Article

Effects of Repeated Administration and Comparability of Alternate Forms for the Global Neuropsychological Assessment (GNA)

Assessment
2023, Vol. 30(1) 160–170

DOI: 10.1177/10731911211045125
journals.sagepub.com/home/asm
SAGE

Alan Smerbeck[1], Lauren T Olson[2], Lindsay F. Morra[3], Jeremy Raines[4], David J. Schretlen[3], and Ralph H. B. Benedict[2]

Abstract

The Global Neuropsychological Assessment (GNA) is an extremely brief battery of cognitive tasks assessing episodic memory, processing speed, working memory, verbal fluency, executive function, and mood. It can be given in under 15 minutes, has five alternate forms, and does not require an examinee to be literate. The purpose of this study was to quantify practice effects over repeated administrations and assess comparability of the GNA's five alternate forms, preparing the battery for repeated administration in research and clinical settings. Forty participants each completed all five GNA forms at weekly intervals following a Latin square design (i.e., each form was administered at every position in the sequence an equal number of times). In a cognitively intact population, practice effects of 0.56 to 1.06 *SD* were observed across GNA measures when comparing the first and fifth administration. Most GNA tests showed nonsignificant interform differences with cross-form means differing by 0.35 *SD* or less, with the exception of modest but statistically significant interform differences for the GNA Story Memory subtest across all five forms. However, post hoc analysis identified clusters of two and three Story Memory alternate forms that were equivalent.

Keywords

practice effects, alternate forms, assessment, psychometrics, reliability

The Global Neuropsychological Assessment (GNA) is a novel battery designed for use in diverse settings to expand access to cognitive testing in both developed and developing nations where personnel or resources required for more comprehensive neuropsychological testing are limited (Olson et al., 2020; Raines et al., 2019). The GNA takes approximately 15 minutes to administer and does not require specialized equipment, making it more suitable for use in situations in which a full neuropsychological evaluation is unavailable. It was designed to be applicable even to examinees with very low education—it does not require literacy, nor the use of a writing implement—allowing it to be administered to an examinee population that may otherwise be unable to complete other neuropsychological tests.

Lower and middle-income countries rarely have sufficient numbers of highly trained clinicians or test materials to provide cognitive assessments for diagnostic, treatment monitoring, or research purposes. In addition, while some screening tests, such as the Mini-Mental State Exam (MMSE; Folstein et al., 1975) and Montreal Cognitive Assessment (MoCA; Nasreddine et al., 2005), have been translated and normed in many countries, the item composition can vary due to cultural or linguistic differences and require harmonization to compare results across countries (Gross et al., 2019). Culture-specific item contents limit the suitability of most widely used cognitive tests for translation and cross-cultural application without the development of international norms. However, even tests whose items minimize culture-specific contents cannot escape the influence of linguistic differences, such as the number of syllables that comprise a string of to-be-repeated digits or words (Jalbert et al., 2011), so the English word "cat" is easier to repeat than its Spanish translation, "gato." If the English list were simply translated to Spanish with no adjustment in the analysis of scores, results would not be comparable, limiting the use of the test in cross-national studies. One approach to this deficiency in the available range of cognitive tests

[1]Rochester Institute of Technology, Rochester, NY, USA
[2]State University of New York at Buffalo, Buffalo, NY, USA
[3]The Johns Hopkins University School of Medicine, Baltimore, MD, USA
[4]University of North Dakota, Grand Forks, ND, USA

Corresponding Author:
Alan Smerbeck, Department of Psychology, Rochester Institute of Technology, 1 Lomb Memorial Drive, Rochester, NY 14623, USA.
Email: amsgss@rit.edu

was pioneered for people with multiple sclerosis (MS) in the Brief International Cognitive Assessment for MS, a set of three tests most sensitive to the cognitive deficits of MS which could be administered by a nonspecialist (Benedict, Amato, et al., 2012; Langdon et al., 2012). Statistical analysis was used to create demographically adjusted norms for each country internally as well as international norms which account for variations due to language differences and other nation-level variables (Smerbeck et al., 2018). This widened the availability of neuropsychological assessment in clinical practice, allowed clinicians to asses populations that had previously lacked reliable norm-referenced assessment, and established a valid means of equating raw scores across national and linguistic boundaries. The GNA was designed to achieve similar aims, but rather than focusing solely on the cognitive deficits found in MS, it was created to assess a broader range of neuropsychological variables. Specifically, the GNA was designed to assess well-defined cognitive functions using tasks that are familiar to most neuropsychologists but based on test stimuli that minimize culture-specific contents for the purpose of translating and norming it in as many languages and countries as possible.

The GNA assesses episodic memory, processing speed, working memory, semantic fluency, executive functioning (set shifting), and mood. These domains overlap considerably with those most frequently tested by established neuropsychological assessment batteries (e.g., Green & Nuechterlein, 2004; Gualtieri & Johnson, 2006; Langdon et al., 2012; Randolph, 1998). Like the MMSE (Folstein et al., 1975) and MoCA (Nasreddine et al., 2005), the GNA is brief, requiring only 15 minutes to administer. However, the MMSE and MoCA were designed for use as screening instruments and were intended to be interpreted as single composites, whereas the GNA subtests are independently interpretable, allowing the battery to provide a level of information between that offered by screening and comprehensive evaluation.

The GNA offers five alternate forms which can be used when repeated testing is required. Note that although the two terms are commonly used interchangeably, *parallel forms* have near identical psychometric properties (a standard that is rarely met), while *alternate forms* are very similar but have modest differences that may require the use of slightly modified norms (Gulliksen, 1950). Repeated testing is often desirable to document decline or recovery, or to assess the efficacy of an intervention, but re-administering neuropsychological tests frequently inflates scores due to practice effects (Bartels et al., 2010), thus obscuring or exaggerating real effects. This makes it very helpful to know the magnitude of any expected practice effects, the degree to which they can be minimized by using alternate forms, and the equivalence of alternate forms.

Practice effects can be caused by the examinee's memory of the exact test stimuli or rehearsal of a specific item response. Use of alternate forms can diminish practice effects due to recall of test stimuli (Calamia et al., 2012). However, practice effects may also represent improved performance due to general familiarity with the test format or procedure—knowing, for example, that there will be an unannounced delayed recall phase—which alternate forms cannot overcome. Such practice effects should be documented before a neuropsychological test is widely used to minimize or control for the effects of repeated testing on test performance.

A meta-analysis of practice effects in neuropsychological testing found that the use of alternate forms can eliminate or reduce practice effects to varying degrees depending on the type of test (Calamia et al., 2012; Gavett et al., 2016). The use of alternate forms greatly diminishes practice effects on verbal memory subtests, but has far less impact on tests of processing speed and executive function (Calamia et al., 2012; Rijnen et al., 2018). Retest gains are not invariant across examinees. Clinical samples show smaller retest gains than healthy controls, emphasizing that maximal learning under ideal conditions is unlikely to be replicated by impaired samples (Jutten et al., 2020; Rijnen et al., 2018). Practice effects can mask decline if not accounted for by clinicians (Elman et al., 2018; Kremen et al., 2020).

