Korean Journal of Clinical Psychology
2023. Vol. 42, No. 4, 118-129
OPEN ACCESS

Original Article
© 2023 Korean Clinical Psychology Association
https://doi.org/10.15842/kjcp.2023.42.4.003
eISSN 2733-4538

Clinical Validity of Neuropsychological Assessment in Dementia: A Univariate and Multivariate Methodological Comparison

Seyul Kwak[1†] Keun You Kim[2] Su Mi Park[3] Da Young Oh[2] Hairin Kim[2] Dasom Lee[2] Jun-Young Lee[2,4†]

[1]Department of Psychology, Pusan National University, Busan;
[2]Department of Psychiatry, Seoul National University College of Medicine & SMG-SNU Boramae Medical Center, Seoul;
[3]Department of Counseling Psychology, Hannam University, Daejeon;
[4]Department of Medical Device Development, Seoul National University College of Medicine, Seoul, Korea

Previous studies have documented validity evidence of neuropsychological measures in the assessment of dementia. However, known cognitive test measures were mostly validated as separate measurements rather than asconstituents of a whole battery. In this study, the neuropsychological battery (CERAD-K) and neuropsychiatric measures were acquired in older adults with Mild Cognitive Impairment, dementia of Alzheimer's Disease (AD), and Vascular dementia (VD). The assessment measures and demographic information were used to predict two validity criteria: dementia severity (CDR) and dementia type (AD or VD). A correlation between a single test measure and the target criteria indicated univariate validity, whereas relative importance among multiple regression models indicated the multivariate validity of a single measure as a constituent of the battery. We identified that test measures including the Boston Naming Test, Trail Making Test, and Word List Recall were predictive of the clinical outcome criteria as univariate validity; however, this strength of association did not remain consistent when evaluated in terms of multivariate validity. Regarding the multivariate validity, measures including Word List recognition, and neuropsychiatric impairment showed robust validity. This contrasting validity indices between univariate and multivariate frameworks may be owing to shared information between other measures, which can distort the conclusions of validity evidence. The findings suggest that the validity of a neuropsychological test differs as a function of the target criteria and whether administered as a whole battery. The findings suggest that the validity of a neuropsychological measure differs as a function of the criteria of clinical context and whether tested under a comprehensive battery.

Keywords: neuropsychological assessment, validity, multivariate, dementia

Introduction

Neuropsychological assessment aims to clarify neurological conditions and describe detailed characteristics of functional impairment (Fields et al., 2011). Typically, combined batteries of ratings and test scores are integrated during interpretation, which subsequently leads to clinical decisions (Vakil, 2012). Individuals with dementia due to specific types of neurological diseases are also administered with a known set of test batteries that were validated to have clinical utility in diagnosing and describing the severity and differential pathology types (Elahi & Miller, 2017).

[†]Correspondence to Jun-Young Lee, Department of Psychiatry, Seoul National University College of Medicine & SMG-SNU Boramae Medical Center, 20 Boramae-ro 5-gil, Dongjak-gu, Seoul 07061, Korea; E-mail: benji@snu.ac.kr

[†]Correspondence to Seyul Kwak, Department of Psychology, Pusan National University, 2 Busandaehak-ro 63bean-gil, Geumjeong-gu, Busan 46241, Korea; E-mail: sykwak@pusan.ac.kr

Received Mar 21, 2023; Revised Jul 11, 2023; Accepted Jul 14, 2023

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2022R1F1A1073289) and the Korea Medical Device Development Fund grant funded by the Korea government (Project Number: 202015X34).

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

Ongoing examinations on the validity of neuropsychological tests have established cornerstones of how clinicians should select and construct a set of tests that can be applied to target illness (Garb & Schramke, 1996). The previous examinations on test validity, however, have shown how individual test evidences their validity as separate measurements. Here, the logical gap lies in the fact that test measures are not validated as a whole set as a battery even when practitioners integrate multiple measures acquired from a set of tests. Thus, the individual tests that constitute a neuropsychological battery should evince their validity under the context of battery composition, rather than as a single independent measure (Russell et al., 2005). Also, a test battery should cover a wide range of clinical purposes since potential clinical conditions are yet unnarrowed. It is likely that previously validated tests examined under a single measurement may not be generalized as valid tests within a whole battery.

The scope of the test validity also regards how the tests are quantified of their validity. One of the main approaches to evidence the test validity is to show the extent of concurrency and predictability to the clinical outcome of interest (Anastasi, 1950; Strauss & Smith, 2009). While the criterion validity of a single test measure can be examined as a univariate association between the test measure and clinical outcome, the validity of a whole set of tests can be examined as a multivariable association between test measures and clinical outcomes. In this case, the validity of a subtest can be evaluated in terms of its accountability of whole subtests (i.e., actuarial method), and this approach tends to provide more reliable and accurate diagnostic findings (Carlew et al., 2023; Fountain-Zaragoza et al., 2021). Specifically, there can be cases when a test is not useful among a whole comprehensive neuropsychological battery yet shows a sound univariate validity because the shared redundancy between the tests is not considered in the univariate examinations. If a test measure shares predictive value with other comprising measures, then the test validity can be undermined, whereas greater uniqueness of a test leads to larger validity as constituents among a battery. This multivariable nature of the test battery can directly be tested with quantifiable metrics and the resulting quantified validity index can aid as the rationale for selecting specific tests that have been determined by experienced clinicians.

Recent studies have utilized how combining multiple measures can aid findings in dementia assessment. For example, previous studies have shown that the reliability of diagnostic classification on mild cognitive impairment can be enhanced when conjointly using multiple measures rather than applying a single test cutoff (Bondi & Smith, 2014; Graves et al., 2020). Other multivariable approaches combined as a predictive model showed the potential to enhance the utility in the assessment of dementing outcomes (Chapman et al., 2010; Kwak et al., 2022; Nation et al., 2019; Stallard et al., 2022). Despite the evidence of the enhanced predictability of multivariate approaches, previous studies lacked detailed examinations of how individual measures contribute to the total criterion validity which would be referenced in the clinician's decision in test selection. In order for clinicians to aid battery construction, the subtests should be subject to validity evaluation (Garb & Schramke, 1996).

Another issue in quantifying the validity of neuropsychological battery is how the test validities can differ by the clinical context of the assessment and validity criterion. In the case of dementia assessment, for example, a particular test can claim its validity under differential diagnosis of dementia types, whereas some other tests may retain validity in predicting overall daily functioning (Bruun et al., 2018; Fields et al., 2010; Kwak et al., 2021; Nyenhuis et al., 2004). Neuropsychological test information may have differing utility in either identifying biological etiology in a medical context or assessing ecological functioning in a rehabilitation context. Indeed, a test with evidence of both aspects would be the most desirable case, the tests can also play a role under a specific validity context. A direct comparison of the two validity contexts has not been thoroughly examined in the studies of dementia assessment instruments.

