Validation of Miro: a mobile, repeatable, self-administrable brain assessment platform

Shenly Glenn, Joel Mefford, Amy E Wright, Shauna Berube, Argye E. Hillis

Miro Overview

About Miro. Miro is a mobile platform that provides in-depth assessment tools to measure a wide range of brain functions including motor, speech, language, attention, cognition, emotion and social function. Rich data types, including audio, video, motion and touch screen data, show a marked improvement in the characterization of brain function over pencil-paper tests and mouse clicks. Miro measurement paradigms were developed in collaboration with leading researchers in respective functional areas. Data-driven analyses further Miro's improvement in the characterization of brain function over human-driven analyses. Miro developed its usability and Artificial Intelligence based on the collection of data from more than 700 users.

How Miro works. Miro assessments run on iPads. Results are wirelessly transmitted to secure servers. Performance data is processed and analyzed by machines and reviewed by human experts. Features, variables and aggregates are flexibly combined and analyzed to discover patterns that may better characterize individual performance, discover phenotypes and disease sub-types, and evaluate therapeutic interventions over time. Clinicians and researchers access Miro scores via a web interface with downloadable .csv files. Miro aims to support improvements in clinical care and research by advancing functional brain assessment in six key areas:
1. Collection of broader data types (e.g., behavior, speech and motor function)
2. Direct recordings of patient performance that are not intermediated by a human clinician
3. Increased range, resolution and precision of captured data
4. Longitudinal performance monitoring (reliable, repeatable, self-administrable assessments)
5. Data-driven analysis (and the continual discovery of informative variables and aggregates)
6. Improved usability for clinicians, researchers and patients

Study Overview

Data is presented on the feasibility and effectiveness of Miro, a neurological health platform and mobile neurological assessment and monitoring app, to support the accurate characterization of a wide range of brain functions. The quantification of brain function can be used for the diagnosis, monitoring and management of brain conditions. **The detection of subtle functional change** is important for the discovery and monitoring of brain disorders as well as the monitoring of the efficacy of therapeutic intervention. Miro's ability to to characterize subtle changes is measured by an "MCI Risk Score". The sensitivity and specificity of Miro's MCI Risk Score is measured in 38 subjects with Mild Cognitive Impairment (MCI) and 32 Normal Control subjects. **The reliable, repeat assessment of neurological function** is a cornerstone for diagnosis and monitoring. The reliability of Miro's patent-pending Infinite Versioning(TM) that supports longitudinal assessment was tested. Test-retest reliability was evaluated over three time points in 23 healthy volunteers. Miro's **correlation to existing clinical practice** is also tested. Concurrent validity, as quantified by correlation of Miro scores to traditional clinician-administered, pencil-paper neuropsychological test scores, was evaluated in 19 healthy volunteers and 33 disordered subjects. The accurate differentiation of functional impairment by domain is critical for precision diagnostics and intervention. An exploratory analysis of Miro's potential to differentiate disorders with significant overlapping symptoms in patient populations with memory impairment, speech and language impairment, executive function impairment, and motor impairment is examined.

Human Subjects: This study received Investigational Device Exemption (IDE) approval from Johns Hopkins University Hospital Institutional Review Board (IRB) and from New England IRB.

Introduction

The assessment and monitoring of brain health is an increasingly critical component of overall health. The proper characterization and management of disorders can lead to more accurate diagnosis, more appropriate treatment options, and improved prognosis. Despite the importanceof accurate brain measures, and the growth of computer-based testing, reliable clinical options are limited. The most advanced assessments lack the ability to accurately distinguish disorders with overlapping symptoms or to characterize sub-types of disorders, contributing to high misdiagnosis rates, mis-medication and increased healthcare costs. The most common approach to the measurement and monitoring of brain function, the pencil-paper brief screen, and recent computer-based brief-screens, present significant clinical challenges:

(1)First, they lack the resolution and range to detect subtle, but debilitating, functional changes and thus cannot be used to detect or monitor mild to moderate conditions nor track therapeutic efficacy.

(2)Second, their data capture is limited only to cognitive function. They neglect emotional, social, motor, speech and language functions.
(3)Third, they lack repeatability and thus are unable to track brain health over time.
(4)Fourth, they lack reliability: administration and scoring vary either by clinician or computer system. Systems with differences in either hardware or software introduce inconsistencies in measurements.
(5)Fifth, they are not self-administrable nor scalable; they require in-office, one-on-one clinician-patient assessments and are logistically and financially cost-prohibitive.
(6)Sixth, humans pre-determine performance scores; the lack of machine-driven exploration of the data prevents the evolution of insights as the data set grows.