The development of alternate forms requires attention to all factors that can affect item difficulty or alter the construct being measured. For example, letter stimuli for a phonemic fluency task must be selected with respect to the relative availability of easily retrievable phonemic clusters (Ross et al., 2006). Alternate forms of a simple processing speed task have been found to be more similar when they control for the fact that examinees who read left-to-right process visual stimuli toward the left more quickly than stimuli toward the right (Benedict, Smerbeck, et al., 2012). The GNA's Perceptual Comparison (PC) task was designed with this confound in mind and does not disproportionately reward right-to-left or left-to-right scanning. Additionally, the symbols were selected to minimize overlap with meaningful writing units (e.g., letters, logograms, etc.).

Verbal memory stimuli, such as meaningful sentences, are especially difficult to equate. Numerous variables affect the difficulty of memory for words, including differences in length (Jalbert et al., 2011), frequency (Miller & Roodenrys, 2009), predictability (Staub et al., 2015; Valian et al., 2006), concreteness (Dye et al., 2013; Roche et al., 2011; Romani et al., 2008), emotional valence (Sutton & Lutz, 2019), and the number of phonologically similar words in the language (Allen & Hulme, 2006). When words are combined into meaningful sentences, difficulty is further influenced by grammatical complexity (Perkins et al., 1986) and conformity to cultural schemata (Harris et al., 1992). Equating all of these features is extraordinarily difficult, and of little use in a test intended

for multinational application because the same semantic content will yield substantially different outcomes across these variables once translated into another language and applied in a different setting. As such, the GNA Story Memory (SM) stimuli were developed based on more straightforward criteria. While no topic is completely universal, the content chosen was likely to be familiar to most examinees. As such, all stories fall at or below the fourth-grade reading level. All target words are content, not function, words and are common nouns, verbs, adjectives, or adverbs. All stories contain 14 target words embedded in a story of 29 to 31 words. All stories are similar in vocabulary and morphological complexity, containing 1.19 to 1.31 syllables per word and 1.17 to 1.28 morphemes per word.

Given the importance of repeat assessment in both clinical and research settings, it is vital to understand how GNA performance is affected when the same individuals complete the test multiple times across relatively brief intervals. This study seeks to establish the extent of practice effects for the GNA battery, using alternate forms as available, and to assess the equivalency of the alternate forms.

Method

Below, we report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Participants

Participants were recruited from a population of university undergraduates, employed community members, and older adults attending a continuing education program. Exclusionary criteria included sensorimotor impairment which precluded unaccommodated testing (other than eyeglasses), primary language other than English, history of language, learning, developmental, or cognitive disorder, medical problem likely to affect cognition, and any history of neurological disorder, traumatic brain injury, seizure, or loss of consciousness for greater than 5 minutes. As the focus of the study was on intraindividual comparison over time, criteria ruled out participants' whose health or cognitive status might change over the course of the study. Thus, participants were excluded for any current mental health disorder, any uncontrolled or unstable medical condition, current use of cognition-affecting medication or drugs, or any change of medication regimen less than 6 weeks prior to the beginning of testing.

Initially, 47 adults enrolled in the study. Two dropped out before completing all of five planned testing sessions, and the test results of one participant were spoiled by examiner error. This left 44 validly completing participants. The research design requires five equal groups, so 40 were retained for analysis. Three cases were selected for exclusion based on minor disruption to test protocol (e.g., noise interruption); the fourth was selected at random.

The counterbalanced set of 40 participants included 18 men and 22 women who ranged from 18 to 75 ($M = 30.6$, $SD = 20.8$) years of age, although most (72.5%) were 18 to 20 years. Their years of education ranged from 12 to 18 ($M = 13.5$, $SD = 1.87$) years. However, most participants with 12 or 13 years of education were current undergraduates who intended to earn (at a minimum) bachelor's degrees. The sample's racial and ethnic makeup was 82.5% Caucasian and 12.5% Asian/Asian-American. One participant (2.5%) identified solely as Hispanic, and one (2.5%) self-identified as a person of Middle Eastern descent. All participants were tested in English, which they described as their dominant language, and 4 (10%) reported being bilingual. 38 participants (95%) reported being right-handed, and 2 reported being left-handed. 19 participants reported using no drugs or medications of any kind during the study and for at least 6 weeks prior. 10 reported drinking alcohol socially, and 26 reported using medications including statins, oral contraceptives, antihypertensives, antihistamines, antiinflammatories, proton pump inhibitors, levothyroxine, and an antihyperuricemic. One participant was taking a selective serotonin reuptake inhibitor. All drug use was stable for at least 6 weeks prior to beginning the study and did not change over the course of the study. Participants were compensated US$50 for their time.

Measures

Administration of the GNA begins with *Story Memory (SM)*. This subtest consists of a brief story that is read aloud to the examinee twice. After each reading, the examinee is asked to repeat the story using identical words, without synonyms or grammatical variations. Each version of SM includes 14 target (to-be-remembered) words. One point is awarded for each target word that the examinee repeats exactly, regardless of context. Thus, scores for each learning trial can range from 0 to 14 points. These are summed as an SM Total Learning score that can range from 0 to 28 points. The examiner next administers and Digit Span (DS) as delay tasks that are not expected to compete with meaningful verbal material. Thereafter, the examinee is asked to repeat the SM narrative again without hearing it a third time. Scores for this SM Delayed Recall trial can range from 0 to 14. Finally, an SM Percent Retained score is computed by dividing the Delayed Recall score by the higher of the two learning trials, then multiplying the quotient by 100. The five forms present different stories with different target words, but are consistent in length and available points.

The *PC* subtest measures processing speed. It requires the examinee to look at 54 sets of paired shapes (e.g., ▷ ✧ — ➪ ➪) and state orally whether the pairs are the same or

different. Following three practice items, the examinee is asked to complete as many items as possible in 45 seconds. This task yields a single score, the number of correct responses within the time limit (PC Total Correct). Alternate forms of PC use the same stimuli, arranged differently.

In *DS*, the examiner reads digit strings aloud at the rate of one digit per second. The examinee must first repeat strings of 1 to 9 digits exactly as presented (DS Forward) and then repeat 2 to 8 digits in reverse order (DS Backward). One trial of each string length is presented, and, to streamline administration, the test is discontinued after two consecutive failures. The longest strings correctly repeated forward and backward are summed as the DS Max Total yielding scores ranging from 0 to 17. Alternate forms of DS use different number strings.

