Maastricht University

A large-scale cross-sectional and longitudinal study into the ecological validity of neuropsychological test measures in neurologically intact people

Please check the document version of this publication:

A large-scale cross-sectional and longitudinal study into the ecological validity of neuropsychological test measures in neurologically intact people

Wim Van der Elst [a,*], Martin P.J. Van Boxtel [a],
Gerard J.P. Van Breukelen [b], Jelle Jolles [a]

[a] *Maastricht Brain and Behavior Institute, European Graduate School of Neuroscience (EURON), Faculty of Health,
Medicine and Life Sciences, Maastricht University, The Netherlands*
[b] *Department of Methodology and Statistics, Maastricht University, The Netherlands*

Abstract

It is often assumed that neuropsychological measures are ecologically valid in 'normal' people, but this assumption has not yet been thoroughly evaluated.

The aim of the present study was to evaluate the cross-sectional and longitudinal ecological validity of individual neuropsychological test scores and their composites in a large sample of neurologically intact people. Three neuropsychological composite measures were established, i.e. a "Memory Quotient", an "Executive functioning and Speed Quotient", and a "General Cognitive Quotient". The ecological validity of the individual neuropsychological measures and their composites was low to moderate. Multivariate models that included both neuropsychological and non-cognitive variables (i.e. demographic variables, depressive symptoms and anxiety) accounted for 4.6–21.4% of the variance in daily life functioning. The General Cognitive Quotient was the neuropsychological measure that was the most consistently related to daily life functioning.
© 2008 National Academy of Neuropsychology. Published by Elsevier Ltd. All rights reserved.

Keywords: Ecological validity; Assessment; Neuropsychology; Older adults; Longitudinal design

1. Introduction

Historically, neuropsychological tests have mainly been used to detect and localize brain pathology (Long & Kibby, 1995). As brain imaging techniques became more widely available in recent decades, the role of neuropsychological assessment in the diagnosis of neuropathology has gradually diminished (Johnstone & Frank, 1995; Rabin, Burton, & Barr, 2007). Today, referral questions in clinical neuropsychology are increasingly more focused on the functional implications of brain damage, such as whether or not a patient is able to follow a rehabilitation program, live independently, or return to work (Rabin, Barr, & Burton, 2005). The degree to which neuropsychological tests can make accurate predictions of a person's behavior in real-world settings has been defined as the ecological validity of a test (Sbordone, 1996).

---

* Corresponding author at: Faculty of Health, Medicine and Life Sciences, Department of Psychiatry and Neuropsychology, Maastricht University, 6200 MD, Maastricht, The Netherlands. Tel.: +31 43 3884086; fax: +31 43 3884092.

E-mail address: Wim_vanderelst@hotmail.com (W. Van der Elst).

Various studies have evaluated the ecological validity of neuropsychological tests. For example, Chaytor, Schmitter-Edgecombe, and Burr (2006) evaluated the relationship between tests that measure executive functioning (such as the Trail Making Test) and everyday executive ability (as measured with informant ratings) in a sample of 46 mixed neurological adult patients. The authors found that the neuropsychological test scores accounted for about 20% of the variance in the measures of everyday executive ability. Farias, Harrell, Neumann, and Houtz (2003) evaluated the ecological validity of composite neuropsychological test scores in a sample of 42 people with Alzheimer's disease. The results showed that these neuropsychological composites accounted for about 20–50% of the variance in performance-based measures of daily living skills (e.g. dialing a telephone, counting currency), and for about 10–30% of the variance in informant-based functional ratings. These and similar studies suggested that the ecological validity of neuropsychological test scores is moderate, at least in clinical populations (for comprehensive reviews refer to Chaytor and Schmitter-Edgecombe (2003), Green, Kern, and Heaton (2004), and Kalechstein, Newton, and Van Gorp (2003)).

Until now, the ecological validity of neuropsychological instruments has not been carefully considered in 'neurologically intact' people (Spooner & Pachana, 2006). This is not surprising, as neuropsychology has traditionally been a discipline with a strong focus on neurologically impaired people. The ecological validity of neuropsychological tests in neurologically intact individuals, however, is an important issue that may have implications for studies in neuropsychology and related fields. For example, studies that evaluate the impact of pharmacological manipulations (such as tryptophan depletion; Evers, Van der Veen, Jolles, Deutz, & Schmitt, 2006), illegal drugs (such as marijuana; Ramaekers et al., 2006), medical variables (such as hypertension; Van Boxtel et al., 2006), or nutritional components (such as fatty acids; De Groot, Hornstra, & Jolles, 2007) on the cognitive abilities of neurologically intact people would be of limited value if the neuropsychological test scores were not related to real-world functioning.

The aim of the present study was to evaluate the cross-sectional and the longitudinal ecological validity of neuropsychological measures in neurologically intact mid-aged to older adults (aged between 49 and 81 years at baseline). Neuropsychological measures (assessed at baseline) were related to daily life functioning as assessed at baseline (i.e. cross-sectional ecological validity) and at several follow-up moments (i.e. longitudinal ecological validity). Both individual neuropsychological test scores and their composites were related to daily life functioning because previous studies have suggested that composite scores had higher ecological validity as compared to individual neuropsychological test scores (see the reviews of Chaytor & Schmitter-Edgecombe, 2003; Green et al., 2004; Kalechstein et al., 2003). As previous studies have suggested that demographical variables (Kalechstein et al., 2003), depressive symptoms (Chaytor, Temkin, Machamer, & Dikmen, 2007), and anxiety (Sbordone, 1996) may account for substantial parts of the variance in daily life measures (in addition to neuropsychological functioning), we also evaluated the contribution of these non-cognitive variables in the prediction of daily life functioning.

2. Method

2.1. Participants

Data were derived from the Maastricht Aging Study (MAAS), a prospective study on the determinants of cognitive aging. MAAS baseline measurements were conducted between 1993 and 1996 and involved four panels of approximately 465 people each (1856 individuals in total). The data were collected in four smaller panels instead of in a single large panel for logistic reasons. All participants were community dwelling, healthy people (aged between 24 and 81 years at baseline) who were without documented medical conditions that could interfere with normal cognitive functioning (i.e. individuals with chronic neurological pathology, mental retardation, psychopathology, or chronic psychotropic drug use were excluded). For the present study, we only included people who were aged 49 years or older at baseline because the follow-up frequency in the MAAS differed as a function of baseline age.

