Published in final edited form as:
Arch Neurol. 2011 March ; 68(3): 343–350. doi:10.1001/archneurol.2010.375.

The SIST-M: Development, reliability and cross-sectional validation of a brief structured Clinical Dementia Rating interview

Olivia I. Okereke[1], **Maura Copeland**[1], **Bradley T. Hyman**[2], **Taylor Wanggaard**[1], **Marilyn S. Albert**[3], and **Deborah Blacker**[1,2,4]

[1] Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

[2] Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

[3] Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD

[4] Department of Epidemiology, Harvard School of Public Health

Abstract

Background—The Clinical Dementia Rating (CDR) and CDR-Sum-of-Boxes (CDR-SB) can be utilized to grade mild but clinically important cognitive symptoms. However, sensitive clinical interview formats are lengthy.

Objective—To develop a brief instrument for obtaining CDR scores, and to assess its reliability and cross-sectional validity.

Methods—Using legacy data from expanded interviews conducted among 347 community-dwelling, older adults in a longitudinal study, we identified 60 questions about cognitive functioning in daily life–out of a possible 131– using clinical judgment, inter-item correlations, and principal components analysis. Items were selected in one cohort (n=147), and a computer algorithm for generating CDR scores was developed in this same cohort and re-run in a replication cohort (n=200) to evaluate how well the 60 items retained information from the original 131. Then, short interviews based on the 60 items were administered to 50 consecutively-recruited elders, with no or mild cognitive symptoms, at an Alzheimer Disease Research Center. CDR scores based on short interviews were compared with those from independent long interviews.

Results—In the replication cohort, agreement between short and long CDR interviews ranged from κ =0.65–0.79, with κ =0.76 for Memory; κ =0.77 for global CDR; ICC (intra-class correlation coefficient) for CDR-SB=0.89. In the cross-sectional validation, short interview scores

Address correspondence to: Dr. Olivia Okereke, MGH Gerontology Research Unit, 149 13th Street, Suite 2691, Charlestown, MA 02129. Tel: (617) 726-5571. Fax: (617) 726-5760. ookereke@partners.org.

The authors have no conflicts of interest pertaining to this manuscript.

were slightly lower than those from long interviews, but good agreement was observed: $\kappa \geq 0.70$ for global CDR and Memory; ICC for CDR-SB=0.73.

Conclusions—The SIST-M is a brief, reliable and sensitive instrument for obtaining CDR scores in persons with symptoms along the spectrum of mild cognitive change.

Keywords

Alzheimer disease; mild cognitive impairment; Clinical Dementia Rating; instrument; questionnaire; clinical interview

BACKGROUND

As potential disease-modifying therapies for Alzheimer disease (AD) enter clinical trials, identifying illness at a prodromal phase takes on growing importance: early cognitive decline may be most amenable to interventions that could slow progression of neuropathology and symptoms[1–3]. Standardized tools, such as the Clinical Dementia Rating (CDR) [4, 5], are effective at distinguishing normal aging from mild cognitive impairment (MCI) and dementia. The CDR features a global rating of impairment, as well as a CDR "Sum-of-Boxes" (CDR-SB) that totals the ratings from each of six cognitive and functional domains (Memory, Orientation, Judgment and Problem-solving [JPS], Community Affairs [CA], Home and Hobbies [HH] and Personal Care [PC]); the CDR-SB can be used to quantify impairment within the range of mild symptoms. The CDR is a mandatory element of the National Institute on Aging-funded Alzheimer's Disease Centers (ADCs) Uniform Data Set (UDS)[6] and the AD Neuroimaging Initiative, and is increasingly used in multi-center trials[7]. There is an in-depth, formal interview protocol[8]; however, many clinicians rate the CDR based on their usual clinical interview. An expanded structured interview[9] is available, and with trained, clinically-skilled interviewers can achieve high reliability and discriminative ability among persons with very mild cognitive change[10] – a population of increasing interest in prevention and early intervention trials. However, this expanded interview takes ~90 minutes to complete – limiting its efficiency in larger-scale research settings.

Thus, there is a need for shorter measures to quantify clinically important change along the spectrum from normal aging to MCI[10, 11]. In this study, our objectives were: 1) to develop an instrument to administer a shortened CDR interview (~25 minutes), 2) to verify its reliability and 3) to conduct a cross-sectional validation by testing concordance of CDR scores from the shorter interview with those obtained using the expanded interview.

METHODS

Participants

The Massachusetts General Hospital Memory and Aging Study (MAS)— Participants were part of a longitudinal study aimed at discriminating prodromal AD from less severe memory impairments[9, 10, 12, 13]. Older adults with and without memory complaints were recruited in three cohorts from the community through advertisements: Cohort 1 (n=165) from 1992–93, Cohort 2 (n=120) from 1997–98, and Cohort 3 (n=95) from 2002–06. To be included in the study, participants needed to be: aged ≥65 years (with the exception of 7 individuals aged 57–64); without dementia; free of significant medical, neurologic, or psychiatric illness; rated as a global CDR ≤ 0.5[5]; and willing to participate in study procedures. Each participant was recruited with a knowledgeable informant – usually an immediate family member (spouse, adult child, or sibling) or close friend.