There is growing consensus among clinicians and researchers of the need for precise, relevant and accessible tools that can better characterize brain function. It is well understood that patients with brain disorders would benefit from efficient, ongoing, in-depth functional brain assessments, just as cardiovascular patients benefit from regular blood pressure and heart rate monitoring. Miro is developing accessible, reliable, efficient and cost-effective methods to accurately characterize brain function. Miro is innovative in a variety of ways:
• Miro assessment combines novel audio, video, gyrometer and touchscreen data capture and analysis with interactive analogues of neurological, psychiatric and neuropsychological exams.
• Miro's data processing pipeline and machine-learning engine support continuous discovery and improvement.
• Miro scores are stored with raw audio, video and touchscreen recordings for future reference.
• Miro is modular and can be tailored to meet the needs of each clinician and researcher.
• Infinite Versioning (TM) (patent-pending) allows Miro modules to be administered over time with little discernible learning effects.
• Miro is self-administrable and available on-demand, permitting in- or out-of-office assessment.

MIRO MODULES

Attention and response inhibition
Measures: Sustained attention, simple reaction time and response inhibition
Time: variable (~120 seconds)
Description: (A) Scary ghosts appear on the screen. The user must tap the ghosts as quickly as possible. (B) Scary and kind ghosts appear on the screen. The user must tap the scary ghosts as quickly as possible and avoid tapping the kind ghosts.

Category fluency
Measures: Word generation, flexibility and working memory (rule monitoring).
Time: 90 seconds
Description: A category (e.g., fruits) appears on the screen. Users must say as many words as possible that belong to that category.

Choice reaction time
Measures: Psycho-motor speed
Time: ~60 seconds (variable)
Description: Two images appear on the screen and the user must decide whether they are the same or different.

Coding
Measures: Processing speed, fine motor function, implicit memory
Time: 90 seconds
Description: The user deciphers as many codes as possible in 90 seconds by matching number/symbol pairs

Design fluency
Measures: Visual generativity, flexibility and working memory (rule monitoring)
Time: 90 seconds
Description: 5 locations appear on a map. Users are instructed to make as many unique paths as possible that connect the locations.

Digit span forward and backward
Measures: Basic auditory attention, auditory memory span and working memory
Time: variable (~120 seconds)
Description: (A) The user is instructed to listen to the number sequences and then to repeat them back loudly and clearly; (B) The user is instructed to listen to the number
sequences and then to repeat them in reverse order loudly and clearly.

Divided attention
Measures: Attention, divided attention
Time: variable (~120 seconds)
Description: An object is hidden under one of three cups. As the cups move, the user must track the hidden object. When the cups stop moving, the user is instructed to tap the cup hiding the object.

Face-name learning and memory
Measures: Face-name learning and memory
Time: variable
Description: People appear on the screen and are introduced to the user. The user listens to their names and recalls them when prompted with pictures. If correct, the user advances to the next level that includes the faces and names on that level and previous levels. The module ends when the user fails to recall at least 80% of words over three repeated trials. Delayed free recall follows a 20 minute delay. The user is prompted to identify as many faces as possible.

Fine motor speed: finger tapping
Measures: Fine motor speed and consistency
Time: ~120 seconds
Description: Hands are anchored to the iPad in specified locations. The user is instructed use their index finger to tap as fast as possible. User performs 3 trials per hand, switching back and forth between hands.

Free speech: Picture description
Measures: Speech, language and grammar, voice, emotion
Time: 120 seconds (max)
Description: Subject sees an animation and is asked to describe the scene out loud in as much detail as possible and in full sentences.

Immediate recall (visual)
Measures: Immediate image memory
Time: Variable (~120 seconds)
Description: images are presented to the user. If the user has seen the image before, the user clicks on the image. If the user has not seen the image before, the user doesn't click on the image. (Display speed can be adjusted for people with motor disorders).

Irregular word reading
Measures: Ability to properly read and pronounce non-phonetic words
Time: ~120 seconds
Description: Ten irregular words and ten regular words are displayed on the screen. The user pronounces each word out loud.

Naming
Measures: Confrontation naming
Time: variable (~60 seconds; max = 225 seconds)
Description: Three objects are displayed on the screen. The user is given up to 15 seconds to name each item.

Repetition
Measures: articulation
Time: variable (~90 seconds; up to 4 minutes for low performers)
Description: Words and phrases are presented one by one in a call and response style. Each of 3 words is said aloud five times and each of three phrases is said once. Users are asked to say and repeat aloud each word and phrase exactly as it is presented.