At baseline, all participants were administered the Stroop Color-Word Test (SCWT; Stroop, 1935), Concept Shifting Test (CST; Van der Elst, Van Boxtel, Van Breukelen, & Jolles, 2006a), Letter Digit Substitution Test (LDST; Van der Elst, Van Boxtel, Van Breukelen, & Jolles, 2006b), Verbal Learning Test of Rey (VLT; Van der Elst, Van Boxtel, Van Breukelen, & Jolles, 2005), Mini-Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 1975), and the depression and anxiety scales of the Symptom Checklist 90 (SCL-90; Derogatis, 1977). Data of people with

Table 1
Descriptive data (*N*, mean (*M*), standard deviation (S.D.), and score range) for the neuropsychological measures, depressive symptoms and anxiety, and the outcome measures at the different measurement moments

Measurement moment	Variable	N	M	S.D.	Score range
Baseline	SCWT Time	859	103.85	25.55	53.6–223.8
	SCWT Errors	859	1.51	2.39	0–22
	CST Shifting	859	14.62	12.63	−9.8 to 121.40
	CST Errors	859	.67	1.13	0–8
	VLT Total recall 1–3	859	21.76	5.38	9–39
	VLT Delayed recall	859	9.07	2.86	1–15
	LDST 60 s	859	29.06	6.76	10–50
	Depressive symptoms	859	20.96	6.31	16–62
	Anxiety	832	12.72	4.24	10–47
	SF-36 PF	442	25.76	3.91	12–30
	CFQ	579	31.84	10.90	2–73
	DECO		Not administered		
First follow-up	SF-36 PF	614	25.43	4.25	10–30
	CFQ	585	30.79	11.42	2–76
	DECO	612	34.99	4.25	5–38
Second follow-up	SF-36 PF	608	24.88	4.62	10–30
	CFQ		Not administered		
	DECO	559	35.11	3.61	17–38
Third follow-up	SF-36 PF	552	24.59	4.78	10–30
	CFQ		Not administered		
	DECO	600	34.70	4.25	10–38

Note. SCWT, Stroop Color-Word Test; CST, Concept Shifting Test; VLT, Verbal Learning Test; LDST, Letter Digit Substitution Test; SF-36 PF, Physical Functioning scale of the Medical Outcomes Study 36-Item Short Form Health Survey; CFQ, Cognitive Failures Questionnaire; DECO, Détérioration Cognitif Observée (*observed deterioration of cognitive functioning*).

missing values in the cognitive measures or with MMSE scores below 24 (a commonly employed cut-off to screen for dementia) were excluded from the analyses. The baseline data of $n = 859$ people were eligible for use in the analyses.

Daily life functioning was measured with self-report and informant questionnaires. The Medical Outcomes Study 36-Item Short Form Health Survey (SF-36; Van der Zee & Sanderman, 1993; Ware & Sherbourne, 1992) was given to two of the four panels of MAAS at baseline ($n = 442$) and to all four panels at the first, second, and third follow-up occasions ($n = 614$, $n = 608$ and $n = 552$, respectively). The test–retest interval between each of the four consecutive measurement moments was 3 years. The Détérioration Cognitif Observée (in English: *observed deterioration of cognitive functioning*) (DECO; Ritchie & Fuhrer, 1996) was given to all four panels at the first ($n = 612$), second ($n = 559$), and third ($n = 600$) follow-up moments, but was not administered to any of the panels at baseline. The Cognitive Failures Questionnaire (CFQ; Broadbent, Cooper, FitzGerald, & Parkes, 1982) was administered to three panels at baseline ($n = 579$) and to four panels at the first follow-up moment ($n = 585$).

The ethnic background of all participants was Caucasian, and all participants were native Dutch speakers. At baseline, 38.4% of the sample was aged between 49 and 59 years, 35.2% was aged between 60 and 69 years, and 26.4% was aged 70+ years. Level of education (LE) was measured by classifying the formal schooling of participants in one of three groups, namely those with at most primary education (LE low; 46.9% of the total sample at baseline), those with at most junior vocational training or high school (LE average; 36.0% of the sample), and those with at most senior vocational or academic training (LE high; 17.1% of the sample). The proportion of males and females was approximately equal (with 48.9% males). The mean (S.D.) IQ of the participants was 115.6 (13.2) at baseline (IQ was measured with the shortened version of the Groningen Intelligence Test; Luteijn & Van der Ploeg, 1983). Table 1 provides descriptive data of the neuropsychological test scores, the depression and anxiety measures, and the daily life measures per measurement moment. The medical ethics committee of Maastricht University approved the study and all participants gave their informed consent.

2.2. Measures

2.2.1. Neuropsychological tests

The SCWT, CST, VLT and LDST were administered individually at the neuropsychological laboratory of the School for Mental Health and Neuroscience, University Maastricht (the Netherlands) by highly trained test assistants.

The SCWT is a measure of executive and related functions (Hammes, 1973; Van der Elst, Van Boxtel, Van Breukelen, & Jolles, 2006c). The three SCWT subtasks display names of colors (red, blue, yellow, green) in random order printed in black ink, solid color patches in one of these four basic colors, and color words printed in an incongruous ink color, respectively. The participants were instructed to read the words, name the colors, and name the ink color of the printed words as quickly and as accurately as possible. The time needed to complete the SCWT subtask three (measured in seconds) and the number of errors that were made in subtask three served as the outcome variables (SCWT Time and SCWT Errors, respectively).

The CST is a Trail Making Type test that measures concept shifting and related cognitive functions (Van der Elst et al., 2006a). In each of five parts, a sheet of paper is presented that contains 16 small circles arranged in a larger circle. The small circles contain digits (part A), letters (part B), both digits and letters (part C), or are empty (parts D and E). Participants were instructed to cross out, as quickly and as accurately as possible, the fixed randomly arranged numbers 1–16 in numerical order (part A), the fixed randomly arranged letters A–P in alphabetical order (part B), the fixed randomly arranged numbers alternating with letters in the appropriate order (1, A, 2, B etc.; part C), and the 16 empty circles in a clockwise fashion (parts D and E). The time needed to complete parts A, B and C of the CST is referred to as CSTa, CSTb, and CSTc, respectively. Outcomes measures were the Shifting score (CST shifting = CSTc − [CSTa + CSTb]/2) and the number of errors made in CST part C (CST Errors).

The VLT is a word learning test that measures diverse components of verbal memory (Schmidt, 1996; Van der Elst et al., 2005). Fifteen words were presented in a fixed order with an inter-stimulus interval of 1 s. As soon as the presentation stopped, the participants were asked to repeat as many words as possible in any order. This procedure was repeated four more times, after which there was a delay of about 20 min. Next, participants were again instructed to recall the words learned. Finally, a recognition trial was administered. Outcome variables were the total number of correctly recalled words summed over the first three trials (VLT Total recall 1–3), and the number of correctly recalled words after the 20-min delay (VLT Delayed recall).

The LDST is a substitution type test that measures general speed of information processing (Van der Elst et al., 2006b). In the LDST, participants were required to replace randomized letters with appropriate digits as quickly and accurately as possible. The number of correct substitutions made in 60 s served as the outcome variable (LDST 60 s).

2.2.2. Outcome measures

The participants were given the SF-36 and the CFQ self-report questionnaires. The SF-36 assesses various health dimensions. Among these dimensions is the SF-36 physical functioning (SF-36 PF) scale, a scale that is highly correlated with ADL measures (e.g. Martin, Irrgang, Burdett, Conti, & Van Swearingen, 2005; Reijneveld, Spijker, & Dijkshoorn, 2007) and which is sensitive for mild functional losses that affect independent living in older people (Anderson, Laubscher, & Burns, 1996). The SF-36 PF scale is composed of 10 items about mobility (e.g. lifting or carrying groceries) and self-care (e.g. bathing or dressing oneself) that are rated on a 3-point Likert scale (score range between 1 and 3). The SF-36 PF scale score range is between 10 and 30, with lower scale scores reflecting less functional independence and lower ADL capacities. The CFQ is an IADL questionnaire that consists of 29 items regarding the frequency of the occurrence of cognitive lapses in daily life (e.g. forgetting appointments). Each item is rated on a 5-point Likert scale (score range between 0 and 4). The CFQ total test score ranges between 0 and 116, with higher scores reflecting a higher frequency of cognitive failures in daily life.