The Massachusetts ADRC (MADRC) Longitudinal Cohort—Participants were part of the MADRC longitudinal research cohort, developed in recent years in response to changes in the ADC program requiring the collection of a UDS on a cohort with normal cognition, MCI, and AD/other dementias. MADRC participants are recruited through both community and clinic populations, and are seen annually. In 2007, we began recruiting MAS participants into the MADRC, and 177 such participants (the great majority of MAS members are still living and able to attend visits) have joined the MADRC. The combined cohort now totals 756 members: 58.2% female; 84.0% Caucasian, 11.2% African American, and 4.8% other race; mean age=74.9 years (SD=9.6; range 46–97).

Methods for Participant Evaluation—MAS cohort members were administered the expanded CDR interview[9]; they also had medical evaluations (i.e., history and physical examination, EKG, and standard laboratory tests), structural and functional neuroimaging tests (MRI and SPECT), comprehensive neuropsychological testing, and blood collection for biomarker and genetic analyses. MAS participants were followed annually with the CDR and brief neuropsychological testing; for those who developed significant decline, a consensus conference was held to determine dementia using standard diagnostic criteria[14]. MADRC cohort participants were evaluated each year according to UDS protocol[6], which includes CDR ratings, a medical history, vital signs, neurological examination, and a standard battery of cognitive tests[15]. The present study was approved by the Institutional Review Board and Human Research Committee of the Massachusetts General Hospital, Boston, MA.

Construction of the Shortened CDR Interview

We developed the shortened CDR interview using legacy data from baseline visits from MAS Cohort 1. This development cohort consisted of 147 participants (18 participants had missing data on the expanded interview items). Expanded interview items had an unequal number of responses – and several items had missing values for many participants; consequently, an automated item selection procedure (e.g., R^2 method in step-wise linear regression) could not be applied to this data, as such procedures exclude any subject with a missing value for even a single item within a domain. Thus, a multi-step, semi-quantitative procedure was used to identify the smallest set of items that could provide information adequate to score the CDR while maintaining sensitivity to the spectrum of mild cognitive change.

Item Selection—The expanded interview consists of 131 items covering the 6 CDR domains. Each item was graded by the original interviewer using CDR categories of 0, 0.5, 1, or 2. In the first step of item selection, item correlations were assessed by domain; exclusions were made if an item: 1) had no variance; 2) had insufficient data to determine correlations with other items; 3) had weak correlations (≤ 0.2) with all or most of the other items, as well as the domain rating; or 4) was redundant, as it tended to be scored identically with a few items in the same cognitive or functional topic area, but was weakly- or un-correlated with the CDR domain rating itself and with many other items – including “core” items of the domain (e.g., the core item “overall more forgetful of recent events” in the Memory domain).

In a second step, some initially excluded items were “forced” back in, as they were considered highly clinically-relevant by experienced clinicians (e.g., a JPS item on whether driving difficulty due to poor cognition had resulted in car accidents) or were helpful in completing other UDS forms (e.g., the Functional Assessment Questionnaire [FAQ])[16], and thus added efficiency (the SIST-M covers all 10 FAQ topic areas). A final set of 60 items

included the following: 14 in Memory, 8 in Orientation, 14 in JPS, 6 in CA, 15 in HH, and 3 in PC.

Creation of a Scoring Algorithm—A SASc (SAS Institute, Cary, NC, USA) computer algorithm was written by one of the study physicians (OO) in order to simulate, in effect, how participants would have been scored if the development cohort interviews had been conducted using only the 60 items. The algorithm was a complex, hierarchical design that used a combination of the grade of each item (e.g., 0, 0.5, 1), the frequency with which different grades of items were observed within a CDR domain, and the relative clinical importance – or "weight" – of each item. We further refined this hierarchical algorithm by addressing whether CDR domains were unitary constructs or composed of key sub-domains using principal components analysis (PCA). As our raw variables were ordinally ranked, we first calculated polychoric correlations and then applied PCA to the polychoric correlation matrix[17], with orthogonal rotation (varimax method). We used the %POLYCHOR macro[18] and FACTOR procedure in SAS. The weighting structure was slightly refined after key sub-domains were identified in two domains: Orientation ("time" and "space") and JPS ("complex decision-making", "finance management", "multi-tasking activities, including driving," and "working memory operations").