Concurring knowledge of neuropsychological test validities is converged mostly based on the examination of each test as a single measure. This could lead to profoundly different conclusions. For example, a test that requires multiple processes (e.g., verbal fluency, trail making test) can exhibit strong utility in predicting clinical outcomes of interest. But the very conclusion may not coincide when tested under multiple sets of measures that are typically administered as a battery. To our knowledge, this discrepancy between univariate versus multivariate validity has not been examined previously.

Thus, the current study aimed to examine how measures of neuropsychological assessment are predictive of (1) overall functional impairment across the spectral population of cognitive impairment, and (2) differential diagnosis between Alzheimer's disease and vascular dementia. In this way, each assessment measure is evaluated for whether retains criterion validity under specific clinical contexts.

Methods

Participants

The older adults with cognitive impairment were retrospectively recruited from SMG-SNU Boramae Medical Center for Dementia from January 2012 to January 2021. The retrospective dataset was extracted from the in-house clinic database available. The participants underwent both neuropsychological assessment and structured clinical interview. This study was conducted under the Declaration of Helsinki, and the protocol was approved by the Institutional Review Board of SMG-SNU Boramae Medical Center for Dementia (IRB No. 10-2020-295). The current study included older adults with Mild Cognitive Impairment (MCI), dementia of Alzheimer's Disease (AD), and Vascular dementia (VD). The clinical diagnosis of the probable or possible AD and MCI was based on the National Institute of Neurological and Communicative Disorders and Stroke and AD and Related Disorders Association (NINCDS-ADRDA) and the core clinical criteria of MCI (Albert et al., 2011; McKhann et al., 1984). The VD was diagnosed according to the National Institute of Neurological Disorders and Stroke/Association Internationale pour la Recherche et l'Enseignementen Neurosciences criteria (Román et al., 1993). The other dementia types were not considered as analyses of interest due to insufficient sample size. In cases of multiple follow-ups, the diagnosis and test measures of the first neuropsychological evaluation were analyzed.

Subjects suspected or diagnosed with dementia types other than AD or VD were not included in the analysis, including Lewy body dementia and frontotemporal lobe dementia. In addition, those identified or suspected of significant neurological or psychiatric conditions including traumatic brain injury, meningioma, subdural hemorrhage, normal pressure hydrocephalus, delirium, intellectual disabilities, and psychotic disorders were excluded. We confined our predictive analysis within the dementia staging of 'moderate' impairment (Clinical Dementia Rating sum of box score ≤15.5) (O'Bryant, 2008). The group size differed across diagnoses (MCI: $n=1{,}025$; AD: $n=1{,}262$; VD: $n=266$), and the validity evaluation set was comprised of two sets: (1) cognitive impairment severity (MCI, AD, and VD; $n=2{,}553$), (2) differential diagnosis (AD and VD; $n=1{,}528$). Descriptive statistics and histograms are shown in Table 1 and Figure 1.

Table 1. *Demographic Characteristics*

	Total	MCI[a]	AD[a, b]	VD[a, b]
Mean (SD)/Frequency				
n	2,553	1,025	1,262	266
Age	76.60 (7.80)	73.26 (7.26)	79.14 (7.15)	77.47 (7.96)
Education	7.37 (5.07)	8.31 (4.68)	6.66 (5.15)	7.17 (5.51)
Sex (M : F)	1618:935	593:432	873:389	152:114
Global CDR	0.80 (0.47)	0.50 (0.07)	0.97 (0.50)	1.14 (0.59)
CERAD-Total	44.35 (15.18)	55.87 (10.44)	36.77 (12.73)	35.95 (13.03)
NPI	6.72 (4.12)	4.43 (3.96)	7.16 (5.67)	7.90 (5.99)
GDS	6.14 (5.28)	6.23 (4.09)	6.93 (4.07)	7.62 (4.27)

Note. CDR = Clinical Dementia Rating; MCI = Mild Cognitive Impairment; AD = dementia of Alzheimer's disease; VD = Vascular dementia; NPI = Neuropsychiatric Inventory; GDS = Geriatric Depression Scale.
[a]*Included in analysis of Validity A (dementia severity),* [b]*Included in analysis of Validity B (dementia type).*

Neuropsychological assessment

All participants were administered the Korean version of the Consortium to Establish a Registry for Alzheimer's Disease neuropsychological battery (CERAD-K) (Lee et al., 2002). The battery measures multiple domains of cognitive function and facilitates the diagnosis of MCI and dementia. The battery contains the following subtests: Semantic fluency (the number of correct animal words; four blocks of 15s interval), Boston Naming Test, Word List Recall (immediate, delayed), Word List Recognition (subtraction of the number of false positives from the number of true positives), and Constructional Praxis (copy, recall). The additional subtests included in CERAD-K was Trail Making Test A/B (TMT-A and B). The TMT measured the total time spent completing the tasks. The test administration had set the maximum time limit at 360 s (TMT-A) and 300 seconds (TMT-B) based on administration instruction in CERAD-K (Seo et al., 2006). The score was interpolat-

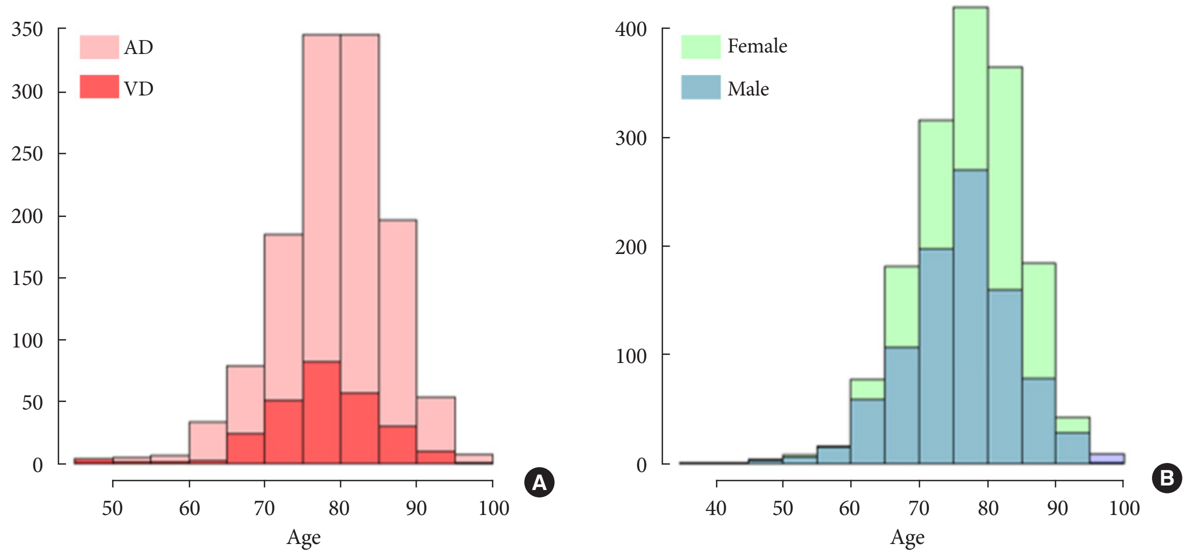

Figure 1. *Age distribution of subgroups.*
Note. (A) AD in light red and VD in dark red. (B) Female in green and male in blue.

