Published in final edited form as:
Neuroepidemiology. 2014 ; 42(3): 144–153. doi:10.1159/000357647.

Calibration and validation of an innovative approach for estimating general cognitive performance

Alden L. Gross, PhD, MHS[a,b,c,*], **Richard N. Jones, ScD**[d], **Tamara G. Fong, MD, PhD**[a,e], **Douglas Tommet, MS**[d], and **Sharon K. Inouye, MD, MPH**[a,e]

[a]Aging Brain Center, Institute for Aging Research, Hebrew SeniorLife, Boston, MA, USA

[b]Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, 02131, USA

[c]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

[d]Departments of Psychiatry and Human Behavior and Neurology, Warren Alpert Medical School, Brown University, Providence, RI, USA

[e]Department of Neurology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, 02131, USA

Abstract

Objective—To evaluate a new approach for creating a composite measure of cognitive function, we calibrated a measure of general cognitive performance from existing neuropsychological batteries.

Methods—We applied our approach in an epidemiologic study and scaled the composite to a nationally representative sample of older adults. Criterion validity was evaluated against standard clinical diagnoses. Convergent validity was evaluated against the Mini-Mental State Examination (MMSE).

Results—The general cognitive performance factor was scaled to have a mean=50 and SD=10 in a nationally representative sample of older adults. A cut-point of approximately 45, corresponding with an MMSE of 23/24, optimally discriminated participants with and without dementia (sensitivity=0.94; specificity=0.90; AUC=0.97). The general cognitive performance factor was internally consistent (Cronbach's alpha=0.91) and provided reliable measures of functional ability across a wide range of cognitive functioning. It demonstrated minimal floor and ceiling effects, which is an improvement over most individual cognitive tests.

Conclusions—The cognitive composite is a highly reliable measure, with minimal floor and ceiling effects. We calibrated it using a nationally representative sample of adults over age 70 in the US and established diagnostically relevant cut-points. Our methods can be used to harmonize neuropsychological test results across diverse settings and studies.

[*]Corresponding author: Alden L. Gross, aldgross@jhsph.edu; Phone: 410 474 3386; Fax: 617-971-5309; Johns Hopkins Center on Aging and Health, 2024 E. Monument Street, Suite 2-700, Baltimore, MD 21205.

None of the authors report any conflicts of interest or received compensation for this work.

Keywords

dementia; cognitive function; measurement; factor analysis; harmonization; calibration

Introduction

Despite its central role in the daily lives of older adults, there is no widely used, standardized method of assessing overall or global cognitive function across a wide range of functioning. Over 500 neuropsychological tests exist for clinical and research purposes (1). This tremendous diversity complicates comparison and synthesis of findings about cognitive functioning across a broad range of performance in multiple samples in which different neuropsychological batteries were administered. Although test batteries are used to examine domain-specific cognitive function, summary measures provide global indices of function (2). Brief global cognitive tests such as the Mini-Mental State Examination (MMSE) and others are limited by prominent ceiling effects and skewed score distributions (3-6), thus evaluating a limited range of cognitive function. These measurement properties substantially hamper the capacity to measure longitudinal change, since score ranges are limited and a point change has different meanings across the range of values (3,7). Summary measures of general cognition, if properly calibrated, may be more sensitive to impairments across a broader range of cognitive function and be more sensitive to changes over time (8).

Approaches to creating summary cognitive measures have been limited and controversial. One approach involves standardizing scores of component cognitive tests, which are then averaged into a composite (9,10). Although widely used, this approach is limited because standardizing variables does not address skewed response distributions, does not allow differential weighting of tests, and ultimately does not facilitate comparisons of findings across studies. An alternative approach uses latent variable methods to summarize tests. The tests are weighted and combined in a composite measure that has more favorable measurement properties including minimal floor and ceiling effects, measurement precision over a wide range of function, and interval-level properties that make the composite optimal for studying longitudinal change (e.g., 8,11).

In a previous study, Jones and colleagues (12) used confirmatory factor analysis to develop a general cognitive performance factor from a neuropsychological battery (13). This measure was shown to be unidimensional and internally consistent (Cronbach's alpha=0.82). The factor was defined by high loadings on six of the ten component tests. The cognitive factor was designed to be sensitive to a wide range of cognitive function from high functioning to scores indicative of dementia. Scores were normally distributed and reliable (reliability index>0.90). With these properties, the general cognitive performance factor provides a robust approach to assess cognitive function over time.

Despite these clear advantages, an important limitation of the general cognitive performance factor is that its scores are not yet clinically interpretable or generalizable across studies. To address this limitation, the aims of the present study were to: (1) calibrate a general cognitive performance factor to a nationally representative sample of adults over age 70 in the US, (2) validate the general cognitive performance factor against reference standard

clinical diagnoses, (3) examine convergent validity of the cognitive performance factor score, and (4) identify clinically meaningful cut-points for the cognitive factor score. Our overall goal was to create a clinically meaningful measurement tool, and importantly, to demonstrate an approach that is generalizable to other different neuropsychological test batteries on a broader scale.

Materials and Methods

Study samples

Participants were drawn from the Successful AGing after Elective Surgery (SAGES) study and the Aging, Demographics, and Memory Study (ADAMS), a substudy of the Health and Retirement Study. SAGES is a prospective cohort study of long-term cognitive and functional outcomes of hospitalization in elective surgery patients. After recruitment, a neuropsychological battery was administered to participants just before surgery and peri-annually for up to three years. Because data collection was ongoing at the time of this study, we used pre-operative data for the first 300 patients enrolled. Eligible participants were at least 70 years of age, English-speaking, and scheduled for elective surgery at one of two academic teaching hospitals in Boston, MA. Exclusion criteria included evidence of dementia or delirium at baseline. Study procedures were approved by the Institutional Review Board at the Beth Israel Deaconess Medical Center.

ADAMS is a nationally representative sample of 856 older adults in the United States interviewed in 2002-2004 (14). Its parent study, the Health and Retirement Study, is a longitudinal survey of over 20,000 community-living retired persons. ADAMS, which began as a population-based study of dementia, initially identified a stratified random sample of 1,770 participants; 687 refused and 227 died before they were interviewed, yielding 856 participants. Participants with probable dementia and minorities were over-sampled. We used survey weights to account for the complex survey design and to make estimates representative of adults over age 70 in the US (15). ADAMS was approved by Institutional Review Boards at the University of Michigan and Duke University.

Measures

Neuropsychological test batteries—In SAGES, a neuropsychological test battery was administered during an in-person evaluation that consisted of 11 tests of memory, attention, language, and executive function. We used ten tests from the ADAMS battery, of which seven were in common with SAGES (Table 1). As explained in more detail in the statistical analysis, our modeling approach allows cognitive ability to be estimated based on responses to any subset of cognitive tests (16).