Creation of the SIST-M and SIST-M-IR—The final instrument is the SIST-M (Structured Interview and Scoring Tool-MADRC), which provides interview prompts representing each of the 60 items and a scoring grid (values of 0, 0.5, 1, or not applicable/unknown for each item). In addition to the SIST-M symptom interview, our clinicians conduct a standard 5-minute objective exam that includes orientation, 3-item registration and delayed recall, abstraction, calculation and serial subtraction. Finally, a separate form – the SIST-M-IR (SIST-M-Informant Report) – was created to obtain reports from a knowledgeable informant. The SIST-M-IR consists of the same 60 items but frames them such that the informant can rate how much the participant has changed, if at all, from 5–10 years earlier. Each item is represented by an introductory question and item-specific response anchors, which can be circled directly on the form. The SIST-M-IR features simple instructions and language, large fonts, and an alternating item shading sequence to enhance readability; pilot work demonstrated that this form is easy for older people to complete in no more than 5–10 minutes. Administration of the SIST-M takes ~25 minutes and involves: 1) performing the structured interview and objective exam with the subject and 2) separately reviewing the SIST-M-IR with the informant. The SIST-M and SIST-M-IR forms are available with this publication as supplemental material.

Methods for evaluating the performance of the SIST-M CDR interview—The SIST-M scoring algorithm was assessed using the legacy data replication cohort, which consisted of 200 participants (15 participants from MAS Cohorts 2 and 3 had missing data on expanded interview items). We cross-sectionally validated the SIST-M in live interviews among MADRC participants: between February 1, 2008 and September 4, 2008, 50 consecutively-recruited participants and their informants were interviewed 1–2 weeks apart (mean=9.7 days, SD=11.6) using the SIST-M and the long (expanded) interview. SIST-M interviews were completed at the MADRC by neurologists and psychiatrists who all had completed online CDR training[19], and 26 of the SIST-M interviews were performed by raters with prior experience with the long interview. Long interviews were conducted via telephone by three experienced MAS raters who were blinded to the SIST-M ratings and algorithm design. Raters for both the SIST-M and expanded interview were unaware of participants' neuropsychological test results. Interviews were assigned such that roughly half the participants were former MAS members (n=24) and half were members of the MADRC cohort only (n=26). Furthermore, approximately half were administered the SIST-M first

(n=27), and half received it second (n=23). Prior to administration of the SIST-M, all informants also completed the SIST-M-IR on their own.

Statistical Analyses

Internal consistency of the SIST-M was initially assessed by calculating Cronbach's α and item-total correlations for each domain in the legacy data replication cohort. To address reliability further, original and algorithm-based CDR ratings for participants were compared using simple or weighted kappa (κ)[20, 21]; CDR-SB agreement was evaluated using intra-class correlation coefficients (ICC)[22, 23].

Internal consistency of the SIST-M was also assessed among the 50 participants of the cross-sectional validation sample, and agreement of the short and long interviews was evaluated using weighted κ for CDR ratings and ICC for CDR-SB. We also assessed whether scores were systematically higher or lower in short vs. long interviews using the Wilcoxon signed rank test for paired observations. Agreement was further scrutinized using Bowker's test of symmetry[24]– to identify patterns among mismatched cells. In addition, differences by cohort type (MAS or MADRC), gender and interview order were assessed using chi-square and Fisher's exact tests as appropriate. Finally, we used κ and ICC to assess agreement between algorithm-based CDR ratings determined using only unguided informant reports on the SIST-M-IR vs. those from the short interview (in which the physician interviewed both subject and informant).

RESULTS

Characteristics of study participants

Table 1 illustrates demographic and clinical characteristics of the SIST-M development and replication cohorts. The cohorts are generally similar, with the exception of race/ethnicity – reflecting assertive recruitment efforts by the MAS to increase minority representation in the later cohorts – as well as greater mean years of education in the replication cohort.

Characteristics of the cross-sectional validation sample are detailed in Table 2. Both sub-cohorts are well-educated, with mean years of education at the baccalaureate level. Notable characteristics among participants recruited directly into the MADRC include younger age, greater proportion of participants with global CDR=0, higher minority participation, and higher prevalence of hypertension and diabetes. Neuropsychological test performance was generally comparable.

Reliability of the SIST-M in legacy data

There was high internal consistency of SIST-M items for each domain (Table 3) except for HH. The relatively low Cronbach's α for all 15 HH items (one item had a negative item-total correlation) was explained by the fact that this domain consists, by definition, of two separate categories ("home" and "hobbies"); the corresponding items were better correlated. Comparing original vs. algorithm-based scores, these were almost identical in the replication cohort; agreement ranged from κ =0.66–0.79 for individual domains; ICC for the CDR-SB=0.9 (Table 4).

Cross-sectional validation of the SIST-M in live interviews

Measures of internal consistency were good to superior (Table 5). Item-total correlations were generally good, but poor correlations were observed for two items. Since one was a core aspect of CA ("decreased participation in social activities"), and the other was an HH item on the FAQ ("difficulty playing a game of skill, such as bridge or chess"), we did not consider removing these items from the SIST-M based on these results.