ed as the maximum time limit (360s or 300s) in the cases when the TMT was aborted or not feasible due to the following reasons: exceeded the time limit, unable to understand the rule, or committed more than five errors. The scores (seconds) were inverted to have the same directional interpretation. The raw scores without demographic adjustment were used in the analyses. The battery was administered by trained clinical psychologists and trainees supervised by board-certified psychologists.

Neuropsychiatric Inventory (NPI) was used to characterize behavioral, social, and affective symptoms that are relevant to dementing illness (Choi et al., 2000; Cummings et al., 1994). The NPI was based on the semi-structured interview administered to the patients' informants or caregivers, if available, and rated by clinical psychologists. It consists of 12 separate items assessing neuropsychiatric disturbances, including delusion, hallucination, agitation/aggression, depression/dysphoria, anxiety, elation/euphoria, apathy/indifference, disinhibition, irritability/lability, aberrant motor behavior, sleep, and appetite. The symptom severity was rated based on observable behaviors that signify each symptom (e.g., expression of sadness and tears in depression/dysphoria item). Each item was rated from 0 to 3 scores across severity levels (0: No symptom, 1: Symptoms causes mild distress, 2: Symptoms are intractable and cause distress, 3: Symptoms are present with major distress). The summed score of 12 items was used to indicate overall behavioral abnormalities.

As a routinely assessed component of depressive symptoms, the self-reported depressive symptoms were assessed with the Korean version of the short-form Geriatric Depression Scale (GDS) (Bae & Cho, 2004). Questions from the original GDS which had the highest correlation with depressive symptoms in validation studies were selected for the short version with 15 items (Sheikh & Yesavage, 1986).

The validity criterion of overall functional impairment was assessed with the sum of box scores of Clinical Dementia Rating (CDR-SB). The CDR is a semi-structured interview developed to provide a global summary of dementia severity. The CDR is useful for staging and tracking the course of neurodegenerative progression (Fillenbaum et al., 1996; Morris et al., 1997; Morris, 1997). In addition, the sum of boxes score was calculated by summing impairment in six domains of daily cognitive categories (memory, orientation, judgment, community affairs, home and hobbies, and personal care), which provides a more fine-grained measure of functional disturbances within the same category of a global score or clinical diagnosis (Lynch et al., 2005; O'Bryant, 2008). The trained clinical psychologists administered the structured interview and the ratings. As noted in the administration standard, the decisions of CDR scoring were based on the information gathered in a structured interview but not on neuropsychological test performance. The global CDR score ranged from 0 to 1 (MCI), 0.5 to 3 (AD and VD). The CDR-SB ranged from 0.5 to 9 (MCI), 0.5 to 15 (AD), and 1 to 15 (VD).

Statistical analysis

The individual measures that comprise the neuropsychological assessment (i.e., CERAD-K subscores, NPI, GDS) were each examined for clinical validity. Criterion validity was evaluated in how each test measure precisely tracks the target criteria. In the regression models, independent (explanatory) variables constitute test measures, and the dependent variable holds as a criterion to be predicted. The first validity criterion (Validity A) was overall impairment severity (i.e., CDR-SB), and was evaluated with total variance explained (R^2) in linear regression models. The second validity criterion (Validity B) was evaluated as the classification performance of the measures on differential diagnosis (i.e., AD-VD). The classification of the logistic regression models at varying cutoffs was summarized as the Area Under Curve (AUC).

The extent of the validity (i.e., validity index) was evaluated with both univariate and multivariate approaches. The univariate approach simply calculates the pairwise correspondence of the measure to the criterion variable. The multivariate approach fits the multiple regression model to the criterion variable as a whole while excluding a specified target test measure. The decreased amount of accuracy metrics (AUC or R^2) after excluding the target test represented indices of multivariate validity. Multivariate validity index aimed to indicate the unique proportion of information among the given set of the total battery.

The validity examination was conducted by two sets of predictors: (1) cognitive test subscores (CERAD-K), and (2) compiled subtest scores of CERAD-K, demographics, and neuropsychiatric measures. This is because several subtest scores in the CERAD-K result from the common source of test stimuli and procedures which thus produce autoregressive scores, leading to underestimated uniqueness of test scores. In Word List Recall, for example, Immediate Recall inherently has high auto-correlations with Delayed Recall and Recognition scores, which would lower the uniqueness of each subtest. Thus, the subsequent multivariate analysis was conducted after summing the scores under the same subtest unit (Word list: immediate recall, delayed recall, delayed recognition; Construction: copy and delayed reproduction; Fluency: 15 seconds interval scores; TMT: Type A and B) in addition to demographics and neuropsychiatric variables (NPI and GDS).

The correspondence with the two validity criteria was examined, which indicated the extent to which the test holds validity in either clinical context. The mismatching order of effect size between univariate and multivariable testing indicated a differing clinical value of the assessment measurements. Lastly, the full regression models examined total explanatory accuracy with the given measures.