Clinical diagnoses—Clinical diagnoses, grouped as normal cognitive function, cognitive impairment-no dementia (CIND), and all-cause dementia, were assigned in ADAMS by an expert clinical consensus panel (14,17). Diagnoses were determined after a review of data collected during in-home participant assessments, which included neuropsychological and functional assessments from participants and proxies. A diagnosis of dementia was based on the Diagnostic and Statistical Manual of Mental Disorders III-R and IV (18,19) and for the

present study included probable and possible AD, probable and possible vascular dementia, dementia associated with other conditions (Parkinson's disease, normal pressure hydrocephalus, frontal lobe dementia, severe head trauma, alcoholic dementia, Lewy Body dementia), and dementia of undetermined etiology. CIND was defined as functional impairment that did not meet criteria for all-cause dementia or as below-average performance on any test (17). CIND included participants with Mild Cognitive Impairment or cognitive impairment due to other causes (e.g., vascular disease, depression, psychiatric disorder, mental retardation, alcohol abuse, stroke, other neurological condition).

MMSE—The MMSE is a brief 30-point cognitive screening instrument used to assess global mental status. The MMSE is widely used in clinical and epidemiologic settings. Its validity as a screening test for all-cause dementia in clinical populations has been previously established (20,21). George (22) recommended cut-points of 23/24 for moderate cognitive impairment and 17/18 for severe cognitive impairment. These cut-points have been widely applied in clinical and research settings (21-25). MMSE 9/10 has also been used to indicate severe impairment (26). Although not a preferred test for identification of Mild Cognitive Impairment (21), cut-points of 26/27 (23,27,28) and 28/29 (29) have been used for that purpose. Although the MMSE has poor measurement properties and ceiling effects, we used is as a standard for comparison in this study because it remains widely used and its scores and cutoffs are well-recognized.

Statistical analyses

We used descriptive statistics to characterize the SAGES and ADAMS samples. Analyses were subsequently conducted in four steps: (1) score the general cognitive performance factor in SAGES and calibrate it to ADAMS using item response theory (IRT) methods, (2) assess criterion validity of the general cognitive performance factor using reference standard clinical ratings in ADAMS, (3) assess convergent validity of the general cognitive performance factor with the MMSE, and (4) identify cut-points on the general cognitive performance factor corresponding to published MMSE cut-points.

Score the general cognitive performance factor in SAGES and calibrate to ADAMS—

We calculated internal consistency of the cognitive tests in SAGES using Cronbach's alpha (13). The Cronbach's alpha statistic has a theoretical range between 0 and 1, with 1 indicating high internal consistency. A generally accepted reliability for analysis of between-person group differences is ≥0.80 and for within-person change is ≥0.90 (13). Next, we calculated the general cognitive performance factor in SAGES from a categorical variable factor analysis of cognitive tests. The general cognitive performance factor score was scaled to have a mean of 50 and standard deviation (SD) of 10 in the US population over age 70. The factor analysis is consistent with the IRT graded response model, and facilitated precise estimation of reliability across the range of performance (30-33). In IRT, reliability is conceptualized as the complement of the squared standard error of measurement (SEM) (34). The SEM is estimated from the test information function, which varies over the range of ability. We described the precision of the measure over the range of general cognitive performance using the standard error of measurement calculated based on the IRT measurement model (35,36).

To scale the general cognitive performance factor in SAGES to the nationally representative ADAMS sample, we took advantage of tests in common between studies to calibrate scores in SAGES to ADAMS (3). We categorized cognitive tests into up to 10 discrete equal-width categories (37) to avoid model convergence problems caused by outlying values (38) and to place all tests on a similar scale (12) (see Supplemental Table 2). Tests in common between studies were categorized based on the sample distribution in ADAMS. We assigned model parameters for anchor tests in the SAGES factor analysis based on corresponding estimates from the ADAMS-only model that used population-based survey weighting. This procedure allowed us to scale the general cognitive performance factor to reflect the general population of US adults aged 70 and older. Importantly, because the IRT model handles missing data under the assumption that cognitive performance data are missing at random conditional on variables in the model, general cognitive performance can be calculated for each participant based on responses to any subset of cognitive tests as long as not all test scores are missing. The factor score is the mean of the posterior probability distribution of the expected a priori latent trait estimate. The posterior probability distribution refers to the conditional density of the latent cognitive performance trait given observed cognitive test scores. Because the factor is computed on the basis of all available information in a participant's response patterns, it can be computed regardless of missing tests (39).

We conducted diagnostic procedures using Monte Carlo simulation by generating 100,001 hypothetical observations with the MMSE and all SAGES and ADAMS cognitive measures. Simulated cognitive test distributions matched those of our empirical samples. This simulation allowed us to rigorously compare SAGES and ADAMS scores to the overall general cognitive performance factor using correlations and Bland-Altman plots to examine systematic differences between the measures (40).

Criterion validity of the general cognitive performance factor—To evaluate criterion validity for distinguishing dementia and CIND, we used logistic regression. We report overall areas under the curve (AUC) for the general cognitive performance factor and diagnostic characteristics for the score that maximized sensitivity and specificity (41).

Convergent validity of the general cognitive performance factor—We correlated the general cognitive performance factor with MMSE using Pearson correlation coefficients.

Link the general cognitive performance factor and MMSE—The MMSE is a widely used screening test for global cognitive status. Because of its widespread use, many clinicians and researchers are familiar with its scores. Thus, the MMSE provides a set of readily recognizable cutpoints which we utilized as guideposts for comparison with the general cognitive performance measure. To produce a crosswalk between the general cognitive performance factor and MMSE, we used equipercentile linking methods to correlate scores (42). This step allowed the direct comparison of general cognitive performance factor scores that correspond to MMSE scores. Equipercentile linking identifies scores on two measures (MMSE and general cognitive performance factor) with the same percentile rank, and assigns general cognitive performance a value from the reference test, MMSE, at that percentile. This approach is appropriate when two tests are on different

scales, and is most useful when the distribution of the reference test is not normally distributed (43).

Results

In SAGES (n=300), most participants were female (56%), white (95%), on average 77 years old (range 70, 92), and had at least a college education (70%)(Table 1). Few had dementia (n=5, 1.7%) or CIND (n=19, 6.3%). By comparison, in ADAMS (n=856), which is representative of persons over age 70, participants were mostly female (61%), white (89%), on average 79 years of age (range 70, 110), and 37% had at least a college education. Relative to SAGES, the ADAMS sample was older (P<0.001), more ethnically diverse (P=0.001), less highly educated (P<0.001), and had higher levels of cognitive impairment (n=308, 13.7% had dementia and n=241, 22.0% had CIND).

Derivation of the general cognitive performance factor in SAGES and calibration to ADAMS

By design, the general cognitive performance factor in ADAMS had a mean of 50 and standard deviation (SD) of 10. The general cognitive performance factor in SAGES was 0.9 SD above the national average, reflecting their higher education and younger average age. The cognitive tests in ADAMS were internally consistent (Cronbach's alpha=0.91). The reliability of the general cognitive performance factor, derived based on the standard error of measurement, was above 90% for scores between 40 and 70, which included 84% of the ADAMS sample. Correlations between the general cognitive performance factor and items from SAGES were above 0.80, with the exception of HVLT delayed recall (r=0.65). The tests represent multiple domains including memory, executive function, language, and attention, suggesting the factor represents general cognitive performance and is not dominated by a particular cognitive domain. Figure 1 demonstrates that general cognitive performance factor scores in SAGES and ADAMS were normally distributed; on average, SAGES participants had higher levels of cognitive function.