Agreement of short and long interviews was generally good, with $\kappa \geq 0.70$ for key ratings of memory and global CDR (κ =0.55–0.75 is considered good, and $\kappa \geq 0.75$ is considered excellent[25]); the ICC of 0.73 for the CDR-SB was also good (Table 6). However, comparison of mean ratings from the short and long interviews showed that the short interviews generated lower scores ($p<0.05$ on Wilcoxon signed rank tests for all ratings except global CDR rating; data not shown in table). Further scrutiny using Bowker's test revealed that a disproportionate number of mismatches occurred in which the short interview rating was lower than that of the long interview. This was especially true for HH (p=0.001); Bowker's test was also statistically significant for JPS (p=0.02) and borderline significant for Orientation (p=0.06). Mismatches did not vary significantly by cohort type, gender or order of interview.

Finally, we compared CDR scores obtained by applying the algorithm only to responses on the SIST-M-IR to actual scores from the SIST-M interviews (i.e., combined information from both subjects and informants). Similarly, we generated CDR scores by applying the algorithm to the long interview and SIST-M items, and compared these to the actual scores from the long and SIST-M interviews, respectively. Results demonstrated that, whether applying the algorithm to the long or short interview, algorithm-based scores agreed strongly with actual scores (Table 7). However, when the algorithm was applied to the informant-only responses, agreement with the SIST-M was substantially lower. For example, the ICC (95% CI) for the CDR-SB was only 0.57 (0.38–0.74); it was even lower when comparing these informant-only ratings to the long interview (ICC=0.39 [95% CI = 0.19–0.64]) (data not shown in table).

COMMENT

The SIST-M is an efficient structured interview that can be used to generate CDR scores that are reliable and discriminate along the spectrum of mild cognitive deficits (i.e., CDR-SB=0.0–4.0)[10]. The SIST-M also provides a scoring grid for each component item, such that a validated algorithm can be applied for generating CDR ratings – a useful application for training purposes. Finally, the 60 items of the SIST-M were adapted to create a convenient informant-report form, the SIST-M-IR. Our results show that the SIST-M produces ratings consistent with those from an expanded CDR interview[9]. We observed strong concordance of CDR scores whether we applied an algorithm based on the SIST-M to legacy data or compared SIST-M scores with those from long interviews among subjects in a cross-sectional validation.

Although there are briefer (5–10 minutes) measures of cognition (e.g., MMSE[26], MoCA[27], Mini-Cog[28]), most are based solely on objective performance and cannot be used to address subtle changes and symptoms. A brief informant interview based on the CDR has been developed (the AD8[29]); it takes ~3 minutes to complete and correlates strongly with the CDR[30]. However, this was designed to achieve rapid yet reliable classification of normal cognition (CDR=0) vs. dementia, including mild dementia (CDR ≥ 0.5); the AD8 cannot be used by itself to obtain the 6 CDR ratings and CDR-SB. By contrast, the SIST-M is an interview method for determining ratings in all CDR categories as well as the graded outcome of the CDR-SB. Thus, the SIST-M "system" makes a unique contribution to the existing repertoire of measures: it is a relatively short interview at ~25 minutes, is easy to administer, and yields both the quantitative and qualitative information of the CDR with sensitivity to very mild symptoms.

Another valuable aspect of this study was the development of the SIST-M-IR. Although other informant-based assessments of cognitive symptoms[31] and dementia[32] are available, these were not designed to map directly to CDR domains. By contrast, the SIST-M-IR yields

information necessary to rate each CDR domain. However, we identified important caveats for its use. Informants tended to endorse fewer symptoms on their own than were identified in the context of clinician-guided interviews covering identical items; furthermore, in early stages of cognitive change, informants may be unaware of subtle symptoms or of compensatory measures that a subject himself/herself has adopted in response to challenges. When the SIST-M scoring algorithm was applied to unguided informant reports, there was fair or poor agreement with clinicians' CDR scores from both short and long interviews. By contrast, when the algorithm was applied to item ratings from the clinician interviews, the ICCs comparing algorithm-based and clinician-rated CDR-SB remained >0.9. This suggests that the algorithm itself was not the primary factor with regard to lack of agreement – but rather the loss of information that occurs when considering only reports from informants. Nevertheless, history is often obtained only from informants in many clinical research settings, for a variety of practical reasons. Our results show that such an approach is likely to underestimate systematically levels of impairment. Obtaining joint information from subject and informant, during a clinician-guided interview, provides the optimal method for detecting early cognitive change.

Limitations of this study must also be recognized. First, our results were likely influenced by differences across CDR interviewers. Although all CDR raters had completed training and certification[19], the clinicians who conducted the long interviews had generally been evaluators in the MAS for longer than those who completed the SIST-M; there may have been some "drift" down in CDR ratings by the newer interviewers. This possibility was suggested by the significant differences on tests of mean differences in scores and concordance asymmetry. Consequently, the overall strong agreement (e.g., $\kappa \geq 0.70$ for global CDR and memory) between the SIST-M and long interview was likely an underestimate of true agreement. Notably, κ statistics were lowest for Orientation (0.51) and HH (0.46); however, this is not surprising, as prior work[33] indicated that these two domains are the most difficult to rate and have the lowest agreement with a "gold standard" rater – even among experienced evaluators. A second limitation is that responses on the SIST-M-IR may have been affected by response biases (e.g., global denial or "naysaying"[34]); thus, future enhancements, such as intermittent reverse-coding of items, will be considered[35, 36]. Finally, the SIST-M and SIST-M-IR were developed in a cohort of well-educated elders; thus, generalizability to less-educated populations has not been established. However, the degree of education of our cohort is consistent with educational attainment observed nationally in other ADC/ADRCs, and it is likely that the instrument calibrated to grading subtle changes in our cohort would perform equally well in other sites.