Response inhibition
Measures: Response inhibition (go/no-go)
Time: ~180 seconds
Description: Objects appear on the screen, half require that the user tap them, half require that the user refrain from tapping them.

Saccades and anti-saccades
Measures: Reaction time, inhibition, eye movement
Time: variable (~120 seconds)
Description: The user sees a stimulus in the center of the screen. A light then appears to either side of the screen. (A) The user is instructed to look at the light and the stimulus that appears on the side of the screen and to verbally acknowledge whether the stimulus is the 'same' or 'different' from the stimulus presented in the middle of the screen. (B) The user is instructed to look away from the light that appears on the side of the screen (a stimulus has appeared on the opposite side of the screen from the light) and to verbally acknowledge whether the stimulus on the opposite side of the screen from the light is the 'same' or 'different' from the stimulus presented in the middle of the screen.

Set shifting
Measures: Mental flexibility, sustained attention, processing speed
Time: 120 secs (max)
Description: The numbers 1-16 and the letters A-P are arranged in a designated order on the screen. The user must switch back and forth between numbers and letters in ascending order as fast as possible. The module ends at completion or 7 consecutive mistakes.

Simple reaction time
Measures: Simple reaction time
Time: ~45 seconds
Description: Objects appear on the screen. The user must tap each object as it appears.

Spatial learning and memory
Measures: Spatial learning and memory
Time: variable (up to 10 minutes for high performers)
Description: The user is introduced to a town map. The user must remember groups of objects, people and places that appear on the map in increasing numbers. If correct, the user advances to the next level that includes the original group plus new additions. The module ends when the user fails to recall at least 80% of the elements over three repeated trials. Delayed free recall follows a 20 minute delay. The user is prompted to locate as many items as they can.

Spatial location and immediate visual memory
Measures: Immediate visual memory, spatial location accuracy
Time: variable (<90 seconds)
Description: A three-letter nonsense word flashes on the screen and then disappears, followed by an object or a group of objects, placed at controlled-random locations. The user is instructed to touch the screen exactly where the objects appeared and is then asked to identify the threeletter nonsense word.

Spatial span forward and backward
Measures: Basic visual attention, spatial memory span and working memory
Time: variable (~120 seconds)
Description: (A) The user is instructed to watch the spatial sequences and then to repeat them on the touchscreen; (B) The user is instructed to watch the spatial sequences and then to repeat them in reverse order on the touchscreen.

Verbal fluency
Measures: Verbal generation, flexibility and working memory (rule monitoring).
Time: 90 seconds
Description: A letter appears on the screen. Users must say as many words as possible that begin with that letter.

Verbal list learning and memory
Measures: Verbal learning and memory
Time: Variable depending on ability (up to 10 minutes for high performers)
Description: The user must remember a passphrase of increasing length. The user listens to the passphrase and repeats it back. If correct, the user advances to the next level that includes the original set of words plus new additions to the list. The module ends when the user fails to recall at least 80% of words over three repeated trials. Delayed free recall follows a 20 minute delay. The user is prompted to repeat as many items as they can, in any order.

Visual search
Measures: Visual search
Time: 90 seconds (max)
Description: The numbers 1-16 are arranged in a designated order on the screen. The user must tap the numbers in ascending order as fast as possible. The module ends at completion or 7 consecutive mistakes.

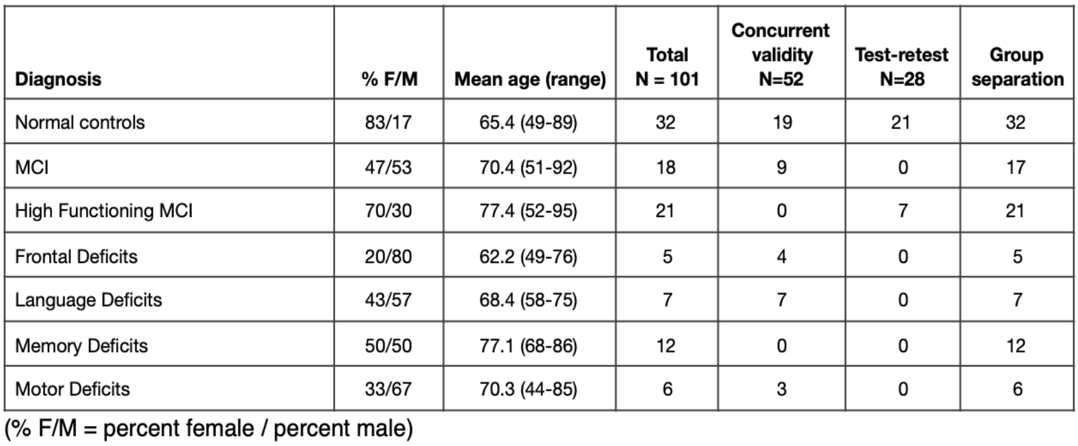
Working memory
Measures: Working memory
Time: variable (~60 seconds)
Description: Objects, such as coins, enter the screen one by one and disappear. The user must track the number of each type of object the enters the screen and enter it at the end of each level.