In addition to the self-report questionnaires, the DECO was filled in by an informant. The DECO assesses changes that have occurred over the past year in a person's IADL functioning (Ritchie, Artero, & Touchon, 2001). The DECO consists of 19 items about changes that occurred in various activities of daily life (e.g. managing money). Items are rated on a 3-point Likert scale with a score range between 0 and 2. The total DECO score ranges between 0 and 38, with lower scores reflecting larger changes in a person's IADL.

Depressive symptoms and anxiety were measured with the depression and anxiety scales of the SCL-90 self-report questionnaire. The SCL-90 items are rated on a 5-point Likert scale with a score range between 1 and 5. The depression scale is composed of 16 items about depression (e.g. feeling lonely). The anxiety scale consists of 10 items about anxiety

(e.g. feeling afraid). The total depression and anxiety scale scores range between 16 and 80 and between 10 and 100, respectively (with higher scale scores reflecting more depressive symptoms and higher anxiety levels, respectively).

2.3. Statistical analyses

The seven neuropsychological test scores were subjected to Principal Components Analyses (PCAs). The first PCA was exploratory, using the eigenvalue >1.0 criterion (Guttman, 1954) and the scree plot (Cattell, 1966) to determine the number of components to be retained. The second PCA forced the extraction of one component, because different theories have suggested that all cognitive tests measure a general factor common to all cognitive tests (factor *g* in the intelligence and cognition literature; Spearman, 1927), as well as test-specific factors.

Next, the individual neuropsychological test scores, the composites, and the non-cognitive variables (demographics, depressive symptoms and anxiety) were correlated with the daily life measures. Such bivariate correlations are informative, but neuropsychologists usually base their judgment of an individual's everyday functioning on a combination of non-cognitive variables and multiple neuropsychological measures – rather than on a single neuropsychological measure. Linear mixed models (LMM) analyses were conducted to evaluate the contribution of the various predictors on daily life functioning. There were three LMM analyses for each daily life outcome measure: one with the seven individual neuropsychological test scores as predictors, another with the neuropsychological composites resulting from the exploratory PCA as predictors, and a third with the neuropsychological composite resulting from the 1-component PCA as a predictor. In addition to the neuropsychological test measures, demographic variables (age, age^2, gender, level of education), depressive symptoms, anxiety, time and all two-way interactions between these predictors and the three-way interactions involving time were included as predictors in the full models. Time was coded with three dummies (first follow-up, second follow-up and third follow-up) and baseline measurement as the reference category. Age was centred (age = calendar age − 65) before computing the quadratic terms and interactions to avoid multicollinearity (Marquardt, 1980). Gender was dummy coded with male = 1 and female = 0. LE was dummy coded with two dummies (LE low and LE high) and LE average as the reference category.

The full models were reduced in a stepwise hierarchical manner by eliminating the least significant predictor if its two-tailed *p*-value was above .01. No predictor was removed from the model as long as it was also included in a higher-order term in the model. All models were estimated with restricted maximum likelihood, rather than with maximum likelihood, because restricted maximum likelihood is better for estimating the covariance matrix and the standard errors of fixed effects. The mixed model assumed fixed effects of all predictors and an unstructured covariance matrix for the repeated measures. This is equivalent to the multivariate method for ANOVA of repeated measures for complete data, but unlike repeated measures ANOVA, LMM can handle missing data without requiring either imputation of missing values or list-wise deletion of persons with a missing value (Verbeke & Molenberghs, 2000). In the present study, 15.6%, 22.9% and 29.2% of the people who were tested at baseline dropped-out from the study at the first, second, and third follow-up moments, respectively. Previous studies with the MAAS data have shown that especially older and lower educated people tend to drop-out (Van Beijsterveldt et al., 2002; Van der Elst, Van Boxtel, Van Breukelen, & Jolles, 2008), but this is not a problem when LMM is used because this method is valid even if the drop-out pattern depends on the observed covariates or previous repeated measures (in contrast to repeated measures ANOVA) (Verbeke & Molenberghs, 2000). Assumptions of linearity, normality of the residuals (normal distribution of the residuals for the various measurement moments), and homoscedasticity (homogenous variance of the residuals over the range of the predicted scores at the different measurement occasions) were examined graphically and analytically for each final model. For technical details of LMM with time and between-subject variables, see Verbeke and Molenberghs (2000) and Van Breukelen (2006).

All analyses were conducted using SPSS 14.0 for Windows, with alpha = .01 for all analyses. A lower alpha level was chosen in order to avoid Type I errors due to multiple testing.

3. Results

3.1. Neuropsychological composites

Unlike the VLT and the LDST test scores, higher SCWT and CST test scores reflect worse performance. For this reason, the signs of the SCWT Time, SCWT Errors, CST Shifting and CST Errors scores were reversed (i.e. their

Table 2
Component loadings that were obtained by Principal Components Analyses in which two (left column) and one (right column) components were extracted

Test score (at baseline)	2-component solution		1-component solution
	Component I	Component II	Component I
SCWT Time	.27	**.61**	**.74**
SCWT Errors	−.11	**.60**	**.42**
CST Shifting	−.12	**.75**	**.54**
CST Errors	−.13	**.61**	**.41**
VLT Total recall 1–3	**.94**	−.07	**.72**
VLT Delayed recall	**.95**	−.11	**.70**
LDST 60 s	.30	**.58**	**.75**
Initial eigenvalues for each component	2.75	1.22	2.75
Percentage of variance accounted for	39.31%	17.40%	39.31%

Note. SCWT, Stroop Color-Word Test; CST, Concept Shifting Test; VLT, Verbal Learning Test; LDST, Letter Digit Substitution Test. The signs of the SCWT Time, SCWT Errors, CST Shifting and CST Errors scores were reversed. The 2-component solution that is presented in this table (left column) is promax-rotated. The correlation between the promax-rotated components of the 2-component solution equaled .37 ($p < .001$). Eigenvalues and percentage of variance explained refer to the unrotated components. Component loadings ≥ 40 are printed in boldface.

original values were multiplied by −1). The reversed SCWT Time, SCWT Errors, CST Shifting and CST Errors scores, and the raw LDST 60 s, VLT Total recall 1–3, and the VLT Delayed recall scores were subjected to PCAs. The eigenvalue >1.0 criterion suggested a 2-component solution, which was confirmed by inspection of the scree plot. These two components accounted for 56.7% of the total variance. The promax-rotated solution is presented in Table 2 (left column). Component I was defined by the VLT Total recall 1–3 and VLT Delayed recall scores, and can be labeled as the (verbal) "memory component". Component II was defined by the SCWT Time, SCWT Errors, CST Shifting, CST Errors and LDST 60 s scores, and can be termed the "executive functioning and speed component". A second PCA was conducted on the individual neuropsychological test scores in which only one component was extracted (see Table 2, right column). This "general cognition component" accounted for 39.3% of the total variance, and all individual neuropsychological test scores had loadings above .40 on this component.