In summary, the SIST-M is a efficient, easily administered and reliable tool for obtaining CDR scores, and provides particular value in clinical and research settings focused on persons with milder cognitive symptoms. Furthermore, we developed a SIST-M algorithm – a tool that could supplement CDR interview training and/or assist with inter-rater score calibration. Finally, we created the SIST-M-IR for rapid attainment of informant input on symptoms. While not sufficient for independent scoring of the CDR, the SIST-M-IR may prove useful for memory and general cognitive screening in large-scale research or primary care clinical settings. Thus, further work in this regard is warranted as well.

Acknowledgments

This study was supported by the Harvard NeuroDiscovery Center, NIA (P01-AG004953), and Massachusetts ADRC (P50-AG005134).

The authors thank Jeanette Gunther and Kelly A. Hennigan for assistance with participant recruitment and visit coordination; Mark Albers, Virginia Hines, Gad Marshall, Carol Mastromauro, Nikolaus McFarland, and Anand Vishwanathan for assistance with clinical evaluations; Laura E. Carroll, Sheela Chandrashekar and Michelle

Schamberg for assistance with data collection, entry and quality checking; and Mary Hyde for assistance with data management and statistical program review. We express special appreciation to all of our study participants.

Supported by the Harvard NeuroDiscovery Center and grants AG004953 and AG005134 from the National Institutes of Health.

Abbreviations

SIST-M	Structured Interview and Scoring Tool - Massachusetts ADRC
SIST-M-IR	SIST-M-Informant Report

References

1. DeKosky ST, Marek K. Looking backward to move forward: early detection of neurodegenerative disorders. Science. 2003; 302:830–834. [PubMed: 14593169]
2. Dickerson BC, Sperling RA. Neuroimaging biomarkers for clinical trials of disease-modifying therapies in Alzheimer's disease. Neurorx. 2005; 2:348–360. [PubMed: 15897955]
3. Lleo A, Greenberg SM, Growdon JH. Current pharmacotherapy for Alzheimer's disease. Annu Rev Med. 2006; 57:513–533. [PubMed: 16409164]
4. Hughes CP, Berg L, Danzinger WL, et al. A new clinical scale for the staging of dementia. Br J Psychiatry. 1982; 140:566–572. [PubMed: 7104545]
5. Morris JC. The Clinical Dementia Rating (CDR): current version and scoring rules. Neurology. 1993; 43:2412–2414. [PubMed: 8232972]
6. The National Alzheimer's Coordinating Center (NACC). http://www.alz.washington.edu
7. Schafer KA, Tractenberg RE, Sano M, et al. Reliability of monitoring the clinical dementia rating in multicenter clinical trials. Alzheimer Dis Assoc Disord. 2004; 18(4):219–222. [PubMed: 15592134]
8. Morris JC, Ernesto C, Schafer K, et al. Clinical Dementia Rating training and reliability in multicenter studies: The Alzheimer's Disease Cooperative Study experience. Neurology. 1997; 48:1508–1510. [PubMed: 9191756]
9. Daly E, Zaitchik D, Copeland M, Schmahmann J, Gunther J, Albert M. Predicting conversion to Alzheimer disease using standardized clinical information. Arch Neurol. 2000; 57:675–680. [PubMed: 10815133]
10. Dickerson BC, Sperling RA, Hyman BT, Albert MS, Blacker D. Clinical prediction of Alzheimer disease dementia across the spectrum of mild cognitive impairment. Arch Gen Psychiatry. 2007; 64(12):1443–1450. [PubMed: 18056553]
11. Storandt M, Grant EA, Miller JP, Morris JC. Longitudinal course and neuropathologic outcomes in original vs revised MCI and in pre-MCI. Neurology. 2006; 67:467–473. [PubMed: 16894109]
12. Blacker D, Lee H, Muzikansky A, et al. Neuropsychological measures in normal individuals that predict subsequent cognitive decline. Arch Neurol. 2007; 64(6):862–871. [PubMed: 17562935]
13. Albert MS, Moss MB, Tanzi R, Jones K. Preclinical prediction of AD using neuropsychological tests. J Int Neuropsychol Soc. 2001; 7:631–639. [PubMed: 11459114]
14. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlam M. Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's disease. Neurology. 1984; 34:939–944. [PubMed: 6610841]
15. Weintraub S, Salmon D, Mercaldo N, et al. The Alzheimer's Disease Centers' Uniform Data Set (UDS): the neuropsychologic test battery. Alzheimer Dis Assoc Disord. 2009; 23(2):91–101. [PubMed: 19474567]
16. Pfeffer RI, Kurosaki TT, Harrah CH Jr, Chance JM, Filos S. Measurement of functional activities in older adults in the community. J Gerontol. 1982; 37(3):323–329. [PubMed: 7069156]
17. Panter AT, Swygert KA, Grant Dahlstrom W, Tanaka JS. Factor analytic approaches to personality item-level data. J Pers Assess. 1997; 68(3):561–589. [PubMed: 16372867]
18. SAS Institute Inc. %POLYCHOR Macro for SAS. Retrieved March 29, 2010 from http://support.sas.com/kb/25/010.html