Study Description

THE DISCRIMINATION BETWEEN SUBJECTS WITH MILD COGNITIVE IMPAIRMENT
AND NORMAL CONTROL

Seventy subjects, comprised of 32 cognitively normal volunteers, and 38 subjects with mild cognitive impairment (MCI) were included in an analysis of the ability of the Miro platform to distinguish subjects with MCI from normal subjects.

Table 1. Demographics



Diagnosis	% F/M	Mean age (range)	Total N = 101	Concurrent validity N=52	Test-retest N=28	Group separation
Normal controls	83/17	65.4 (49-89)	32	19	21	32
MCI	47/53	70.4 (51-92)	18	9	0	17
High Functioning MCI	70/30	77.4 (52-95)	21	0	7	21
Frontal Deficits	20/80	62.2 (49-76)	5	4	0	5
Language Deficits	43/57	68.4 (58-75)	7	7	0	7
Memory Deficits	50/50	77.1 (68-86)	12	0	0	12
Motor Deficits	33/67	70.3 (44-85)	6	3	0	6

(% F/M = percent female / percent male)

The subjects with MCI were a heterogeneous group with varying levels of performance. Some were referred by community neuropsychologists in Northern California and others identified by researchers at Johns Hopkins University. To account for the functional range in the MCI subjects, experts identified 21 MCI subjects as being “High Functioning MCI” and 17 subjects as “MCI”. “High functioning MCI” includes subjects who perform within the normal range on standard tests of cognitive function, but who present with complaints of perceived cognitive deficits. “MCI” includes subjects whose performance on standardized tests fall within the range of Mild Cognitive Impairment. Automated MCI Risk Score classification was developed with data from 32 Normal Control subjects and the 17 subjects with MCI. Both MCI and High Functioning MCI groups were analyzed according to the MCI Risk Score.

Methods. Standardized versions of basic variable scores were combined to form an MCI Risk Score. This score is designed to specifically distinguish performance of normal subjects from the performance of subjects with Mild Cognitive Impairment. Basic variable scores were standardized based on the Normal Control data set to have means set to zero and standard deviations set to one. For a minimal subset of Miro modules with non-equivalent versions, basic scores were standardized independently, per version. These included: 1. Picture Description, wherein each picture to be described produces a unique lexicon; Category Fluency, wherein each category to be explored produces an independent word list; and Letter Fluency, wherein each letter to initiate word-production varies in difficulty and produces a unique word-list length. Prior to combining variables into aggregate scores, each standardized variable score was quantile-normalized to mitigate undue influence of outliers or peculiar distributions of any individual scores. Missing values were imputed using low rank matrix completion[1]. If a subject participated in multiple assessments (as for test-retest 1 reliability), only initial (T1) assessment results were included in the discrimination analysis. The process for combining normalized variables to form an MCI risk score is described below:

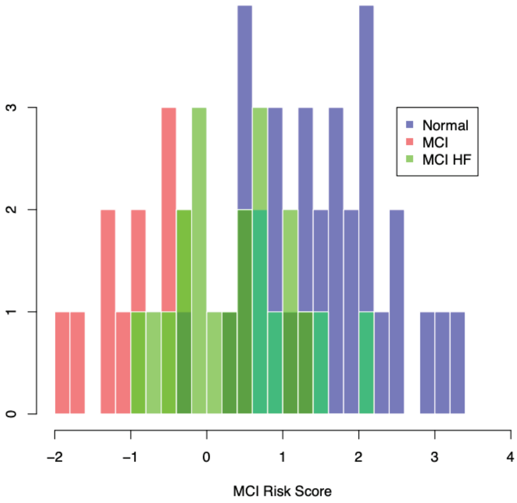
[1] Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen A Singular Value Thresholding Algorithm for Matrix Completion. SIAM J. Optim., 20(4), 1956–1982. (27 pages)
[2] Jerome Friedman, Trevor Hastie and Rob Tibshirani. (2008). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 2Vol. 33(1), 1-22 Feb 2010.

The risk score was developed using L1-, L2-regularized “elastic net” logistic regression[2] which is a modified version of logistic regression. In this application of L1-, L2- elastic net regression, the combination of normalized input variables were optimized based on the log-odds estimates of each individual’s performance pattern correlating with a predefined MCI performance pattern.

