may appear to function normally in the community, cognitive changes put them at higher risk of progressing toward dementia (Albert et al, 2011; Ravaglia et al, 2008). Early, accurate symptom-identification offers the opportunity to provide appropriate interventions, collaborative support, and life planning that can help maximize health outcomes and reduce clinical burden. Cognitive assessments performed by licensed clinical neuropsychologists are the gold-standard for the assessment and diagnosis of MCI (Petersen et al, 2018). However, MCI assessments typically begin with non-specialists, often primary care physicians, who ask a brief set of questions to test a patient's memory or work from a brief cognitive screen such as the Mini Mental State Examination (MMSE) to assess cognitive status.

The MMSE and other brief screens like the MoCA (Montreal Cognitive Assessment) and the TICS (Telephone Interview for Cognitive Status) were designed for brevity, not for precision measurement (Folstein et al, 1975; Nasreddine et al, 2005; Seo et al, 2011). However, the characterization of MCI involves the quantification of subtle shifts away from healthy cognitive function and thus requires the most sensitive cognitive measurements that also possess the ability to track change over time. Longitudinal assessment remains a challenge for legacy neuropsychological testing due to the lack of alternate test versions (Calamia et al, 2012), labor costs, and limited availability of specialists.

Here we present validity studies for a new approach to cognitive assessment using Miro, a mobile, self-administered, interactive tool. Miro provides a potential alternative to the current care protocols. By assessing first symptoms with comprehensive cognitive and behavioral assessments that can detect subtle shifts away from healthy functioning and can distinguish amnestic MCI (aMCI) from non-amnestic MCI (naMCI), clinicians gain the opportunity to deliver effective interventions and support as early in the disease process as possible. Early planning and support can help prolong independence and decrease clinical burden.

Miro captures and quantifies video, voice, and touchscreen recordings of user interactions with Miro modules. Features from these recordings are extracted using proprietary digital signal processing methods that allow Miro to derive variables based on ‘observed’ patient behaviors. While many observations of patient behaviors are clinically relevant, their use is limited to research settings due to the time and labor involved in their capture and quantification. Miro’s patented methods deliver new, equivalent modules at each assessment which supports reliable status monitoring over time and eliminates costs associated with in-person assessment (Glenn and Mefford, 2020).

Other computerized neuropsychological assessments exist, including the NIH Toolbox (Weintraub et al, 2014) and CogState (Maruff et al, 2013. However, neither measures fine motor function, voice, speech, or language performance. Winterlight Labs (Zhu et al, 2018) has developed a mobile app for speech analysis in the detection of dementia, however it lacks repeatability and the cognitive and behavioral measurements that are

required to differentiate Alzheimer's dementia from the many other neurological and psychiatric disorders that also bring changes to speech and language.

The validation of the Miro platform discussed in this paper includes (1a) a concurrent validity study of Miro Version 3 (V.3) and (1b) a concurrent validity study of Miro Version 2 (V.2) that demonstrate the correlation of variable scores common to legacy tests and Miro mobile assessments and (2) a test-retest reliability study to evaluate the stability of Miro's measurement properties. The practical significance of Miro variables is demonstrated by quantifying each variable's performance as a classifier that (3) separates performance of a healthy control from that of a person with mild cognitive impairment (MCI), amnestic MCI (aMCI) and non-amnestic MCI (naMCI). Lastly, (4) individual variable scores are combined and weighted using machine learning methods to generate an "MCI Risk Score" that reliably distinguishes participants with MCI, aMCI, and naMCI from Healthy Controls and distinguishes participants with aMCI and naMCI.

Miro has been designated as a Breakthrough Device by the FDA.

Method

Human Subject Data

This study was approved by the Johns Hopkins University Institutional Review Board (protocol 00088299) and New England IRB (protocols 120180208, 120180211, 120180209 and 12080253) for the use of human subjects. Capacity was determined and consent was obtained from each participant.

Verification of study participants' cognitive statuses

In group-to-group analyses, the practical significance of a score lies in its ability to accurately identify the cohort with which an individual shares the greatest similarity, i.e., how well the score functions as a 'classifier'. When a score or a combination of scores demonstrates different distributions by diagnosis, that score or combination of scores may serve as a clinically valuable classifier. This study analyzed Miro's ability to classify Healthy participants and those with aMCI and naMCI. The numbers of observations used in the classification analysis, together with age and sex distributions are shown in Table 1.