The results of the PCAs were used to establish the neuropsychological composites. This required four steps. First, the test scores that had loadings of at least .40 on the component to-be-computed were converted into z-scores (z-score$_i$ = [raw test score − mean raw score in MAAS]/[S.D.(raw score)]) by means of Table 1 (note that the signs of the raw and the mean SCWT and CST scores have to be reversed). Second, these z-scores were multiplied with the component coefficients (c_i's) and aggregated (aggregated score$_i$ = $\Sigma[z$-score$_i \times c_i]$). Third, the aggregated scores were divided by the S.D.s of the aggregated scores in the total MAAS sample to obtain component scores (component score$_i$ = aggregated score$_i$/S.D.[aggregated scores in MAAS]). The S.D. (aggregated scores in MAAS) values equaled 1.77, 2.10 and 2.64 in the computation of the memory, executive functioning and speed, and general cognition components, respectively. Fourth, the component scores were linearly transformed and placed on a scale with a mean of 100 and an S.D. of 15 (i.e. an IQ scale). The transformed scores will be referred to as the Memory Quotient (MQ = [Memory Component score$_i$ × 15] + 100), Executive functioning and Speed Quotient (ESQ = [Executive functioning and Speed Component score$_i$ × 15] + 100) and the General Cognitive Quotient (GCQ = [General Cognitive Component score$_i$ × 15] + 100) measures, respectively. These transformations were conducted to increase the ease of interpretation of the neuropsychological composites.

3.2. Correlation analysis

The results of a series of correlational analyses are presented in Table 3. The SCWT Time and LDST 60 s scores were significantly correlated with all outcome measures (with the exception of the baseline CFQ score for the SCWT Time measure, and the baseline CFQ and second follow-up DECO scores for the LDST 60 s score, respectively). The VLT Total recall 1–3 and VLT Delayed recall measures were significantly correlated with the SF-36 PF scores at the second and third follow-up moments, and with the DECO scores at each measurement moment. The CST Shifting score was significantly correlated with the SF-36 PF scores at the first, second, and third follow-up moments. The CST

Table 3
Correlations between the individual neuropsychological test scores, their composites, demographic variables, depressive symptoms, anxiety, and the measures of daily life functioning

	SF-36 PF baseline	SF-36 PF first follow-up	SF-36 PF second follow-up	SF-36 PF third follow-up	DECO first follow-up	DECO second follow-up	DECO third follow-up	CFQ baseline	CFQ first follow-up
SCWT Time	**−.221**	**−.193**	**−.225**	**−.277**	**−.164**	**−.179**	**−.174**	.041	**.107**
SCWT Errors	−.028	**−.111**	−.064	−.060	−.093	−.116	−.070	.065	.046
CST Shifting	−.109	**−.165**	**−.135**	**−.188**	−.062	−.005	−.055	.037	.079
CST Errors	**−.155**	**−.108**	−.102	**−.142**	−.066	−.099	**−.162**	.073	.039
VLT Total recall 1–3	.101	.067	**.154**	**.183**	**.191**	**.166**	**.189**	−.014	−.067
VLT Delayed recall	.075	.065	**.151**	**.157**	**.253**	**.184**	**.137**	−.001	−.046
LDST 60 s	**.220**	**.216**	**.206**	**.234**	**.146**	.102	**.131**	−.082	**−.127**
ESQ	**.232**	**.248**	**.236**	**.285**	**.161**	**.152**	**.181**	−.095	**−.132**
MQ	.094	.072	**.166**	**.184**	**.241**	**.192**	**.176**	−.007	−.060
GCQ	**.218**	**.211**	**.265**	**.314**	**.248**	**.219**	**.229**	−.063	**−.126**
Depressive symptoms	**−.192**	**−.178**	**−.156**	**−.166**	**−.170**	**−.145**	**−.123**	**.333**	**.301**
Anxiety	**−.169**	**−.150**	**−.172**	**−.185**	**−.149**	**−.111**	**−.131**	**.322**	**.266**
Age	**−.288**	**−.314**	**−.380**	**−.419**	**−.205**	**−.244**	**−.216**	−.047	−.001
Gender	.093	**.159**	.097	**.112**	−.074	−.103	−.099	−.087	**−.167**
LE low	**−.199**	**−.165**	**−.202**	**−.249**	**−.136**	−.085	**−.170**	.093	**.126**
LE high	.048	.070	.102	.145	.031	.094	**.139**	−.016	−.065

Note. SF-36 PF, Physical Functioning scale of the Medical Outcomes Study 36-Item Short Form Health Survey; DECO, Détérioration Cognitif Observée (*observed deterioration of cognitive functioning*); CFQ, Cognitive Failures Questionnaire; SCWT, Stroop Color-Word Test; CST, Concept Shifting Test; VLT, Verbal Learning Test; LDST, Letter Digit Substitution Test; ESQ, Executive functioning and Speed Quotient; MQ, Memory Quotient; GCQ, General Cognitive Quotient; LE, level of education. Coding of the predictors: Gender: 1 if male, 0 if female; LE low: Low Education = 1, Average or High Education = 0; LE high: High Education = 1, Low or Average Education = 0. Correlations in boldface are significant ($p < .01$).

Errors score was significantly correlated with the baseline, first-follow-up, and third follow-up SF-36 PF scores, and with the DECO score at the third follow-up. The SCWT Errors score was only significantly correlated with the SF-36 PF first follow-up score. The ESQ and GCQ scores were significantly correlated with all outcome measures (with the exception of the CFQ score at baseline). The MQ measure was significantly correlated with the SF-36 PF scores at the second and third follow-up moments, and with the DECO scores at all measurement moments.

Age was significantly correlated with the SF-36 PF and DECO scores at all measurement moments. LE was significantly correlated to most, but not all, measures of daily life functioning (especially LE low, see Table 3). Correlations between gender and the daily life measures were low and mostly non-significant. Depressive symptoms and anxiety were significantly related to all daily life outcome measures at each measurement moment.

