The Clinical Neuropsychologist
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ntcn20

Psychometric Equivalence of a Paper-Based and Computerized (iPad) Version of the Memory Orientation Screening Test (MOST®)

Mitchell Clionsky[a] & Emily Clionsky[a]
[a] Clionsky Neuro Systems, Inc., Springfield, MA, USA
Published online: 12 May 2014.

To cite this article: Mitchell Clionsky & Emily Clionsky (2014): Psychometric Equivalence of a Paper-Based and Computerized (iPad) Version of the Memory Orientation Screening Test (MOST®), The Clinical Neuropsychologist, DOI: 10.1080/13854046.2014.913686

To link to this article: http://dx.doi.org/10.1080/13854046.2014.913686

The Clinical Neuropsychologist, 2014
http://dx.doi.org/10.1080/13854046.2014.913686

Routledge
Taylor & Francis Group

Psychometric Equivalence of a Paper-Based and Computerized (iPad) Version of the Memory Orientation Screening Test (MOST®)

Mitchell Clionsky, and Emily Clionsky
Clionsky Neuro Systems, Inc., Springfield, MA, USA

The Memory Orientation Screening Test (MOST®) is a 29-point scale for identifying and following mild and major neurocognitive disorders in older patients. Previous research demonstrated validity in separating patients with normal vs. impaired cognition and high correlations with tests of memory and attention. This study compares the original paper-based MOST® with a computerized (iPad app) version, the MOST®-96120, to determine the equivalence of formats. A total of 98 consecutive older patients were administered identical versions of the MOST and MOST-96120 in a random order, separated by 1 hour of interspersed testing, in a 3-hour neuropsychological evaluation. MOST and MOST-96120 scores were compared with each other, with global cognitive ratings, and with standardized tests of memory and attention. Both versions had equivalent means and standard deviations, very high inter-test correlation (r = .92, p < .001), and equal correlations with outcome measures. Both versions separated patients into normal vs. mild NCD vs. major NCD categories with equal accuracy. ANOVA showed no significant difference between versions or presentation order. Both versions correlated very highly with cognitive level and neuropsychological endpoints, confirming previous research. The MOST-96120 is a computerized neuropsychological assessment device that demonstrates equivalence with its paper-based original, allowing for confident reliance on the findings of previous research.

INTRODUCTION

The Memory Orientation Screening Test (MOST®) is a 5-minute test of cognition in older patients which evaluates memory, attention, and executive function (Clionsky & Clionsky, 2010). The MOST was developed to improve the identification and ongoing assessment of Mild Cognitive Impairment (MCI) and dementia, now respectively reclassified as Mild Neurocognitive Disorder (Mild NCD) and Major Neurocognitive Disorder (Major NCD) in the *Diagnostic and Statistical Manual of Mental Disorders 5th edition* (American Psychiatric Association, 2013).

The MOST—an acronym derived from four interwoven tasks which comprise the test: Memory (three-word recall), Orientation (to time, day, date, month, season, year), Sequence (recall of 12 pictured household items initially named by the patient), and Time (clock drawing and hand setting on a predrawn outline)—offers a scoring range from 0 to 29, with higher scores representing better cognition.

Previous studies have found consistently high correlations between the MOST and standardized neuropsychological tests of memory, attention, and executive function

Address correspondence to: Mitchell Clionsky, Ph.D. ABPP (CN), Clionsky Neuro Systems, Inc., 155 Maple Street, Suite 203, Springfield, MA 01105, USA. E-mail: mitch@cns-neuro.com
(*Received 9 December 2013; accepted 7 April 2014*)