The GNA *Animal Naming* subtest asks examinees to name as many different animals as possible in 1 minute. Variations of the same animal (e.g., *puppy* and *dog*, *cow* and *bull*), type and subtypes (e.g., *fish* and *salmon*), and mythical animals (e.g., *dragon*) are scored as correct. However, the addition of an adjective not forming a species, breed, or type is not sufficiently distinct (e.g., *dog* and *big dog* would earn one point). Repetitions and nonanimal intrusion errors are recorded but not credited. The total number of correct words is recorded. Animal Naming is administered the same way across all five forms of the GNA.

For *Category Switching* (CS), the examinee is asked to alternate between saying body parts and foods/drinks (e.g., arm—mushroom—nose—Pepsi . . .) as quickly as possible (Delis et al., 2001). Again, scoring is liberal—examinees can be credited for a whole and its parts (e.g., *arm*, *elbow*, *wrist* or *pizza*, *cheese*, *tomato sauce*); and similar variations (e.g., *left eye*, *right eye* or *birthday cake*, *coffee cake*, *cookie cake*); but not for the addition of a descriptor which fails to describe a specific food or body part (e.g., *hair*, *long hair* would only be worth one point). This subtest yields two scores: total correct switches (CS Switches) and total correct words (CS Words). Repetitions and intrusion errors (nonfood or nonbody part words) are scored as incorrect, but switches can be scored as correct even if a body part or food is repeated. The CS subtest is administered the same way across all five forms of the GNA.

Administration of the GNA concludes with the *Patient Health Questionnaire–4* (PHQ-4; Löwe et al., 2010), which screens for syndromal depression and generalized anxiety by asking the examinee to indicate on a Likert-type scale from 0 to 3 how frequently each symptom was experienced during the preceding 2 weeks (not at all, several days, more than half the days, or nearly every day). The PHQ-4 includes two symptoms of anxiety (i.e., feeling nervous, anxious, or on edge and not being able to stop or control worrying) and two symptoms of depression (i.e., little interest or pleasure in doing things and feeling down, depressed, or hopeless). The PHQ-4 is sufficiently reliable for screening ($\alpha = .78$; Löwe et al., 2010). The PHQ-4 items were unchanged across the experiment's five sessions.

Procedures and Analysis

Tests were administered by a licensed psychologist or trained graduate assistant. A detailed administration checklist was developed, and examiners were required to demonstrate compliance with all elements before they could conduct study assessments. Note that this checklist merely provided a highly detailed task analysis of common neuropsychological test administration procedures such as reading instructions verbatim, refraining from providing performance feedback, and discontinuing once a ceiling was achieved. Illustrating the GNA's applicability to less trained examiners, the graduate students were able to meet this standard following 2 to 5 hours of preparation despite lacking prior training in psychological testing. Follow-up examiner observations and record-form checks were performed by the lead investigator to ensure test fidelity. Only one score-affecting administrative error was detected; that examinee was removed from analysis.

Five administration orders were generated according to a Latin square design such that every form was equally represented at each time point. The target interval between visits was one week and this was achieved in 74% of cases, although actual delays ranged from 3 to 14 ($M = 7.7$, $SD = 2.0$) days. In completing a Latin square design with forty participants across five conditions, there were a total of 160 interform delay intervals. Of these, the overwhelming majority (118 intervals) were precisely 7 days. However, to accommodate participant schedules, some deviation occurred. The next most frequent interval was 8 days, with 10 cases. All other intervals occurred less frequently, with a 3-day interval occurring only once (0.6% of intervals). Bivariate correlations were calculated between the length of each participant's delay interval and their performance on the tests following the delay—that is, for example, the interval between Times 1 and 2 correlated with performance at Time 2. This was conducted for all major outcome measures across each of the four delay intervals. All correlations were found to be nonsignificant at an alpha level of .05. Thus, although more rigid constraints on delay intervals would have been preferable, in this case there seems to have been no effect on performance. Of participants who completed the GNA only, first session completion times ranged from 8 to 13 minutes ($M = 10.2$, $SD = 1.5$). At each visit, participants were asked about changes in their health and drug use. None were reported, with the exception of a single participant who used an asthma inhaler between Visits 2 and 3.

All analyses were conducted using SPSS 26.0. The Latin square design ensures that each form is equally represented at each time point so that when data are

Table 1. Raw Scores Obtained by Repeat Administration of the Global Neuropsychological Assessment, all Forms Equally Represented in Each Session.

Subtest	Session 1	Session 2	Session 3	Session 4	Session 5	*F*	*p*	Partial η^2	Cohen's *d*
SM Total Learning	19.50 (3.98)[a]	21.10 (3.53)	20.98 (3.75)	21.20 (3.38)	22.60 (3.14)[a]	5.36	<.001	.12	0.87
SM Delayed Recall	10.20 (2.46)	10.77 (1.99)	10.90 (1.92)	11.03 (1.91)	11.37 (2.11)	2.32	.059	.06	0.56
SM Percent Retained	88.78 (16.77)	92.49 (13.48)	91.83 (12.28)	93.70 (11.47)	91.17 (10.98)	0.93	.448	.02	0.18
PC Total Correct	42.63 (7.59)[a,b,c]	45.00 (7.19)	46.40 (6.90)[a]	47.28 (7.40)[b]	47.27 (6.83)[c]	19.56	<.001	.33	0.65
DS Max Total	11.50 (2.18)	12.03 (2.43)	12.45 (2.58)	12.63 (2.27)	13.03 (2.33)	5.55	<.001	.13	0.65
Animal Naming	25.73 (5.12)	27.35 (6.15)	28.37 (5.87)	28.28 (4.63)	29.25 (4.80)	6.24	<.001	.14	0.66
CS Switches	20.62 (4.48)[a,b]	23.00 (5.08)	23.90 (5.44)	24.38 (4.83)[a]	25.90 (5.07)[b]	17.34	<.001	.31	1.06
CS Words	21.80 (4.31)[a,b,c]	24.28 (4.79)	24.93 (5.54)[a]	25.45 (5.04)[b]	26.95 (5.04)[c]	17.89	<.001	.31	1.04
PHQ-4	1.38 (1.51)	0.93 (0.94)	0.55 (0.93)	0.60 (0.98)	0.63 (1.21)	9.56	<.001	.20	−0.66

