Drug Treatments for Alzheimer’s Disease. II. A Review of Outcome Measures in Clinical Trials
Drug Treatments for Alzheimer’s Disease.
II. A Review of Outcome Measures in Clinical Trials

Christina Wolfson, PhD\textsuperscript{1,2}
Yola Moride, PhD\textsuperscript{1,3}
Anne Perrault, MSc\textsuperscript{1}
Franco Momoli, MSc\textsuperscript{1,2}
Louise Demers, PhD\textsuperscript{1}
Mark Oremus, MA\textsuperscript{1,2}

May 2000

Technical assistance provided by:
Carole Bohbot, Irina Lazariciu, Alexander Tsertsvadze

1. Centre for Clinical Epidemiology and Community Studies, S.M.B.D. Jewish General Hospital, Montréal, Québec
2. Joint Departments of Epidemiology & Biostatistics and Occupational Health, McGill University, Montréal, Québec
3. Faculté de Pharmacie, Université de Montréal, Montréal, Québec
REVIEWERS

These individuals kindly provided comments on this report. CCOHTA takes sole responsibility for the final form and content.

CCOHTA Scientific Advisory Panel

Dr. Gina Bravo
Sherbrooke University
Geriatric Institute
Sherbrooke, Quebec

Dr. David Persaud
Dalhousie University
Halifax, Nova Scotia

External Reviewers

Dr. Richard Gladstone
University of Toronto
Toronto, Ontario

Dr. Willie Malloy
McMaster University
Geriatric Research Group
Hamilton, Ontario

ACKNOWLEDGMENTS

The Canadian Coordinating Office for Health Technology Assessment (CCOHTA) provided the funding for this study. Partial salary support for Carole Bohbot and Alexander Tsertsvadze was provided by the “Fonds de la Recherche en Santé du Québec” (FRSQ), through an Equipe Prioritaire “Épidémiologie de la démence chez les personnes âgées”.
EXECUTIVE SUMMARY

Alzheimer’s disease is a devastating disorder of the elderly, which is expected to become a major public health problem in Canada over the next two decades. Recent projections indicate that the number of cases of Alzheimer’s disease in Canada will increase from 160,000 to more than 380,000 by the year 2020. These projections are based on the assumption that the incidence of Alzheimer’s disease will remain unchanged.

Significant research efforts in Canada and around the world are seeking to determine the cause of AD and to find a cure. Until this has been achieved, however, there remain thousands of Canadian elderly suffering from AD. A major research focus in AD, at the present time, is the development of therapies that could alter the course of the disease in those already afflicted.

In a companion report (Drug Treatments for Alzheimer’s Disease. I. A Comparative Analysis of Clinical Trials), we review those published clinical trials that we believe have the methodological integrity to provide the best evidence on the efficacy of donepezil, metrifonate, rivastigmine, selegiline, vitamin E, lecithin, linopirdine, propentofylline and gingko biloba for the treatment of Alzheimer’s disease.

Twenty-seven randomized clinical trials were retrieved from the literature and found to meet appropriate methodological standards.

We conclude that for selegiline, vitamin E, lecithin, linopirdine, and propentofylline the published data do not provide support for efficacy.

Based on the evidence we reviewed, it is our conclusion that donepezil, metrifonate and rivastigmine, however, all provide statistically significant modest benefit on cognitive performance and global functioning to the elderly with probable AD who are eligible for inclusion in clinical trials. The magnitude of the effect is similar for all of the medications. The results from the trials of gingko biloba are promising but the effects are smaller than those from the above mentioned therapies.

Although all of these medications appear to be well tolerated, in terms of the occurrence of adverse events, dropout rates are sometimes high and may have resulted in overestimation of apparent treatment effects.
It is important to note that this report is based on the results of published trials only and as a result may be subject to publication bias. In particular, if a bias exists such that trials showing no effect are less likely to be published then our findings may overestimate efficacy through the selective inclusion of positive trials. The exclusion of trials that were not published due to poor methodology, however, would not have resulted in a bias as it is unlikely that such trials would have met our standards of methodological rigour.

In the current report (Drug Treatments for Alzheimer’s Disease. II. A Review of Outcome Measures in Clinical Trials), we review the psychometric properties of the primary and secondary outcome measures used in these trials and we remain concerned about the wide variety of scales used that do not have adequate psychometric assessment. Although a great number of the scales used have evidence of reliability and validity, responsiveness to change has generally not been adequately assessed. For this reason, the clinical significance of the treatment: placebo differences remain unclear.