(Clionsky & Clionsky, 2010, 2011, 2013) as well as between the MOST and stages of cognitive decline. In a sample of more than 700 patients (Clionsky & Clionsky, 2010), the MOST had significantly higher receiver operating characteristics and relationships with neuropsychological endpoints than either the Mini Mental State Exam (Folstein, Folstein, & McHugh, 1975) or the Mini-Cog™ (Borson, Scanlon, Brush, Vitaliano, & Dokmak, 2000). MOST scores demonstrated high inter-rater reliability ($r \geq .90$), internal consistency (Cronbach alpha = .79) and test–retest reliability ($r \geq .90$) across 1-month and 6- to 12-month intervals. A change of 3 or more MOST points accurately (CI = .95) reflects a change in the underlying cognitive level. More than 90% of geriatric psychiatry patients successfully completed the test. Administration and scoring time averaged 4.5 ± 1 minutes in a demented sample. The MOST outperformed patient self-ratings and informant ratings of cognitive levels (Clionsky & Clionsky, 2011) while demonstrating high relationships with neuropsychological testing and dementia ratings. In a stratified patient sample which was designed to reflect the prevalence of varying cognitive levels seen in a typical primary care medical population (Clionsky & Clionsky, 2013), the MOST correctly classified patients as having normal cognition, Mild NCD, or Major NCD with 80% accuracy. In this study, MOST scores separated cognitively normal vs. impaired elders with a sensitivity of 88% and a specificity of 83%.

An iPad version of the MOST was designed to meet the criteria for a computerized neuropsychological assessment device (CNAD), and released as the MOST-96120 in June 2012. Its End User License Agreement (EULA) conforms to APA guidelines regarding limitations of the test and qualifications of end users (American Psychological Association, 2010) and is presented when downloading the app to alert users to the requirements for using the test.

The MOST-96120 provides an easy administrative interface, semi-automated scoring of the clock drawing, and fully-automated calculation of the total score. Interactive administration with a test giver is required in its current form. Using a programmed, research-based algorithm, involving age and education level, the app places the score within the most likely level of cognitive functioning and lists the percentage of patients with that score who have been judged as normal vs. Mild NCD vs. Major NCD. The app then produces a detailed report, modified by the health care professional, who chooses from a menu of score-specific treatment recommendations and can add additional recommendations specific to the setting or the patient. Comparison to previous MMSE or MOST scores is also available.

The final report is signed by the clinician within the app and can be saved, printed, or emailed as a pdf document. A companion family report, written in lay language, is also generated.

While the iPad version has potential advantages in administration, scoring, plan development, and documentation, its equivalence with the paper-based test remained unknown. The 2012 joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology on computerized neuropsychological assessment devices (CNADs) (Bauer et al., 2012, p. 179) advises that "when a traditional examiner-administered test is programmed for computer administration, *it becomes a new and different test*" (emphasis in the original). This position prompted us to investigate the psychometric equivalence of the iPad and the paper-based MOST to each other, as well as to the patient's general cognitive level and

other neuropsychological tests. This study's counterbalanced randomized design also permitted a test of very brief (1-hour) test–retest reliability on a clinical sample to assess the influence of practice effect on the outcome of this test.

METHOD

Participants

A total of 104 consecutive patients in an outpatient neuropsychology practice in Springfield, MA served as participants in this study. The purpose and methodology of the study were approved by the Institutional Review Board of the Sisters of Providence Health Care System (SPHS2012-02) and were judged not to require specific informed consent, as the MOST was already in clinical use as part of the routine neuropsychological test battery. Of these 104 patients, 6 were excluded because of visual impairments or failure to complete the second administration of the test. Of the remaining 98 patients, 53 were female and 45 were male, with an age range of 51 to 92 and a mean age of 76.8 (± 7.4) years. Mean education was 14.0 (± 3.3) years. Diagnostic distribution, based on all neuropsychological testing, family history, and medical records included: 8 with normal cognition, 17 with Mild NCD, 27 with Major NCD due to Alzheimer's Disease, 31 with Major NCD due to Vascular Disorder, 14 with Major NCD due to Alzheimer's Disease and Vascular Disorder, and 1 with Major NCD due to Parkinson's Disease. Dementia severity was typically mild to moderate (mean CDR 1.3 ± 0.8), with normal cognition classified as 0, Mild NCD as 0.5, and Major NCD severity ranging from 1 to 3 (Morris, 1993).

