A Neuropsychological Test Battery for Use in Alzheimer Disease Clinical Trials

John Harrison, C. Psychol, PhD; Sonia L. Minassian, DrPH; Lisa Jenkins, PhD; Ronald S. Black, MD; Martin Koller, MD, MPH; Michael Grundman, MD, MPH

Objective: To report the psychometric properties of an alternative instrument to the cognitive subscale of the Alzheimer’s Disease Assessment Scale, a neuropsychological test battery (NTB) for measuring drug efficacy in Alzheimer disease clinical trials.

Design: The NTB was evaluated in a randomized, double-blind, placebo-controlled trial of AN1792(QS-21) (synthetic β-amyloid plus an adjuvant) (300 patients) and isotonic sodium chloride solution (72 patients). The test-retest reliability of the NTB was examined, and the NTB was correlated with other cognitive (cognitive subscale of the Alzheimer’s Disease Assessment Scale and Mini-Mental State Examination) and functional (Disability Assessment Scale for Dementia and Clinical Dementia Rating Sum of Boxes) measures. In addition, a factor analysis was performed on NTB components. Finally, the sensitivity of the NTB to change was assessed as a function of Mini-Mental State Examination performance.

Results: The NTB had high test-retest reliability at 6 (Pearson product moment correlation \(r=0.92\)) and 12 (\(r=0.88\)) months. Internal consistency was high (Cronbach \(\alpha=0.84\)). The correlations between the NTB \(z\) score and scores on traditional measures of cognition and function were significantly different from 0 \((P<.001)\). A factor analysis yielded “memory” and “executive function” factors. The NTB \(z\) score declined linearly over 1 year in patients receiving placebo and, in contrast to the Alzheimer’s Disease Assessment Scale cognitive subscale, demonstrated similar declines in patients with high (21-26) and low (15-20) Mini-Mental State Examination scores at baseline.

Conclusions: The NTB exhibits excellent psychometric properties and seems to be a reliable and sensitive measure of cognitive change in patients with mild to moderate Alzheimer disease. The psychometric properties of the NTB suggest that it may have particular utility in evaluating drug efficacy in clinical trials in which patients with mild Alzheimer disease are included.
suggested this may not be true. Wilcock et al\(^8\) reported that the 6-month placebo group decline on the ADAS-Cog in patients scoring 18 or higher on the MMSE was only slightly more than 1 point. Other recent studies have also reported similar modest changes in ADAS-Cog scores over time. For example, Seltzer et al\(^7\) report that the AD placebo group (mean baseline MMSE score, 24.3; SD, 1.3) in a 24-week study of donepezil hydrochloride showed less than a 1-point change. In addition, Kurz et al,\(^8\) in a study that examined treatment effects of rivastigmine by disease stage in AD, reported that the placebo group of 288 patients scoring 22 or higher on the MMSE at baseline changed by just 1.13 ADAS-Cog points in 26 weeks. The ADAS-Cog change from baseline scores in patients with mild cognitive impairment has been even more modest.\(^7,\(^9\)\) These findings suggest that it would be prudent to identify cognitive outcome measures that are more sensitive to small changes in function in studies of patients with mild AD (ie, MMSE score > 20) and mild cognitive impairment.

In a previously reported, randomized, double-blind, placebo-controlled trial of AN1792(QS-21) (synthetic \(\beta\)-amyloid plus an adjuvant),\(^10\) a neuropsychological test battery (NTB), a novel combination of 6 well-known, validated, cognitive tests, yielding 9 measures of patient performance, was used. Scores for each of the individual NTB components from baseline to month 12 in the antibody responder (anti-AN1792 IgG [total] serum titer \(\geq 1: 2200\) at any postinjection visit) and placebo groups were standardized into \(z\) scores and converted into an overall composite \(z\) score across all NTB components. The study found less worsening in the change from baseline NTB \(z\) scores at month 12 in the antibody responder group compared with \(z\) scores in the placebo group.\(^10\) The mean NTB \(z\) score improvement from baseline to month 12 was 0.03 (SD, 0.37) in the antibody responder group, while the NTB decline was 0.20 (SD, 0.45) in the placebo group. There were no significant differences in the decline on other traditional AD assessments (eg, ADAS-Cog, Disability Assessment Scale for Dementia [DAD], Clinical Dementia Rating Sum of Boxes [CDR-SOB], and Alzheimer Disease Cooperative Study Clinical Global Impression of Change [ADCS-CGIC]) between the placebo and responder groups. Concentration-response analysis indicated significant relationships between geometric mean titer and the NTB \(z\) score; greater improvements from baseline in the NTB \(z\) scores were associated with higher IgG antibody titers.\(^10\)