Using simulated data, the correlation between the study-specific general cognitive performance factor for SAGES and ADAMS was above 0.97. Bland-Altman plots further revealed no systematic bias across the range of general cognitive performance scores, suggesting the general cognitive performance factor was not measured differently across the two studies (Supplemental Figure 1).

Criterion validity of the general cognitive performance factor

Figure 2 shows receiver operating curves for distinguishing dementia and CIND in ADAMS. The general cognitive performance factor score that best discriminated dementia participants from cognitively normal participants was less than 44.8 (sensitivity = 0.94; specificity = 0.90; Figure 2, right panel). This cut-point correctly classified 94% of the sample. The area under the curve (AUC) was 0.97. The general cognitive performance factor score that best discriminated CIND participants from cognitively normal participants was less than 49.5 (sensitivity = 0.80, specificity = 0.76; Figure 2, left panel). This cut-point correctly classified 79% of the sample (AUC=0.84).

The AUC for the general cognitive performance factor was significantly greater than the AUC for each constituent test (Supplemental Table 1). The only exception was for immediate word recall, which was superior for predicting dementia.

Convergent validity of the general cognitive performance factor

The correlation between the general cognitive performance factor and the MMSE was 0.91 in ADAMS (P<0.001), indicating strong evidence of convergent validity.

Crosswalk between general cognitive performance factor and MMSE

The equipercentile linking procedure is illustrated in Figure 3. Scores for the general cognitive performance factor and MMSE were matched based on percentile ranks. For example, a score of 24 on the MMSE, for example, has the same percentile rank as a score of 45 on the general cognitive performance factor.

After equipercentile-linking the general cognitive performance factor with the MMSE, the score corresponding to an MMSE cut-point of 23/24 was <45.2, or 0.48SD below the national average (Table 2). This cut-point was nearly identical to the score that best discriminated participants with and without dementia (Figure 2). General cognitive performance factor scores of <50.9 and <40.7 corresponded to MMSE scores of 26/27 and 17/18, respectively. Table 2 provides sensitivity, specificity, positive and negative predictive values, and likelihood ratio positive and negative statistics for predicting dementia and CIND for general cognitive performance factor scores corresponding to MMSE scores of 29, 27, 24, 18, and 10. The general cognitive performance factor cut-point of 45.2 correctly classified 90% of persons with dementia (sensitivity) and 93% of persons without dementia (specificity). This cut-point is moderately strong for confirming dementia (positive predictive value = 73%; likelihood ratio positive = 12.8) and has excellent negative predictive value (98%).

We constructed a crosswalk to show scores on the general cognitive performance factor and corresponding scores on the MMSE (Figure 4). Irregular levels between each score on the MMSE and the limited observable range of MMSE evident in this figure underscores the broader range and better interval scaling properties of the general cognitive performance factor.

Discussion

We developed a unidimensional factor of cognitive performance in the SAGES study, scaled it to national norms for adults over 70 years of age living in the US, and evaluated its criterion and convergent validity. When validated against reference standard diagnoses for dementia, the score of approximately 45 has a sensitivity of 94% and specificity of 90% (AUC=0.97), indicating outstanding performance. Convergent validity with the MMSE was excellent (correlation=0.91, P<0.001). Cognitive tests comprising the general cognitive performance factor are internally consistent (Cronbach's alpha=0.91). The factor is highly precise across most of the score range (reliability>0.90). To enhance the clinical relevance of the scores, we provided correlations with widely used scores for the MMSE. Notably, the

score of 45 was the optimal cut-point for dementia and also corresponded to an MMSE of 23/24.

General cognitive performance factors previously have been shown to have minimal floor and ceiling effects, measurement precision over a wide range of cognitive function, and interval scaling properties that make it an ideally suited measure of change (12,31). We replicated these findings, and identified meaningful cut-points to further enhance the potential utility of the measure. Strengths of this study include calibration of the cognitive composite to a large, nationally representative sample of older adults in which rigorous reference standard clinical diagnoses were available. With this sample, we were able to evaluate criterion and convergent validity of the general cognitive performance factor for detecting cognitive impairment and demonstrate its favorable test performance characteristics.

Several caveats merit attention. First, the general cognitive performance factor is not intended to diagnose dementia or MCI. It simply provides a refined cognitive measure, which like any cognitive measure represents only one piece of the necessary armamentarium for establishing a clinical diagnosis. Second, seven tests were in common to calibrate cognitive composites in ADAMS and SAGES, and Bland-Altman plots confirmed they were similar between studies. Although we are convinced that the general cognitive performance factors developed in the two samples were equivalent, further research is needed to determine minimum sample sizes, number of cognitive tests available in common between studies, and the degree of correlation among tests needed to estimate a reliable composite in a new sample. Existing research suggests that fewer than five anchor items in an IRT analysis such as ours is enough to produce reasonably accurate parameter estimates (44); one previous study used one item in common to calibrate different scales (45). Third, some criterion contamination is potentially present in our evaluation of criterion validity because the general cognitive performance factor in ADAMS is a unique combination of shared variability among test items that were available to clinicians when assigning clinical diagnoses. However, comparison of AUC's in Supplemental Table 1 for the general factor and individual cognitive tests revealed the former performed better than most of its constituent parts. Fourth, positive and negative predictive values in our study are dependent on base rates, and may differ in other samples. Fifth, while reliability of the general cognitive performance factor was excellent across a range of performance that included 84% of the ADAMS population, results suggest it is less reliable at more impaired levels. Future calibration efforts should consider cognitive tests that are sensitive to impairment in populations with more severe degrees of dementia. A final caveat is that our approach is not intended to replace examination and interpretation of individual neuropsychological tests. Such examination remains an important approach to examine domain-specific cognitive changes to assist with clinical diagnosis and to understand pathophysiologic correlates of various cognitive disorders.

An important implication of the present work is the potential of deriving the general cognitive performance factor to other samples with neuropsychological batteries that have overlap with the battery used in ADAMS. Extrapolation of these methods holds the potential for harmonizing batteries to enhance comparability and even synthesize results across

studies through integrative data analysis. This method would thus address a substantial limitation for combining existing studies using disparate neuropsychological batteries (46). Harmonizing samples together with different research designs and demographic characteristics provides opportunities to make findings more generalizable. Without this common metric, or at least one test in common across studies, to conduct integrative analysis, one must resort to comparing multiple data points from normative tests that potentially measure diverse cognitive domains.

The need for uniform measures of cognitive function, derived using rigorous psychometric methods, has been recognized by national groups (47). Uniform, psychometrically sound measures are a central focus of the NIH PROMIS and Toolbox initiatives. Our study is consistent with these goals. The innovative approach demonstrated here used psychometric methods to generate a unidimensional general cognitive performance composite with excellent performance characteristics that can be used to measure cognitive change over time and across study. We established clinically meaningful, population-based cut-points. This measure, and the methods used to create it, holds substantial promise for advancing work to evaluate progression of cognitive functioning over time. Perhaps most importantly, our methods can facilitate future strategies to integrate cognitive test results across epidemiologic and clinical studies of cognitive aging.