The ability to carry out a comparative analysis of therapies for AD depends not only on the comparability of design, duration, and outcome measures used but also on the methods of reporting the results of the trials. There was no consistent method of reporting the results of the AD trials even when emanating from the same group of investigators. Although journals do have different guidelines for authors, we recommend that reports of clinical trials of therapies for AD follow a standardized format for presenting results, which would enhance the ability to compare results across studies.
# TABLE OF CONTENTS

1 INTRODUCTION........................................................................................................... 1

1.1 Method for Selecting Scales for Review ........................................................................ 2

1.2 Psychometric properties of scales .................................................................................. 3
  1.2.1 Reliability .................................................................................................................. 3
  1.2.2 Validity ...................................................................................................................... 4

1.3 Scope and Mode of Presentation .................................................................................... 5

2 CRITICAL EVALUATION OF THE GLOBAL OUTCOME MEASUREMENT SCALES ........................................................................................................... 7

2.1 Introduction: Global Scales ............................................................................................. 7

2.2 Clinician Global Impression of Change (CGIC) Scales ......................................................... 8
  2.2.1 Introduction ................................................................................................................ 8
  2.2.2 Description ............................................................................................................... 8
  2.2.3 Evaluation ............................................................................................................. 9
  2.2.4 Critical Appraisal ................................................................................................. 11

2.3 Clinical Dementia Rating (CDR) ..................................................................................... 13
  2.3.1 Introduction ............................................................................................................ 13
  2.3.2 Description ........................................................................................................... 13
  2.3.3 Evaluation ........................................................................................................... 14
  2.3.4 Critical Appraisal ............................................................................................... 17

2.4 Global Deterioration Scale (GDS) ................................................................................... 17
  2.4.1 Introduction ............................................................................................................ 17
  2.4.2 Description ........................................................................................................... 17
  2.4.3 Evaluation ........................................................................................................... 18
  2.4.4 Critical Appraisal ............................................................................................... 20

2.5 Gottfries-Bråne-Steen (GBS) .......................................................................................... 21
  2.5.1 Introduction ........................................................................................................... 21
  2.5.2 Description ........................................................................................................... 21
  2.5.3 Evaluation ........................................................................................................... 22
  2.5.4 Critical Appraisal ............................................................................................... 24

2.6 Mini-Mental State Examination (MMSE) .................................................................... 24

2.7 Global Scales: Strengths and Limitations ..................................................................... 24

3 COGNITIVE OUTCOME MEASUREMENT SCALES ........................................... 26

3.1 Introduction .................................................................................................................. 26

3.2 Alzheimer’s Disease Assessment Scale-cognitive (ADAS-cog) ........................................ 26
  3.2.1 Introduction .......................................................................................................... 26
  3.2.2 Description .......................................................................................................... 27
3.14 Strengths and Weaknesses for Individual Neuropsychological Tests ......................................................... 55
  3.14.1 Strengths ................................................................. 55
  3.14.2 Weaknesses ............................................................ 55
  3.14.3 Responsiveness to Change ........................................ 56

3.13 Abstraction ......................................................................................................................................................... 55
  3.13.1 Category Retrieval Test ............................................... 55

3.12 Attention ............................................................................................................................................................. 52
  3.12.1 Trail Making Test ....................................................... 52
  3.12.2 Digit Symbol – Subtest from the WAIS-R .................. 53
  3.12.3 Symbol Digit Modalities Test ................................... 54
  3.12.4 Digit Span subtest of the WAIS-R ............................. 54

3.11 Visuo-Perceptual Function ............................................. 50
  3.11.1 Hooper Visual Organization Test (VOT) ..................... 50
  3.11.2 Rey-Osterrieth Complex Figure Test (RF) .................. 51

3.10 Language........................................................................................................................................................... 48
  3.10.1 Controlled Oral Word Association (COWA) .............. 48
  3.10.2 Boston Naming Test (BNT) ....................................... 49

3.9 Memory ................................................................................................................................................................. 45
  3.9.1 Buschke Selective Reminding Test (BSRT) ................. 45
  3.9.2 Wechsler Memory Scale–Revised (WMS-R) ............... 46

3.8 Psychomotor ........................................................................................................................................................... 44
  3.8.1 Finger Oscillation Test ................................................ 45
  3.8.2 Choice Reaction Time ............................................... 45
  3.8.3 Evaluation ................................................................. 45

3.7 Individual Neuropsychological Tests ................................................................................................................ 44