The NTB components were selected because they indexed memory and/or executive function. Tests of visual and verbal memory were selected for inclusion. Delayed verbal recall has been repeatedly shown to be a sensitive measure of cognitive decline in patients with AD.\(^11\) To maximize the chances of capturing delayed verbal memory effects, the Rey Auditory Verbal Learning Test was included in the NTB. However, to reduce the number of outcome variables to a manageable number, we combined total words recalled at delayed testing with total number of words correctly recognized. No correction for false-positive recognition was made in this analysis. Impaired executive function, defined as the inability to plan, organize, set, and adapt current and past knowledge to future behavior, is one of the core features of AD. In one study,\(^12\) 64% of patients had executive dysfunction and these patients performed worse on tests of cognition, dementia severity, and activities of daily living compared with patients with normal executive function. Neuro-psychological tests are typically classified according to the primary area of cognition believed to be necessary for successful completion of the task. However, this does not suggest that other cognitive skills are not required, because even a word recall task will require satisfactory attentional, working memory, language, and other skills.\(^13\)

The NTB was also designed to incorporate tests of paired associative learning. Impairment on these tasks can be induced through scopolamine hydrobromide administration,\(^14\) and successful paired associative learning task performance is known to rely on the integrity of medial temporal lobe structures. These brain areas are reported to be among the first to experience degeneration in AD.\(^15\) This class of test has been shown to be a sensitive measure of cognitive decline in patients with AD.\(^16\)

In this study, we performed a factor analysis of the individual NTB components to determine whether a “memory” factor and an “executive function” factor could be identified. We also characterized the internal consistency and test-retest reliability of the NTB and, to address the issue of clinical relevance, correlated the NTB with cognitive and functional measures. The concurrent validity of the NTB was also assessed by correlating the NTB \(z\) scores with scores from the ADAS-Cog, MMSE, and CDR-SOB. Finally, the sensitivity to change of the NTB was assessed in relation to MMSE baseline performance.

**METHODS**

**PATIENTS**

Eligible patients were aged 50 to 85 years and met the criteria for a diagnosis of probable AD as defined by the National Institute of Neurological and Communicative Disorders and Stroke—Alzheimer’s Disease and Related Disorders Association\(^17\); magnetic resonance imaging of the brain was conducted to rule out other brain pathology to support the clinical diagnosis of AD. Additional inclusion and exclusion criteria and other details of patient disposition are described elsewhere.\(^10\)

**STUDY DESIGN AND TREATMENT**

The NTB and other cognitive and functional measures were evaluated using data from a randomized, placebo-controlled, double-blind trial conducted at 28 centers in the United States and Europe between September 24, 2001, and December 18, 2002. A total of 372 patients with mild to moderate AD were randomized in a double-blind manner to receive treatment with AN1792, 225 µg (Elan Pharmaceuticals), and QS-21, 50 µg (Antigenics, Framingham, Massachusetts), containing 0.4% polysorbate 80 or isotonic sodium chloride solution in a 4:1 ratio. However, dosing was discontinued after reports of encephalitis in 6.0% of AN1792(QS-21)-treated patients.

**SELECTION OF THE NTB COMPONENT TESTS**

The NTB is composed of “paper-and-pencil” assessments and takes on average approximately 40 minutes to complete. The
The neuropsychological test battery (NTB) test-retest reliability at month 6. The Pearson product moment correlation coefficient (r) was 0.92 (P<.001) between baseline and month 6.