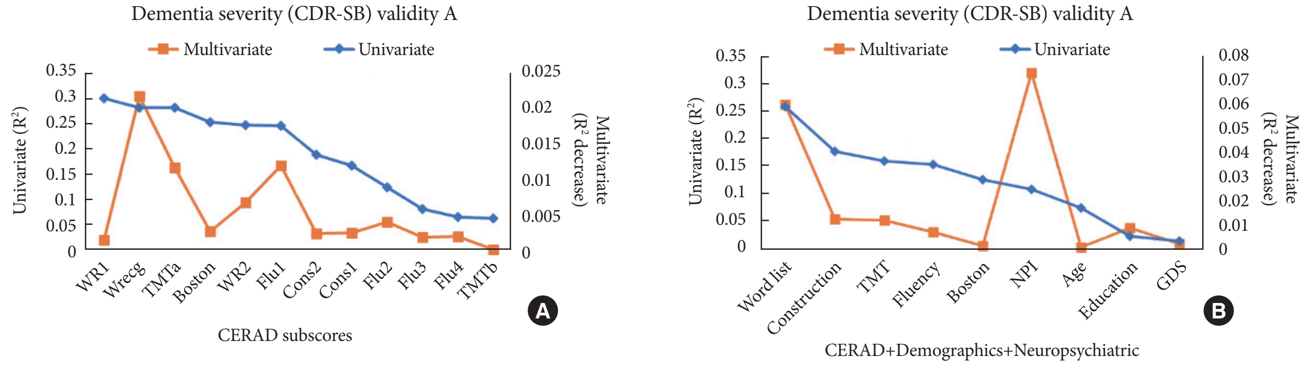

Figure 2. *Univariate and multivariate validity of neuropsychological tests on dementia severity (CDR-SB).*
Note. (A) Results with neuropsychological subtest set. (B) Results combined with demographics and neuropsychiatric measures. Y-axis (A-sided scale) and blue line indicate univariate associations (R^2) between the assessment variables and CDR-SB. Y-axis (B-sided scale) and orange line indicate multivariate contribution (decrease in R^2 when excluded) in the multiple regression model. The subtests are sorted in descending order of univariate validity. CDR-SB = Clinical rating scale-sum of boxes; WR1/2 = Word list immediate/delayed recall; Wrecg = Word list recognition; Cons1/2 = Constructional praxis copy/delayed; TMTa/b = Trail making test A and B; Flu1-4 = Animal fluency (four blocks of 15s interval); Boston = Boston naming test; NPI = Neuropsychiatric Inventory; GDS = Geriatric Depression Scale.

Results

When examining the difference between univariate and multivariate correspondence with CDR-SB (Validity A), the overall tendency of consistency showed that the measures of high univariate association also showed higher multivariate importance (Figure 2). Specifically, however, Word Recall immediate (WR1) and Boston Naming Test showed a large discrepancy, showing generally low validity in the multivariate approach (Figure 2A). That is, the test scores were associated with CDR-SB individually, but the unique explanatory information was minimally provided. When examined while including demographics and neuropsychiatric variables with more summarized cognitive scores, NPI showed the most distinctive validity difference (Figure 2B). NPI was not the strongest predicting feature as a single univariate score but the multivariate index was superior to other cognitive test measures.

With the same analytic approach, the difference between univariate and multivariate classification accuracy on dementia types (AD vs. VD, Validity B) (Figure 3). Again, the overall tendency showed a general correspondence in that measures of high univariate validity also showed higher multivariate validity. However, specific patterns showed notable discrepancies. While TMT-A worked as a relatively superior univariate classifier of dementia types, it provided almost nonexistent information in the multivariate model (Figure 3A). Moreover, Word Recall delayed (WR2) was not uniquely informative in the multivariate model contrary

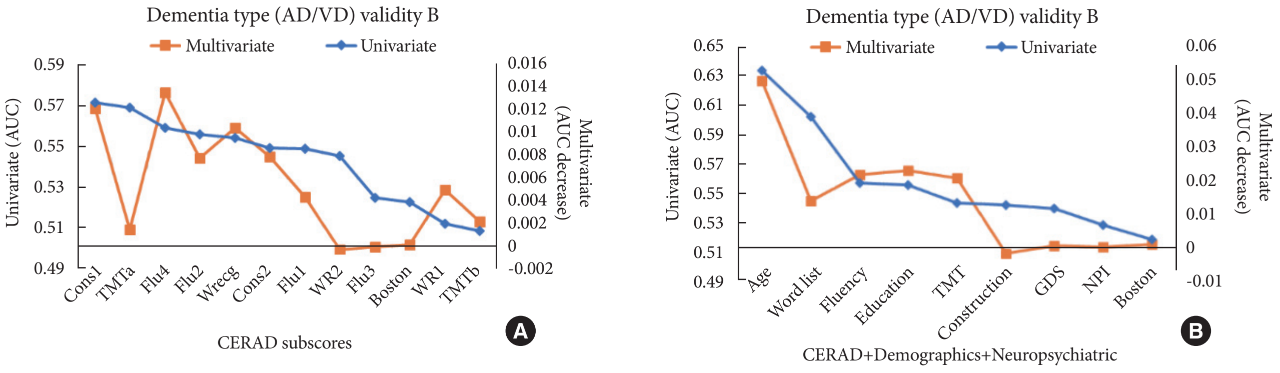

Figure 3. *Univariate and multivariate validity of neuropsychological tests on the differential diagnosis of dementia types (AD vs. VD).*
Note. (A) Results with neuropsychological subtest set. (B) Results combined with demographics (age) and neuropsychiatric measures (NPI). Y-axis (A-sided scale) and blue line indicate univariate classification accuracy (AUC) on dementia types. Y-axis (B-sided scale) and orange line indicate multivariate contribution (decrease in AUC when excluded) in the multiple regression model. The subtests are sorted in descending order of univariate validity.

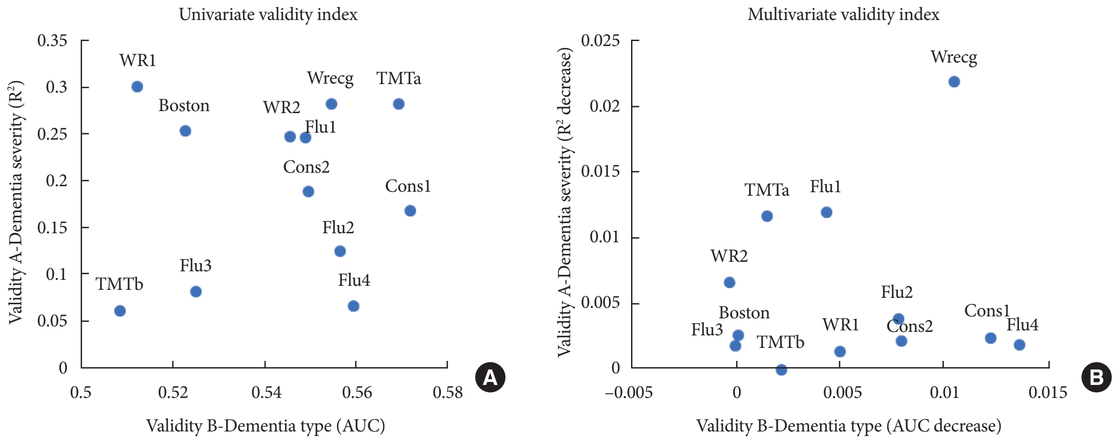

Figure 4. *Correspondence between two different types of validity indices.*
Note. Y-axis: Validity A indicates the amount of unique information of the measures in predicting dementia severity (CDR-SB). X-axis: Validity B indicates unique information in classifying Alzheimer's versus vascular dementia.

to its strong association in the univariate model. When examined by including demographics and neuropsychiatric variables, age showed the strongest validity on the classification in both univariate and multivariate models (Figure 3B).