Enrolled participants were assigned to a cohort using two distinct methods: (1) brief screen score; and (2) interpretation of standard neuropsychological test results by a licensed clinical neuropsychologist. Brief screen assignments include 76 healthy controls, and 27 MCI participants. Neuropsychologist assignment numbers for the same participants show 65 healthy controls and 38 MCI participants (Table 2).

Brief screen cohort assignment used the following criteria: Healthy controls were identified by TICS score (Knopman et al, 2010; Seo et al, 2011) of 35-41, or an MMSE score (Mitchell, 2009; O'Bryant, 2008) of 27-30. MCI participants were identified by an MMSE score of 20-26. After assessment with the standard neurocognitive tests used in the concurrent validity study, a clinical neuropsychologist assigned participants to study cohorts based on neurocognitive test performance scores. The neuropsychologist had

access to eligibility and cohort-assignment brief screen scores, demographics, medical history, current status, and legacy test scores.

Of the 76 Healthy Controls identified by the brief screen, 15 were labeled differently by a clinical neuropsychologist (20%). Participants assigned to 'naMCI' exhibited abnormal performance in 1-2 non-memory domains or low-normal performance across 3 or more functional domains. MMSE scores for the naMCI cohort ranged from 23 to 30.

Classification of Users using Miro Variables for Cohort Labeling

The Miro versions of legacy variable scores were evaluated for performance differences in the aMCI and healthy control cohorts based on cohort assignments made by neuropsychologists. The statistical significance in the difference of Miro variable distributions by cohort was tested with t-tests and rank-sum tests. The practical significance of the differences in variable scores across cohorts was evaluated by using each Miro score as a classifier and quantifying the performance of the classifier with the Area Under the Receiver-Operator Curve.

The aMCI Risk Score. Combining and Weighting Variable Scores using machine learning to Identify aMCI

Improvement in the stability and accuracy of score-based classifiers can be achieved through the weighted combination of individual variable scores. When variable scores are combined, a new score is generated that may more robustly and reliably separates groups than can an individual variable score. Statistical and machine learning

approaches can find optimal weighted combinations of variable scores to produce combined scores that effectively associate participants with specific cohorts.

In this study, 110 additional Miro variable scores, derived from sensor-based audio and touchscreen recordings, were added to the variable set for analysis to capture observational data that may be noted on clinical examination but not quantified. Previously, in the concurrent validity and test-retest reliability studies, analyses were limited to Miro variables for which there exists a legacy test variable.

The variables were quantile normalized to eliminate outliers. Missing values were imputed using the SoftImpute algorithm prior to optimization (Hastie et al, 2013). The described method used low-rank matrix completion to impute missing values. The OrdinalNet regression algorithm (Wurm et al, 2017) was used to define a new MCI Risk Score as a linear combination of a subset of those variables. The MCI Risk Score is optimized to separate the aMCI, naMCI, and control groups. The parameters alpha (the ratio of L1 to L2 penalties) and lambda (the scaling factor for the penalties) were optimized for classification performance as measured by AUROC with 5-fold cross-validation. Classification models were not adjusted for age, sex, or education due to the small MCI sample size and the risk of overfitting the model. The new score is a weighted combination of input scores that produced a better classifier for separating aMCI from healthy controls in this data set. This paper will refer to an individual's combined score derived by this method as their "aMCI Risk Score."

Because the data used to select and weigh Miro variable scores also served as inputs to the algorithm, the algorithm was evaluated by leave-one-out cross-validation. The

selected formula for the risk score was applied to the left-out individual's data set to get an out-of-sample predicted risk score. The AUROC was calculated using the risk score as a classifier.

Results

Concurrent Validity

Table 3 shows the legacy tests and paired Miro modules considered for the concurrent validity analysis. Tables 4 and S1 show correlation of Miro results with familiar variables derived from legacy tests. Correlations of scores from Miro V.2 and standard variables displayed in Table S1 show concurrent validities as estimated by Spearman correlations ranging from 0.36 to 0.60. Spearman correlations of displayed scores from Miro V.3 shown in Table 4 fall between 0.27 and 0.66. All correlations shown have 95% confidence intervals that do not include zero. Data for this concurrent validity analysis is included in the supplemental file openMiro_ConcurrentValidity.xlsx.

Miro required less than twenty-five percent of the time (< 1 hour) to administer and score compared with a legacy test battery (4-4.5 hours).