Conclusions

We created a composite factor for general cognitive performance from widely used tests for neuropsychological functioning using psychometrically sophisticated methods. We used publicly available neuropsychological performance data from the Aging, Demographics, and Memory Study to calibrate general cognitive performance to a nationally representative sample of adults age 70 and older in the US. This calibration enabled us to describe cognitive functioning in our study on a nationally representative scale. The general cognitive performance factor was internally consistent, provided reliable measures of functional ability across a wide range of cognitive functioning, and demonstrated minimal floor and ceiling effects. It also demonstrated criterion validity: a cut-point of approximately 45, corresponding with an MMSE of 23/24, optimally discriminated participants with and without dementia (sensitivity=0.94; specificity=0.90; AUC=0.97). Our approach has broad applicability and usefulness to directly compare cognitive performance in new and existing studies when overlapping items with the ADAMS neuropsychological battery are present. These methods can facilitate interpretation and synthesis of findings in existing and future research studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by a grant from the National Institute on Aging (P01AG031720, Inouye). Dr. Gross was supported by a National Institutes of Health Translational Research in Aging post-doctoral fellowship (T32AG023480). Dr. Inouye holds the Milton and Shirley F. Levy Family Chair in Alzheimer's Disease. Dr. Fong

was supported by NIA Career Development Award, K23AG031320. The contents do not necessarily represent the views of the funding entities.

Dr. Gross had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Author contributions are as follows. Study concept and design: Inouye, Gross, Jones. Acquisition of data: Inouye, Gross, Jones. Analysis and interpretation of data: Gross, Jones, Inouye, Tommet, Fong. Drafting the manuscript: Gross, Jones, Inouye. Critical revision of the manuscript for important intellectual content: Inouye, Gross, Jones, Fong. Statistical analysis: Gross, Jones. Obtained funding: Inouye. Administrative, technical, and material support: Inouye, Tommet.

References

1. Lezak, MD.; Howieson, DB.; Loring, DW. Neuropsychological assessment. New York: Oxford University Press; 2004.
2. Chandler MJ, Lacritz LH, Hynan LS, Barnard HD, Allen G, Deschner M, Weiner MF, Cullum CM. A total score for the CERAD neuropsychological battery. Neurology. 2005; 65(1):102–106. [PubMed: 16009893]
3. Crane PK, Narasimhalu K, Gibbons LE, Mungas DM, Haneuse S, Larson EB, et al. Item response theory facilitated calibrating cognitive tests and reduced bias in estimated rates of decline. Journal of Clinical Epidemiology. 2008; 61:1018–1027. [PubMed: 18455909]
4. Folstein MF, Folstein SE, McHugh PR. “Mini–mental state”: A practical guide for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research. 1975; 12:189–198. [PubMed: 1202204]
5. Schultz-Larsen K, Kreiner S, Lomholt RK. Mini-Mental State Examination: Mixed Rasch model item analysis derived two different cognitive dimensions of the MMSE. Journal of Clinical Epidemiology. 2007; 60:268–279. [PubMed: 17292021]
6. Simard M. The Mini-Mental State Examination: Strengths and weaknesses of a clinical instrument. The Canadian Alzheimer Disease Review. 1998; 12:10–12.
7. Wouters H, van Gool WA, Schmand B, Zwinderman AH, Lindeboom R. Three sides of the same coin: measuring global cognitive impairment with the MMSE, ADAS-cog and CAMCOG. International Journal of Geriatric Psychiatry. 2010; 25(8):770–779. [PubMed: 19946861]
8. McArdle JJ, Grimm KJ, Hamagami F, Bowles RP, Meredith W. Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. Psychological Methods. 2009; 14:126–149. [PubMed: 19485625]
9. Tuokko H, Woodward TS. Development and validation of a demographic correction system for neuropsychological measures used in the Canadian Study of Health and Aging. Journal of Clinical and Experimental Neuropsychology. 1996; 18:479–616. [PubMed: 8877629]
10. Wilson RS, Mendes de Leon CF, Barnes LL, Schneider JA, Bienias JL, Evans DA, et al. Participation in cognitively stimulating activities and risk of incident Alzheimer's disease. JAMA. 2002; 287:742–748. [PubMed: 11851541]
11. Strauss ME, Fritsch T. Factor structure of the CERAD neuropsychological battery. Journal of the International Neuropsychological Society. 2004; 10:559–565. [PubMed: 15327734]
12. Jones RN, Rudolph JL, Inouye SK, Yang FM, Fong TG, Milberg WP, et al. Development of a unidimensional composite measure of neuropsychological functioning in older cardiac surgery patients with good measurement precision. Journal of Clinical and Experimental Neuropsychology. 2010; 32:1041–1049. [PubMed: 20446144]
13. Nunnally, JC.; Bernstein, IH. Psychometric Theory. New York: McGraw-Hill; 1994.
14. Langa KM, Plassman BL, Wallace RB, Herzog AR, Heeringa SG, Ofstedal MB, et al. The Aging, Demographics, and Memory Study: study design and methods. Neuroepidemiology. 2005; 25:181–91. [PubMed: 16103729]
15. Juster FT, Suzman R. An overview of the Health and Retirement Study. Journal of Human Resources. 1995; 30(Suppl):7–56.
16. Bjorner, JB.; Kosinski, M.; Ware, JE, Jr. Computerized adaptive testing and item banking. In: Fayers, PM.; Hays, RD., editors. Assessing quality of life. Oxford: Oxford University Press; 2004.

17. Plassman BL, Langa KM, Fisher GG, Heeringa SG, Weir DR, Ofstedal MB, et al. Prevalence of dementia in the United States: The Aging, Demographics, and Memory Study. Neuroepidemiology. 2007; 29:125–132. [PubMed: 17975326]
18. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (3th edn) (DSM–III-R). Washington DC: APA; 1987.
19. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (4th edn) (DSM–IV). Washington DC: APA; 1994.
20. Feher EP, Mahurin RK, Doody RS, Cooke N, Sims J, Pirozzolo FJ. Establishing the limits of the Mini-Mental State: Examination of 'subtests'. Archives of Neurology. 1992; 49:87–92. [PubMed: 1728269]
21. Mitchell AJ. A meta-analysis of the accuracy of the Mini-Mental State Examination in the detection of dementia and mild cognitive impairment. Journal of Psychiatric Research. 2009; 43:411–431. [PubMed: 18579155]
22. George, L.; Landerman, R.; Blazer, D.; Anthony, J. Cognitive impairment In Psychiatric Disorders in America. Robins, L.; Regier, D., editors. The Free Press; New York: 1991. p. 291-327.
23. Crum RM, Anthony JC, Bassett SS, Folstein MF. Population-Based Norms for the Mini-Mental State Examination by Age and Educational Level. Journal of the American Medical Association. 1993; 269:2386–2391. [PubMed: 8479064]
24. Huppert FA, Cabelli ST, Matthews FE, MRC CFAS. Brief cognitive assessment in a UK population sample – distributional properties and the relationship between the MMSE and an extended mental state examination. BMC Geriatrics. 2005; 5:1–14. [PubMed: 15627403]
25. Moylan T, Das K, Gibb A, Hill A, Kane A, Lee C. Assessment of cognitive function in older hospital inpatients: is the Telephone Interview for Cognitive Status (TICS-M) a useful alternative to the Mini Mental State Examination? International Journal of Geriatric Psychiatry. 2004; 19:1008–1009. [PubMed: 15449371]
26. Mungas D. In-office mental status testing: a practical guide. Geriatrics. 1991; 46(7):54–58. [PubMed: 2060803]
27. Kilada S, Gamaldo A, Grant EA, Moghekar A, Morris JC, O'Brien RJ. Brief screening tests for the diagnosis of dementia: comparison with the mini-mental state exam. Alzheimer's Disease & Associated Disorders. 2005; 19:8–16.
28. Xu G, Meyer JS, Thornby J, Chowdhury M, Quach M. Screening for mild cognitive impairment (MCI) utilizing combined mini-mental-cognitive capacity examinations for identifying dementia prodrome. International Journal of Geriatric Psychiatry. 2002; 17:1027–1033. [PubMed: 12404652]
29. Tang-Wai DF, Knopman DS, Geda YE. Comparison of the short test of mental status and the mini-mental state examination in mild cognitive impairment. Archives of Neurology. 2003; 60:1777–1781. [PubMed: 14676056]
30. Jöreskog KG, Moustaki I. Factor analysis of ordinal variables: A comparison of three approaches. Multivariate Behavioral Research. 2001; 36(3):347–387.
31. McHorney CA. Ten recommendations for advancing patient-centered outcomes measurement for older persons. Annals of Internal Medicine. 2003; 139:403–409. [PubMed: 12965966]
32. Mislevy RJ. Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics. 1986; 11(1):3–31.
33. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph. 1969; 17
34. Green BF, Bock RD, Humphreys LG, Linn RL, Reckase MD. Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement. 1984; 21:347–360.
35. Embretson, SE.; Reise, SP. Item response theory for psychologists. Mahwah, NJ: Erlbaum; 2000.
36. Hambleton, RK.; Swaminathan, H.; Rogers, HJ. Fundamentals of item response theory. Newbury Park, CA: Sage; 1991.
37. Kotsiantis S, Kanellopoulos D. Discretization Techniques: A recent survey, GESTS. International Transactions on Computer Science and Engineering. 2006; 32(1):47–58.
38. Heywood HB. On finite sequences of real numbers. Proceedings of the Royal Society of London Series A, Containing Papers of a Mathematical and Physical Character. 1931; 134:486–501.