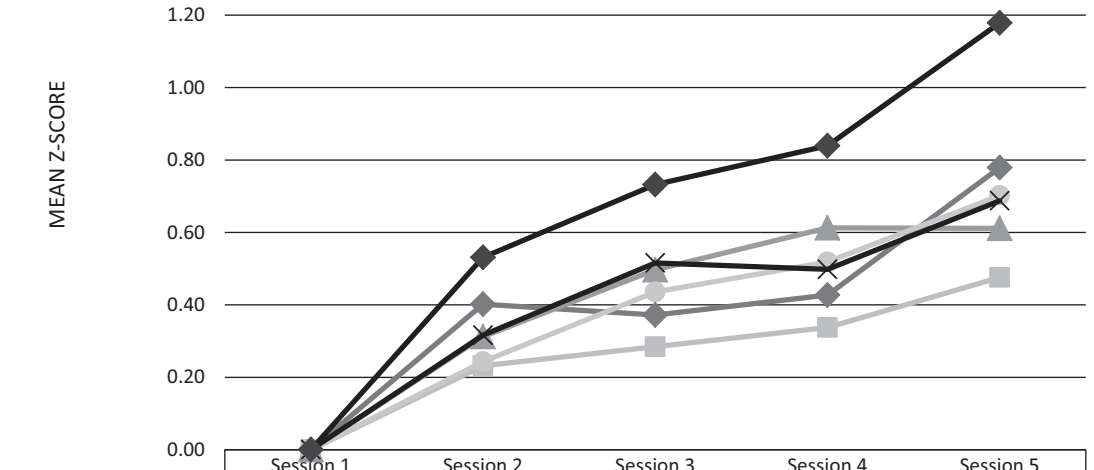
Note. Forms were administered according to a Latin square design, so all forms were equally weighted at each session. Uncorrected two-tailed *p* values are shown. Controlling for multiple comparisons, $p < .006$ is significant for an omnibus *F* test at tablewise α of .05. Cohen's *d* is based the difference between Session 5 and Session 1 means divided by the pooled *SD* for all sessions. Statistically significant post hoc contrasts are identified by assigning the same superscript letter to each statistically significantly different pair. Given the large number of post hoc tests performed, only contrasts with $p < .001$ are highlighted. The GNA Spatial Span subtest was not administered as part of this study. SM = Story Memory; PC = Perceptual Comparison; DS = Digit Span; CS = Category Switching; PHQ-4 = Patient Health Questionnaire–4 Item.

Table 2. Raw Scores for Each Form Obtained by Repeat Administration of the Global Neuropsychological Assessment, Controlling for Order of Administration.

Subtest and measure	Form 1	Form 2	Form 3	Form 4	Form 5	*F*	*p*	Partial η^2	Cohen's *d*
SM Total Learning	21.70 (3.17)	19.18 (3.71)[a]	23.15 (3.36)[a,b]	21.55 (3.15)	19.83 (3.64)[b]	13.16	<.001	.25	1.16
SM Delayed Recall	11.23 (2.07)	10.17 (2.50)[a]	11.93 (1.72)[a,b]	10.67 (1.91)	10.38 (1.85)[b]	7.24	<.001	.16	0.87
SM Percent Retained	92.49 (12.45)	92.42 (16.69)	93.23 (10.51)	87.89 (12.52)	92.58 (12.56)	1.33	.261	.03	0.41
PC Total Correct	45.55 (7.38)	46.15 (7.04)	47.05 (7.21)	45.38 (7.99)	44.45 (7.10)	3.42	.010	.08	0.35
DS Max Total	12.23 (2.78)	12.35 (2.30)	12.75 (2.54)	11.95 (2.07)	12.35 (2.28)	1.22	.304	.03	0.33
Animal Naming	27.30 (5.54)	27.10 (4.72)	28.08 (5.83)	28.05 (5.81)	28.53 (5.40)	1.09	.365	.03	0.26
CS Switches	24.00 (5.81)	23.30 (5.31)	23.85 (5.14)	23.38 (4.84)	23.43 (5.12)	0.32	.867	.01	0.13
CS Words	25.20 (5.50)	24.33 (5.35)	24.85 (5.23)	24.58 (4.96)	24.53 (5.08)	0.40	.809	.01	0.17
PHQ-4	0.83 (1.13)	0.75 (1.01)	0.85 (1.42)	0.95 (1.15)	0.70 (1.14)	0.60	.661	.02	0.21

Note. Forms were administered according to a Latin square design, so each form was administered at each time in the sequence an equal number of times. Uncorrected two-tailed *p* values are shown. Controlling for multiple comparisons, $p < .006$ is significant for an omnibus *F* test at tablewise α of .05. Cohen's *d* is calculated based on the largest interform pair difference divided by the pooled standard deviation across forms. Statistically significant post hoc contrasts are identified by assigning the same superscript letter to each statistically significantly different pair. Given the large number of post hoc tests performed, only contrasts with $p < .001$ are highlighted. The GNA Spatial Span subtest was not administered as part of this study. SM = Story Memory; PC = Perceptual Comparison; DS = Digit Span; CS = Category Switching; PHQ-4 = Patient Health Questionnaire–4 Item.

analyzed for the effects of repeated administration, there is no need to statistically control for form differences; the design controls for form differences. Similarly, when analyzing form effects, the design controls for session effects. Both analyses were performed for all scores listed below. All omnibus tests were conducted using the General Linear Model repeated-measures technique with unadjusted two-tailed *p* values reported. However, because the GNA yields nine scores, unadjusted analysis would lead to an increased risk of false positive findings. As such, within both session and form effect analyses, the alpha level should be divided by nine, the total number of GNA scores analyzed by this article (Tables 1 and 2), making $p < .006$ the standard for significance. If the omnibus test was statistically significant at the $p < .006$ level, post hoc pairwise comparisons were examined. Given the very large number of post hoc comparisons run over the whole analysis, only comparisons statistically significant at the $p < .001$ level were flagged as significant. To communicate omnibus test effect size, partial η^2 values were provided reflecting the percentage of variance attributable to the main effect under investigation (i.e., session or form). As the partial η^2 statistic is used less frequently, the more common Cohen's *d* was provided as well. In analyses of the main effect of time, Cohen's *d* reflects the difference between the first and final administration divided by the pooled *SD* across all sessions. In analyses of the main effect of form, Cohen's *d* reflects the largest interform difference divided by the pooled *SD* across all forms.



	Session 1	Session 2	Session 3	Session 4	Session 5
SM Total Learning	0.00	0.40	0.37	0.43	0.78
SM Delayed Recall	0.00	0.23	0.28	0.34	0.48
PC Total Correct	0.00	0.31	0.50	0.61	0.61
DS Max Total	0.00	0.24	0.44	0.52	0.70
Animal Naming	0.00	0.32	0.52	0.50	0.69
CS Switches	0.00	0.53	0.73	0.84	1.18

Figure 1. Change in *Z*-score by session, controlling for form.
N = 40 in all cases. *Z*-scores were calculated by comparing each session mean to the mean and standard deviation for Session 1. SM = Story Memory; PC = Perceptual Comparison; DS = Digit Span; CS = Category Switching.

Results

Session Effects

Performance on all tasks showed improvement over time for cognitive measures (Figure 1). Table 1 shows descriptive statistics for each time point, test statistics, post hoc comparisons, and effect size measures. In all cases, unadjusted two-tailed *p* values are provided for the omnibus test associated with each variable.