Table S2 includes results of both Pearson and Spearman correlations for the concurrent validity analysis of Miro Version 3 and demonstrates that the two results are strongly in agreement. Spearman and Pearson correlations with 95% confidence intervals and p-values for tests with the null hypothesis that the Spearman correlation is zero were

calculated using R 4.0.3 (R Core Team, 2020) and the DescTools package (Signorell et al, 2020).

Test-Retest Reliability

Table 5 shows Miro variables that possess legacy correlates and ICC estimates between 0.79 and 0.40 with 95 percent confidence intervals that do not include zero. Table S3 shows the estimated trends or learning effects for Miro variables, calculated as the mean slope through a plot of Miro scores for each of three time points in the test-retest study. We also calculated the mean difference in slopes between the first and second assessments, and between the second and third assessments. We call this difference in slopes the ‘progression’. The statistical significance of the trends and of progression was tested with Wald-tests and the resulting p-values are reported in Table S3. No variable evidenced significant trends or progressions. We expect larger datasets to have more power to detect minimal learning effects.

Miro scores as classifiers for cohort labeling

Many Miro variable scores act as effective classifiers for distinguishing paired cohorts. Some that are correlated with legacy testing are shown in Tables 6 and S4. Mean values and standard deviations of Miro scores in the healthy normal and aMCI cohorts are summarized in supplement Table S4. The statistical significance in the difference of Miro score distributions by cohort was tested with t-tests and rank-sum tests, also shown in Table S4.

The practical significance of the differences in Miro scores across cohorts was evaluated by using Miro scores as a classifiers and calculating AUROCs for the separation participants with MCI and healthy controls in Table 6. Performance of the scores-as-classifiers was evaluated using four sets of study participants. Using the diagnoses of aMCI, naMCI, and healthy normal as determined by *neuropsychologists*, classification performances were evaluated for separation of (1) aMCI vs. healthy controls, (2) naMCI vs. healthy controls, and (3) aMCI and naMCI vs healthy controls. Using group assignments based on the *brief screen* scores, classification performance was evaluated for (4) MCI vs control.

<u>The comparison of Miro Risk Score classifier performance when MCI is labeled by brief screen results vs standard neuropsychological testing and neuropsychologist interpretation</u>

Table 6 also shows the classification performance of the aMCI risk score produced by applying the OrdinalNet algorithm to the set of Miro variable scores.

In this data set, when MCI and normal are labeled by a brief screen, Miro's aMCI risk score has a classification performance measured as an AUROC of 0.83. When aMCI, naMCI, and normal labels are determined by a neuropsychologist, Miro's aMCI risk score has an AUROC of 0.97 for classification of aMCI vs normal, and an AUROC of 0.80 for classification of naMCI vs normal. When the study participants labeled aMCI or naMCI are combined into a single MCI group, classification of MCI vs normal by Miro's

aMCI risk score has an AUROC of 0.89. Findings demonstrate the limited ability of brief screens to characterize MCI or designate the amnestic MCI sub-type.

The MCI Risk Score results presented here were conducted on V.3 of the Miro platform. A similar study was conducted with Miro V.2. In the Version 2 study, aMCI Risk Scores were generated using the elastic net algorithm (Zou et al, 2005) rather than OrdinalNet to distinguish aMCI participants and controls. The V.2 aMCI Risk Score classified the aMCI and control participants with an AUROC of 0.92 and separated the combined aMCI and naMCI participants from controls with an AUROC of 0.87.

Discussion

The growth of the aging population and the difficulties associated with obtaining in-person neuropsychological testing highlight the need for accurate, remote, self-administered testing to assess cognitive abilities in this population. Accurate data on cognitive performance, followed over time, can assist in determining the pathology, following the progression, and determining a prognosis of the underlying neurological disease (Oltra-Cucarella et al, 2018).

Our study presents findings of validity of Miro, an accessible, cost-effective, mobile testing modality, compared to legacy testing. The study included concurrent validity, test-retest reliability as well as classification of performance into categories of cognition: healthy control from that of a person with MCI, aMCI, or naMCI. Concurrent validity analysis of variable scores from a battery of legacy neuropsychological assessments and Miro's self-administered assessments shows modest correlation. Modest

correlations are expected given limited test-retest reliability of standard tests (Iverson, 2001; Snow et al, 1989) while positive correlations suggest that Miro and legacy tests quantify equivalent functional variables. Concurrent validity data are included in the supplemental materials.

Many Miro variables demonstrate significant test-retest reliability, as indicated by positive intraclass correlation estimates whose 95% confidence intervals are entirely above zero. The test-retest results across three time points demonstrate that learning effects are minimal and quantifiable across most Miro modules and can be accounted for in longitudinal score calculation.