Procedure

Each patient was administered a 3-hour neuropsychological evaluation. Tests of interest for this study included Full Scale IQ from the Wechsler Adult Intelligence Scale-Fourth Edition (Wechsler, 2008), Logical Memory-II and Visual Reproduction-II subtests of the Wechsler Memory Scale-Fourth Edition (Wechsler, 2009), delayed list recall of the 12-item Shopping List Test (Clionsky, 1995), Trailmaking A and B (Reitan, 1955), The Verbal Absurdities test—adapted from the Verbal Absurdities Subtest of the Stanford-Binet Intelligence Test, Form L-M—(Sherman & Merrill, 1973), an 8-item Proverb Interpretation Test—modeled after the Proverb test from Delis-Kaplan Executive Function System—(Delis, Kaplan, & Kramer, 2001), the Folstein Mini-Mental State Exam (Folstein et al., 1975), and the 15-item Geriatric Depression Scale (Sheikh & Yesavage, 1986). These tests were selected as endpoints because of their frequent clinical use as measures of memory, attention, executive function, intelligence, and depression in this population and our reliance on them in forming clinical diagnoses.

Each patient was administered identical versions of the paper-based MOST and the iPad MOST-96120, 1 hour apart, separated by other neuropsychological tests in the battery. Each patient was tested by two different clinicians, first by the psychologist and second by a trained psychometrician who was blind to the results of the first test. The

presentation order of paper or iPad test was randomly assigned and counterbalanced, resulting in 50 patients in the paper first condition and 48 taking the iPad version first.

Statistical analyses

A two-way repeated-measures analysis of variance (ANOVA) with order of test presentation as a between-participants variable and test version (paper or iPad) as a within-participants variable was conducted to determine if there was a main effect of version, order of presentation, or interaction between version and presentation order. Pearson correlations for each variable and their levels of significance vs. 0 were computed. Tests for significant differences between correlations used Fisher's *z*-transformation and tested the normalized difference vs. 0. Additional analyses compared means and standard deviations for each presentation order and concordance of each version's diagnostic outcome (normal vs. MCI vs. dementia) with that determined by the neuropsychological evaluation.

RESULTS

The MOST (paper) had a mean score of 16.9 (± 5.5) and the MOST-96120 app had a mean score of 17.1 (± 5.4). This was not a statistically significant difference (p = .88). Both mean scores fell were within the range typically found in patients with Major NCD of mild severity (Clionsky & Clionsky, 2010). When separated by order of administration, first administrations of the test yielded mean scores of 17.2 (± 5.5) and second administrations yielded mean scores of 16.8 (± 5.3) This was not a statistically significant difference (p = .94). Table 1 shows the means of each version by presentation order. There was also not a significant difference for administration order based on the level of patient cognition (t = .85, p = .40). Higher-functioning patients (normal plus Mild NCD, n = 25) had initial means of 22.36 and subsequent means of 22.6, while lower-functioning patients (Major NCD, n = 73) had means of 14.99 and 15.28.

ANOVA revealed no significant differences between the paper-based MOST and the iPad MOST-96120, $F(1, 96) = 1.939$, p = .167, the order in which they were presented, $F(1, 96) = 2.419$, p = .123 or the interaction of version and order, $F(1, 96) = 3.368$, p = .07. Because the interaction effect was trending toward significance, follow-up comparisons were made using Scheffe's test for unplanned comparisons. These did not reveal any systematic order by version relationships, although there was a trend toward significance in one interaction, which appears artifactual.