For SM, Total Learning, $F(4, 156) = 5.36$, $p < .001$, partial $\eta^2 = .121$, $d = 0.87$, showed a large, statistically significant improvement over repeated sessions, with the largest gains between Session 1 and Session 2. SM Delayed Recall trended toward, but did not reach, statistically significant improvement over sessions ($p = .059$), and SM Percent Retained showed minimal change in either direction across sessions. Because the latter is derived from the relationship between two other scores, its pattern of change in response to repeated administration is complex. Statistically significant practice effects were also observed on PC, $F(4, 156) = 19.56$, $p < .001$, partial $\eta^2 = .334$, $d = 0.65$, where the mean score increased by 4.7 points from Session 1 to Session 5, although this represents a "moderate" effect size overall. When the longest spans forward and backward were summed on DS, participants scored 1.5 points higher at Session 5 than they did at Session 1, an improvement that is statistically significant and of similar magnitude, $F(4, 156) = 5.55$, $p < .001$, partial $\eta^2 = .125$, $d = 0.65$. Participants showed larger practice effects on the CS measures Correct Switches, $F(4, 156) = 17.34$, $p < .001$, partial $\eta^2 = .308$, $d = 1.06$; and Total Words, $F(4, 156) = 17.89$, $p < .001$, partial $\eta^2 = .314$, $d = 1.04$; than on Animal Naming, $F(4, 156) = 6.24$, $p < .001$, partial $\eta^2 = .138$, $d = 0.66$, even though both subtests are identical across forms. Of note, all practice effects detected in this analysis would be retained even after correcting for multiple comparisons. Last, all participants completed the PHQ-4 questionnaire, for which no practice effects were expected. Nonetheless, participants scored lower on the PHQ-4 at Session 5 than they did at Session 1, $F(4, 156) = 9.56$, $p < .001$, partial $\eta^2 = .197$, $d = -0.66$, though the mean score always remained quite low. Individuals with mood or anxiety disorders were excluded from the study at recruitment, making this finding likely a measurement artifact.

Form Effects

Table 2 provides participants' mean scores by form. Again, unadjusted two-tailed *p* values were provided for

Table 3. Equivalence of Story Memory Form Subsets.

SM Total Learning forms							SM delayed forms						
Form 1	Form 2	Form 3	Form 4	Form 5	*F*	*p*	F1	F2	F3	F4	F5	*F*	*p*
Four-form subsets													
1[a]	2[a,b]	3[b]	4		14.84	<.001	1	2[a]	3[a]	4		7.61	<.001
1[a]	2[a,b]	3[b,c]		5[c]	17.47	<.001	1	2[a]	3[a,b]		5[b]	9.32	<.001
1[a]	2[a]		4	5	7.77	<.001	1	2		4	5	2.78	.044
1		3[a]	4	5[a]	9.69	<.001	1		3[a]	4	5[a]	7.04	<.001
	2[a]	3[a,b]	4	5[b]	16.46	<.001		2[a]	3[a,b]	4	5[b]	10.09	<.001
Three-form subsets													
1[a]	2[a,b]	3[b]			23.37	<.001	1	2[a]	3[a]			9.79	<.001
1[a]	2[a]		4		10.73	<.001	1	2		4		3.35	.040
1[a]	2[a]			5	8.65	<.001	1	2			5	4.02	.022
1		3	4		4.33	.016	1		3	4		5.53	.006
1		3[a]		5[a]	14.67	<.001	1		3[a]		5[a]	8.90	<.001
1			4	5	5.15	.008	1			4	5	2.65	.077
	2[a]	3[a]	4		21.08	<.001		2[a]	3[a]	4		12.74	<.001
	2[a]	3[a,b]		5[b]	24.17	<.001		2[a]	3[a,b]		5[b]	16.52	<.001
	2		4	5	6.96	.002		2		4	5	0.90	.411
		3[a]	4	5[a]	15.01	<.001			3[a]	4	5[a]	12.26	<.001
Two-form subsets													
1[a]	2[a]				20.47	<.001	1	2				5.81	.021
1		3			5.84	.020	1		3			2.85	.099
1			4		0.06	.809	1			4		2.09	.156
1				5	7.85	.008	1				5	5.01	.031
	2[a]	3[a]			43.02	<.001		2[a]	3[a]			26.92	<.001
	2		4		13.11	.001		2		4		1.56	.219
	2			5	0.98	.329		2			5	0.31	.582
		3	4		7.51	.009			3	4		14.38	.001
		3[a]		5[a]	33.63	<.001			3		5	26.79	<.001
			4	5	6.83	.013				4	5	0.68	.415

Note. Uncorrected two-tailed *p* values are provided. Controlling for multiple comparisons, $p < .001$ is significant for an omnibus test at a tablewise α of .05. Statistically significant post hoc contrasts are identified by assigning the same superscript letter to each statistically significantly different pair. Given the large number of post hoc tests performed, only contrasts with $p < .001$ are notated. Subsets of forms which show no statistically significant differences at the .001 level are highlighted with gray shading. SM = Story Memory.

omnibus tests; to adjust for multiple comparisons, only *p* values < .006 should be treated as statistically significant. Due to the Latin square design, each form was administered equally often at each point in the sequence, so no form was advantaged by order effects.

SM showed large interform differences. The constraints of grammar and sense prevent these stories from being exactly parallel to one another. As such, it was unsurprising that statistically significant differences were found among the forms for Total Learning, $F(4, 156) = 13.16$, $p < .001$, partial $\eta^2 = .252$, $d = 1.16$; and Delayed Recall, $F(4, 156) = 7.24$, $p < .001$, partial $\eta^2 = .157$, $d = 0.87$; measures, though not for Percent Retained, $F(4, 156) = 1.33$, $p = .261$, partial $\eta^2 = .033$, $d = 0.41$. Post hoc testing attempted to identify a set of three or four SM forms which would be equivalent (see Table 3). All possible four-form sets yielded statistically significant omnibus test results. Among the three-form subsets, there were two groupings that would be statistically insignificantly different on both SM Total Learning and SM Delayed Recall if a correction were applied for multiple comparisons: {1,3,4} and {1,4,5}. In addition, there were two-form subsets which were statistically insignificant even without correcting for multiple comparisons: {1,3}; {1,4}; and {2,5}, with the addition of {1,5} and {4,5} after correcting for multiple comparisons. Considering all subsets, the grouping associated with the lowest mean partial η^2 is {2;5} followed closely by {1;4}.

Relatively small interform differences were found for PC Total Correct overall, but they did not survive Bonferroni correction, $F(4, 156) = 3.42$, $p = .010$, partial $\eta^2 = .081$, $d = 0.35$. Post hoc testing revealed no statistically significant differences between forms. No statistically significant interform differences were found for DS Max Total, $F(4, 156) = 1.22$, $p = .276$, partial $\eta^2 = .030$,

Table 4. GNA Practice Effects at Following a Single Administration.