Further, this study demonstrates that Miro automatically classifies the performance of healthy controls, and those with aMCI and naMCI as well as licensed clinical neuropsychologists for participants within this data set.

One limitation of this study is the small sample sizes. Increasing the sample sizes will allow more stable estimates of the weights for combined scores and evaluation of score performance as a classifier by cross-validation and hold-out sets. Another limitation revolved around the COVID-19 pandemic and difficulties with data collection. In response, study protocols were updated to remove in-person assessment. Because legacy testing requires in-person assessment, the concurrent validity study will be discontinued. Future studies will proceed remotely.

While the translation of legacy tests into touchscreen applications is common, the touchscreen fails to measure key patient behaviors that are observed and accounted for by neuropsychologists during a clinical evaluation. Miro mobile assessments gather

patient behaviors using built-in cameras, microphones, and touch screen sensors. This data is then processed and analyzed to quantify relevant patient behaviors that are subjectively observed by clinicians. To examine the value of specific variables extracted from rich data streams, larger sample sizes are needed.

.

Conclusion

Scores from legacy neuropsychological tests are significantly correlated with 12 Miro variables demonstrating Spearman correlations ($P < 0.05$) with their estimates between 0.27 and 0.68. Thirteen Miro scores have significant ($P < 0.05$) test-retest reliabilities with intra-class correlations between 0.40 and 0.79. A machine-learning derived risk score distinguished aMCI participants from controls with an Area Under the Receiver Operator Curve (AUROC) of 0.97, while the naMCI participants and controls were separated with an AUROC of 0.80, and the combined (aMCI + naMCI) group was separated from healthy controls with an AUROC of 0.89.

Miro allows self-administration of a battery of neuropsychological test assessments with no clinical supervision required. Miro also collects data streams and conducts novel analyses including motor, acoustic, speech, and language. These familiar test scores and novel sensor-based features were combined to make an effective classifier to detect both aMCI and naMCI.

This study considers healthy controls and participants with MCI. Ongoing and future work will collect Miro assessment data from participants with a broad range of potential causes of MCI, including major depression, stroke, non-Alzheimer's neurodegenerative

disorders, and prescriptions medications that may negatively affect cognition. A longitudinal study for predictive prognosis would improve our understanding of Miro's ability to predict symptom and disease progression.

References

1. Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., ... & Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia*, *7*(3), 270-279.
2. American Academy of Neurology (AAN). (2017). *Update: Mild Cognitive Impairment*. Practice Guideline.
3. Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, *26*(4), 543-570.
4. Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, *12*(3), 189-198.
5. Glenn, S., & Mefford, J. (2020). *U.S. Patent No. 10,748,439*. Washington, DC: U.S. Patent and Trademark Office.
6. Hastie, T., Mazumder, R., & Hastie, M. T. (2013). Package 'softImpute'. https://cran.r-project.org/web/packages/softImpute/index.html.

7. Iverson, G. L. (2001). Interpreting change on the WAIS-III/WMS-III in clinical samples. *Archives of Clinical Neuropsychology*, *16*(2), 183-191.
8. Knopman, D. S., Roberts, R. O., Geda, Y. E., Pankratz, V. S., Christianson, T. J., Petersen, R. C., & Rocca, W. A. (2010). Validation of the telephone interview for cognitive status-modified in subjects with normal cognition, mild cognitive impairment, or dementia. *Neuroepidemiology*, *34*(1), 34-42.
9. Maruff, P., Lim, Y. Y., Darby, D., Ellis, K. A., Pietrzak, R. H., Snyder, P. J., ... & Masters, C. L. (2013). Clinical utility of the cogstate brief battery in identifying cognitive impairment in mild cognitive impairment and Alzheimer's disease. *BMC psychology*, *1*(1), 1-11.
10. Miro Health (2017). https://www.mirohealth.com/science/study/mci_2017#content.
11. Mitchell, A. J. (2009). A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. *Journal of psychiatric research*, *43*(4), 411-431.
12. Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., ... & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*(4), 695-699.
13. O'Bryant, S. E., Humphreys, J. D., Smith, G. E., Ivnik, R. J., Graff-Radford, N. R., Petersen, R. C., & Lucas, J. A. (2008). Detecting dementia with the mini-mental state examination in highly educated individuals. *Archives of neurology*, *65*(7), 963-967.