Table 1. Means and standard deviations of each test version by presentation order

Order	*N*	Mean (*SD*) Paper	Mean (*SD*) iPad
Paper-iPad	50	16.24 (4.90)	16.14 (4.66)
iPad-Paper	48	17.50 (6.11)	18.23 (6.01)

Table 2. Correlations of Memory Orientation Screening Test (MOST) and MOST-96120 with dementia severity and cognitive tests

Test	MOST (paper)	*p*	MOST-96120 (iPad)	*p*	*Z*	*p*
MOST-96120	0.915	<.001				
Severity	–0.842	<.001	–0.814	<.001	0.61	*ns*
LM-II	0.753	<.001	0.741	<.001	0.19	*ns*
VR-II	0.619	<.001	0.622	<.001	0.03	*ns*
SLT-R	0.809	<.001	0.807	<.001	0.02	*ns*
MMSE	0.819	<.001	0.763	<.001	1.05	*ns*
WAIS-IV, FSIQ	0.509	<.001	0.481	<.001	0.26	*ns*
Proverbs	0.515	<.001	0.499	<.001	0.15	*ns*
Verbal Absurdities	0.396	<.001	0.423	<.001	0.22	*ns*
Trailmaking A	–0.463	<.001	–0.408	<.001	0.47	*ns*
Trailmaking B	–0.489	<.001	–0.486	<.001	0.03	*ns*
GDS-15 item	0.111	.27 ns	0.098	0.33 ns	0.08	*ns*

MOST-96120: iPad Memory Orientation Screening Test; LM-II: Wechsler Memory Scale-Fourth Edition, Delayed Logical Memory; VR-II: Wechsler Memory Scale-Fourth Edition, Delayed Visual Reproduction; SLT-R: Shopping List Test, Delayed Recall; MMSE: Mini-Mental State Exam; WAIS-IV FSIQ: Wechsler Adult Intelligence Scale-Fourth Edition, Full Scale IQ; GDS-15: Geriatric Depression Scale, 15-item version.

Table 2 shows the correlation matrix of the MOST and MOST-96120 with all outcome variables. The two MOST versions correlated very highly with each other ($r = 0.91$, $p < .001$). Each version correlated highly ($r = . –84$ vs. $r = . –81$) and significantly ($p < .001$) with general cognition level, the primary clinical endpoint, and with the MMSE ($r = .82$ vs. $r = .76$). There was no significant difference between the correlations of each version with those markers. Similarly, each version correlated strongly and significantly ($p < .001$) with delayed story memory (LM-II), visual design memory (VR-II), and list memory (SLT-R). Each test version had significant, moderate strength correlations with measures of attention (Trailmaking A), executive function (Trailmaking B, Proverb Interpretation, Verbal Absurdities) and intelligence (WAIS FSIQ). Neither version correlated significantly with depression. There were no significant differences between the correlations of paper or iPad test versions with any of these variables.

To compare the effect of test version on clinical decision making (normal vs. mild NCD vs. Major NCD) we classified each score according to cutoffs established in the three previous published studies using the MOST: a MOST score of 0–18 is regarded as reflecting Major NCD, a MOST score of 19–22 as indicating Mild NCD, and a MOST score of 23–29 as normal cognition. Using these classifications, there was 83% agreement (82 of 98 cases) in classification. Of the remaining 16 cases, the iPad yielded a higher score in 7 and a lower score in 9. In 15 of the 16, the classification difference represented a discrepancy of one category, either between normal vs. Mild NCD or Mild NCD vs. Major NCD. In only one case did this difference represent a classification difference of normal vs. Major NCD.

Diagnostic classification agreement was high between both versions (Table 3) and the clinical criterion diagnosis determined by full neuropsychological testing, integrated with medical record review and family report. The paper version yielded a 73.5% overall agreement (Normal vs. Mild NCD vs. Major NCD) and the iPad yielded 77.6%

Table 3. Agreement rates of Memory Orientation Screening Test (MOST) and MOST-96120 with diagnostic criterion, using 3-category and 2-category placements

3-category placement	Paper-severity agree	Paper-severity disagree	iPad-severity agree	iPad-severity disagree
Normal	6	2	7	1
Mild NCD	9	8	12	5
Major NCD	57	16	57	12
Sum col.	72	26	76	22
% of total	73.5%	26.5%	77.6%	22.4%
2-category placement				
Normal + Mild NCD	23	2	25	0
Major NCD	57	16	47	16
Sum col.	80	18	82	16
% of total	81.6%	18.4%	83.7%	16.3%

agreement. Collapsing Mild NCD and Normal categories, to compare accuracy of separating non-demented from demented individuals, yielded nearly identical accuracies (paper = 81.6%, iPad = 83.7%).