The MCI performance pattern was defined by the MCI subject group mean and standard deviation per variable. The effect of the L1- penalty was to exclude input variables from the risk score if they were not particularly useful for inferring the odds of individuals being categorized as MCI. The effect of the L2- penalty was seen in situations where there are several highly correlated variables that each predict the odds of being categorized as MCI. With an L2- penalty, rather than picking a single input from the set of correlated input scores, a weighted combination was used — possibly smoothing out noise or measurement error. When datasets were limited, rather than using a portion of the data to optimize the weights put on the two penalties, the relative weights put on the two penalties were set equal (alpha=0.5) and the overall strength of the penalties was set to 0.2 (lambda=0.2). Results describe the maximum likelihood that the data was subject to a penalty on both a) the sum of the absolute values of the weights put on each of the input variables (the L1- penalty) and b) on the sum of the squares of the weights (L2- penalty). For an individual, the weighted combination of input variables corresponding to the learned penalized logistic regression is treated as a new score — that individual’s risk score for being categorized as MCI. To evaluate these risk scores for the Normal and MCI subjects used to learn the model (the weights on the input variables to combine into the risk score), leave-one-out crossvalidation was used. Each of the Normal or MCI subjects in turn was left out of the analysis, the weights on the input variables for the risk score were reevaluated in the remaining subjects, and the resulting model was used to calculate the risk score for the left-out subject. MCI Risk Scores were calculated for all subjects.

Results. The high intra-class correlation (0.79) for the MCI Risk Score indicates that this is a reliable measure of individuals’ performances. The AUC of the MCI Risk Score to separate normal subjects from clinically impaired MCI subjects is (0.94). When High Functioning MCI and MCI subjects were combined into a single group, the AUC is (0.87). When all impaired study subjects were combined (including High Functioning MCI, Alzheimer’s Disease, Parkinson’s Disease, aphasia, and frontotemporal disorders) and compared to Normal Controls, the AUC for the MCI Risk Score to separate impaired subjects from Normal Controls was (0.92).

Table 2. Discrimination between MCI, High Functioning MCI and Normal Controls



For the 32 Normal Control subjects, the cross-validated MCI risk scores had mean values of (1.49), with a standard deviation of (0.87). For the 17 MCI subjects the risk scores had a mean value of (-0.44) with a standard deviation of (0.92). For the 21 high functioning MCI subjects, The risk score had a mean value of (0.41) with a standard deviation of (0.77). For the combined group of MCI subjects, the risk score had a mean value of (0.03) with a standard deviation of (0.92). For a pool of 67 impaired subjects, the risk score had a mean value of (0.52) with a standard deviation of (1.14).

Discussion. Preliminary results indicate that the combination of precision data capture combined with machine-learning approaches shows notable improvement to sensitivity and specificity of MCI vs. Normal Controls as compared to traditional clinical and research practices. While the leave-one-out approach curtails potential over-fitting of the model, the collection of larger data sets will allow further exploration of alternative approaches, for example the use of training and test sets. It is expected that increased data set size will generally correspond to improved performance.

THE DISCRIMINATION OF DISTINCT FUNCTIONAL DEFICITS

A preliminary analysis of the use of Miro scores to separate subjects with particular functional deficiencies was conducted using the 32 normal subjects, 5 subjects with frontal stroke or other frontal deficits, 12 subjects with Alzheimer's Disease and characterized as having memory deficits, 7 subjects with primary progressive aphasia or other language deficits, and 6 subjects with Parkinson's disease or other motor deficits. Non-normal subjects included in the analysis received gold-standard clinical diagnoses from specialty academic research centers.

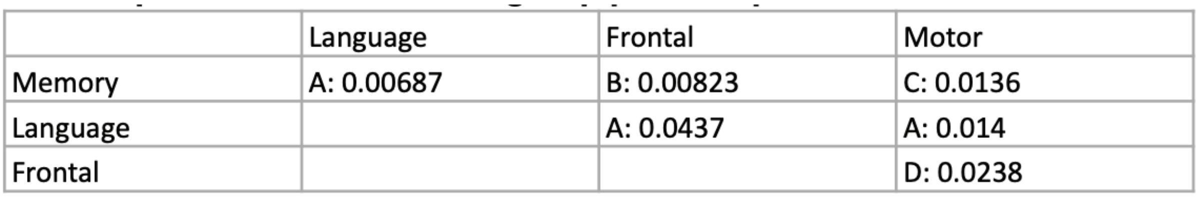
Methods. Each subject was assessed with Miro. For subjects who participated in multiple assessments, only initial (T1) assessments were used in this analysis. Variable scores were calculated with available subject data; missing data was not imputed given the limited sample size per diagnostic group. Raw scores were standardized to mean values of 0 and standard deviations of 1 in the Normal Control subgroup. Standardization for each unique stimulus prompt for Picture Description, Category Fluency, and Letter Fluency were incorporated. For each standardized score and each subject group, the statistical significance of the separation of the group mean score from the normal mean score was assessed using a Mann-Whitney test. Miro variables were screened for their ability to separate each paired group of functional deficits including: motor-memory, motor-language, motor-executive, memory-language, memoryexecutive, and language executive.