14. Oltra-Cucarella, J., Ferrer-Cascales, R. , Alegret, M., Gasparini, R., Díaz-Ortiz, L. M., Ríos, R., ... & Sanchez-SanSegundo, M. (2018). Risk of progression to Alzheimer's disease for different neuropsychological Mild Cognitive Impairment subtypes: A hierarchical meta-analysis of longitudinal studies. *Psychology and aging*, *33*(7), 1007.
15. Petersen, Ronald C., Lopez, O, Armstrong, M. J., Getchius, T. S. D., Ganguli, M., ...Rae-Grant, A.. (2018). Practice guideline update summary: Mild cognitive impairment. Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology. *Neurology*, January 16, 2018; 90 (3).
16. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
17. Ravaglia, G., Forti, P., Montesi, F., Lucicesare, A., Pisacane, N., Rietti, E., ... & Mecocci, P. (2008). Mild cognitive impairment: epidemiology and dementia risk in an elderly Italian population. *Journal of the American Geriatrics Society*, *56*(1), 51-58.
18. Seo, E. H., Lee, D. Y., Kim, S. G., Kim, K. W., Kim, D. H., jo Kim, B., ... & Kim, J. W. (2011). Validity of the telephone interview for cognitive status (TICS) and modified TICS (TICSm) for mild cognitive impairment (MCI) and dementia screening. *Archives of gerontology and geriatrics*, *52*(1), e26-e30.

19. Signorell, A., Aho, K., Alfons, A., Anderegg, N., Aragon, T., Arppe, A., ... & Borchers, H. W. (2020). DescTools: tools for descriptive statistics. R package version 0.99.34.
20. Snow, W. G., Tierney, M. C., Zorzitto, M. L., Fisher, R. H., & Reid, D. W. (1989). WAIS-R test-retest reliability in a normal elderly sample. *Journal of clinical and experimental neuropsychology*, *11*(4), 423-428.
21. Trzepacz, P. T., Hochstetler, H., Wang, S., Walker, B., Saykin, A. J., & Alzheimer's Disease Neuroimaging Initiative. (2015). Relationship between the Montreal Cognitive Assessment and Mini-mental State Examination for assessment of mild cognitive impairment in older adults. *BMC geriatrics*, *15*(1), 107.
22. Qin, S., Nelson, L., McLeod, L., Eremenco, S., & Coons, S. J. (2019). Assessing test–retest reliability of patient-reported outcome measures using intraclass correlation coefficients: recommendations for selecting and documenting the analytical formula. *Quality of Life Research*, *28*(4), 1029-1033.
23. Weintraub, S., Dikmen, S.S., Heaton, R.K., Tulsky, D.S., Zelazo, P.D., Slotkin, J., Carlozzi, N.E., Bauer, P.J., Wallner-Allen, K., Fox, N. and Havlik, R., 2014. The cognition battery of the NIH toolbox for assessment of neurological and behavioral function: validation in an adult sample. *Journal of the International Neuropsychological Society: JINS*, *20*(6), p.567.
24. Wolak, M. (2016). Package 'ICC'. https://cran.r-project.org/web/packages/ICC/ICC.pdf.

25. Wurm, M. J., Rathouz, P. J., & Hanlon, B. M. (2017). Regularized ordinal regression and the ordinalNet R package. *arXiv preprint arXiv:1706.05003*.
26. Zhu, Z., Novikova, J., & Rudzicz, F. (2018). Detecting cognitive impairments by agreeing on interpretations of linguistic features. *arXiv preprint arXiv:1808.06570*.
27. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301-320.

Tables

Table 1. Study participant demographics

Cohort	Participants	Female / Male	Avg Age (range) in yrs
Participant demographics for Concurrent validity analysis of Miro Version 2			
Normal Control	19	15 / 4	73.2 (64 - 89)
Affected	27	16 / 11	78.5 (68 - 95)
Participant demographics for Concurrent validity analysis of Miro Version 3			
Normal Control	103	58 / 45	70.4 (61 - 82)
Affected	58	18 / 40	68.9 (48 - 83)
Test-retest reliability study participants			
	26	16 / 10	72.0 (64 - 82)
Participants used in classification analyses			
Healthy Controls	65	36 / 29	70.2 (61-82)
aMCI	21	6 / 15	71.1 (49-83)
naMCI	17	6 / 11	71.1 (60-82)

MCI=mild cognitive impairment, aMCI = amnestic MCI, naMCI = non-amnestic MCI

Table 2. Identification of MCI by brief screen vs. clinical neurocognitive assessment