Subtest	Session 1	Session 2	Raw score Δ	Cohen's *d*
SM Total Learning	19.50 (3.98)	21.10 (3.53)	+1.60	0.42
SM Delayed Recall	10.20 (2.46)	10.77 (1.99)	+0.57	0.25
SM Percent Retained	88.78 (16.77)	92.49 (13.48)	+3.71	0.24
PC Total Correct	42.63 (7.59)	45.00 (7.19)	+2.37	0.31
DS Max Total	11.50 (2.18)	12.03 (2.43)	+0.53	0.23
Animal Naming	25.73 (5.12)	27.35 (6.15)	+1.62	0.28
CS Switches	20.62 (4.48)	23.00 (5.08)	+2.38	0.48
CS Words	21.80 (4.31)	24.28 (4.79)	+2.48	0.52
PHQ-4	1.38 (1.51)	0.93 (0.94)	−0.45	−0.35

Note. Cohen's *d* is derived from difference between Session 2 and Session 1 means divided by the pooled *SD* for both sessions. The GNA Spatial Span subtest was not administered as part of this study. SM = Story Memory; PC = Perceptual Comparison; DS = Digit Span; CS = Category Switching; PHQ-4 = Patient Health Questionnaire–4 Item.

$d = 0.33$. Because the Animal Naming, CS, and PHQ-4 subtests are identical across forms, no interform differences were expected. Consistent with this, none were found (all omnibus p values were $>.3$).

Discussion

Session Effects

Performance on the GNA improved over five successive administrations. The largest practice effects were expected and found for CS. Not only is this subtest the same across all forms of the GNA but it also lends itself to strategy development with experience more than other GNA subtests. From Session 1 to Session 5, mean scores for both CS Switches and CS Words improved by a full standard deviation. Large practice effects also were found for SM Total Learning, but not SM Delayed Recall or SM Percent Retained, despite the fact that each GNA form has a different story. This likely reflects the development of a general learning strategy since the story contents differ across forms. Smaller but still statistically significant practice effects were seen on the PC and DS subtests. Given that the PC and DS stimuli differ across forms of the GNA, these effects also likely represent general strategy development. Conversely, participants reported fewer symptoms of anxiety and depression on the PHQ-4 from Session 1 to Session 5, although the mean score at every session was normal. One possible explanation of this finding is that some participants felt mildly anxious about an unknown testing experience at the first session but less so thereafter once the procedures became familiar.

Overall, these results indicate that, even when using alternate forms, repeated administration of the GNA is associated with practice effects, at least when administered five times over 5 weeks to cognitively intact adults. This study was designed to amplify practice effects. Administering the GNA to cognitively impaired persons would likely reduce such effects (Calamia et al., 2012; Hassenstab et al., 2015; Jutten et al., 2020). Additionally, Table 1 shows that practice effects are primarily concentrated between Sessions 1 and 2, with much smaller gains across the remaining administrations. As gains following a single administration represent the most common area of concern in practice, Table 4 reports the raw and standardized score changes from Session 1 to Session 2. When the GNA is used for repeated measures research designs, appropriate experimental and statistical techniques should be applied to control for practice effects. When the GNA is administered repeatedly in a clinical setting, practitioners should use alternate forms to diminish practice effects, increase the time delay between assessments, and consider adjusting obtained scores for the second or subsequent administration(s), especially over short test–retest intervals to account for practice effects.

Form Effects

Some GNA subtests were identical across forms, some had modest changes, and one (SM) varied markedly. As expected, the three subtests that are identical across forms (Animal Naming, CS, and PHQ-4) showed small to negligible interform differences. Their interform Cohen's *d* effect sizes ranged from 0.13 to 0.26 for a mean of 0.19. These can be seen as a kind of "baseline" for random performance variation across (identical) forms. The other three GNA subtests (SM, PC, and DS) vary across forms. Since PC and DS alternate forms were generated by rearranging a small set of possible stimuli, interform similarity was expected to be high. Consistent with this expectation, the omnibus tests were not statistically significant after the Bonferroni correction was applied and the Cohen's *d* effect sizes were 0.33 to 0.35. The SM subtest differs the most across forms. As such, the largest across-form differences are expected. Our analysis found statistically significant differences across forms in SM Total Learning ($d = 1.16$) and SM Delayed Recall ($d = 0.87$),

though the five forms were statistically indistinguishable on SM Percent Retained. To begin to understand the form differences, post hoc tests were run which show elevated scores on Form 3 as compared with Forms 2 and 5. However, merely removing Form 3 was not enough to guarantee the remaining forms are equivalent. When Forms 1, 2, 4, and 5 were compared on the SM Total Learning Variable with an omnibus test, the result was statistically significant ($p < .001$) indicating that those four forms are unequal. In fact, there was no four-form subset that showed no statistically significant inequality among the forms on SM Total Learning and SM Delayed Recall. After correcting for multiple comparisons, there were two three-form subsets which met these criteria: {1, 3, 4} and {1,4,5}, as well as the two-form subsets {1,3}, {1,4}, {1,5}, {2,5}, and {4,5}, with the greatest similarity found in pairs {1,4} and {2,5}.

It is challenging to create meaningful verbal stimuli that are equal in the ease with which they can be perceived, understood, stored, retrieved, and reproduced (Cunje et al., 2007). Many factors affect the difficulty of verbal memory stimuli, including the frequency of the word in usage (Miller & Roodenrys, 2009), the degree to which the stimulus evokes an affective response (Sutton & Lutz, 2019), and the extent to which a part of the stimulus can be predicted by other parts (Staub et al., 2015; Valian et al., 2006). As an example of the latter effect, consider a narrative which uses the phrase "eat cookies and drink milk." If the participant recalls the word "cookies," doing so is likely to cue the word "milk" by virtue of their semantic relatedness, and both of these are likely to cue the words "eat" and "drink" for the same reason. The story for GNA Form 3 contains an analogous sequence of highly linked nouns with predictable verbs, which might explain why participants recalled more of its details than they did for the remaining four stories.

In short, the five GNA forms are nearly equivalent for all subtests except SM which nonetheless offers several options for equivalent retesting. Based on the findings of this study, if the GNA is to be used in a repeated measures research design, using alternate forms with counterbalancing is recommended. If a protocol calls for only two administrations of SM, our results suggest using Forms 1 and 4 or 2 and 5. If a protocol calls for three administrations, the most similar forms are 2, 4, and 5 or 1, 4, and 5. If four administrations or more administrations are required, the use of form-adjusted scores should be considered.

Strengths, Limitations, and Future Directions

A strength of this study was its use of a Latin square design that counterbalanced the order in which five parallel forms of the GNA were administered across five test sessions. This type of design is rarely implemented, especially when so many alternate forms are investigated, due to the difficulty recruiting and retaining healthy control participants over so many visits. However, this design also introduces a limitation—because the Latin square design is comparatively onerous, it necessitates a smaller sample size, which increases the risk of random error. In the same vein, the intensive research design necessitated the shortest possible test session, to minimize the burden on participants. As such, we elected to omit the GNA's optional Spatial Span subtest and will therefore need to rely on future research to analyze its alternate forms.