39. Muraki E, Engelhard G. Full-Information Item Factor Analysis: Applications of EAP Scores. Applied Psychological Measurement. 1985; 9:417.
40. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 327:307–310. [PubMed: 2868172]
41. Coffin M, Sukhatme S. Receiver Operating Characteristic Studies and Measurement Errors. Biometrics. 1997; 53:823–837. [PubMed: 9333348]
42. Kolen, M.; Brennan, R. Test equating: Methods and practices. New York: Springer; 1995.
43. Gross AL, Inouye SK, Rebok GW, Brandt J, Crane PK, Parisi JM, et al. Parallel But Not Equivalent: Challenges and Solutions for Repeated Assessment of Cognition over Time. Journal of Clinical and Experimental Neuropsychology. 2012; 34:758–772. [PubMed: 22540849]
44. Wang WC. Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. Journal of Experimental Education. 2004; 72(3):221–261.
45. Jones RN, Fonda SJ. Use of an IRT-based latent variable model to link different forms of the CES-D from the Health and Retirement Study. Social Psychiatry and Psychiatric Epidemiology. 2004; 39:828–835. [PubMed: 15669664]
46. Curran PJ, Hussong AM. Integrative data analysis: The simultaneous analysis of multiple data sets. Psychological Methods. 2009; 14:81–100. [PubMed: 19485623]
47. Hendrie HC, Albert MS, Butters MA, Gao S, Knopman DS, Launer LJ, Wagster MV. The NIH Cognitive and Emotional Health Project: Report of the Critical Evaluation Study Committee. Alzheimers Dement. 2006; 2:12–32. [PubMed: 19595852]
48. Reitan R. Validity of the trail making test as an indicator of organic brain damage. Perceptual and Motor Skills. 1958; 8:271–276.
49. Wechsler, D. Wechsler Memory Scale-Revised. San Antonio, Texas: Psychological Corporation; 1987.
50. Benton, A.; Hamsher, K. Multilingual Aphasia Examination. Iowa City, IA: University of Iowa; 1976.
51. Williams BW, Mack W, Henderson VW. Boston naming test in Alzheimer's disease. Neuropsychologia. 1989; 27(8):1073–1079. [PubMed: 2797414]
52. Smith, A. Symbol Digits Modalities Test. Los Angeles: Western Psychological Services; 1973.
53. Morris JC, Heyman A, Mohs RC, Hughes JP, van Belle G, Fillenbaum G, Mellits ED, Clark C, the CERAD investigators. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. Neurology. 1989; 39:1159–1165. [PubMed: 2771064]
54. Brandt, J.; Benedict, RHB. Hopkins Verbal Learning Test–Revised: Professional manual. Odessa, FL: Psychological Assessment Resources; 2001.
55. Trenerry, MR.; Crosson, B.; DeBoe, J.; Leber, WR. Visual Search and Attention Test (VSAT). Odessa, FL: Psychological Assessment Resources, Inc; 1990.

Abbreviations

SAGES	Successful AGing after Elective Surgery
ADAMS	Aging, Demographics, and Memory Study
CIND	Cognitive impairment without dementia
MMSE	Mini-Mental State Examination
AUC	area under the curve