A total of 372 patients were randomly assigned to receive study treatment, 300 received AN1792(QS-21) and 72 received placebo. Included patients were approximately 72 years old, with a slightly greater preponderance of women (55.1%), and had a mean MMSE score of 20. Of the patients, 83.3% were at steady-state use of acetylcholinesterase inhibitors at study enrollment, and inclusion criteria required that acetylcholinesterase inhibitors at baseline were continued throughout the study. No patient was withdrawn from the study because of changes in the doses of acetylcholinesterase inhibitor medication.

The NTB z scores for patients receiving placebo at baseline were correlated with their z scores at 6 and 12 months to assess test-retest reliability. The NTB had high test-retest reliability at 6 (r = 0.92; 95% confidence interval, 0.87-0.95) and 12 (r = 0.88; 95% confidence interval, 0.79-0.93) months. The NTB also had high levels of internal consistency, with a Cronbach α of 0.84.

There was a correlation between the NTB z score and scores on traditional measures of cognition (MMSE and ADAS-Cog) and function (DAD and CDR-SOB) (P < .001) at baseline, month 6, and month 12 for patients receiving placebo and for all patients included in the study. Given the negligible differences between correlations calculated for all patients and for patients receiving placebo, only the correlations for all patients are presented in Table 1. In particular, correlations with the ADAS-Cog
Table 1. Pearson Product Moment Correlation Coefficients Between the NTB and Other Measures at Baseline and at 6 and 12 Months

<table>
<thead>
<tr>
<th>Measure</th>
<th>Baseline Data</th>
<th>6-mo Data</th>
<th>12-mo Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAS-Cog</td>
<td>−0.73 (−0.67 to −0.68)</td>
<td>[n = 367]</td>
<td>−0.78 (−0.72 to −0.73)</td>
</tr>
<tr>
<td>DAD</td>
<td>0.37 (0.28 to 0.46)</td>
<td>[n = 367]</td>
<td>0.48 (0.39 to 0.56)</td>
</tr>
<tr>
<td>CDR-SOB</td>
<td>−0.49 (−0.56 to −0.41)</td>
<td>[n = 367]</td>
<td>−0.56 (−0.63 to −0.48)</td>
</tr>
<tr>
<td>MMSE</td>
<td>0.69 (0.63 to 0.74)</td>
<td>[n = 367]</td>
<td>NA</td>
</tr>
<tr>
<td>ADCS-CGIC</td>
<td>NA</td>
<td></td>
<td>−0.31 (−0.41 to −0.20)</td>
</tr>
</tbody>
</table>

Abbreviations: ADAS-Cog, Alzheimer’s Disease Assessment Scale cognitive subscale; ADCS-CGIC, Alzheimer Disease Cooperative Study Clinical Global Impression of Change; CDR-SOB, Clinical Dementia Rating Sum of Boxes; DAD, Disability Assessment Scale for Dementia; MMSE, Mini-Mental State Examination; NA, data not available; NTB, neuropsychological test battery.

a Data are given as mean (95% confidence interval) correlation with change from the baseline NTB. The number of participants is given in brackets. All coefficients are statistically significantly different from 0 (P < .001).

Table 2. Pearson Product Moment Correlation Coefficients for the Change From Baseline Between the NTB and Other Measures at 6 and 12 Months

<table>
<thead>
<tr>
<th>Measure</th>
<th>6-mo Change From Baseline</th>
<th>12-mo Change From Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAS-Cog</td>
<td>0.34 (0.24 to 0.44)</td>
<td>[n = 309]</td>
</tr>
<tr>
<td>DAD</td>
<td>0.26 (0.15 to 0.36)</td>
<td>[n = 309]</td>
</tr>
<tr>
<td>CDR-SOB</td>
<td>0.23 (0.12 to 0.33)</td>
<td>[n = 305]</td>
</tr>
<tr>
<td>MMSE</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>ADCS-CGIC</td>
<td>−0.25 (−0.35 to −0.14)</td>
<td>[n = 292]</td>
</tr>
</tbody>
</table>

Abbreviations: See Table 1.

a Data are given as mean (95% confidence interval) correlation with change from the baseline NTB. The number of participants is given in brackets. All coefficients are statistically significantly different from 0 (P < .001).