3.6 Blessed Dementia Scale (BDS) ........................................ 41
  3.6.1 Introduction .............................................................. 41
  3.6.2 Description .............................................................. 41
  3.6.3 Evaluation ............................................................... 42
  3.6.4 Critical Appraisal .................................................... 44

3.5 MMSE as a Staging Instrument ......................................... 40

3.4 Mini-Mental State Examination (MMSE) .......................... 37
  3.4.1 Introduction ............................................................ 37
  3.4.2 Description ............................................................ 37
  3.4.3 Evaluation ............................................................. 38
  3.4.4 Critical Appraisal .................................................. 40

3.3 Syndrome Kurtz Test (SKT) .............................................. 32
  3.3.1 Introduction ........................................................... 32
  3.3.2 Description ........................................................... 32
  3.3.3 Evaluation ............................................................ 33
  3.3.4 Critical Appraisal .................................................. 36

3.2 Attention ................................................................................................................................................................. 30
  3.2.1 Trail Making Test ..................................................... 30
  3.2.2 Digit Symbol – Subtest from the WAIS-R ................ 31
  3.2.3 Critical Appraisal .................................................... 33

3.1 Visuo-Perceptual Function .............................................. 27
  3.1.1 Hooper Visual Organization Test (VOT) ..................... 27
  3.1.2 Rey-Osterrieth Complex Figure Test (RF) .................. 29
  3.1.3 Critical Appraisal .................................................... 30

3.0 Introduction ........................................................................................................................................................... 27
### 3.15 Cognitive Scales: Strengths and Limitations

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>FUNCTIONAL AND QUALITY OF LIFE OUTCOME MEASUREMENT SCALES</td>
<td>58</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction: Functional Disability Scales</td>
<td>58</td>
</tr>
<tr>
<td>4.2</td>
<td>The Geriatric Evaluation by Relative's Rating Instrument (GERRI)</td>
<td>59</td>
</tr>
<tr>
<td>4.3</td>
<td>Physical Self-Maintenance Scale (PSMS)</td>
<td>61</td>
</tr>
<tr>
<td>4.4</td>
<td>Instrumental Activities of Daily Living (IADL)</td>
<td>63</td>
</tr>
<tr>
<td>4.5</td>
<td>Blessed Dementia Scale, Part 1 (BDS Part 1)</td>
<td>65</td>
</tr>
<tr>
<td>4.6</td>
<td>The Nüremberg Geriatric Observation Scale (NAB)</td>
<td>67</td>
</tr>
<tr>
<td>4.7</td>
<td>The Nüremberg-Alters-Alltags-Skala (NAA)</td>
<td>68</td>
</tr>
<tr>
<td>4.8</td>
<td>Activities of Daily Living Checklist (ADLC)</td>
<td>68</td>
</tr>
<tr>
<td>4.9</td>
<td>The Interview for Deterioration in Daily Living in Dementia (IDDDD)</td>
<td>69</td>
</tr>
<tr>
<td>4.10</td>
<td>Unified Activities of Daily Living Form (Unified ADL)</td>
<td>70</td>
</tr>
<tr>
<td>4.11</td>
<td>Dependence Scale (DS)</td>
<td>70</td>
</tr>
<tr>
<td>4.12</td>
<td>Overall Conclusion on the Functional Disability Scales</td>
<td>72</td>
</tr>
<tr>
<td>4.13</td>
<td>Introduction: Quality of Life Scales</td>
<td>72</td>
</tr>
<tr>
<td>4.14</td>
<td>The Quality of Life Scale (QoL)</td>
<td>73</td>
</tr>
</tbody>
</table>
8 BIBLIOGRAPHY (BY AUTHOR) ................................................................. 118

9 BIBLIOGRAPHY: ALZHEIMER DISEASE TRIALS REVIEWED IN
REPORT I ................................................................. 135
1 INTRODUCTION

Alzheimer's disease (AD) is a neurodegenerative disease that is characterized by deterioration in three major domains: cognition, function ability, and behavior and mood (1-4). Cognitive decline occurs in three major clinical phases: an early *forgettable* phase, when the deficit is primarily subjective and usually verifiable with neuropsychological testing; an intermediate *confusional* phase, when the deficit is apparent to an objective observer; and a late *dementia* phase, when a patient can no longer survive without some assistance. As cognitive decline becomes apparent, the ability to function in daily life (e.g., perform common tasks such as walking outdoors or handling money) is gradually lost. Later, as cognitive decline reaches overt dementia, the ability to perform basic self-maintaining activities is also gradually lost. Substantial assistance from caregivers may be needed to compensate for these losses. Personality and emotional changes also occur during the course of AD and produce behavior and mood problems (e.g., anxiety, agitation, violent or delusional behavior and obsessive symptoms). These problems may become apparent at any stage of the disease, although they are more frequent in later stages.