Cohort	**Label assigned by** brief screen	**Label assigned by** clinical neuropsychologist	**Label change by** clinical neuropsychologist
Healthy Control	76	65	(-13 participant / +2 participants)
aMCI	27	21	(-13 participants/ +7 participants)
naMCI	N/A	17	(+19 participants)

Table 3. Concurrent validity tests

Clinical Neuropsychological Tests Administered in-person by a licensed clinical neuropsychologist	**Miro** Self-Administered mobile assessment
Hopkins Verbal Learning Test (HVLT-R) Immediate Recall	Spy Games
HVLT Free Recall + Cued Recall	Spy Report
Iowa Trail Making Test (A and B)	Bolt Bot
Digit Span Forward and Backward (WAIS-IV)	Hungry Bees
Coding (WAIS-IV)	Treasure Tomb
Design Fluency (D-KEFS) (Filled Dots, Filled + Empty Dots)	Chart-A-Course
Phonemic Fluency (D-KEFS)	Lucky Letters
Category Fluency (D-KEFS)	Categories
Spatial Span Forward and Backward (WNSA)	Follow the Glow
Stroop Test (PAR)	Colors and Words
Confrontation Naming (BNT-15)	Wordy Goat
Finger Tapping Test	Take Flight
Sentence Repetition (V2.0 NACC)	Repeat After Me

Table 4. Concurrent validity of Miro Version 3 vs. comparative neurocognitive tests

Variable	**Spearman (95% CI)**	**P-value**

Digit Span Backward: Longest correct span	0.68 (0.54, 0.79)	1.2E-11***
Phonemic Fluency: Total correct	0.67 (0.55, 0.76)	1.8E-15***
Category Fluency: Total correct	0.65 (0.53, 0.75)	3.1E-14***
Digit Span Forward: Longest correct span	0.57 (0.43, 0.69)	3.6E-11***
Trails B: Total time	0.55 (0.40, 0.67)	3.8E-09***
Design Fluency A: Total correct	0.46 (0.28, 0.61)	4.4E-06***
Trails A: Total time	0.46 (0.29, 0.60)	9.5E-07***
Coding: Total correct	0.40 (0.21, 0.57)	1.1E-04***
Finger Tapping: Average taps nondominant	0.35 (0.15, 0.52)	8.4E-04***
Spatial Span Backward: Longest correct span	0.33 (0.15, 0.50)	6.3E-04***
Finger Tapping: Average taps dominant	0.32 (0.11, 0.49)	2.7E-03**
VLM: Largest correct set spoken	0.27 (0.06, 0.47)	0.013**

Spearman correlation estimates and p-values for tests to reject the assumption that the true correlations are zero; VLM = Verbal Learning and Memory
**Significant at P<0.05 **Significant at P<0.01 ***Significant at P<0.001*

Table 5. Intraclass correlations for Miro versions of legacy scores

Variable	**ICC (95% CI)**
Design Fluency Switching: Total correct	0.79 (0.65, 0.89)*
Trails B: Total time	0.76 (0.62, 0.86)*
Phonemic Fluency: Total correct	0.75 (0.61, 0.85)*
Trails A: Total time	0.74 (0.60 0.85)*
DVLM: Total correct	0.72 (0.47, 0.86)*
Coding: Total correct	0.61 (0.42, 0.76)*
Digit Span Backward: Longest correct span	0.58 (0.23, 0.78)*

Digit Span Forward: Longest correct span	0.58 (0.23, 0.78)*
VLM: Largest correct set spoken	0.56 (0.34, 0.74)*
DVLM: Total errors	0.54 (0.19, 0.76)*
Design Fluency A: Total correct	0.53 (0.33, 0.70)*
Finger Tapping: Average taps dominant	0.42 (0.21, 0.63)*
Spatial Span Forward: Total correct	0.40 (0.19, 0.61)*

VLM = Verbal Learning and Memory; DVLM = Delayed Verbal Learning and Memory.
*significantly different than zero at the 0.05 level

Table 6. Performance of the MCI risk score and ten legacy test scores on separation of healthy controls from MCI participants using diagnoses from either a Neuropsychologist or a brief screen