Ceiling effects were found for all GNA subtests, especially SM (Form 3) and PC (Sessions 3-5). No participant achieved the maximum score for SM Total Learning on any form except Form 3 which yielded 3 ceiling scores. However, participants earned the maximum score on SM Delayed Recall in 19 instances, out of the 200 total administrations of the GNA. PC yielded the greatest number of ceiling scores, with 2 to 9 participants earning the highest possible raw score at each session, for a total of 28 maximum scores. There were also six instances of performance at the ceiling for DS Max Total. The most likely reason for this is that the majority of study participants were young, healthy university students. Ceiling effects are less likely to constrain the performance of children, older adults, persons with cognitive disorders, and persons of average intellectual ability (Hassenstab et al., 2015; Kremen et al., 2020). The GNA was not designed to distinguish superior or better performance, even though it likely can do so in many demographic subgroups. Rather, the GNA was designed to assess cognitive dysfunction and changes in function due to treatment effects or disease progression. While ceiling-level performance was always in the minority, statistical analysis can be distorted by even a handful of scores restricted by ceiling effects, particularly given a small sample size. These high performers may have distorted our findings.

In contrast, floor effects may or may not have limited the interpretability of PHQ-4 data. 55.5% of the 200 times the PHQ-4 was administered for this study, the raw score total was zero. Mood disorders were an exclusionary criterion for the study, so this is unsurprising. Nonetheless, the lack of variation still limited statistical analysis. The unexpected finding that PHQ-4 scores increased slightly over the course of the study may simply reflect the fact that no test could detect decreases in anxious or depressive symptoms in a population already at minimum. In both future studies which include broader populations and applied clinical use, a significant relationship between PHQ-4 scores and cognitive performance will likely emerge. Thus, it may become desirable to statistically control for the influence of affect.

This study's sample was largely homogeneous in terms of race/ethnicity, language, and national origin. Future research on the comparability of the GNA forms should include a more heterogeneous sample in terms of age, educational background, language, race/ethnicity, intellectual

ability, and health status. Greater diversity would likely reduce both practice and ceiling effects, thereby allowing more variable test performance. In addition, marginal effects are less likely to be replicated in larger studies, diminishing possibly false positive findings.

A major goal of the GNA is to offer accessible neuropsychological assessment to low- and middle-income countries, which will obviously require the translation of the instructions and stimuli. As this paper has shown, developing comparable SM stimuli is extremely difficult. When these passages are then translated, some difficulty factors (e.g., imageability) are unlikely to change (Dye et al., 2013; Roche et al., 2011; Romani et al., 2008), but others (e.g., phonemic neighborhood size) will be completely different (Allen & Hulme, 2006). Word length, in letters or syllables, may be understood differently in languages with different rhythmic or orthographic structures. Cultural diversity will require attention to differences in affective content and conformity to cultural scripts (Harris et al., 1992). For example, we hypothesized that SM Form 3 yielded higher scores than other forms because it included an unusually predictable sequence of target words. In another linguistic or cultural setting, that sequence may not be as familiar or expected (Staub et al., 2015; Valian et al., 2006). Local or adjusted norms will need to be developed.

This article establishes the general comparability of the five forms of the Global Neuropsychological Assessment and provides a baseline for the degree of change expected across repeated administrations in a healthy, well-educated sample. It lays the groundwork for future studies establishing the validity and utility of the GNA as a brief, accessible neurocognitive battery ready for worldwide adaptation.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors would like to gratefully acknowledge the financial support of the American Academy of Clinical Neuropsychology Relevance 2050 initiative for this project.

ORCID iDs

Alan Smerbeck https://orcid.org/0000-0001-8087-4661
Lauren T Olson https://orcid.org/0000-0001-7707-9325

References

Allen, R., & Hulme, C. (2006). Speech and language processing mechanisms in verbal serial recall. *Journal of Memory and Language*, *55*(1), 64-88. https://doi.org/10.1016/j.jml.2006.02.002

Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., & Ehrenreich, H. (2010). Practice effects in healthy adults: A longitudinal study on frequent repetitive cognitive testing. *BMC Neuroscience*, *11*(1), Article 118. https://doi.org/10.1186/1471-2202-11-118

Benedict, R. H., Amato, M. P., Boringa, J., Brochet, B., Foley, F., Fredrikson, S., Hamalainen, P., Hartung, H., Krupp, L., Penner, I., Reder, A. T., & Langdon, D. (2012). Brief International Cognitive Assessment for MS (BICAMS): International standards for validation. *BMC Neurology*, *12*(1), Article 55. https://doi.org/10/gb495n

Benedict, R. H., Smerbeck, A., Parikh, R., Rodgers, J., Cadavid, D., & Erlanger, D. (2012). Reliability and equivalence of alternate forms for the Symbol Digit Modalities Test: Implications for multiple sclerosis clinical trials. *Multiple Sclerosis Journal*, *18*(9), 1320-1325. https://doi.org/10/fzcrjk

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, *26*(4), 543-570. https://doi.org/10/ggr4f5

Cunje, A., Molloy, D. W., Standish, T. I., & Lewis, D. L. (2007). Alternate forms of logical memory and verbal fluency tasks for repeated testing in early cognitive changes. *International Psychogeriatrics*, *19*(1), 65-75. https://doi.org/10/bw3phk

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan Executive Function System*. Psychological Corporation.

Dye, C. D., Walenski, M., Prado, E. L., Mostofsky, S., & Ullman, M. T. (2013). Children's computation of complex linguistic forms: A study of frequency and imageability effects. *PLOS ONE*, *8*(9), e74683. https://doi.org/10.1371/journal.pone.0074683

Elman, J. A., Jak, A. J., Panizzon, M. S., Tu, X. M., Chen, T., Reynolds, C. A., Gustavson, D. E., Franz, C. E., Hatton, S. N., Jacobson, K. C., Toomey, R., McKenzie, R., Xian, H., Lyons, M. J., & Kremen, W. S. (2018). Underdiagnosis of mild cognitive impairment: A consequence of ignoring practice effects. *Alzheimer's & Dementia*, *10*(1), 372-381. https://doi.org/10/ghzpt4

Folstein, M. F., Folstein, S., & McHugh, P. R. (1975). Mini-Mental State: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189-198. https://doi.org/10.1016/0022-3956(75)90026-6

Gavett, B. E., Gurnani, A. S., Saurman, J. L., Chapman, K. R., Steinberg, E. G., Martin, B., Chaisson, C. E., Mez, J., Tripodis, Y., & Stern, R. A. (2016). Practice effects on story memory and list learning tests in the neuropsychological assessment of older adults. *PLOS ONE*, *11*(10), e0164492. https://doi.org/10/f9q4bb