Measurement instruments that have achieved widespread use in AD drug trials often seek to quantify a patient’s status in one or more of these three domains. Other scales attempt to quantify a patient’s overall (or global) health state with respect to AD. Since AD is closer to a ‘construct’ (syndrome) than an actual entity, definitive clinical tests for the disease’s presence and severity in living patients do not exist. Therefore, measurement scales have come to play an important role in the diagnosis and staging of AD in drug trials. These instruments have also become prominent outcome measures in the search for treatment effects.

The present report complements the research conducted by Wolfson and her colleagues on drug treatments for AD (I. *A Comparative Analysis of Clinical Trials*). The overall research has three main objectives:

1) To assess the evidence for the clinical efficacy of individual therapies for Alzheimer’s disease.

2) To compare the clinical efficacy of individual therapies for AD as appropriate.
3) To review the reliability and validity data on many of the primary and secondary outcome measures used in the trials selected in objective 1.

The third objective listed above, that is the review of the measurement properties of the outcomes used in the trials, is the subject of this report.

1.1 Method for Selecting Scales for Review

A list of scales that were used in the trials included in report I (*A Comparative Analysis of Clinical Trials*) was compiled as shown in Tables 1 and 2. Each scale was then classified into one of six domains: global, cognition, functional disability, quality of life, behavior and mood or other. Some scales were easily classified because the instrument’s name and/or supporting literature provided the domain of interest (e.g., Alzheimer Disease Assessment Scale-cognitive, Activities of Daily Living, etc.). Other scales were classified according to descriptions provided in the literature and/or the trial(s) under review. For example, an instrument measuring a patient’s ability to dress or climb a flight of stairs was classified as being part of the functional disability domain. Staging scales that attempt to provide an indication of a patient’s overall disease severity were placed in the global domain because they ideally examine several aspects of dementia.

Once the list of scales was devised, they were reviewed according to the extent of available information and literature. The only exceptions were the behavior and mood scales. Because of the great diversity found in this domain, secondary outcome measures received only a limited examination. Instruments that were described in languages other than English or French were not reviewed.

Literature for the scale assessments was obtained from numerous sources. First, the bibliography of the specific trial being reviewed was consulted for any references to the scales of interest. Second, articles acquired from this bibliography had their own bibliographies scanned for more references. Third, database searches were conducted. The databases were Medline (1966-present) and PsycINFO (1967-present). The full name of each scale and its abbreviation were entered as keywords. Occasionally, when this strategy yielded more than 20 references, the name and abbreviation were separately combined with the individual keywords ‘Alzheimer’, ‘reliability’ and ‘validity’ to narrow
down the scope of the literature search. Abstracts were then consulted in order to select
the articles that would be used for the scale review.

1.2 Psychometric properties of scales

The importance of scales implies that they should be designed and validated in a
manner analogous to tests for other constructs (e.g., psychological attributes) (5). Indeed,
scales designed to measure or quantify AD must be both reliable and valid. For clinical
trials, this means that a scale should measure AD, not other forms of dementia or other
illnesses. The scale should also be able to measure AD in a repeatable manner so that the
same results are obtained across different raters. These two requirements are difficult to
achieve because AD cannot be observed or measured directly, and the multidimensional
nature of the condition requires the use of many instruments. Measurement instruments
should also be responsive to change. That is, they should be capable of detecting changes
in disease status. More complete definitions of reliability and validity are presented
below.

1.2.1 Reliability

A reliable scale is one that produces the same result each time it is used, given
that the phenomenon addressed is unchanged. There are three major approaches to the
estimation of reliability, depending on the sources of errors that are being considered.
The first, referred to as test-retest, is concerned with the stability of the measure at two
different points in time. According to this approach, subjects are evaluated twice, using
the same scale, and the two sets of scores thereby obtained are compared. The second
approach, known as interrater reliability, assesses the reproducibility of the measure
when the phenomenon under study is measured by two or more raters at the same point in
time. The collected sets of scores are correlated to produce a reliability coefficient. The
third approach is referred to as internal consistency, whose focus is on the extent to which
items of a composite score substantively measure the same concept. It is a measurement
property that characterizes the scale or the subscales in their entirety.

It is important to note that both reliability and validity are situation and population
specific traits. Therefore, whenever an instrument is to be used in a new setting or for a
different population from that which it was tested in, both reliability and validity need to be reassessed.