Subject numbers per functional domain are not yet large enough to effectively use machine learning paradigms to determine the best approach for group separation and characterization. Instead, for this exploratory analysis, 4 variable relationships were identified by human experts to characterize typical functional differences between groups of subjects with particular functional deficits:

A: Relative performance (standardized score) on Category Fluency
B: Combined relative performances on Trails B (time) and Symbol Coding
C: Relative performance on Verbal Learning
D: Difference between relative performances on Verbal Fluency and Design Fluency

Results. Four key variable relationships separate functional domain pairs with the following p-values for Mann-Whitney tests.

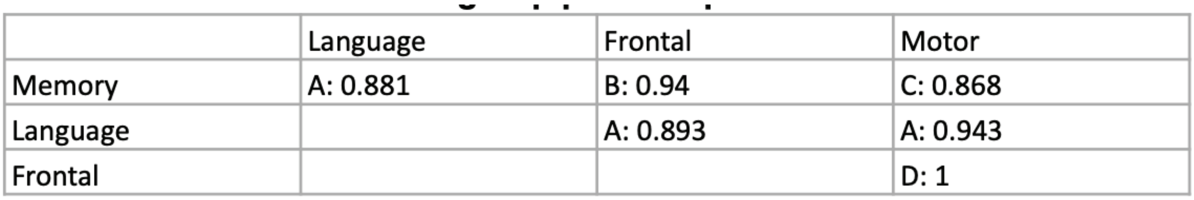
Table 3. p-values for functional group pair-comparisons



	Language	Frontal	Motor
Memory	A: 0.00687	B: 0.00823	C: 0.0136
Language		A: 0.0437	A: 0.014
Frontal			D: 0.0238

The corresponding AUCs for separating the groups are strong, often over 0.9.

Table 4. AUCs for functional group pair-comparisons



	Language	Frontal	Motor
Memory	A: 0.881	B: 0.94	C: 0.868
Language		A: 0.893	A: 0.943
Frontal			D: 1

Many individual Miro variables separate the groups with specific deficits (Memory, Frontal, Language, Motor) from Normals or from each other. Of 155 basic Miro variables, the following are the counts of variables that separate a pair of groups of subjects with p-values from Mann-Whitney test of below 0.0001.

Frontal vs Memory: 35
Frontal vs Language: 16
Frontal vs Motor: 11
Memory vs Language: 34
Memory vs Motor: 12
Language vs Motor: 12

Discussion. Preliminary analysis on a small data set shows strong separation of subject groups by deficits across four functional domains: Memory, frontal-executive, language, motor. It is hypothesized that improved data capture methods combined with the extraction of features from rich data types, like voice, movement and timing, provided the improved precision and functional diversity that is needed to accurately characterize and separate groups. As the data set grows, it is expected that machine-learning models will perform equally well or better than human-driven models. The precise characterization of a broad range of functions is expected to better support accurate diagnosis and monitoring of patient status over time.

Construct validity

Subjects and normative data

Thirty two normal volunteers participated in the normative study. Normal volunteers were recruited from a local retirement community; they were without past or present psychiatric or neurological disorders or head injuries and were free of medications that might affect the central nervous system. Normal subjects ranged in age from 49 to 89. One-hundred and one total subjects have been assessed for the current study as of January 2017. Patients were recruited from Johns Hopkins University School of Medicine and local neurology and neuropsychology practices.

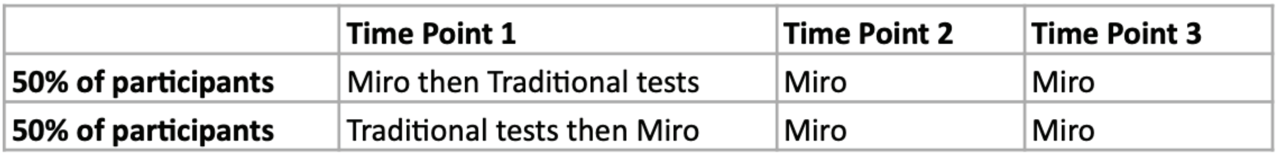
TEST-RETEST RELIABILITY AND LEARNING EFFECTS

Miro's reliability was investigated through a test-retest reliability study that assessed performance in normal controls at three time points over three months.