3.3. The prediction of daily life functioning based on neuropsychological measures and non-cognitive variables

For each outcome variable, three full models were constructed. Each full model included the neuropsychological measures (i.e. the seven individual neuropsychological test scores, the ESQ and MQ measures, or the GCQ measure),

Table 4
Coefficients, standard errors, T-values and their significance levels for the final mixed models that included the individual neuropsychological test scores, time, age, age^2, gender, level of education, depressive symptoms, anxiety, and the relevant interactions as predictors in the full models, and the SF-36 PF, CFQ and DECO scores as outcome variables

Outcome	Variable	Estimate	S.E.	T
SF-36 PF	(Constant)	29.002	.528	54.928[**]
	Age	−.134	.020	−6.713[**]
	Age^2	−.005	.002	−2.850[**]
	LE low	−.928	.294	−3.154[**]
	LE high	.260	.375	.692
	Depressive symptoms	−.135	.023	−5.915[**]
	First follow-up	−.352	.199	−1.765
	Second follow-up	−1.336	.209	−6.376[**]
	Third follow-up	−2.069	.233	−8.875[**]
	Age × first follow-up	−.042	.018	−2.290
	Age × second follow-up	−.112	.019	−5.839[**]
	Age × third follow-up	−.151	.021	−7.070[**]
	SCWT Error	.150	.077	1.939
	SCWT Errors × first follow-up	−.255	.082	−3.124[**]
	SCWT Errors × second follow-up	−.196	.085	−2.310
	SCWT Errors × third follow-up	−.160	.094	−1.705
CFQ	(Constant)	19.003	2.086	9.109[**]
	Depressive symptoms	.343	.099	3.462[**]
	Anxiety	.452	.156	2.895[**]
	First follow-up	−5.636	1.609	−3.504[**]
	SCWT Time	−.0001	.016	−.006
	SCWT Time × first follow-up	.048	.016	3.042[**]
DECO	(Constant)	35.883	.735	48.830[**]
	Age	−.103	.015	−6.702[**]
	Gender	−.885	.258	−3.426[**]
	Depressive symptoms	−.118	.021	−5.499[**]
	Second follow-up	−.078	.169	−.461
	Third follow-up	−.621	.181	−3.433[**]
	VLT Delayed Recall	.193	.049	3.970[**]

Note. Estimate, B regression weight; SF-36 PF, Physical Functioning scale of the Medical Outcomes Study 36-Item Short Form Health Survey; CFQ, Cognitive Failures Questionnaire; DECO, Détérioration Cognitif Observée (*observed deterioration of cognitive functioning*); SCWT, Stroop Color-Word Test; VLT, Verbal Learning Test. Coding of the predictors: Age = calendar age − 65; Age^2 = (calendar age − 65)^2; Gender: 1 if male, 0 if female; LE low: Low Education = 1, Average or High Education = 0; LE high: High Education = 1, Low or Average Education = 0; First follow-up: 1 if first follow-up moment, 0 if other measurement moment; Second follow-up: 1 if second follow-up moment, 0 if other measurement moment; Third follow-up: 1 if third follow-up moment, 0 if other follow-up moment.
  [**] $p < .005$.

time, age, age^2, gender, level of education, depressive symptoms and anxiety (and all two-way interactions between these predictors and the three-way interactions involving time). The final models that resulted from the hierarchical stepwise procedure are referred to as the 'final individual test score models', the 'final ESQ and MQ models' and the 'final GCQ models'. These models are presented in Tables 4–6. Model assumptions of linearity, normality of the residuals, and homoscedasticity were met for all models with the exception of the small tendencies to heteroscedasticity for the models that predicted the SF-36 PF and DECO scores (i.e. the variance of the residuals decreased as a function of increasing predicted scores at all measurement moments).

The final individual test score model of the SF-36 PF score showed that lower levels of education and more depressive symptoms affected the SF-36 PF scores negatively at all measurement moments. The decline in SF-36 PF scores over time was stronger for older people and for people who obtained higher SCWT Error scores (at baseline). People with more depressive symptoms and higher anxiety scores (at baseline) were predicted to have higher CFQ scores at the various measurement moments. The increase in CFQ scores over time was more pronounced for people with higher SCWT Time scores (at baseline) (note that higher CFQ and higher SCWT Time scores reflect worse functioning). Higher age, being female, more depressive symptoms and lower VLT Delayed recall scores (at baseline) were associated with lower DECO scores at all measurement moments.

The final ESQ and MQ models showed that the decline over time in SF-36 PF scores was more pronounced for older people. More depressive symptoms and lower ESQ scores (at baseline) were negatively associated with the SF-36 PF scores at all measurement moments. The MQ score did not affect the SF-36 PF score at any measurement moment. More depressive symptoms and higher anxiety scores (at baseline) were associated with higher CFQ scores at all

Table 5

Coefficients, standard errors, *T*-values and their significance levels for the final mixed models that included the ESQ, MQ, time, age, age^2, gender, level of education, depressive symptoms, anxiety, and the relevant interactions as predictors in the full models, and the SF-36 PF, CFQ and DECO scores as outcome variables

Outcome	Variable	Estimate	S.E.	*T*
SF-36 PF	(Constant)	24.980	1.186	21.063**
	Age	−.095	.019	−4.912**
	Depressive symptoms	−.133	.023	−5.885**
	First follow-up	−.711	.164	−4.328**
	Second follow-up	−1.611	.173	−9.309**
	Third follow-up	−2.304	.193	−11.929**
	Age × first follow-up	−.047	.018	−2.560*
	Age × second follow-up	−.115	.019	−5.979**
	Age × third follow-up	−.154	.021	−7.199**
	ESQ	.035	.010	3.391**
CFQ	(Constant)	21.983	3.359	6.544**
	Depressive symptoms	.336	.099	3.400**
	Anxiety	.449	.156	2.880**
	First follow-up	6.835	2.932	2.331
	ESQ	−.028	.029	−.969
	ESQ × first follow-up	−.075	.028	−2.646**
DECO	(Constant)	33.669	1.155	29.150**
	Age	−.100	.016	−6.456**
	Gender	−.865	.258	−3.349**
	Depressive symptoms	−.120	.021	−5.607**
	Second follow-up	−.083	.169	−.494
	Third follow-up	−.619	.181	−3.423**
	MQ	.040	.010	4.160**

Note. Estimate, B regression weight; SF-36 PF, Physical Functioning scale of the Medical Outcomes Study 36-Item Short Form Health Survey; CFQ, Cognitive Failures Questionnaire; DECO, Détérioration Cognitif Observée (*observed deterioration of cognitive functioning*); MQ, Memory Quotient; ESQ, Executive functioning and Speed Quotient. Coding of the predictors: Age = calendar age − 65; Gender: 1 if male, 0 if female; First follow-up: 1 if first follow-up moment, 0 if other measurement moment; Second follow-up: 1 if second follow-up moment, 0 if other measurement moment; Third follow-up: 1 if third follow-up moment, 0 if other follow-up moment.

* $p < .01$.
** $p < .005$.