19. Washington University St. Louis Alzheimer's Disease Research Center. Clinical Dementia Rating Training Application. http://alzheimer.wustl.edu/cdr/Application/Step1.htm
20. Fleiss, JL. Statistical Methods for Rates and Proportions. 2. New York, NY: John Wiley & Sons Inc; 1981.
21. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968; 70:213–220. [PubMed: 19673146]
22. Rosner, BA. Fundamentals of Biostatistics. 4. Belmont, CA: Duxbury Press; 1995.
23. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas. 1973; 33:613–619.
24. Bowker AH. A test for symmetry in contingency tables. J Am Stat Assoc. 1948; 43(244):572–574. [PubMed: 18123073]
25. Fleiss JL, Chilton NW. The measurement of interexaminer agreement on periodontal disease. J Periodontal Res. 1983; 18(6):601–606. [PubMed: 6230433]
26. Folstein MF, Folstein SE, McHugh PR. Mini-Mental State: a practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res. 1975; 12:189–198. [PubMed: 1202204]
27. Nasreddine ZS, Phillips NA, Bedirian V, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc. 2005; 53(4):695–699. [PubMed: 15817019]
28. Borson S, Scanlan JM, Watanabe J, Tu SP, Lessig M. Simplifying detection of cognitive impairment: comparison of the Mini-Cog and Mini-Mental State Examination in a multiethnic sample. J Am Geriatr Soc. 2005; 53(5):871–874. [PubMed: 15877567]
29. Galvin JE, Roe CM, Powlishta KK, et al. The AD8: a brief informant interview to detect dementia. Neurology. 2005; 65:559–564. [PubMed: 16116116]
30. Galvin JE, Roe CM, Xiong C, Morris JC. Validity and reliability of the AD8 informant interview in dementia. Neurology. 2006; 67(11):1942–1948. [PubMed: 17159098]
31. Jorm AF, Jacomb PA. The informant questionnaire on cognitive decline in the elderly (IQCODE): socio-demographic correlates, reliability, validity and some norms. Psychol Med. 1989; 19:1015–1022. [PubMed: 2594878]
32. Kawas C, Segal J, Stewart WF, Corrada M, Thal LJ. A validation study of the Dementia Questionnaire. Arch Neurol. 1994; 51(9):901–906. [PubMed: 8080390]
33. Tractenberg RE, Schafer K, Morris JC. Interobserver disagreements on Clinical Dementia Rating assessment: interpretation and implications for training. Alzheimer Dis Assoc Disord. 2001; 15(3): 155–161. [PubMed: 11522933]
34. Babor TF, Stephens RS, Marlatt GA. Verbal report methods in clinical research on alcoholism: response bias and its minimization. J Stud Alcohol. 1987; 48(5):410–424. [PubMed: 3312821]
35. Bradburn, NM. Response effects. In: Rossi, PH.; Wright, JD., editors. Handbook of Survey Research. New York, NY: Academic Press; 1985. p. 289-328.
36. Bradburn, NM.; Sudman, S., editors. Improving Interview Method and Questionnaire Design: Response Effects to Threatening Questions in Survey Research. San Francisco: Jossey-Bass, Inc., Pubs; 1979.

Table 1

Characteristics at study entry of participants in legacy data sample.

Demographic, Clinical and Genetic	Development cohort (n=147)	Replication cohort (n=200)
Mean (SD) age	72.0 (5.6)	73.0 (5.6)
Mean (SD) years of education	14.8 (2.8)	16.0 (2.9)
Female (%)	61.2	54.0
Non-white race (%)	5.4	11.0
Marital status (%)		
- married	60.5	61.5
- widowed	27.9	17.0
- separated/divorced	6.8	13.5
- never married	4.8	8.0
History of hypertension (%)	36.1	42.5
History of diabetes (%)	4.1	4.5
Current or past smoking (%)	57.5	56.0
APOE ε4 carrier status (%)	29.2	29.3
Neuropsychological, in Mean (SD)		
MMSE	29.2 (1.1)	29.2 (1.1)
CVLT Total Score	50.9 (10.5)	46.7 (12.0)
CVLT Percent Retention	82.6 (18.3)	80.8 (23.5)
Time to Complete Trails B	111.3 (67.3)	98.4 (48.9)
Phonemic Fluency (total of F, A, S)	45.6 (12.3)	43.7 (13.6)
Digit Span Backwards	5.4 (1.5)	5.2 (1.4)
CDR Global Rating (n [%])		
CDR = 0	37 (25.2)	75 (37.5)
CDR = 0.5	110 (74.8)	125 (62.5)

Table 2

Characteristics at study visit of participants in cross-sectional validation sample.