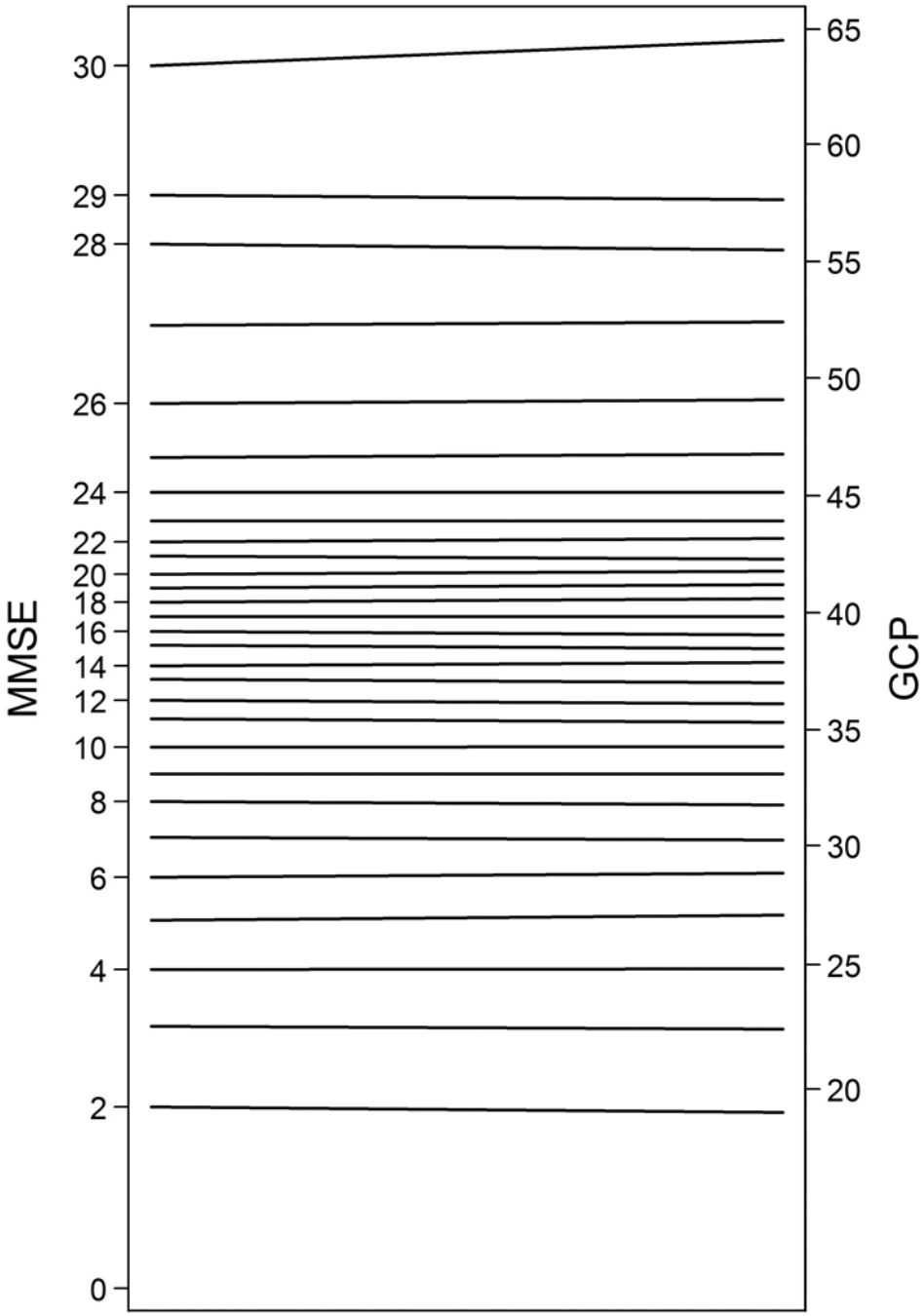

Figure 1. Distribution of General Cognitive Performance Score in SAGES and ADAMS

ADAMS: Aging, Demographics, and Memory Study; SAGES: Successful AGing after Elective Surgery

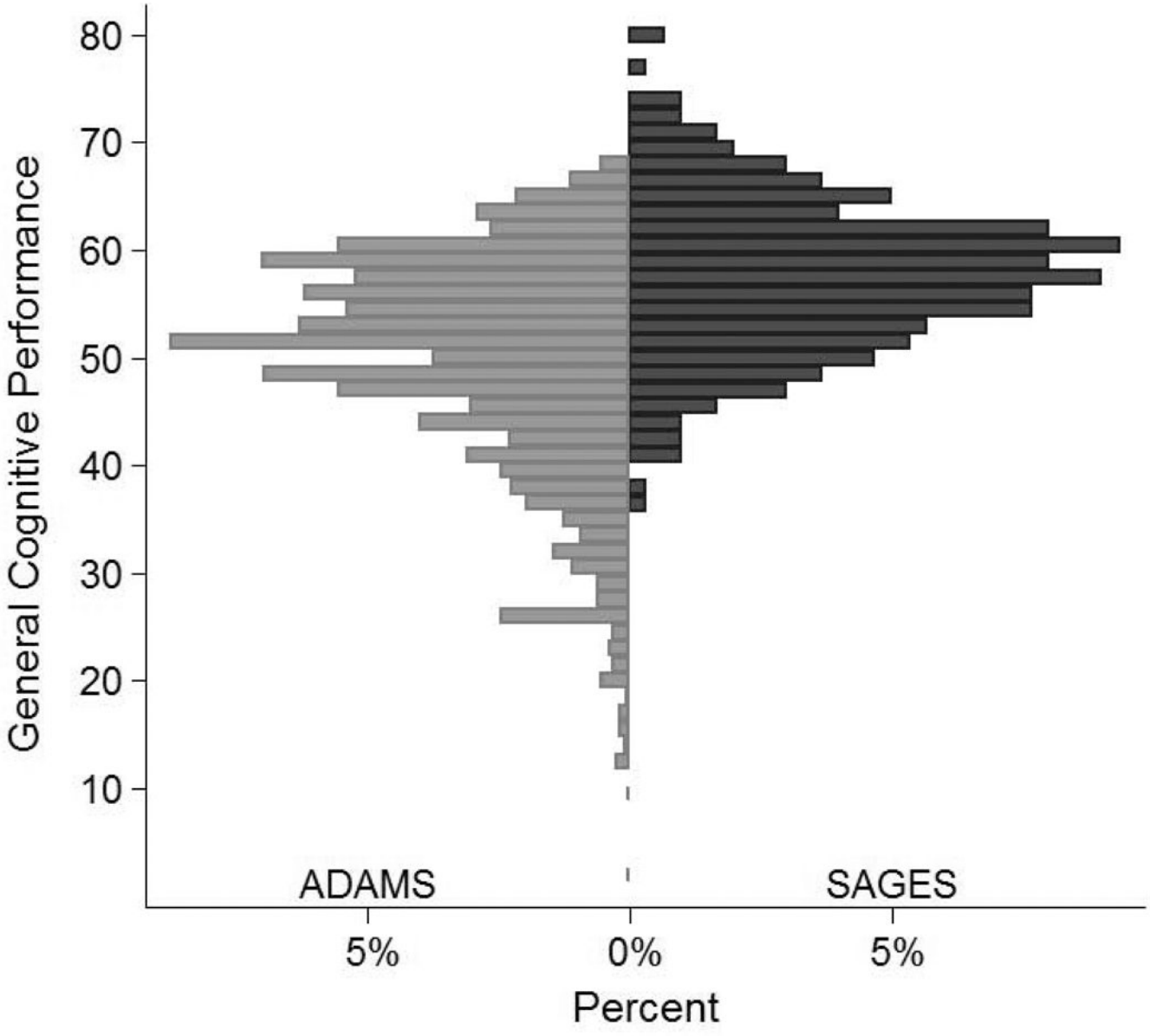

Figure 2. Receiver Operator Characteristic Curves for General Cognitive Performance Predicting CIND and Dementia: Results from ADAMS (n=856)

Legend. In right panel, the general cognitive performance score that best discriminated dementia participants from other participants was less than 44.8 (sensitivity = 0.94; specificity = 0.90; right panel). This cut-point correctly classified 93.8% of the sample. The area under the curve (AUC) was 0.97. In the left panel, the general cognitive performance score that best discriminated CIND participants from cognitively normal participants was less than 49.5 (sensitivity = 0.80, specificity = 0.76; Figure 1, left panel). This cut-point correctly classified 78.5% of the sample. The AUC was 0.84.

CIND: Cognitive impairment without dementia.