Table 6

Coefficients, standard errors, *T*-values and their significance levels for the final mixed models that included the GCQ, time, age, age^2, gender, level of education, depressive symptoms, anxiety, and the relevant interactions as predictors in the full models, and the SF-36 PF, CFQ and DECO scores as outcome variables

Outcome	Variable	Estimate	S.E.	*T*
SF-36 PF	(Constant)	25.808	1.199	21.532**
	Age	−.096	.020	−4.799**
	Depressive symptoms	−.136	.023	−5.966**
	First follow-up	−.703	.164	−4.286**
	Second follow-up	−1.604	.173	−9.272**
	Third follow-up	−2.300	.193	−11.913**
	Age × first follow-up	−.047	.018	−2.551*
	Age × second follow-up	−.115	.019	−5.982**
	Age × third follow-up	−.155	.021	−7.214**
	GCQ	.027	.010	2.593*
CFQ	(Constant)	19.751	3.273	6.035**
	Depressive symptoms	.343	.099	3.468**
	Anxiety	.445	.156	2.850**
	First follow-up	7.211	2.733	2.638**
	GCQ	−.007	.028	−.240
	GCQ × first follow-up	−.079	.026	−2.987**
DECO	(Constant)	33.425	1.183	28.249**
	Age	−.091	.016	−5.565**
	Gender	−1.043	.252	−4.139**
	Depressive symptoms	−.116	.021	−5.420**
	Second follow-up	−.083	.169	−.492
	Third follow-up	−.616	.181	−3.406**
	GCQ	.043	.010	4.263**

Note. Estimate, B regression weight; SF-36 PF, Physical Functioning scale of the Medical Outcomes Study 36-Item Short Form Health Survey; CFQ, Cognitive Failures Questionnaire; DECO, Détérioration Cognitif Observée (*observed deterioration of cognitive functioning*); GCQ, General Cognitive Quotient. Coding of the predictors: Age = calendar age − 65; Gender: 1 if male, 0 if female; First follow-up: 1 if first follow-up moment, 0 if other measurement moment; Second follow-up: 1 if second follow-up moment, 0 if other measurement moment; Third follow-up: 1 if third follow-up moment, 0 if other follow-up moment.
* $p < .01$.
** $p < .005$.

measurement moments (note that higher CFQ scores reflects worse functioning). The increase in CFQ scores over time was especially pronounced for people with lower ESQ scores (at baseline). The MQ score did not affect CFQ scores at any of the measurement moments. Higher age, being female, more depressive symptoms and lower MQ scores (at baseline) were associated with lower DECO scores at all measurement moments.

The final GCQ models showed that the decline over time in SF-36 PF scores was more pronounced for older people. More depressive symptoms and lower GCQ scores (at baseline) were associated with lower SF-36 PF scores at all measurement moments. More depressive symptoms and higher anxiety scores (at baseline) were associated with higher CFQ scores at all measurement moments (note that higher CFQ scores reflect worse functioning). Higher GCQ scores (at baseline) predicted less increase in CFQ scores over time. Higher age, being female, more depressive symptoms, and lower GCQ scores (at baseline) predicted lower DECO scores at all measurement moments.

As shown in Table 7, the final individual test score models, the final ESQ and MQ models, and the final GCQ models explained between 4.6% and 21.4% of the variance in the daily life measures.

4. Discussion

The aim of the present study was to evaluate the cross-sectional and the longitudinal ecological validity of neuropsychological test scores and their composites in a large sample of neurologically intact mid-aged to older adults. We first established neuropsychological composites by means of PCAs. A 2-component solution accounted for 56.7% of the total variance, and a 1-component solution accounted for 39.3% of the variance. These components were referred

Table 7
Proportion of explained variance in the SF-36 PF, CFQ and DECO scores by the final neuropsychological test score models, the final ESQ and MQ models, and the final GCQ models (see Tables 4–6)

		Final individual test score models (%)	Final ESQ and MQ models (%)	Final GCQ models (%)
SF-36 PF	Baseline	11.6	11.4	11.9
	First follow-up	13.5	12.0	12.8
	Second follow-up	21.4	19.8	20.0
	Third follow-up	19.5	17.1	17.2
CFQ	Baseline	14.4	14.2	14.4
	First follow-up	6.0	6.1	6.2
DECO	First follow-up	10.1	9.6	10.0
	Second follow-up	4.7	4.6	4.6
	Third follow-up	5.8	7.5	6.6

Note. SF-36 PF, Physical Functioning scale of the Medical Outcomes Study 36-Item Short Form Health Survey; CFQ, Cognitive Failures Questionnaire; DECO, Détérioration Cognitif Observée (*observed deterioration of cognitive functioning*); GCQ, General Cognitive Quotient; MQ, Memory Quotient; ESQ, Executive functioning and Speed Quotient.

to as the ESQ and MQ measures, and the GCQ measure, respectively. The neuropsychological composites consist of linear combinations (=weighted sums) of individual test scores, a method that produces statistically sound summary measures (Miller & Rohling, 2001). Many studies have used neuropsychological composites that consist of unweighted rather than weighted sums of standardized test scores (e.g. Van Hooren et al., 2005). Both approaches yield highly correlated composites when all the component loadings of the individual test scores that comprise the composite of interest are roughly equal to each other (the latter was the case for the MQ score, see Table 2).

Correlations between the neuropsychological scores and the daily life measures ranged between 0 and .32 (see Table 3). There is no agreement on how strong the association between neuropsychological scores and measures of daily life functioning should be in order to consider a neuropsychological measure "ecologically valid". In general, Pearson correlations of .10, .30, and .50 are considered to correspond to small, medium, and large effects, respectively (Cohen & Cohen, 1983). Given these conventional effect sizes, the results suggest that the ecological validity of the neuropsychological measures was low to moderate (i.e. each neuropsychological measure achieved a correlation of at least .10 with at least one daily life measure, see Table 3).

Neuropsychologists in real-world situations base their judgment regarding the everyday functioning of an individual on a *combination* of neuropsychological measures (individual test scores or composites) and non-cognitive variables. The contribution of neuropsychological measures and non-cognitive variables in the prediction of daily life functioning was evaluated with LMM analyses. Higher age was shown to be associated with lower SF-36 PF and lower DECO scores. Being female was associated with higher DECO scores. Depressive symptoms affected all daily life measures negatively and higher levels of anxiety were associated with higher CFQ scores. Independently from these non-cognitive variables, the SF-36 PF, CFQ and DECO scores were affected by the SCWT Errors, the SCWT Time, and the VLT Delayed recall scores, respectively. The ESQ measure was associated with the SF-36 PF and CFQ scores, and the MQ score predicted the DECO score. The GCQ measure was the only neuropsychological measure that was consistently associated with all daily life measures (after the relevant non-cognitive variables were taken into account). The LDST and CST scores were not related to any of the daily life measures in the final individual neuropsychological test score models (see Table 4). It must be stressed out that this finding does not imply that the LDST and CST scores were unrelated to daily life functioning (see also Table 3). Rather, this finding suggests that the LDST and CST scores did not increase the amount of explained variance in the daily life measures given the non-cognitive variables, the SCWT Time, the SCWT Errors, and/or the VLT delayed recall scores. Or, in applied terms, the CST and LDST test scores do not improve the prediction of daily life functioning when the non-cognitive variables and the SCWT and the VLT test scores are already considered by the neuropsychologist.

The final individual neuropsychological test score models, the final ESQ & MQ models, and the final GCQ models explained about the same amounts of variance in daily life functioning (i.e. 4.6–21.4%, see Table 7). This finding suggests that the seven individual neuropsychological test scores can be compressed in one or two neuropsychological composites (i.e. the GCQ measure, and the ESQ and MQ measures) without lowering the ecological validity. As

the GCQ measure was the most consistently related to the various daily life measures, it can be considered to be the measure of first choice when an ecologically valid neuropsychological measure is needed. Note that models that only included the neuropsychological measures accounted for less than 7% of the variance in daily life functioning (models available by contacting the corresponding author). This finding clearly shows that the ecological validity of the neuropsychological assessment can be substantially improved when non-cognitive variables are taken into account together with the neuropsychological measures.