DISCUSSION

Accurately identifying cognitive impairment in the elderly is of great importance in primary care and in many specialty medical settings because dementia is a growing problem and cognitive decline represents a major comorbidity for an already vulnerable population. Because of time limitations and demands for attention to other medical problems and preventive care protocols, such evaluations need to rely increasingly on brief, well-validated tests which can be administered economically and easily. The true utility of such a test is realized when its results are available immediately, are provided in a clear and understandable manner, and when providers can use the results to formulate and initiate a treatment plan immediately. The ability to document the chart and make patients and their families aware of the findings and plan further enhances the value of an automated approach.

The CNAD guidelines have wisely cautioned against making paper to computer translations of tests without undergoing a rigorous comparison of the test formats. We believe that the current study fulfills this equivalency requirement. This study demonstrates that the paper MOST and the iPad MOST-9120 are, in fact, equivalent measures and do not differ significantly from each other in any way that can be demonstrated statistically. The two tests correlate at an extremely high level, have nearly identical means and standard deviations, separate patients into diagnostic groups with similar accuracies, and show no significant differences in their correlations with a variety of well-validated cognitive tests measuring essential mental functions.

This study has also substantiated previous research findings, showing a high correlation of the MOST-96120 with the patient's independently determined level of cognitive functioning (r = –.81) with three primary tests of delayed memory

(stories: r = .74, designs: r = .62, and lists: r = .81) as well as neuropsychological measures of executive function, intelligence, and attention. Also, confirming the findings of the initial MOST validation study (Clionsky & Clionsky, 2010), neither version correlated significantly with depression level, making it useful in clinical situations questioning dementia vs. depression.

This study also did not find a significant practice effect, even when the MOST was repeated with only an hour between administrations, even for higher-functioning patients who might show a greater improvement due to test experience. This was an unexpected finding, but reduces concerns that giving the test too frequently would yield invalid scores because of practice or foreknowledge effects.

An extensive discussion of the role of brief evaluation or screening of cognition in medical practices is beyond the scope of this paper, but a few comments are worthwhile. A recent US Preventive Services Task Force Review (Lin, O'Connor, Rossom, Perdue, & Eckstrom, 2013) saw limited benefit for dementia screening, but their argument was based largely on the belief that screening is not useful until dementia interventions become more effective. Roundtree, Atri, Lopez, and Doody (2013) disagree with Lin et al.'s approach, concluding that currently available medications slow cognitive and functional decline and delay nursing home treatment. Clionsky (2012) agrees with Roundtree et al., and argues that brief cognitive evaluation, in the primary care medical setting can objectively and efficiently separate patients with normal cognition from those with probable Mild NCD or Major NCD, encourages earlier identification and treatment of patients with cognitive loss, and can be used to gauge the effectiveness of medical or environmental interventions over time. Earlier identification and periodic follow-up from initial findings of even Mild NCD follows logically from the findings of Roberts et al. (2014) that Mild NCD progresses frequently to Major NCD even in patients who initially improve to normal after the Mild NCD diagnosis.

While comprehensive neuropsychological assessment remains the gold standard for diagnosis, brief evaluation in the medical office offers a reasonable alternative, particularly when using a measure that has demonstrated high agreement with more extensive testing. Brief testing provides specific value to the medical provider who may be working in settings with limited referral options, time, or financial resources.

A weakness of the present study is its reliance on a memory center sample and its single geographic location. It, and other studies employing the MOST, would benefit from replication by other investigators and in other settings. Nevertheless, the methodological approach used in this study serves as a model for other test producers who are adapting paper-validated neuropsychological tests to computer and other digital device formats.

AUTHOR NOTES

Both authors contributed equally to this article. Dr. Mitchell Clionsky had full access to the data and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Both authors hold a copyright and trademark on the Memory Orientation Screening Test (MOST®) and its iPad apps, the MOST®, MOST-96120®, and mdMOST®.