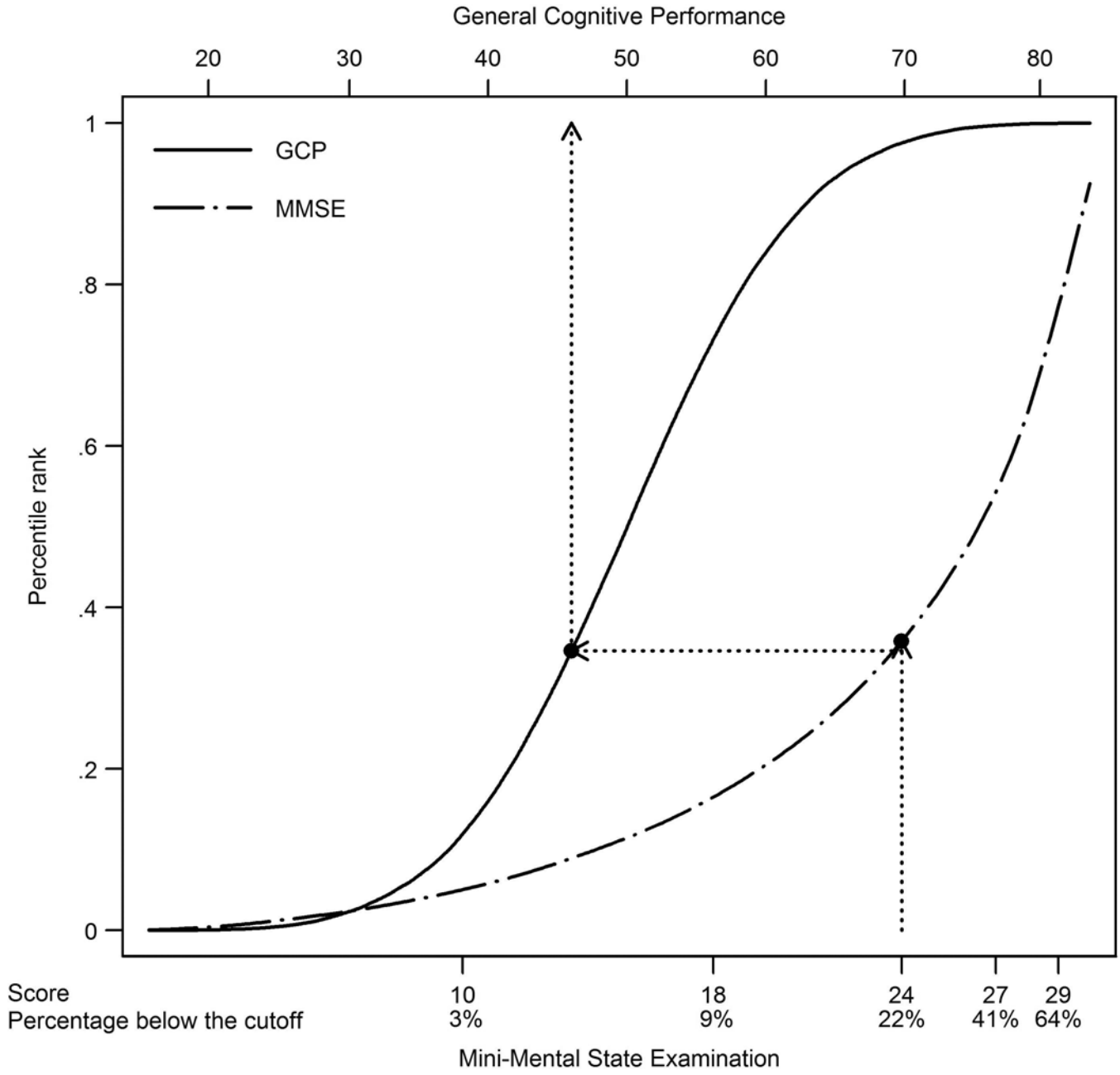

Figure 3. Corresponding Scores and Percentile Ranks for the General Cognitive Performance Factor and MMSE

Legend. General cognitive performance scores are shown on the top axis. MMSE scores with the proportion of participants in ADAMS falling below each score are shown on the bottom x axis. Dotted lines show that approximately 25% of participants have below a 45 on the general cognitive performance factor and below 24 on the MMSE.

MMSE: Mini-Mental State Examination.

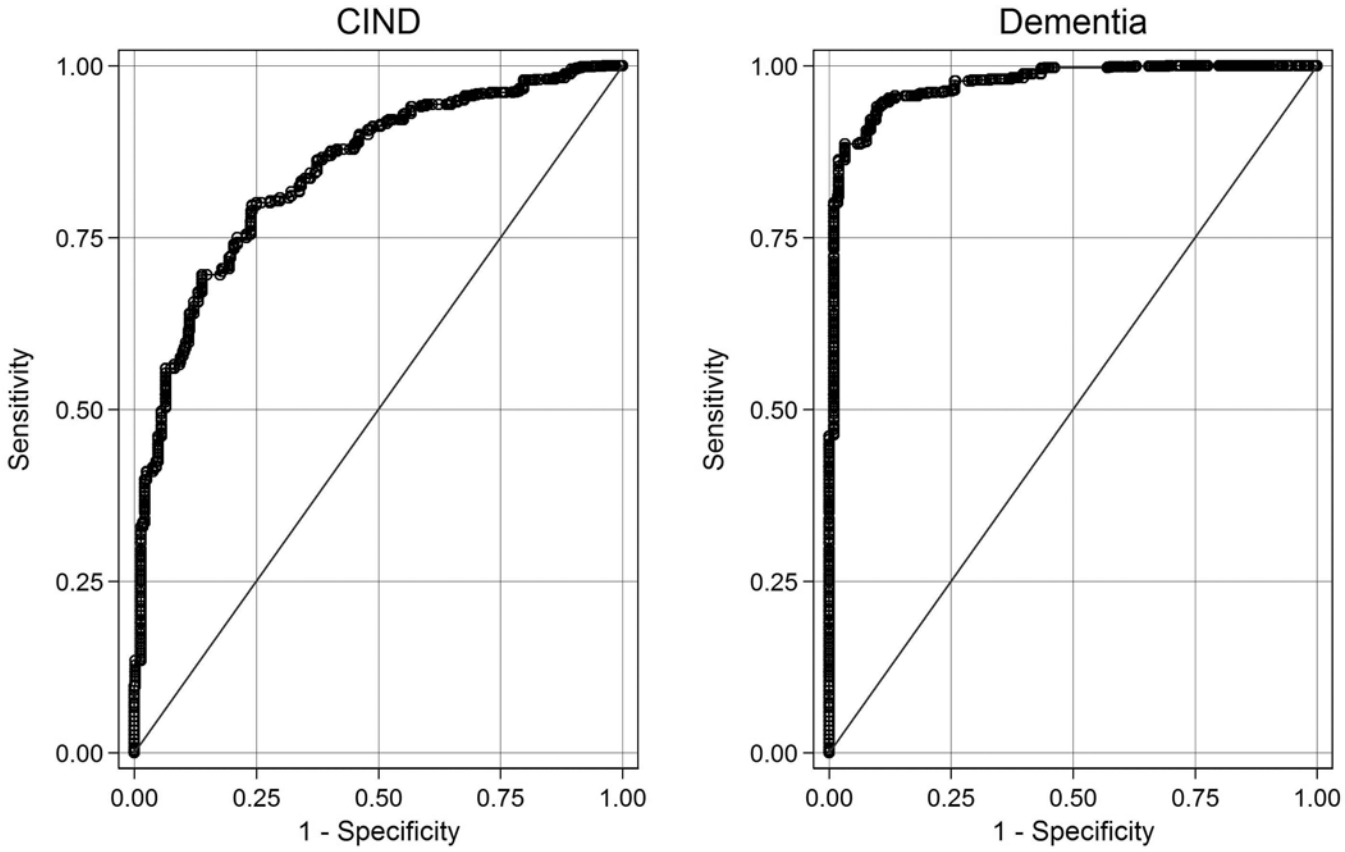

Figure 4. Crosswalk Between MMSE and General Cognitive Performance Factor: Results from ADAMS (n=856)

Legend. Crosswalk of scores on the MMSE to the general cognitive performance factor. The left side of the crosswalk shows MMSE scores (range 0 to 30) and the right side shows corresponding general cognitive performance scores. To facilitate comparison of the distribution of general cognitive performance scores against MMSE scores, y-axes are plotted on an inverse normalized percentile scale. Original units are labeled. We used data from the Monte Carlo simulation because the distribution of the observed data did not permit a finely graded linkage between the MMSE and general cognitive performance factor.