Table 3. Factor Analysis Results (z Scores)

<table>
<thead>
<tr>
<th>Measure According to Test Type</th>
<th>Factor 1</th>
<th>Factor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WMVer-D</td>
<td>0.75</td>
<td>0.28</td>
</tr>
<tr>
<td>WMVer-I</td>
<td>0.70</td>
<td>0.41</td>
</tr>
<tr>
<td>RAVLT-D</td>
<td>0.68</td>
<td>−0.11</td>
</tr>
<tr>
<td>WMVis-D</td>
<td>0.69</td>
<td>0.15</td>
</tr>
<tr>
<td>WMVis-I</td>
<td>0.63</td>
<td>0.28</td>
</tr>
<tr>
<td>RAVLT-I</td>
<td>0.49</td>
<td>0.66</td>
</tr>
<tr>
<td>Executive function</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CFT</td>
<td>0.33</td>
<td>0.63</td>
</tr>
<tr>
<td>COWAT</td>
<td>0.15</td>
<td>0.84</td>
</tr>
<tr>
<td>WMDS</td>
<td>−0.02</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Abbreviations: CFT, Category Fluency Test; COWAT, Controlled Word Association Test; RAVLT-D, Rey Auditory Verbal Learning Test delayed; RAVLT-I, Rey Auditory Verbal Learning Test immediate; WMDS, Wechsler Memory Digit Span; WMVer-D, Wechsler Memory Scale verbal delayed; WMVer-I, Wechsler Memory Scale verbal immediate; WMVis-D, Wechsler Memory Scale visual delayed; WMVis-I, Wechsler Memory Scale visual immediate.

a Although RAVLT-I is considered a measure of memory and is, therefore, included in the memory z score, in the factor analysis, RAVLT-I loaded more heavily on factor 2 than factor 1.

tients were also different from 0 (P < .001) (Table 2). These correlations were smaller than correlations between scores obtained for each measure at baseline and 6 and 12 months (Table 1).

The factor analysis yielded 2 factors with eigenvalues greater than 1.0, indicating the presence of 2 factors. The factor loadings for both factors are given in Table 3 and are shown graphically in Figure 2. The factors for the Wechsler Memory Digit Span, COWAT, CFT, and RAVLT immediate tests load heavily on factor 2 (factor loadings, >0.6). The rest of the NTB tests load primarily on factor 1. Of the 4 tests that primarily load on factor 2, 3 measure executive function, and all of the tests that primarily load on factor 1 measure memory. Thus, factor 1 can be viewed as a memory factor and factor 2 can be viewed as an executive function factor.

Finally, the average 12-month NTB and ADAS-Cog changes from baseline were examined for patients with mild and moderate AD receiving placebo. Patients with mild AD had MMSE scores of 21 to 26 at baseline, and patients with moderate AD had MMSE scores of 15 to 20 at baseline (Figure 3). The mean (SD) NTB change from baseline for patients with a high MMSE score (−0.21 [0.47]) was similar to that for patients with a low MMSE score (−0.18 [0.42]). In contrast, the mean ADAS-Cog change from baseline to month 12 in the placebo group was different for patients with high and low MMSE scores. Patients with high baseline MMSE scores declined by a mean (SD) of 0.63 (4.32) ADAS-Cog points, and patients with low baseline MMSE scores and more severe AD declined by a mean (SD) of 4.92 (7.61) points on the ADAS-Cog. As expected, the mean ADAS-Cog z score...
change from baseline to month 12 was also different for patients with high and low MMSE scores. Patients with high baseline MMSE scores declined by a mean (SD) of 0.07 (0.46) ADAS-Cog \( z \) score points, and patients with low baseline MMSE scores with more severe manifestations of AD declined by a mean (SD) of 0.52 (0.80) ADAS-Cog \( z \) score points.

**COMMENT**

Knowing the limitations of the ADAS-Cog, the primary aim of including the NTB in the AN1792(QS-21) phase 2 study\(^{10} \) was to establish a sensitive easy-to-administer test battery to assess cognitive change in patients with mild to moderate AD. The NTB \( z \) score composite assessing cognitive change in patients in the antibody responder group vs the placebo group yielded statistically significant evidence of reduced impairment in the treated group. This result was in contrast to the analyses conducted with the ADAS-Cog and other functional and cognitive measures, which did not find evidence of reduced impairment in the treatment group. Furthermore, a high correlation was found between performance on the NTB and traditional measures at baseline, month 6, and month 12, suggesting that the NTB exhibits high concurrent validity vs measures such as the ADAS-Cog and MMSE. In addition, we also observed statistically significant correlations with functional measures, such as the CDR-SOB and DAD. These correlations suggest that these scales predominately measure the same constructs. The correlations between the NTB change from baseline at 6 and 12 months and the other functional and cognitive measures were all different from 0.