Score	**HC vs aMCI**	**HC vs naMCI**	**HC vs MCI combined**	**HC vs MCI**
Trails A: Total time	0.94	0.69	0.83	0.79
Category Fluency: Total correct	0.92	0.69	0.82	0.78
Coding: Total correct	0.91	0.68	0.80	0.79
Trails B: Total correct	0.89	0.69	0.80	0.80
Phonemic Fluency: Total correct	0.88	0.79	0.84	0.84
Trails B: Total time	0.85	0.72	0.80	0.79
Digit Span Forward: Longest correct span	0.85	0.55	0.70	0.63
Picture Description: Total prepositions	0.84	0.76	0.76	0.68
Digit Span Backward: Longest correct span	0.82	0.57	0.70	0.72
Design Fluency: Total correct	0.81	0.63	0.72	0.72
MCI Risk score	0.97	0.80	0.89	0.83

Appendix

Table S1. Concurrent validity of Miro Version 2 vs. comparison neurocognitive tests
**Significant at P < 0.05 **Significant at P < 0.01 ***Significant at P < 0.001*

Score	Spearman (95% CI)	P-value
Digit Span Backward: Longest correct span	0.60 (0.37, 0.76).	1.5E-05***
Coding: Total correct	0.58 (0.34, 0.75)	3.1E-05***
Design Fluency A: Total correct	0.58 (0.34, 0.75)	4.3E-05***
Digit Span Forward: Longest correct span	0.49 (0.23, 0.68)	5.4E-04***
Category Fluency: Total correct	0.49 (0.23, 0.68)	5.9E-04***
Finger Tapping: Avg taps dominant	0.50 (0.21, 0.71)	1.5E-03**
Phonemic Fluency: Total correct	0.42 (0.14, 0.63)	3.9E-03**
Finger Tapping: Avg taps nondominant	0.36 (0.04, 0.61)	0.027*

Table S2. Comparison of Spearman and Pearson correlations for concurrent validity

Score	Spearman (95%CI)	p-value	Pearson (95% CI)	p-value
Phonemic Fluency: Total correct	0.67 (0.55, 0.76)	1.8E-15***	0.67 (0.56, 0.77)	8.5E-16***
Category Fluency: Total correct	0.65 (0.53, 0.75)	3.1E-14***	0.57 (0.43, 0.69)	1.4E-10***
Digits Backward: Longest correct span	0.68 (0.54, 0.79)	1.2E-11***	0.70 (0.56, 0.80)	1.9E-12***
Digits Forward: Longest correct span	0.57 (0.43, 0.69)	3.6E-11***	0.62 (0.49, 0.72)	2.3E-13***
Trails B: Total time	0.55 (0.40, 0.67)	3.8E-09***	0.49 (0.32, 0.63)	2.6E-07***
Trails A: Total time	0.46 (0.29, 0.60)	9.5E-07***	0.33 (0.15, 0.49)	6.0E-04***
Coding: Total correct	0.40 (0.21, 0.57)	1.1E-04***	0.40 (0.21, 0.56)	1.1E-04***
Spatial Span Backward: Total correct	0.33 (0.15, 0.50)	6.3E-04***	0.41 (0.23, 0.56)	2.4E-05***

Finger Tapping: Avg taps nondominant	0.35 (0.15, 0.52)	8.4E-04***	0.26 (0.06, 0.44)	0.013*
Finger Tapping: Avg taps dominant	0.32 (0.11, 0.49)	2.7E-03**	0.27 (0.06, 0.45)	0.011*
VLM: Largest correct set spoken	0.27 (0.06, 0.47)	0.013*	0.25 (0.04, 0.45)	0.023*
Design Fluency Switching: Total correct	0.28 (-0.02, 0.54)	0.064	0.18 (-0.13, 0.46)	0.25
Phonemic Fluency: Error set loss	-0.09 (-0.29, 0.11)	0.38	-0.08 (-0.28, 0.13)	0.47
Phonemic Fluency: Error repetitions	0.09 (-0.11, 0.29)	0.39	0.29 (0.09, 0.46)	4.8E-03**
Design Fluency Switching: Error repeat	0.09 (-0.12, 0.30)	0.39	0.20 (-0.02, 0.40)	0.07
Category Fluency: Error repeat	0.06 (-0.15, 0.26)	0.57	0.01 (-0.19, 0.21)	0.94
Design Fluency Switching: Total errors	-0.08 (-0.37, 0.22)	0.59	-0.14 (-0.42, 0.17)	0.38
Category Fluency: Error set loss	-0.04 (-0.25, 0.16)	0.68	0.01 (-0.20, 0.21)	0.94
Design Fluency A: Total errors	-0.02 (-0.24, 0.20)	0.85	-0.06 (-0.27, 0.16)	0.61

VLM = Verbal Learning and Memory; DVLM = Delayed Verbal Learning and Memory
**Significant at P<0.05 **Significant at P<0.01 ***Significant at P<0.001*