When mapping the validity indices on the two dimensions of the validity criterion, measures were dispersed as having relatively higher and lower validity in each context (Figure 4). In the univariate approach, the measures did not represent specific validity, and multiple measures reflected both of the validity criteria. When examined as a multivariate approach, the Word List Recognition score showed the highest validity in both contexts of validity indices (Figure 4B). Furthermore, TMT-A and the initial performance of animal fluency (Flu1; performance in 1-15s) showed high contributions in predicting dementia severity, while constructional praxis and relatively later phases of animal fluency (Flu2, 4; performances in 16-30s, 46-60s) showed higher discriminating validity. Boston Naming Test, Trail Making B, Word List, and interim phase of animal fluency (Flu3; performance in 31-45s) showed minimal validity indices under both validity criteria. On the contrary, the univariate validity index shows that TMT-A has superior validity over other tests when examined as a univariate approach (Figure 4A). Similarly, the univariate approach was less dependent on the type of validity, showing that tests were high in both of the validity types.

Lastly, the full regression model that included all of the predictors showed moderate levels of correspondence (adjusted $R^2 = 0.588$, AUC = 0.680; Table 2).

Discussion

The current study examined the validity of individual measures that comprise neuropsychological assessment of dementia. The validity of assessment measures was evaluated in predicting dementia severity and differentiating dementia types (AD versus VD). The result generally showed that tests with high correspondence to the validity criterion as a single measure also showed high unique contributions among a whole assessment battery, indicating that univariate validity partly reflects multivariate validity. However, there were also notable discrepancies in the validity indices. Although the test has shown a strong association with the

Table 2. *Multivariate Models that Explain the Clinical Outcome of Dementia Severity (CDR-SB) and Dementia Type (AD vs. VD).*

Outcome	Dementia severity (CDR-SB)			Dementia type (AD < VD)		
(Dependent variable)	B	SE	*p*-value	B	SE	*p*-value
Age	0.013	0.006	.032	-0.028	0.010	.005
Education	0.089	0.010	$< 10^{-16}$	0.063	0.016	$< 8 \times 10^{-5}$
NPI	0.186	0.009	$< 10^{-16}$	0.013	0.013	.307
GDS	-0.032	0.011	.003	0.018	0.019	.336
WR 1	-0.042	0.013	.002	0.030	0.023	.185
WR 2	-0.156	0.030	2×10^{-7}	0.030	0.060	.614
Wrecg	-0.161	0.017	$< 10^{-16}$	0.074	0.027	.007
Cons 1	-0.142	0.024	3×10^{-9}	-0.124	0.038	.001
Cons 2	-0.071	0.020	3×10^{-4}	0.079	0.041	.053
TMT-A	0.005	0.001	$< 10^{-16}$	0.002	0.001	.007
TMT-B	0.000	0.001	.831	-0.001	0.002	.593
Fluency 1	-0.164	0.022	3×10^{-13}	-0.061	0.040	.128
Fluency 2	-0.101	0.028	4×10^{-4}	-0.113	0.058	.052
Fluency 3	-0.071	0.033	.031	-0.009	0.063	.884
Fluency 4	-0.099	0.033	.003	-0.231	0.077	.003
Boston	-0.073	0.017	2×10^{-5}	0.013	0.029	.652
N		2,553			1,528	
Full model		Adjusted $R^2 = 0.588$			AUC = 0.680	

Note. CDR-SB = Clinical rating scale-sum of boxes; AD = dementia of Alzheimer's disease; VD = Vascular dementia; WR1/2 = Word list immediate/delayed recall; Wrecg = Word list recognition; Cons1/2 = Constructional praxis copy/delayed; TMT-A/B = Trail making test A and B; Flu1-4 = Animal fluency (four blocks of 15s interval); Boston = Boston naming test; NPI = Neuropsychiatric Inventory; GDS = Geriatric Depression Scale.

criterion as a single score, fewer test scores remained valid in the multivariate model. Specifically, Word List Recognition showed the highest multivariate validity in both of the clinical criteria, whereas other subtest measures including TMT, construction, and fluency tests showed validity under either specific validity criteria (i.e., dementia severity or differential diagnosis of AD and VD).

One of the main purposes of the study was to examine whether the validity metric of assessment measures depends on the framework of univariate or multivariate testing. When comparing the validity index between univariate and multivariate approaches, there was a tendency for a corresponding pattern. In other words, the test measures with high accountability as a single test tend to contain larger unique information among the multiple regression model that accounts for the validity criterion. For example, the Word List memory test showed favorable validity in both univariate and multivariable validity, and the subsequent measures tend to follow the ranks correspondingly. Based on the high proportional weight of AD, verbal memory measures were also indicative of functional impairment due to neurological disease (Belleville et al., 2017).

However, there are several notable discrepancies that the ordering of univariate importance does not map into multivariate importance. In the criterion of dementia severity (Validity A), Word Recall (Immediate) and Boston Naming Test scores showed a moderate level of bivariate association with the CDR-SB, whereas their accountability became nullified among the total battery set. In the criterion of dementia type (Validity B), the discrepancy was profound in TMT-A and Word Recall delayed scores. This indicates that the Word Recall (Immediate) and BNT measures contained redundant information that was mostly shared by other test measures in predicting AD spectrum severity.

Another notable finding was the unique contribution of NPI in predicting dementia severity. NPI qualitatively differs from other cognitive test performances in that the measured domain of socio-affective function is distinct from classical neurocognitive domains and that the source of information comes from the behavioral disturbances observed by clinicians, caregivers, and informants (Delgado et al., 2019; Sachdev et al., 2014). Such uniqueness of the information may have led to relatively higher multivariate validity compared to the univariate index. Our findings support an indispensable role of acquiring neuropsychiatric symptoms in characterizing the progression of dementia that are not replaced with classical cognitive tests (Ismail et al., 2016, 2017). Although the significance of NPI in the assessment of dementia severity is not a novel finding itself, the conspicuous discrepancy between the univariate and multivariate indices indicates an irreplaceable value the NPI measure can provide.

It was also examined whether the test measures can effectively distinguish prevalent types of dementia. Previous findings have shown that VD is more sensitively detected by tests of frontal lobe function or executive/speed domain under the time-limited protocol, whereas AD is more subject to specific memory processing which in turn leads to noticeable differences in subtest validity indices (Jang et al., 2017; Mathias & Burke, 2009; Oosterman & Scherder, 2006; Vasquez & Zakzanis, 2015). In the current evaluation of diagnostic validity, well-documented measures of episodic memory and executive/speed function also showed favorable multivariate validity. For example, the low score in Word List recognition was specific to the deficit in AD, and the low fluency test score was specific to VD, which was indicated by indices of high multivariate validity.