MMSE: Mini-Mental State Examination; GCP: General Cognitive Performance

Table 1
Neuropsychological Test Batteries in the SAGES and ADAMS Samples

	SAGES (n=300) Mean ± SD	ADAMS (n=856*) Mean ± SD	Cognitive domain	Test description or effect size for group difference†
Demographic				
Age, *mean ± SD*	76.9 ± 5.1	79.0 ± 6.0		0.35
Age range, *years*	70, 92	70, 110		
Race, White, *n (%)*	286 (95.0)	658 (89.3)		0.02
Years of education, *n (%)*				0.96
High school or less	91 (30.3)	638 (62.9)		
College	124 (41.3)	167 (28.7)		
Graduate	86 (28.4)	51 (8.4)		
Sex, female, *n (%)*	167 (55.6)	501 (60.6)		0.2
Cognitive status, *n (%)*				
Cognitively normal	276 (92.0)	307 (50.6)		
Cognitive impairment without dementia	19 (6.3)	241 (35.7)		
Alzheimer s disease	5 (1.7)	308 (13.7)		
Neuropsychological test variables				
Trails A (Time to complete, seconds) (48)	42.1 ± 15.1	65.0 ± 45.3	Processing speed	Connect a series of numbers
Trails B (Time to complete, seconds) (48)	183.7 ± 57.2	154.6 ± 72.7	Processing speed, task-switching	Connect an alternating series of letters and numbers
Digit Span Forwards (Total number of digits) (49)	9.9 ± 2.2	8.6 ± 2.1	Attention	Repeat a series of pre-specified random digits forwards
Digit Span Backwards (Total number of digits) (49)	6.3 ± 2.1	5.4 ± 2.2	Attention	Repeat a series of pre-specified random digits backwards
Semantic Fluency (Number of items) (50)	21.4 ± 6.0	14.5 ± 5.3	Language, executive function	Name as many items as possible in one minute from a preselected category
Phonemic Fluency (Number of items) (50)	34.7 ± 12.5	29.3 ± 11.8	Language, executive function	Name as many items as possible in one minute that begin with letters F, A, or S (3 trials).
Boston Naming Test - 15 items (Number of items) (51)	13.3 ± 2.1	13.3 ± 2.2	Language	Name objects in a series of drawings
Symbol Digit Modalities (Total number of correct numbers) (52)	--	30.3 ± 11.3	Attention, executive function	Match numbers to a series of symbols
Digit Symbol Substitution (Total number of correct symbols) (49)	36.0 ± 10.2	--	Attention, executive function	Match symbols to a series of numbers
CERAD word recall – immediate (Number of words) (53)	--	16.0 ± 5.8	Episodic memory	Sum of 3 trials of recalled words from a 10-word list of unrelated nouns
HVLT sum of recall – immediate (Number of words) (54)	21.3 ± 5.1	--	Episodic memory	Sum of 3 trials of recalled words from a 12-word list of 4 groups of 3 semantically related nouns

	SAGES (n=300) Mean ± SD	ADAMS (n=856*) Mean ± SD	Cognitive domain	Test description or effect size for group difference†
CERAD word recall – delayed (Number of words recalled) (53)	--	0.7 ± 0.4	Episodic long-term memory	Recalled words from a delayed recall trial
HVLT sum of recall – delayed (Number of words) (54)	0.8 ± 0.3	--	Episodic long-term memory	Recalled words from a delayed recall trial
Visual Search and Attention (Number of targets) (55)	43.0 ± 9.5	--	Visuospatial function	Search for a target letter amidst other letters

*In ADAMS, means and percentages calculated using population weights to account for complex sampling. Raw numbers are shown for race, education, sex, and cognitive status.

†Cohen's d for mean differences and Cohen's h for proportions.

SD: Standard deviation; WMS: Wechsler Memory Scale; HVLT: Hopkins Verbal Learning Test; SD: standard deviation; CERAD: Consortium to Establish a Registry for Alzheimer's Disease; MMSE: Mini Mental State Examination; SAGES: Successful Aging after Elective Surgery; ADAMS: Aging, Demographics, and Memory Study.

Missing data. In SAGES, 0.3% of the sample was missing Trails B, Digit Symbol Substitution, and Boston Naming; there was no missingness in any other cognitive variables. In ADAMS, 10% were missing Trails A, 19% were missing Trails B, 6% were missing Digit Span Forward, 7% were missing Digit Span Backward, 3% were missing Semantic Fluency, 2% were missing Boston Naming, 16% were missing Symbol Digit Modalities, 3% were missing CERAD word recall, 4% were missing delayed CERAD word recall, and 8% were missing Phonemic Fluency.

Table 2

Clinically Important Cut-points on the MMSE and Corresponding General Cognitive Performance Scores

MMSE score	General Cognitive Performance cut-point (95% interval)	Participants below cut-point in ADAMS, n (%) *	Reference standard diagnosis	Test performance characteristics of the general cognitive performance factor†					
				Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)	Likelihood ratio positive	Likelihood ratio negative
< 10	34.0 (33.8, 34.2)	79 (3.3)	Dementia	38.5	99.9	98.7	88.4	367.0	0.6
			CIND	7.2	99.9	95.9	75.9	68.9	0.9
< 18	40.7 (40.6, 40.7)	252 (9.4)	Dementia	74.9	96.3	81.3	94.7	20.5	0.3
			CIND	21.3	96.3	66.7	78.1	5.8	0.8
< 24	**45.2 (44.9, 45.5)**	**449 (22.4)**	**Dementia**	**90.4**	**92.9**	**73.1**	**97.9**	**12.8**	**0.1**
			CIND	44.4	92.9	68.3	83.0	6.3	0.6
< 27	50.9 (50.1, 51.5)	598 (41.1)	Dementia	98.9	75.0	45.7	99.7	4.0	0.0
			CIND	77.0	75.0	51.3	90.5	3.1	0.3
< 29	59.1 (57.7, 60.6)	717 (64.0)	Dementia	100.0	27.5	22.7	100.0	1.4	0.0
			CIND	98.7	27.5	31.8	98.4	1.4	0.0

Legend. General cognitive performance scores corresponding to clinical cut-points on the MMSE in ADAMS. 95% intervals are the 2.5th and 97.5th percentiles of general cognitive performance scores for the corresponding equipercentile- linked MMSE score.

CIND: Cognitive impairment without dementia.

*Unweighted sample sizes (weighted percentages) are shown.

†Validated against reference standard clinical diagnoses of either dementia or CIND