Subjects and methods. Miro's test-retest reliability was evaluated in 21 normal volunteers and 7 High Functioning MCI subjects who were assessed in their homes by a clinical neuropsychologist on 3 occasions. Assessments were separated on average by 22 days. The test-retest interval ranged from 2 to 54 days, with a median interval of 16 days.

Intra-class correlations were calculated for the MCI Risk Score with data from 3 time points (T1, T2, T3) in order to quantify test-retest reliability, or the ability to consistently identify performance levels for specific individuals. To assess learning effects, the changes in scores across test administrations for each subject (trends) were calculated. Permutation tests were used to identify mean trends and their significance.

The calculation of test-retest reliability and learning effects included the standardization of raw scores. Raw scores were standardized by the distribution of scores in the normal subject population upon initial Miro assessment (T1). T1 observations for each score were standardized with a mean value of 0 and standard deviation of 1. Subsequent observations (T2, T3) were standardized using the initial (T1) reference distribution. Mean standardized scores in T2 or T3 are measured in standard deviation units relative to T1 scores. Slopes have units of standard deviations per assessment (SD/A).



	Time Point 1	Time Point 2	Time Point 3
50% of participants	Miro then Traditional tests	Miro	Miro
50% of participants	Traditional tests then Miro	Miro	Miro

Results. The test-retest reliability intra-class correlation coefficient (ICC) for the MCI Risk Score was (0.79), with a 95% confidence interval of (0.65, 0.89). This shows the stability or reliability of measurements of individuals' functional abilities.

Discussion. Correlation between individuals' scores on each administration. The intra-class correlation (ICC) is the ratio of inter-individual variance to total variance of measurements across time points. A narrow range of functional ability as captured in this normal sample set demonstrate a low inter-individual variance relative to the variance in a mixed clinical population whose results include scores outside the normal range. Results show a significant ICC, even within this narrow range of normal performance. This not only suggests Miro's ability to consistently quantify individual performance, but also Miro's ability to distinguish performance signatures of individual normal subjects.

Learning effects. Analysis of repeated Miro assessments in a cognitively normal population shows minimal learning effects or other trends over three sequential test administrations. Across Miro variables, the mean subject performance at time point 3 shifted less than 0.09 standard deviations (SD) from initial performance.

CONCURRENT VALIDITY

Miro's construct validity was investigated through a concurrent validity study comparing Miro scores to standard neuropsychological tests in normal and impaired populations.

Subjects. Fifty-two subjects were tested on a battery of standard neuropsychological tests and analogous Miro modules. Analysis is based on 19 normal subjects and 33 subjects with brain impairment.

Methods. Assessment. Eighty-eight percent of subjects were assessed in their homes by clinical neuropsychologists, 12% were assessed at Johns Hopkins

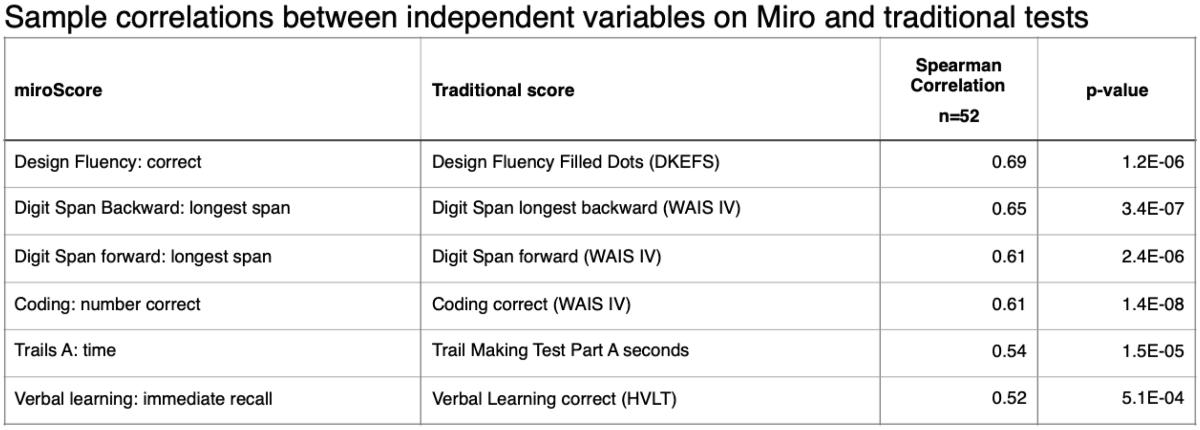
University Medical Center. Fifty percent of subjects were assessed with traditional tests prior to Miro assessment; 50% were assessed with Miro prior to traditional assessment. Traditional tests were hand-scored by the administering neuropsychologist and entered into a spreadsheet. Miro performance data was automatically uploaded from the iPad to Miro's HIPAA-compliant cloud-based server.