Table S3. Test-retest analysis of learning effects

Score	**Avg learning effect (sd)**	**Learning effect P-value**	**Avg progression (sd)**	**Progression P-value**
Category Fluency: Total correct	-0.019 (0.55)	0.99	-0.12 (1.8)	0.99
Category Fluency: Error set loss	0.000 (0.50)	1.00	0.35 (1.9)	0.97
Coding: Total correct	-0.089 (0.40)	0.97	0.27 (1.6)	0.97
Design Fluency A: Total correct	0.100 (0.51)	0.97	-0.25 (1.4)	0.97
Design Fluency Switch: Error repeat	0.025 (0.72)	0.99	0.37 (1.7)	0.97
Design Fluency A: Total errors	-0.130 (0.57)	0.96	-0.10 (1.9)	0.99
Design Fluency Switch: Total correct	0.043 (0.31)	0.98	-0.13 (1.1)	0.98
Design Fluency Switch: Error repeat	-0.063 (0.39)	0.97	-0.55 (1.7)	0.95

Finger Tapping: Avg taps				
nondominant	0.002 (0.47)	1.00	-0.031 (1.1)	1.00
Finger Tapping: Average taps				
dominant	-0.025 (0.46)	0.99	-0.43 (1.2)	0.95
Phonemic Fluency: Total correct	-0.060 (0.30)	0.97	-0.29 (1.1)	0.96
Phonemic Fluency: Error repetitions	0.063 (0.50)	0.98	0.26 (1.6)	0.97
Trails A: Total time	-0.053 (0.32)	0.97	0.54 (1.2)	0.93
Trails B: Total time	-0.080 (0.32)	0.96	0.20 (1.1)	0.97
VLM: Largest correct set spoken	0.036 (0.51)	0.99	0.39 (1.3)	0.95
VLM: Error set loss	-0.035 (0.44)	0.99	-0.06 (1.1)	0.99
VLM: Error repeat	-0.048 (0.27)	0.97	-0.22 (2.4)	0.99

The difference in the trend from assessment 2 to 3 minus the trend from assessment 1 to 2 is shown in Table S1 and is called, 'progression'. Estimates of progression with their standard errors and p-values from a t-test for significance are included. VLM = Verbal Learning and Memory; DVLM = Delayed Verbal Learning and Memory.

Table S4. Means, standard deviations, and tests of differences in distribution of Miro scores in Healthy Control and aMCI subjects

Score	**Average(sd) in Healthy Controls**	**Average(sd) in aMCI patients**	**Rank sum test P-value**	**t-test P-value**
Trails A: Total time	16 (4.0)	28 (8.7)	1.6E-09***	8.9E-13***
Category Fluency: Total correct	13 (5.0)	4.1 (3.7)	7.5E-09***	6.8E-11***
Phonemic Fluency: Total correct	17 (5.0)	8.3 (5.4)	2.0E-07***	1.3E-09***
Trails B: Total time	87 (21)	110 (14)	1.3E-06***	2.7E-06***
VLM: Error set loss	0.05 (0.21)	0.62 (1.0)	4.7E-04***	1.4E-04***
VLM: Total errors	1.1 (2.6)	2.3 (4.1)	1.1E-03**	0.19
VLM: Error repeat	1.1 (2.6)	1.7 (3.5)	1.8E-03**	0.46
Phonemic Fluency: Error repetitions	1.8 (1.7)	1.1 (1.6)	0.05	0.12
Design Fluency A: Total correct	14 (3.8)	9.4 (3.7)	0.07	8.6E-05***
Phonemic Fluency: Error set loss	0.67 (1.7)	1.2 (1.5)	0.09	0.18
Design Fluency A: Error set loss	2.8 (2.5)	2.5 (1.8)	0.33	0.65
Category Fluency: Error repeat	0.67 (1.0)	0.52 (1.0)	0.34	0.58
Design Fluency A: Total errors	5.1 (2.8)	4.8 (3.1)	0.37	0.72
Coding: Total correct	37 (11)	21 (8.2)	0.41	4.7E-06***
VLM: Largest correct set spoken	7.5 (5.0)	4.9 (1.5)	0.66	0.07
Category Fluency: Error set loss	1.7 (3.5)	1.9 (2.6)	0.67	0.88

VLM = Verbal Learning and Memory; DVLM = Delayed Verbal Learning and Memory
**Significant at P<0.05 **Significant at P<0.01 ***Significant at P<0.001*