No external funding was obtained for this study. This paper has not been published and is not under consideration for publication elsewhere.

REFERENCES

American Psychological Association. (2010). *Ethical principles of psychologists and code of conduct (2010 amendments)*. Retrieved from: http://www.apa.org/ethics/code/index.aspx

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Bauer, R., Iverson, G., Cernich, A., Binder, L., Ruff, R., & Naugle, R. (2012). Computerized neuropsychological assessment devices: Joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. *The Clinical Neuropsychologist, 26*, 177–196. doi:10.1080/13854046.2012.663001

Borson, S., Scanlon, J., Brush, M., Vitaliano, P., & Dokmak, A. (2000). The Mini-Cog: A cognitive vital signs measure for dementia screening. *International Journal of Geriatric Psychiatry, 15*, 1021–1027.

Clionsky, M. (1995). *The Shopping List Test: A list learning and recall test for geriatric patients*. Unpublished manuscript, Springfield, MA.

Clionsky, M. (2012). iPad screening for dementia holds great promise. *Journal of Alzheimer's Disease and Parkinsonism*, 2. http://dx.doi.org/10.4172/2161-0460.1000e112

Clionsky, M., & Clionsky, E. (2010). Development and validation of the Memory Orientation Screening Test (MOST™): A better screening test for dementia. *American Journal of Alzheimers Disease and Other Dementias, 25*, 650–656. doi:10.1177/1533317510386216

Clionsky, M., & Clionsky, E. (2011). Identifying cognitive impairment in the annual wellness visit: Who can you trust? *Journal of Family Practice, 60*, 653–659.

Clionsky, M., & Clionsky, E. (2013). The Memory Orientation Screening Test (MOST®) accurately separates normal from MCI and demented elders in a prevalence-stratified sample, *Journal of Alzheimer's Disease and Parkinsonism*, *3*. http://dx.doi.org/10.4172/2161-0460.1000109

Delis, D., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan Executive Function System*™. San Antonio, TX: Psychological Corporation.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). 'Mini-Mental State': A practical method for grading cognitive states of patients for the clinician. *Journal of Psychiatric Research, 12*, 189–198.

Lin, J. S., O'Connor, E., Rossom, R. C., Perdue, L. A., & Eckstrom, E. (2013). Screening for Cognitive Impairment in Older Adults: A Systematic Review for the U.S. Preventive Services Task Force. *Annals of Internal Medicine*, 159, 601–612. http://annals.org/on 10/30/2013

Morris, J. (1993). The Clinical Dementia Rating Scale (CDR): Current version and scoring rules. *Neurology, 43*, 2412–2414.

Reitan, R. M. (1955). The relation of the trail making test to organic brain damage. *Journal of Consulting Psychology, 19*, 393–394.

Roberts, R. O., Knopman, D. S., Mielke, M. M., Cha, R. H., Pankratz, V. S., Christianson, T. J. H., … Peterson, R. C. (2014). Higher risk of progression to dementia in mild cognitive impairment cases who revert to normal. *Neurology, 82*, 317–325.

Roundtree, S., Atri, A., Lopez, O., & Doody, R. (2013). Effectiveness of antidementia drugs in delaying Alzheimer's disease progression. *Alzheimer's & Dementia, 9*, 338–345.

Sheikh, J. I., & Yesavage, J. A. (1986). Geriatric Depression Scale (GDS): Recent evidence and development of a shorter version. *Clinical Gerontology, 5*, 165–173. doi:10.1300/J018v05n01_09

Sherman, L. M., & Merrill, M. A. (1973). *Stanford-Binet Intelligence Scale. Manual for the Third Revision, Form L-M*. Boston, MA: Houghton Mifflin.

Wechsler, D. (2008). *The Wechsler Adult Intelligence Scale-IV Manual*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2009). *The Wechsler Memory Scale-Fourth Edition (WMS-IV)*. San Antonio, TX: Psychological Corporation.