Green, M. F., & Nuechterlein, K. H. (2004). The MATRICS initiative: Developing a consensus cognitive battery for clinical trials. *Schizophrenia Research*, *72*(1), 1-3. https://doi.org/10.1016/j.schres.2004.09.006

Gross, A. L., Kueider-Paisley, A. M., Sullivan, C., Schretlen, D., & International Neuropsychological Normative Database Initiative. (2019). Comparison of approaches for equating different versions of the Mini-Mental State Examination administered in 22 studies. *American Journal of Epidemiology*, *188*(12), 2202-2212. https://doi.org/10/ghqmr6

Gualtieri, C. T., & Johnson, L. G. (2006). Reliability and validity of a computerized neurocognitive test battery, CNS Vital Signs. *Archives of Clinical Neuropsychology*, *21*(7), 623-643. https://doi.org/10.1016/j.acn.2006.05.007

Gulliksen, H. (1950). *Theory of Mental Tests*. Wiley & Sons. https://doi.org/10.1037/13240-000

Harris, R. J., Schoen, L. M., & Hensley, D. L. (1992). A cross-cultural study of story memory. *Journal of Cross-Cultural Psychology*, *23*(2), 133-147. https://doi.org/10.1177/0022022192232001

Hassenstab, J., Ruvolo, D., Jasielec, M., Xiong, C., Grant, E., & Morris, J. C. (2015). Absence of practice effects in preclinical Alzheimer's disease. *Neuropsychology*, *29*(6), 940-948. https://doi.org/10/f7zbjd

Jalbert, A., Neath, I., Bireta, T. J., & Surprenant, A. M. (2011). When does length cause the word length effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(2), 338-353. https://doi.org/10.1037/a0021804

Jutten, R. J., Grandoit, E., Foldi, N. S., Sikkes, S. A. M., Jones, R. N., Choi, S., Lamar, M. L., Louden, D. K. N., Rich, J., Tommet, D., Crane, P. K., & Rabin, L. A. (2020). Lower practice effects as a marker of cognitive performance and dementia risk: A literature review. *Alzheimer's & Dementia*, *12*(1), e12055. https://doi.org/10/ghzczp

Kremen, W. S., Sanderson-Cimino, M. E., Elman, J. A., Tu, X. M., Gross, A. L., Panizzon, M. S., Eglit, G. M. L., Jak, A. J., Edmonds, E. C., Thomas, K. R., Eppig, J. S., Williams, M. E., Bondi, M. W., Lyons, M. J., & Franz, C. E. (2020). Accounting for cognitive practice effects results in earlier detection and more accurate diagnosis of MCI. *Alzheimer's & Dementia*, *16*(S6), e044883. https://doi.org/10/gkpwkk

Langdon, D. W., Amato, M. P., Boringa, J., Brochet, B., Foley, F., Fredrikson, S., Hamalainen, P., Hartung, H. P., Krupp, L., Penner, I. K., Reder, A. T., & Benedict, R. H. (2012). Recommendations for a Brief International Cognitive Assessment for Multiple Sclerosis (BICAMS). *Multiple Sclerosis*, *18*(6), 891-898. https://doi.org/10.1177/1352458511431076

Löwe, B., Wahl, I., Rose, M., Spitzer, C., Glaesmer, H., Wingenfeld, K., Schneider, A., & Brähler, E. (2010). A 4-item measure of depression and anxiety: Validation and standardization of the Patient Health Questionnaire-4 (PHQ-4) in the general population. *Journal of Affective Disorders*, *122*(1-2), 86-95. https://doi.org/10/fvz53b

Miller, L. M., & Roodenrys, S. (2009). The interaction of word frequency and concreteness in immediate serial recall. *Memory & Cognition*, *37*(6), 850-865. https://doi.org/10/dkj5x4

Nasreddine, Z. S., Phillips, N. A., Bedirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*(4), 695-699. https://doi.org/10/dmt678

Olson, L. T., Smerbeck, A., Figueroa, C. M., Raines, J. M., Szigeti, K., Schretlen, D. J., & Benedict, R. H. B. (2020). Preliminary validation of the Global Neuropsychological Assessment in Alzheimer's disease and healthy volunteers. *Assessment*. Advance online publication. https://doi.org/10.1177/1073191121991221

Perkins, K., Brutten, S. R., & Angelis, P. J. (1986). Derivational complexity and item difficulty in a sentence repetition task. *Language Learning*, *36*(2), 125-141. https://doi.org/10.1111/j.1467-1770.1986.tb00375.x

Raines, J., Carroll, A., Morra, L., & Schretlen, D. (2019). Global Neuropsychological Assessment (GNA): Preliminary evidence of clinical utility. *Archives of Clinical Neuropsychology*, *34*(6), 940-940. https://doi.org/10/ggsc3r

Randolph, C. (1998). The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): Preliminary clinical validity. *Journal of Clinical and Experimental Neuropsychology*, *20*(3), 310-319. https://doi.org/10.1076/jcen.20.3.310.823

Rijnen, S. J. M., van der Linden, S. D., Emons, W. H. M., Sitskoorn, M. M., & Gehring, K. (2018). Test-retest reliability and practice effects of a computerized neuropsychological battery: A solution-oriented approach. *Psychological Assessment*, *30*(12), 1652-1662. https://doi.org/10.1037/pas0000618

Roche, J., Tolan, G. A., & Tehan, G. (2011). Concreteness effects in short-term memory: A test of the item-order hypothesis. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *65*(4), 245-253. https://doi.org/10.1037/a0024693

Romani, C., Mcalpine, S., & Martin, R. C. (2008). Concreteness effects in different tasks: Implications for models of short-term memory. *Quarterly Journal of Experimental Psychology*, *61*(2), 292-323. https://doi.org/10.1080/17470210601147747

Ross, T. P., Furr, A. E., Carter, S. E., & Weinberg, M. (2006). The psychometric equivalence of two alternate forms of the Controlled Oral Word Association Test. *The Clinical Neuropsychologist*, *20*(3), 414-431. https://doi.org/10/frz3tr

Smerbeck, A., Benedict, R. H. B., Eshaghi, A., Vanotti, S., Spedo, C., Blahova Dusankova, J., Sahraian, M. A., Marques, V. D., & Langdon, D. (2018). Influence of nationality on the Brief International Cognitive Assessment for Multiple Sclerosis (BICAMS). *The Clinical Neuropsychologist*, *32*(1), 54-62. https://doi.org/10/ggxpvk

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, *82*(July), 1-17. https://doi.org/10.1016/j.jml.2015.02.004

Sutton, T. M., & Lutz, C. (2019). Attentional capture for emotional words and images: The importance of valence and arousal. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *73*(1), 47-54. https://doi.org/10.1037/cep0000154

Valian, V., Prasada, S., & Scarpa, J. (2006). Direct object predictability: Effects on young children's imitation of sentences. *Journal of Child Language*, *33*(2), 247-269. https://doi.org/10.1017/S0305000906007392