Demographic and Clinical	**All (n=50)**	**Former MAS (n=24)**	**MADRC only (n=26)**
Mean (SD) days between interviews	9.7 (11.6)	10.3 (14.2)	9.2 (8.8)
Mean (SD) age	76.1 (7.8)	78.9 (5.3)	73.6 (8.8)
Mean (SD) years of education	16.6 (3.1)	17.3 (2.9)	15.9 (3.1)
Female (%)	62.0	58.3	65.4
Non-white race (%)	10.0	4.2	15.4
Marital status (%)			
- married	70.0	75.0	65.4
- widowed	14.0	8.3	19.2
- separated/divorced	2.0	4.2	0
- never married	14.0	12.5	15.4
History of hypertension (%)	56.0	45.8	65.4
History of diabetes (%)	10.0	4.2	15.4
Current or past smoking (%)	58.0	54.2	61.5
Neuropsychological, in Mean (SD)			
MMSE	28.7 (2.0)	29.3 (1.1)	28.4 (2.5)
Logical Memory Immediate Recall	13.5 (3.6)	13.6 (3.5)	13.4 (3.8)
Logical Memory Delayed Recall	12.4 (4.1)	12.5 (3.6)	12.3 (4.5)
Time to Complete Trails B	85.8 (42.4)	89.8 (40.3)	81.9 (44.8)
Semantic Fluency			
- Animals	19.2 (5.1)	20.0 (5.7)	18.4 (4.4)
- Vegetables	14.4 (3.3)	15.1 (3.7)	13.8 (2.8)
Digit Span Backwards	4.9 (1.3)	5.0 (1.3)	4.8 (1.3)
CDR Global Rating (n [%])			
CDR = 0	23 (46.0)	6 (25.0)	17 (65.4)
CDR = 0.5	27 (54.0)	18 (75.0)	9 (34.6)

Table 3

Internal consistency of SIST-M items in the legacy data replication cohort (n=200).

CDR Domain	Cronbach's coefficient alpha	Range of item-total correlations
Memory	0.89	0.47 – 0.73
Orientation	0.87	0.35 – 0.81
Judgment and Problem-solving	0.90	0.36 – 0.81
Community Affairs	0.87	0.64 – 0.74
Home and Hobbies*	0.58	−0.16 — 0.53
- Home†	0.67	0.15 – 0.50
- Hobbies†	0.65	0.04 – 0.66
Personal Care‡	N/A	N/A

*All 15 Home and Hobbies items were assessed together.

†The 6 Home and 9 Hobbies items were assessed separately.

‡Reliability coefficients were not applicable for Personal Care, as only three items are used to rate this domain in the SIST-M, and nearly all participants received a score of 0 in this domain.

Table 4

CDR scores in the legacy data replication cohort (n=200), by original vs. algorithm-based rating.

	Original rating: N (%)	**Algorithm-based rating: N (%)**	**Agreement by κ or ICC (95% CI)**
CDR Sum-of-Boxes	Mean (SD), range = 0.96 (1.00), 0–3.5	Mean (SD), range = 0.94 (1.01), 0–4.0	0.89 (0.86–0.91)
- 0	74 (37.0)	74 (37.0)	
- 0.5	31 (15.5)	31 (15.5)	
- 1.0, 1.5	52 (26.0)	51 (25.5)	
- 2.0, 2.5	30 (15.0)	32 (16.0)	
- 3.0, 3.5	13 (6.5)	9 (4.5)	
- ≥ 4.0	0 (0)	3 (1.5)	
CDR Global rating			0.77 (0.68–0.86)
- 0	75 (37.5)	83 (41.5)	
- 0.5	125 (62.5)	117 (58.5)	
Memory			0.76 (0.69–0.83)
- 0	76 (38.0)	88 (44.0)	
- 0.5	106 (52.0)	93 (46.5)	
- 1.0	8 (4.0)	19 (9.5)	
Orientation			0.76 (0.67–0.85)
- 0	145 (72.5)	143 (71.5)	
- 0.5	50 (25.0)	50 (25.0)	
- 1.0	5 (2.5)	7 (3.5)	
Judgment and Problem-solving			0.70 (0.61–0.79)
- 0	127 (63.5)	137 (68.5)	
- 0.5	71 (35.5)	56 (28.0)	
- 1.0	2 (1.0)	7 (3.5)	
Community Affairs			0.66 (0.55–0.78)
- 0	158 (79.0)	158 (79.0)	
- 0.5	40 (20.0)	36 (18.0)	
- 1.0	2 (1.0)	6 (3.0)	
Home and Hobbies			0.79 (0.70–0.87)
- 0	138 (69.0)	139 (69.5)	
- 0.5	61 (30.5)	60 (30.0)	
- 1.0	1 (0.5)	1 (0.5)	
Personal Care*			N/A
- 0	200 (100)	199 (99.5)	
- 1.0	0 (0)	1 (0.5)	

* κ was not calculated, as only one participant was rated as impaired (>99% absolute agreement).