Despite the general alignment with the previous findings, however, some of the inconsistent findings need detailed discussion. The notable finding was shown in BNT score and TMT-A. In the previous meta-analysis of the neurocognitive difference between AD and VD, the picture naming test showed a moderated effect size (d = -0.4) (Mathias & Burke, 2009) but our findings showed that the picture naming test (BNT) included in the CERAD-K battery was moderately valid only as a single test score but not as a combined element among the battery. This rather discordant finding may be due to the way a test contributes to the prediction as a common or specific variance. The components of expressive language function reflected in the picture naming test may be more abundantly measured from the fluency test, leaving the test utility redundant (Greenaway et al., 2009). It is also possible that the subtest included in the BNT was more reflective of content-based semantic knowledge rather than process-based efficiency (Ackerman, 2022), and this tendency may be especially so in the population of a wide range of education levels (Kim et al., 2017).

Another notable measure in AD-VD differing measure was

TMT-A, which showed minimalized multivariate validity contrary to univariate validity in distinguishing AD from VD. This result contrast with the previous literatures that notes TMT as useful measure in detecting presence of subcortical or frontal lesions (Bagnoli et al., 2012; Ghafar et al., 2019). Since the previously summarized meta-analysis has examined the test measure as a single predictor, the finding may not generalize to the validity under the constituent of the whole neuropsychological battery. These contrasting results indicate that test measures with similar univariate utility may show disagreeing multivariate utility when composed as a total battery set.

Overall, the current study is suggestive of the cautious perspective in interpreting the validity evidence of neuropsychological measures. Most of the widespread research design reports the group comparison result between AD versus MCI, AD versus healthy controls, or AD versus other dementia types which in turn aggregated as univariate meta-analytic analyses. Though the univariate validity evidence is intuitive in determining the utility of a test, we have suggested that some of the instruments do not retain their validity as a comprehensive set. In practice, clinicians acknowledge the inter-mixture of the information provided by each measure but the decisions regarding the selection and construction of neuropsychological battery heavily rely on the qualitative aptness of the clinician. Our framework suggests that such clinical decision-making of battery construction can be aided by more direct quantification of test validity. This actuarial approach may not always coincide with clinical intuition but can buttress the possible human bias made under univariate research findings.

There are several limitations that require future investigation. First, there are varying sets of neuropsychological batteries used for the assessment of dementia, and the current study lacked some of the popular instruments that have shown robust validity (e.g., Digit Symbol Substitution, vegetable/fruit fluency, logical memory). The indices of multivariate indices can be easily affected by the presence of homologous tests with high collinearity, and adding these tests can alter the main conclusions. Since the CERAD battery was initially developed for diagnosing AD with additional extensions of executive/speed subtests (Seo et al., 2006, 2008), more flexible utilization of the battery may be required in identifying other dementia types.

Another critical limitation of the study was the scope of validity criterion that targets dementia types. There are other varying types of dementia that require differential diagnosis in practice other than AD and VD, and current examinations suggest little information regarding the validity criterion in other contexts of differential diagnosis. Since the current study only compared the two types of dementia, the index of Validity B may or may not reflect the pathology of certain dementia pathology. This issue is also relevant in that the index of Validity A more likely to reflect the severity of the AD spectrum rather than the severity of general dementing pathology. The populational characteristics and proportions of dementia types that the current dataset covers can affect the relative weight of the validity criterion toward specific diagnoses. Moreover, the resulting Validity B may only reflect the prototypic correspondence to the single category side of AD (i.e., presence or absence of AD-specific pattern) rather than reflecting the effect of VD. Further data on other dementia types should be integrated in order to develop the validity index that differentiates 'unspecified non-AD' or 'specific non-AD.'

Lastly, another remaining issue of test selection regards that the validity index does not fully represent the cost-efficiency of neuropsychological measures. Indeed, the number of administered tests is proportional to the time and effort required for the total assessment, and a test can be excluded from the battery if it does not provide incremental information regarding clinical criteria (Donders, 2020; Hunsley & Meyer, 2003). There are, however, there are more complex issues, in that not every measure requires the same amount of time and effort to administer. Some test produces a score with minimal time while some test measures require anteceding procedures (e.g., delayed recall and recognition). In the case of NPI, the single sum score may require a huge amount of cost including the semi-structured interview with an informant while the Boston Naming Test requires a shorter time to administer, which suggests that the measures should not be compared on an equal starting line. Thus, a test can be justified as valid if it requires minimal cost, whereas a costly test should prove its expensive utility accordingly.

Author contributions statement

SK, Formal analysis, writing-original draft; KYK, SMP, and DYO, Data curation, supervision; HK and DL, Data curation; JL, Supervision, funding acquisition, investigation.

References

Ackerman, P. L. (2022). Intelligence process vs. content and academic performance: a trip through a house of mirrors. *Journal of Intelligence, 10*, 128. https://doi.org/10.3390/jintelligence10040128

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., Gamst, A., Holtzman, D. M., Jagust, W. J., Petersen, R. C., Snyder, P. J., Carrillo, M. C., Thies, B., & Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia, 7*, 270-279. https://doi.org/10.1016/j.jalz.2011.03.008

Anastasi, A. (1950). The concept of validity in the interpretation of test scores. *Educational and Psychological Measurement*. https://doi.org/10.1177/001316445001000105

Bae, J. N., & Cho, M. J. (2004). Development of the Korean version of the Geriatric Depression Scale and its short form among elderly psychiatric patients. *Journal of Psychosomatic Research, 57*, 297-305. https://doi.org/10.1016/j.jpsychores.2004.01.004

Bagnoli, S., Failli, Y., Piaceri, I., Rinnoci, V., Bessi, V., Tedde, A., Nacmias, B., & Sorbi, S. (2012). Suitability of neuropsychological tests in patients with vascular dementia (VaD). *Journal of the Neurological Sciences, 322*, 41-45. https://doi.org/10.1016/J.JNS.2012.05.045

Belleville, S., Fouquet, C., Hudon, C., Zomahoun, H. T. V., & Croteau, J. (2017). Neuropsychological measures that predict progression from mild cognitive impairment to Alzheimer's type dementia in older adults: A systematic review and meta-analysis. *Neuropsychology Review, 27*, 328-353. https://doi.org/10.1007/S11065-017-9361-5

Bondi, M. W., & Smith, G. E. (2014). Mild cognitive impairment: a concept and diagnostic entity in need of input from neuropsychology. *Journal of the International Neuropsychological Society, 20*, 129-134. https://doi.org/10.1017/S1355617714000010