Missing data. Correlations for each variable were calculated with subjects whose results included both Miro and traditional scores. Subjects were excluded from the correlation of individual variables when missing either Miro or traditional scores (or both) for that variable.

Standardization of scores. Standardization of scores occurred via a two-step process: 1. Subtracting the mean from the normal subject reference set, 2. Rescaling centered scores by the standard deviation of reference set scores. The mean score of the normal reference set was set to 0 and the standard deviation was set to 1.

Results. Miro module scores demonstrate significant correlation with traditional scores (Table 3). Estimated Spearman correlations for most Miro and traditional scores are greater than 0.5 and are significantly different than zero with p-values of 0.05 or lower. Statistical results provide preliminary evidence that Miro scores quantify brain function comparably to traditional, in-depth, clinician-administered neuropsychological assessment methods.

Table 3. Concurrent Validity



Sample correlations between independent variables on Miro and traditional tests

miroScore	Traditional score	Spearman Correlation n=52	p-value
Design Fluency: correct	Design Fluency Filled Dots (DKEFS)	0.69	1.2E-06
Digit Span Backward: longest span	Digit Span longest backward (WAIS IV)	0.65	3.4E-07
Digit Span forward: longest span	Digit Span forward (WAIS IV)	0.61	2.4E-06
Coding: number correct	Coding correct (WAIS IV)	0.61	1.4E-08
Trails A: time	Trail Making Test Part A seconds	0.54	1.5E-05
Verbal learning: immediate recall	Verbal Learning correct (HVLT)	0.52	5.1E-04

Discussion. As expected, concurrent validity between independent variables from Miro's selfadministered, clinician-supervised tablet assessment and traditional clinician-administered pencil-paper based testing was moderate, ranging from (0.42) to (0.69). These results are similar to test-retest reliability for standard, in-depth neuropsychological test scores[3]. Relatively modest correlation is expected given traditional test challenges with inter- and intrarater reliability, blunt scores, and low levels of sensitivity and specificity for differentiating disorders. Statistically significant Spearman correlations between Miro scores and their traditional analogues suggest that Miro and traditional exams quantify equivalent functional abilities. Low p-values indicate a high degree of confidence in the correlations. This is notable given the study's limited sample size, the large proportion of normals who demonstrate a narrow range of functional ability, and the large proportion of mildly impaired subjects with near-normal abilities.

It is important to note that the sample set of Normal Controls exhibits a narrow range of performance on Miro modules. The narrow range of scores from the normal controls dampens potential correlation with traditional scores, whereas the broader performance range of impaired subjects strengthens potential correlation. Also noteworthy is the fact that more than half of the impaired subjects volunteered to participate as Normal Controls but failed the screening test by a slim margin (1-2 points below the normal-group inclusion threshold of 26 on the MoCA[4]). Many of these mildly impaired subjects participated in the study not as Normal Controls, but as MCI subjects.

Conclusion. While traditional assessment methods have been useful in confirming moderate to severe impairment, they have struggled to characterize mild, clinically meaningful functional differences. Preliminary findings show promise for machine-driven methods like rich data capture, digital signal processing, and machine-learning to precisely characterize brain function. Early results on a small sample size indicate that machine-driven approaches support the separation of overlapping groups of mildly impaired subjects from normal controls and from each other. Test-retest reliability results demonstrate the potential to track the signature performance of each individual over time. It is expected that with larger data sets collected over time, these capabilities could be used to predict disease course, monitor therapeutic effects, support differential diagnosis, describe disease sub-types, and find phenotypic markers.

[3] Grant L. Iverson. Interpreting change on the WAIS-III/WMS-III in clinical samples. Archives of Clinical Neuropsychology. 2001; Snow WG. WAIS-R test-retest reliability 3 in a normal elderly sample. Journal of Clinical Experimental Neuropsychology. 1989

[4] Damian AM, The Montreal Cognitive Assessment and the mini-mental state examination as screening instruments for cognitive impairment: item analyses and threshold 4 scores. Dementia and Geriatric Cognitive Disorders. March 2011