Table 5

Internal consistency of SIST-M items in the cross-sectional validation sample (n=50).

CDR Domain	Cronbach's coefficient alpha	Range of item-total correlations
Memory	0.92	0.48 – 0.85
Orientation	0.83	0.37 – 0.85
Judgment and Problem-solving	0.93	0.27 – 0.91
Community Affairs	0.62	−0.12 — 0.83
Home and Hobbies	0.90	−0.10 — 0.87
Personal Care*	N/A	N/A

*Reliability coefficients were not applicable for Personal Care, as only three items are used to rate this domain in the SIST-M, and all participants received a score of 0 is this domain.

Table 6

CDR scores in the cross-sectional validation sample (n=50), by long vs. short (SIST-M) interview format.

	Long format: N (%)	**Short format: N (%)**	**Agreement by κ or ICC (95% CI)**
CDR Sum-of-Boxes	Mean (SD), range = 1.34 (1.32), 0–4.5	Mean (SD), range = 0.81 (1.05), 0–4.0	0.73 (0.58–0.84)
- 0	17 (34.0)	20 (40.0)	
- 0.5	4 (8.0)	11 (22.0)	
- 1.0, 1.5	11 (22.0)	12 (24.0)	
- 2.0, 2.5	10 (20.0)	2 (4.0)	
- 3.0, 3.5	6 (12.0)	4 (8.0)	
- ≥ 4.0	2 (4.0)	1 (2.0)	
CDR Global rating			0.70 (0.52–0.88)
- 0	18 (36.0)	23 (46.0)	
- 0.5	31 (62.0)	27 (54.0)	
- 1.0	1 (2.0)	0 (0)	
Memory			0.71 (0.58–0.84)
- 0	18 (36.0)	23 (46.0)	
- 0.5	22 (44.0)	21 (42.0)	
- 1.0	10 (20.0)	6 (12.0)	
Orientation			0.51 (0.29–0.74)
- 0	33 (66.0)	41 (82.0)	
- 0.5	16 (32.0)	9 (18.0)	
- 1.0	1 (2.0)	0 (0)	
Judgment and Problem-solving			0.61 (0.43–0.79)
- 0	26 (52.0)	35 (70.0)	
- 0.5	20 (40.0)	14 (28.0)	
- 1.0	4 (8.0)	1 (2.0)	
Community Affairs			0.66 (0.45–0.87)
- 0	35 (70.0)	39 (78.0)	
- 0.5	13 (26.0)	9 (18.0)	
- 1.0	2 (4.0)	2 (4.0)	
Home and Hobbies			0.46 (0.24–0.67)
- 0	26 (52.0)	41 (82.0)	
- 0.5	20 (40.0)	8 (16.0)	
- 1.0	4 (8.0)	1 (2.0)	
Personal Care*			N/A
- 0	50 (100)	50 (100)	
- 1.0	0 (0)	0 (0)	

* κ was not calculated, as no participants were rated as impaired (100% absolute agreement).

Table 7

Agreement of algorithm-based ratings with actual ratings.

	Agreement by κ or ICC (95% CI)*	Agreement by κ or ICC (95% CI)†	Agreement by κ or ICC (95% CI)‡
CDR Sum-of-Boxes	0.95 (0.91 – 0.97)	0.93 (0.89 – 0.96)	0.57 (0.38 – 0.74)
CDR Global rating	0.81 (0.66 – 0.96)	0.73 (0.56 – 0.91)	0.44 (0.22 – 0.66)
Memory	0.83 (0.72 – 0.92)	0.84 (0.73 – 0.95)	0.55 (0.31 – 0.80)
Orientation	0.76 (0.59 – 0.94)	0.73 (0.54 – 0.93)	0.47 (0.23 – 0.71)
Judgment and Problem-solving	0.85 (0.72 – 0.97)	0.71 (0.47 – 0.94)	0.49 (0.16 – 0.83)
Community Affairs	0.87 (0.72 – 0.99)	0.86 (0.71 – 0.99)	0.42 (0.07 – 0.77)
Home and Hobbies	0.81 (0.67 – 0.94)	0.78 (0.60 – 0.96)	0.28 (−0.05 – 0.62)
Personal Care	N/A	N/A	N/A

*Algorithm-based ratings from long interview vs. actual ratings from long interview.

†Algorithm-based ratings from SIST-M interview vs. actual ratings from SIST-M interview.

‡Algorithm-based ratings from spontaneous informant reports on the SIST-M-IR vs. actual ratings from SIST-M interview.