Bruun, M., Rhodius-Meester, H. F. M., Koikkalainen, J., Baroni, M., Gjerum, L., Lemstra, A. W., Barkhof, F., Remes, A. M., Urhemaa, T., Tolonen, A., Rueckert, D., Gils, M., Frederiksen, K. S., Waldemar, G., Scheltens, P., Mecocci, P., Soininen, H., Lötjönen, J., Hasselbalch, S. G., … van der Flier, W. M. (2018). Evaluating combinations of diagnostic tests to discriminate different dementia types. *Alzheimer's & dementia: Diagnosis, Assessment & Disease Monitoring, 10*, 509-518. https://doi.org/10.1016/j.dadm.2018.07.003

Carlew, A. R., Kaser, A., Schaffert, J., Goette, W., Lacritz, L., & Rossetti, H. (2023). A critical review of neuropsychological actuarial criteria for mild cognitive impairment. *Journal of Alzheimer's Disease, 91*, 169-182. https://doi.org/10.3233/JAD-220805

Chapman, R. M., Mapstone, M., Porsteinsson, A. P., Gardner, M. N., McCrary, J. W., DeGrush, E., Reilly, L. A., Sandoval, T. C., & Guillily, M. D. (2010). Diagnosis of Alzheimer's disease using neuropsychological testing improved by multivariate analyses. *Journal of Clinical and Experimental Neuropsychology, 32*, 793-808. https://doi.org/10.1080/13803390903540315

Choi, S. H., Na, D. L., Kwon, H. M., Yoon, S. J., Jeong, J. H., & Ha, C. K. (2000). The Korean version of the Neuropsychiatric Inventory: A scoring tool for neuropsychiatric disturbance in dementia patients. *Journal of Korean Medical Science, 15*, 609-615. https://doi.org/10.3346/jkms.2000.15.6.609

Cummings, J. L., Mega, M., Gray, K., Rosenberg-Thompson, S., Carusi, D. A., & Gornbein, J. (1994). The neuropsychiatric inventory: Comprehensive assessment of psychopathology in dementia. *Neurology, 44*, 2308-2314. https://doi.org/10.1212/wnl.44.12.2308

Delgado, C., Vergara, R. C., Martínez, M., Musa, G., Henríquez, F., & Slachevsky, A. (2019). Neuropsychiatric symptoms in Alzheimer's disease are the main determinants of functional impairment in advanced everyday activities. *Journal of Alzheimer's Disease, 67*, 381-392. https://doi.org/10.3233/JAD-180771

Donders, J. (2020). The incremental value of neuropsychological assessment: A critical review. *Clinical Neuropsychologist, 34*, 56-87. https://doi.org/10.1080/13854046.2019.1575471

Elahi, F. M., & Miller, B. L. (2017). A clinicopathological approach to the diagnosis of dementia. *Nature Reviews Neurology, 13*, 457-476. https://doi.org/10.1038/nrneurol.2017.96

Fields, J. A., Ferman, T. J., Boeve, B. F., & Smith, G. E. (2011). Neuropsychological assessment of patients with dementing illness. *Nature Reviews Neurology, 7*, 677-687. https://doi.org/10.1038/nrneurol.2011.173

Fields, J. A., Machulda, M., Aakre, J., Ivnik, R. J., Boeve, B. F., Knopman, D. S., Petersen, R. C., & Smith, G. E. (2010). Utility of the DRS for predicting problems in day-to-day functioning. *The Clinical Neuropsychologist, 24*, 1167-1180. https://doi.org/10.1080/13854046.2010.514865

Fillenbaum, G. G., Peterson, B., & Morris, J. C. (1996). Estimating the validity of the Clinical Dementia Rating scale: The CERAD experience. *Aging, 8*, 379-385. https://doi.org/10.1007/bf03339599

Fountain-Zaragoza, S., Braun, S. E., Horner, M. D., & Benitez, A. (2021). Comparison of conventional and actuarial neuropsychological criteria for mild cognitive impairment in a clinical setting. *Journal of Clinical and Experimental Neuropsychology, 43*, 753-765. https://doi.org/10.1080/13803395.2021.2007857

Garb, H. N., & Schramke, C. J. (1996). Judgment research and neuropsychological assessment: A narrative review and meta-analyses. *Psychological Bulletin, 120*, 140-153. https://doi.org/10.1037/0033-2909.120.1.140

Ghafar, M. Z. A. A., Miptah, H. N., & O'Caoimh, R. (2019). Cognitive screening instruments to identify vascular cognitive impairment: A systematic review. *International Journal of Geriatric Psychiatry, 34*, 1114-1127. https://doi.org/10.1002/gps.5136

Graves, L. V., Edmonds, E. C., Thomas, K. R., Weigand, A. J., Cooper, S., Bondi, M. W., & Loewenstein, D. (2020). Evidence for the utility of actuarial neuropsychological criteria across the continuum of normal aging, mild cognitive impairment, and dementia. *Journal of Alzheimer's Disease, 78*, 371-386. https://doi.org/10.3233/JAD-200778

Greenaway, M. C., Smith, G. E., Tangalos, E. G., Geda, Y. E., & Ivnik, R. J. (2009). Mayo older americans normative studies: factor analysis of an expanded neuropsychological battery. *The Clinical Neuropsychologist, 23*, 7-20. https://doi.org/10.1080/13854040801891686

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment, 15*, 446-455. https://doi.org/10.1037/1040-3590.15.4.446

Ismail, Z., Elbayoumi, H., Fischer, C. E., Hogan, D. B., Millikin, C. P., Schweizer, T., Mortby, M. E., Smith, E. E., Patten, S. B., & Fiest, K. M. (2017). Prevalence of depression in patients with mild cognitive impairment: A systematic review and meta-analysis. *JAMA Psychiatry, 74*, 58-67. https://doi.org/10.1001/JAMAPSYCHIATRY.2016.3162

Ismail, Z., Smith, E. E., Geda, Y., Sultzer, D., Brodaty, H., Smith, G., Agüera-Ortiz, L., Sweet, R., Miller, D., & Lyketsos, C. G. (2016). Neuropsychiatric symptoms as early manifestations of emergent dementia: Provisional diagnostic criteria for mild behavioral impairment. *Alzheimer's & Dementia, 12*, 195-202. https://doi.org/10.1016/j.jalz.2015.05.017

Jang, H., Ye, B. S., Woo, S., Kim, S. W., Chin, J., Choi, S. H., Jeong, J. H., Yoon, S. J., Yoon, B., Park, K. W., Hong, Y. J., Kim, H. J., Lockhart, S. N., Na, D. L., & Seo, S. W. (2017). Prediction model of conversion to dementia risk in subjects with amnestic mild cognitive impairment: A longitudinal, multi-center clinic-based study. *Journal of Alzheimer's Disease*. https://doi.org/10.3233/JAD-170507

Kim, B. S., Lee, D. W., Bae, J. N., Kim, J. H., Kim, S., Kim, K. W., Park, J. E., Cho, M. J., & Chang, S. M. (2017). Effects of education on differential item functioning on the 15-item modified Korean version of the Boston Naming Test. *Psychiatry Investigation, 14*, 126. https://doi.org/10.4306/pi.2017.14.2.126