A number of limitations in the present study suggest possibilities for future research. First, a critical element in ecological validity research is how everyday functioning is defined. Everyday functioning is a broad construct that can be defined as ADL functioning, IADL functioning, functioning at work, or social functioning. Several of the neuropsychological measures that were considered in the present study were significantly related to the ADL and IADL measures, but it remains to be determined to what extent that these neuropsychological measures have predictive power with respect to the other aspects of daily life functioning (e.g. functioning at work or social functioning). Moreover, the different aspects of everyday functioning can be measured in different ways. We used self-report and informant questionnaires to assess ADL and IADL functioning, but both methods are 'subjective' and indirect. Other researchers have used laboratory-based simulations of everyday tasks (e.g. shopping and financial management) to obtain more objective and direct measures of daily living (Farias et al., 2003; Heaton et al., 2004). It may be valuable to include both behavior-based and questionnaire-based measures of daily living in future studies, as this could provide converging evidence regarding the ecological validity of neuropsychological measures (especially because neuropsychological measures seem to be more strongly related to behavior-based measures of everyday functioning than to questionnaire-based measures; Farias et al., 2003; Heaton et al., 2004). Note, however, that certain domains of daily living may be difficult to simulate in laboratory settings (e.g. daily life social functioning; Chaytor et al., 2007), and that practical complications may hamper the use of behavior-based measurements of daily life functioning in large-scale studies (e.g. considerations of time and cost-efficiency).

Second, all participants were cognitively healthy at baseline, but a proportion of the sample developed dementia during the follow-up measurements. In our sample, $n = 33$ people (3.8% of the total baseline sample) were diagnosed with dementia at the first ($n = 8$), second ($n = 14$), and third ($n = 11$) follow-up moments. As expected, the baseline GCQ, ESQ and MQ scores of the participants who developed dementia during follow-up was substantially lower as compared to the baseline scores of people who did not develop dementia during follow-up (i.e. 85.8 vs. 100.6 for the GCQ, 92.5 vs. 100.3 for the ESQ, and 86.7 vs. 100.6 for the MQ). The data of people who developed dementia during follow-up were not excluded from the analyses because the aim of the present study was to evaluate the ecological validity of neuropsychological measures that were administered at the moment that a person was neurologically intact. The question nevertheless rises whether exclusion of the data of people who developed dementia during the follow-up measurements would have affected the results. Additional LMM analyses showed that this was not the case (data not shown).

Third, we evaluated the ecological validity of 'traditional' neuropsychological tests (i.e. tests that were originally developed to detect brain damage). An alternative approach to address the problem of ecological validity consists of abandoning the existing traditional neuropsychological tests and creating new ones with optimal ecological validity held explicitly in mind during the development of the instrument (i.e. the verisimilitude approach; Wilson, Alderman, Burgess, Emslie, & Evans, 1996). Studies with neurologically impaired people have suggested that tests with verisimilitude may have higher ecological validity as compared to traditional neuropsychological tests, though Chaytor and Schmitter-Edgecombe (2003) noted in their review that the evidence is limited and far from universal. It therefore remains important to evaluate the ecological validity of traditional neuropsychological tests, also because the majority of clinical neuropsychologists continue to use traditional tests to ground their judgments about the everyday functioning of an individual rather than the newly developed ecologically oriented instruments (Rabin et al., 2007). Nevertheless, future studies should also evaluate the ecological validity of verisimilitude tests in neurologically intact populations.

Fourth, it has been suggested that participants in longitudinal studies (especially older participants) may not be representative of their peers who do not participate in longitudinal studies due to a bias towards greater health. Compliance to participate in the MAAS was also found to be affected by health status in the group of people aged above 70 years (see Jolles, Houx, Van Boxtel, & Ponds, 1995). This bias toward greater health may affect the generalizability of the results of the present study, but this problem exists for all longitudinal studies.

In summary, the results of the present study suggested that neuropsychological measures, demographic variables, depressive symptoms and anxiety were associated with present and future daily life functioning (3, 6, and/or 9 years

following the neuropsychological testing) in neurologically intact people. Between 4.6% and 21.4% of the variance in daily life measures was accounted for by the neuropsychological measures and the non-cognitive variables. Future studies should evaluate how the ecological validity of the neuropsychological assessment can be increased (e.g. by including additional neuropsychological measures that assess other cognitive domains such as conceptual reasoning, visuospatial perception, and language). The GCQ measure was the neuropsychological measure that was the most consistently related to the various daily life measures.

Acknowledgements

References

Anderson, C., Laubscher, S., & Burns, R. (1996). Validation of the Short Form 36 (SF-36) health survey questionnaire among stroke patients. *Stroke*, *27*, 1812–1816.

Broadbent, D. E., Cooper, P. F., FitzGerald, P., & Parkes, K. R. (1982). The Cognitive Failures Questionnaire (CFQ) and its correlates. *The British Journal of Clinical Psychology*, *21*, 1–16.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 629–637.

Chaytor, N., & Schmitter-Edgecombe, M. (2003). The ecological validity of neuropsychological tests: a review of the literature on everyday cognitive skills. *Neuropsychology Review*, *13*, 181–197.

Chaytor, N., Schmitter-Edgecombe, M., & Burr, R. (2006). Improving the ecological validity of executive functioning assessment. *Archives of Clinical Neuropsychology*, *21*, 217–227.

Chaytor, N., Temkin, N., Machamer, J., & Dikmen, S. (2007). The ecological validity of neuropsychological assessment and the role of depressive symptoms in moderate to severe traumatic brain injury. *Journal of the International Neuropsychological Society*, *13*, 377–385.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

De Groot, R. H. M., Hornstra, G., & Jolles, J. (2007). Exploratory study into the relationship between plasma phospholipid fatty acid status and cognitive performance, Prostaglandins. *Leukotrienes and Essential Fatty Acids*, *76*, 165–172.

Derogatis, L. D. (1977). *The SCL-90-R*. Baltimore, MD: Clinical Psychometric Research.

Evers, E. A. T., Van der Veen, F. M., Jolles, J., Deutz, N. E. P., & Schmitt, J. A. J. (2006). Acute tryptophan depletion improves performance and modulates the BOLD response during a Stroop task in healthy females. *Neuroimage*, *32*, 248–255.

Farias, S. T., Harrell, E., Neumann, C., & Houtz, A. (2003). The relationship between neuropsychological performance and daily functioning in individuals with Alzheimer's disease: Ecological validity of neuropsychological tests. *Archives of Clinical Neuropsychology*, *18*, 655–672.

Folstein, M. F., Folstein, S. E., & 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*, 189–198.

Green, M. F., Kern, R. S., & Heaton, R. K. (2004). Longitudinal studies of cognition and functional outcome in schizophrenia: Implications for MATRICS. *Schizophrenia Research*, *72*, 41–51.

Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika*, *19*, 149–161.

Hammes, J. (1973). *De Stroop Kleur-Woord Test: Handleiding [The Stroop Color-Word Test: Manual]*. Amsterdam: Swets & Zeitlinger.

Heaton, R. K., Marcotte, T. D., Mindt, M. R., Sadek, J., Moore, D. J., Bentley, H., et al., & and the HNRC group. (2004). The impact of HIV-associated neuropsychological impairment on everyday functioning. *Journal of the International Neuropsychological Society*, *10*, 317–331.

Johnstone, B., & Frank, R. G. (1995). Neuropsychological assessment in rehabilitation: Current limitations and applications. *NeuroRehabilitation*, *5*, 75–86.

Jolles, J., Houx, P. J., Van Boxtel, M. P. J., & Ponds, R. W. H. M. (1995). *Maastricht aging study: Determinants of cognitive aging*. Maastricht, The Netherlands: Neuropsych Publishers.

Kalechstein, A. D., Newton, T. F., & Van Gorp, W. G. (2003). Neurocognitive functioning is associated with employment status: A quantitative review. *Journal of Clinical and Experimental Neuropsychology*, *25*, 1186–1191.

Long, C. J., & Kibby, M. Y. (1995). Ecological validity of neuropsychological tests: A look at neuropsychology's past and the impact that ecological issues may have on its future. *Advances in Medical Psychotherapy*, *8*, 59–78.

Luteijn, F., & Van der Ploeg, F. A. E. (1983). *Handleiding Groninger Intelligentietest (GIT) [Manual Groningen Intelligence Test]*. Lisse, The Netherlands: Swets and Zeitlinger.

Marquardt, D. W. (1980). You should standardize the predictor variables in your regression models. *Journal of the American Statistical Association*, *75*, 87–91.

Martin, R. L., Irrgang, J. J., Burdett, R. G., Conti, S. F., & Van Swearingen, J. M. (2005). Evidence of validity for the Foot and Ankle Ability Measure (FAAM). *Foot & Ankle International*, *26*, 968–983.

Miller, L. S., & Rohling, M. L. (2001). A statistical interpretative method for neuropsychological test data. *Neuropsychology Review*, *11*, 143–169.

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology*, *20*, 33–65.

Rabin, L., Burton, L., & Barr, W. (2007). Utilization rates of ecologically oriented instruments among clinical neuropsychologists. *The Clinical Neuropsychologist*, *21*, 727–743.

Ramaekers, J. G., Kauert, G., Van Ruitenbeek, P., Theunissen, E. L., Schneider, E., & Moeller, M. R. (2006). High-potency marijuana impairs executive function and inhibitory motor control. *Neuropsychopharmacology*, *31*, 2296–2303.

Reijneveld, S. A., Spijker, J., & Dijkshoorn, H. (2007). Katz' ADL index assessed functional performance of Dutch, Moroccan, and Dutch elderly. *Journal of Clinical Epidemiology*, *60*, 382–388.

Ritchie, K., Artero, S., & Touchon, J. (2001). Classification criteria for mild cognitive impairment: A population-based validation study. *Neurology*, *56*, 37–42.

Ritchie, K., & Fuhrer, R. (1996). The validation of an informant screening test for irreversible cognitive decline in the elderly: Performance characteristics within a general population sample. *International Journal of Geriatric Psychiatry*, *11*, 149–156.

Sbordone, R. J. (1996). Ecological validity: Some critical issues for the neuropsychologist. In R. J. Sbordone & C. J. Long (Eds.), *Ecological validity of neuropsychological testing* (pp. 15–42). Delray Beach, Florida: GR Press/St. Lucie Press.

Schmidt, M. (1996). *Rey Auditory Verbal Learning Test: a handbook*. Los Angeles: Western Psychological Services.

Spearman, C. E. (1927). *The abilities of man, their nature and measurement*. New York: Macmillan.

Spooner, D. M., & Pachana, N. A. (2006). Ecological validity in neuropsychological assessment: A case for greater consideration in research with neurologically intact populations. *Archives of Clinical Neuropsychology*, *21*, 327–337.

Stroop, J. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.

Van Hooren, S. A. H., Van Boxtel, M. P. J., Valentijn, S. A. M., Bosma, H., Ponds, R. W. H. M., & Jolles, J. (2005). Influence of cognitive functioning on functional status in an older population: 3- and 6-year follow-up of the Maastricht Aging Study. *International Journal of Geriatric Psychiatry*, *20*, 838–888.

Van Beijsterveldt, C. E. M., Van Boxtel, M. P. J., Bosma, H., Houx, P. J., Buntinx, F., & Jolles, J. (2002). Predictors of attrition in a longitudinal cognitive aging study: The Maastricht Aging Study (MAAS). *Journal of Clinical Epidemiology*, *55*, 216–223.

Van Boxtel, M. P. J., Henskens, L. H. G., Kroon, A. A., Hofman, P. A. M., Gronenschild, E. H. B. M., Jolles, J., et al. (2006). Ambulatory blood pressure, asymptomatic cerebrovascular damage and cognitive function in essential hypertension. *Journal of Human Hypertension*, *20*, 5–13.

Van Breukelen, G. J. P. (2006). ANCOVA vs. change from baseline: More power in randomized studies, more bias in nonrandomized studies. *Journal of Clinical Epidemiology*, *59*, 920–925.

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2005). Rey's Verbal Learning Test: Normative data for 1855 healthy participants aged 24–81 years and the influence of age, sex, education, and mode of presentation. *Journal of the International Neuropsychological Society*, *11*, 290–302.

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2006). The Concept Shifting Test: Adult normative data. *Psychological Assessment*, *18*, 424–432.

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. P. J., & Jolles, J. (2006). The Letter Digit Substitution Test: Normative data for 1,858 healthy participants aged 24–81 from the Maastricht Aging Study (MAAS): Influence of age, education, and sex. *Journal of Clinical and Experimental Neuropsychology*, *28*, 998–1009.

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2006). The Stroop Color-Word Test: Influence of age, sex, and education; and normative data for a large sample across the adult age range. *Assessment*, *13*, 62–79.

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2008). Assessment of change in performance on the Stroop Color-Word Test, Verbal Learning Test of Rey, and Letter Digit Substitution Test after a test–retest interval of three years: The regression-based change approach. *Journal of the International Neuropsychological Society*, *14*, 71–80.

Van der Zee, K. I., & Sanderman, R. (1993). *Het meten van de gezondheidstoestand met de RAND-36: een handleiding [The measurement of health with the RAND-36: Manual]*. Groningen: Noordelijk Centrum voor Gezondheidsvraagstukken.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.

Ware, J. J. E., & Sherbourne, C. D. (1992). The MOS 36-item short-form healthy survey (SF-36): I. Conceptual framework and item selection. *Medical Care*, *30*, 473–483.

Wilson, B. A., Alderman, N., Burgess, P., Emslie, H., & Evans, J. J. (1996). *Behavioural Assessment of the Dysexecutive Syndrome (BADS)*. Bury St. Edmunds, England: Thames Valley Test Company.