Good test-retest reliability was a key requirement of the position paper published by the International Working Group on Harmonization of Dementia Drug Guidelines.\(^2 \) Further requirements included validity (the instrument must measure the intended disease-relevant cognitive functions), reliability (test-retest, also interrater and intrarater if scoring is subjective), appropriate sensitivity range (the absence of ceiling and floor effects), longitudinal data (information should be available on expected change during the trial), practice effects (determines the need for practice sessions before starting a trial), and equivalent forms (for repeated testing during the trial). Of these requirements, the most difficult to satisfy when selecting from the available cognitive tests are the requirements for practice effects and equivalent forms. Standard texts, such as that of Lezak,\(^{13} \) reveal that evidence regarding practice effects and the availability of genuinely equivalent cognitive tests is difficult to find. While the guidelines of Ferris et al\(^2 \) represent excellent guidance, it could be argued that cognitive tests with high levels of test-retest reliability are sufficiently psychometrically robust to preclude the necessity for parallel forms.

The NTB was initially designed to emphasize assessing cognition in the domains of memory and executive function. When all 9 NTB components were subjected to factor analysis, the model yielded evidence to support 2 clear factors. One factor is best characterized as memory, and the other factor represents executive function. However, while the NTB was designed to emphasize the memory and executive function domains, it also provides an index of global cognitive function by drawing on the many cortical areas required to support language (COWAT and CFT), attention (digit span forward), visual perception (Wechsler Memory Scale visual), verbal memory (Wechsler Memory Scale verbal and RAVLT), working memory (digit span backward), and list learning (RAVLT).

The outcome measures that contributed to the executive function factor were all derived from tests generally...
acknowledged to index this cognitive domain. Further interpretation of the executive function factor is challenging largely because of the lack of a clear definition of the term. There are a number of definitions, and while some “family resemblance” between them is evident, some are couched in neuroanatomical terms, typically with reference to the prefrontal cortex, while others are more functional. An alternative view is that all cognitive tasks require some contribution from executive function; what varies from task to task is the extent to which this resource is taxed. Tasks with demanding working memory components seem to require significant contributions from executive function. Consequently, those tasks that require planning, the development of strategy, or the online monitoring of task components are all likely to make robust demands on executive function resources. Thus, given the need to search semantic memory for appropriate items in the NTB 2-word fluency tasks (CFT and COWAT), and the need to ensure that words are not repeated, it is not surprising that performance on these tests loads on the putative executive function factor. The need to maintain and order items in working memory would also explain why digit span loads on the executive function factor. The same need to utilize working memory when attempting to hold in memory the 15 items of the RAVLT could also account for why this metric seems to load on both factors. We suspect that executive function could be profitably indexed using other putative tests of this resource. Further studies might usefully measure indicators that index executive function through assessing set-shifting abilities and abstraction.

The measures of cognition that compose the NTB were selected to demonstrate sensitivity to cognitive deterioration and to exhibit robust psychometric properties. The NTB performed well, exhibiting a test-retest reliability correlation coefficient of 0.92 at 6 months and 0.88 at 12 months. This is among the higher levels of reliability for cognitive measures and suggests that the NTB is a highly reliable index of cognitive function and that it is well suited for repeated assessment. An analysis of the NTB's internal consistency also yielded satisfactory levels of performance. The observed Cronbach α of 0.84 represents high levels of consistency, exceeding the desirable threshold of 0.80 proposed by Kline.

The sensitivity of the NTB z score was compared with traditional measures, including the ADAS-Cog. The NTB seems to be sensitive to cognitive change independently of MMSE status when stratified by baseline score. The NTB showed similar decline in patients with mild (high baseline MMSE score [21-26]) and moderate (low baseline MMSE score [15-20]) AD. In contrast, performance on the ADAS-Cog shows systematic change depending on disease severity at baseline. The ADAS-Cog showed cognitive decline for patients with moderate AD but did not show decline in patients with mild AD. This is consistent with previous reports of comparatively small annual decline on the ADAS-Cog in those with mild AD.