Kwak, S., Oh, D. J., Jeon, Y. J., Oh, D. Y., Park, S. M., Kim, H., & Lee, J. Y. (2022). Utility of machine learning approach with neuropsychological tests in predicting functional impairment of Alzheimer's disease. *Journal of Alzheimer's Disease, 85*, 1357-1372. https://doi.org/10.3233/JAD-215244

Kwak, S., Park, S. M., Jeon, Y. J., Ko, H., Oh, D. J., & Lee, J. Y. (2021). Multiple cognitive and behavioral factors link association between brain structure and functional impairment of daily instrumental activities in older adults. *Journal of the International Neuropsychological Society*, 1-14. https://doi.org/10.1017/S1355617721000916

Lee, J. H., Lee, K. U., Lee, D. Y., Kim, K. W., Jhoo, J. H., Kim, J. H., Lee, K. H., Kim, S. Y., Han, S. H., & Woo, J. I. (2002). Development of the Korean version of the Consortium to Establish a Registry for Alzheimer's Disease Assessment Packet (CERAD-K): Clinical and neuropsychological assessment batteries. *Journals of Gerontology - Series B Psychological Sciences and Social Sciences, 57*, 47-53. https://doi.org/10.1093/geronb/57.1.P47

Lynch, C. A., Walsh, C., Blanco, A., Moran, M., Coen, R. F., Walsh, J. B., & Lawlor, B. A. (2005). The clinical dementia rating sum of box score in mild dementia. *Dementia and Geriatric Cognitive Disorders, 21*, 40-43. https://doi.org/10.1159/000089218

Mathias, J. L., & Burke, J. (2009). Cognitive functioning in Alzheimer's and vascular dementia: A meta-analysis. *Neuropsychology, 23*, 411-423. https://doi.org/10.1037/a0015384

McKhann, G. M., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology, 34*, 939-944. https://doi.org/10.1212/WNL.34.7.939

Morris, J. C., Ernesto, C., Schafer, K., Coats, M., Leon, S., Sano, M., Thal, L. J., & Woodbury, P. (1997). Clinical dementia rating training and reliability in multicenter studies: The Alzheimer's Disease Cooperative Study experience. *Neurology, 48*, 1508-1510. https://doi.org/10.1212/WNL.48.6.1508

Morris, J. C. (1997). Clinical dementia rating: a reliable and valid diagnostic and staging measure for dementia of the Alzheimer type. *International Psychogeriatrics, 9*, 173-176. https://doi.org/10.1017/S1041610297004870

Nation, D. A., Ho, J. K., Dutt, S., Han, S. D., Lai, M. H. C., & Alzheimer's Disease Neuroimaging Initiative. (2019). Neuropsychological decline improves prediction of dementia beyond Alzheimer's disease biomarker and mild cognitive impairment diagnoses. *Journal of Alzheimer's Disease, 69*, 1-12. https://doi.org/10.3233/JAD-180525

Nyenhuis, D. L., Gorelick, P. B., Geenen, E. J., Smith, C. A., Gencheva, E., Freels, S., & DeToledo-Morrell, L. (2004). The pattern of neuropsychological deficits in Vascular Cognitive Impairment-No Dementia (Vascular CIND). *The Clinical Neuropsychologist, 18*, 41-49. https://doi.org/10.1080/13854040490507145

O'Bryant, S. E. (2008). Staging dementia using clinical dementia rating scale sum of boxes scores. *Archives of Neurology, 65*(8), 1091. https://doi.org/10.1001/archneur.65.8.1091

Oosterman, J. M., & Scherder, E. J. A. (2006). Distinguishing between Vascular dementia and Alzheimer's disease by means of the WAIS: A meta-analysis. *Journal of Clinical and Experimental Neuropsychology, 28*, 1158-1175. https://doi.org/10.1080/13803390500263543

Román, G. C., Tatemichi, T. K., Erkinjuntti, T., Cummings, J. L., Masdeu, J. C., Garcia, J. H., Amaducci, L., Orgogozo, J. M., Brun, A., Hofman, A., Moody, D. M., O'Brien, M. D., Yamaguchi, T., Grafman, J., Drayer, B. P., Bennett, D. A., Fisher, M., Ogata, J… Scheinberg, P. (1993). Vascular dementia. *Neurology, 43*, 250. https://doi.org/10.1212/WNL.43.2.250

Russell, E. W., Russell, S. L. K., & Hill, B. D. (2005). The fundamental psychometric status of neuropsychological batteries. *Archives of Clinical Neuropsychology, 20*, 785-794. https://doi.org/10.1016/J.ACN.2005.05.001

Sachdev, P. S., Blacker, D., Blazer, D. G., Ganguli, M., Jeste, D. V., Paulsen, J. S., & Petersen, R. C. (2014). Classifying neurocognitive disorders: The DSM-5 approach. *Nature Reviews Neurology, 10*, 634-642. https://doi.org/10.1038/nrneurol.2014.181

Seo, E. H., Lee, D. Y., Choo, I. H., Kim, S. G., Kim, K. W., Youn, J. C., Jhoo, J. H., & Woo, J. I. (2008). Normative study of the Stroop Color and Word Test in an educationally diverse elderly population. *International Journal of Geriatric Psychiatry, 23*, 1020-1027. https://doi.org/10.1002/gps.2027

Seo, E. H., Lee, D. Y., Kim, K. W., Lee, J. H., Jhoo, J. H., Youn, J. C., Choo, I. H., Ha, J., & Woo, J. I. (2006). A normative study of the Trail Making Test in Korean elders. *International Journal of Geriatric Psychiatry, 21*, 844-852. https://doi.org/10.1002/gps.1570

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

Stallard, E., Kociolek, A., Jin, Z., Ryu, H., Lee, S., Cosentino, S., Zhu, C., Gu, Y., Fernandez, K., Hernandez, M., Kinosian, B., Stern, Y., & Peters, J. J. (2022). Validation of a multivariate prediction model of the clinical progression of Alzheimer's disease in a community-dwelling multiethnic cohort. *Med Rxiv*, 2022.06.28.22277006. https://doi.org/10.1101/2022.06.28.22277006

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5*, 1-25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

Vakil, E. (2012). Neuropsychological assessment: Principles, rationale, and challenges. *Journal of Clinical and Experimental Neuro-Psychology, 34*, 135-150. https://doi.org/10.1080/13803395.2011.623121

Vasquez, B. P., & Zakzanis, K. K. (2015). The neuropsychological profile of vascular cognitive impairment not demented: A meta-analysis. *Journal of Neuropsychology, 9*, 109-136. https://doi.org/10.1111/jnp.12039