The relative lack of decline in ADAS-Cog scores in patients with mild AD becomes problematic in the design of clinical trials for disease-modifying AD therapies. For trials that are powered to detect a reduction in the rate of decline over 1 year in the drug-treated patients compared with placebo, the number of subjects with mild AD required to detect a reduction in ADAS-Cog would be excessive. For example, sample size estimates were calculated to determine the number of patients with mild AD (MMSE score, 21-26) necessary to detect a drug-placebo difference of 50% of the annual placebo decline using the 12-month NTB and ADAS-Cog changes for patients with mild AD based upon the results of the AN1792 (QS-21) study. To detect a change from baseline in the treatment group that is 50% less than that in the placebo group, with 80% power, a significance level of .05, and assuming equal sample sizes in the treatment and placebo groups, 1478 patients per group would be needed using the ADAS-Cog (mean placebo group decline, 0.63; SD, 4.32). In contrast, only 159 patients per group would be necessary to detect a change from baseline in the treatment group that is 50% less than that in the placebo group for the NTB (mean placebo group decline, 0.21; SD, 0.47). While the difference in sample sizes required for 80% power using the ADAS-Cog and NTB is accentuated by the particularly low rate of decline on the ADAS-Cog in the placebo group of the AN1792(QS-21) trial, the insensitivity of the ADAS-Cog is also apparent from other recent studies7,8 that found relatively small declines on this measure in patients with mild AD.

In summary, the NTB has exhibited excellent psychometric properties, especially in the areas of concurrent validity and test-retest reliability. In contrast to the ADAS-Cog, the NTB z score showed sensitivity to change across the range of MMSE scores in this trial and was the only measure of cognitive efficacy to yield a statistically significant effect in favor of the AN1792(QS-21) immunotherapeutic treatment. The NTB may represent a useful cognitive measure for clinical trials to assess cognitive change in patients with mild to moderate AD or potentially mild cognitive impairment and may be more effective than traditional measures in detecting cognitive changes in the clinical assessment of a therapy designed to reverse or halt disease pathological features.

Accepted for Publication: January 24, 2007.

Correspondence: John Harrison, C. Psychol, PhD; CPC Pharma Services, PO Box 3223, Warminster, Wiltshire BA12 8XA, England (Johnhpc@btinternet.com).

Author Contributions: Dr Harrison had full access to all of the data and takes responsibility for the integrity of

©2007 American Medical Association. All rights reserved.
the data and the accuracy of the data analysis. **Study concept and design:** Harrison, Minassian, Jenkins, Black, Koller, and Grundman. **Acquisition of data:** Black. **Analysis and interpretation of data:** Harrison, Minassian, Jenkins, Black, and Grundman. **Drafting of the manuscript:** Harrison, Minassian, Jenkins, and Grundman. **Critical revision of the manuscript for important intellectual content:** Harrison, Minassian, Jenkins, and Grundman. **Statistical analysis:** Harrison, Minassian, Jenkins, and Grundman. **Obtained funding:** Grundman. **Administrative, technical, and material support:** Harrison, Black, and Koller. **Study supervision:** Black, Koller, and Grundman. **Financial Disclosure:** Dr Minassian is a salaried employee of and has been awarded stock options from Elan Pharmaceuticals; Dr Black is a full-time employee of and holds stock in Wyeth Research; and Drs Koller and Grundman are salaried employees of and have stock options and ownership in Elan Pharmaceuticals.

**Funding/Support:** The AN1792 clinical study was supported by Elan Pharmaceuticals and Wyeth Research; however, no support was provided for the analyses described herein.

**Role of the Sponsor:** Researchers from Elan Pharmaceuticals and Wyeth Research participated in the design, conduct, and analysis of the research described herein.

**Disclaimer:** The intent of the present study is not to provide the clinical results of the AN1792 trial (previously published) but to describe the properties of instruments used in that trial.

**REFERENCES**