1.2.2 Validity

Appraisal of an assessment instrument's validity consists of evaluating its capacity to measure the concept under study. Validity refers to the adequacy between the theoretical concept and its operationalization at the variable level. Classically, test developers are concerned with three types of validity: content, criterion and construct validity. In our review of AD, we are also particularly concerned about an instrument’s ability to measure AD states and to be responsive to change - specifically changes that fall under the domains of cognition, behavior and function.

Content validity refers to the extent to which the instrument covers the scope of the construct under study. It includes evaluating the relevance of selected components (content relevance) and their capacity to represent every facet of the measured concept (concept coverage) (6). In AD, for example, one would like to see scale items measure aspects of the domains of cognition, function and behavior. To have items measuring only one of these domains would be an incomplete measure AD as a whole. However, this may be desirable when examining only one aspect of the disease (i.e. the effect of a given drug on cognition).

Construct validity refers to the theoretical linkage between a scale and the more abstract conceptual entity under study. Because the underlying disease process in AD is not directly observable, only its hypothetical manifestations (i.e. reduced cognitive performance, behavioral problems, inability to perform certain tasks) can be measured. Testing and improving an instrument’s construct validity is an ongoing process that involves refining the understanding of the components of the underlying condition through posing and testing a variety of hypotheses. For instance, one may hypothesize that patients who are assessed clinically as having “severe” AD will score more impaired on the scale than those thought to have “mild” AD. If the hypothesized relationships are confirmed as predicted by the theory, both the instrument and the theory are supported. However, in the opposite case, the test developer cannot know whether the test or the theoretical construct is flawed (7).
Criterion validity refers to how well the results of the measurement scale correlate with some performance criterion that cannot be directly measured by a test. Predictive validity refers to the degree to which test scores predict criterion measurements that will be made at some time in the future. Concurrent validity refers to the relationship between test scores and criterion measurements made at the time the test was given. In AD, the only gold standard is diagnosis at autopsy, thus concurrent validity cannot be assessed. However, predictive validity can be tested.

Before it is used for assessing differences in response to treatment, a scale must be capable of detecting change. Responsiveness to change refers to an instrument's ability to accurately detect clinically meaningful change (8). Although responsiveness to change is part of the validation process of an instrument, evaluating it separately may be needed (9). Ideally, evaluation should involve the prospective examination of within-person change in the setting of a clinical trial (10). However, both longitudinal and cross-sectional studies have been used to provide some evidence of responsiveness to change (9). Longitudinal studies of demented subjects allow for the monitoring of disease progression (change within subjects), while cross-sectional studies allow the comparison of subjects at different levels of disease severity (change across subjects).

1.3 Scope and Mode of Presentation

The review of scales used in this report is organized by domain, with a chapter devoted to each domain. Accordingly, global scales are examined in Chapter 2, followed by cognitive scales and tests in Chapter 3. Chapter 4 addresses functional scales including the closely related quality of life instruments. Chapter 5 introduces behavior and mood scales, some only briefly. Within each domain, scales are discussed individually according to a uniform format starting with an overview of the constructs measured, administration/application procedures, and scoring. Psychometric properties are then examined, including reliability, validity and responsiveness to change. Lastly, a critical appraisal is provided in conclusion to every scale. The critique highlights strengths, weaknesses, suggested improvements and usefulness in the context of AD drug trials. Chapter 6 concludes with a summary of some of the common problems that were
encountered in the scale review with respect to the impact on the measurement of drug efficacy.

Literature searches did not always uncover information about all forms of reliability or validity. Information on content validity and responsiveness to change was particularly scarce. Some discussion of content validity was occasionally possible through inferences from the scale development process. In cases where information could not be obtained, the affected headings were either simply removed from the text, or a note indicating a dearth of information was placed into the text.
2 CRITICAL EVALUATION OF THE GLOBAL OUTCOME MEASUREMENT SCALES

2.1 Introduction: Global Scales

The global outcome domain relies on measurement instruments that attempt to summarize a patient’s overall condition at a given point in time. This domain has taken on great practical importance since the US Food and Drug Administration (FDA) stated in 1991 that global assessments must be included as a primary outcome measure in anti-dementia drug trials (11). Typically, global assessment instruments are semi-structured and require evaluators to rank the level of patient improvement on an ordinal scale. Improvement is graded relative to some earlier reference point, which is usually the time of randomization. Evaluators consider cognition, function, behavior and mood when globally assessing patients.

Staging methods for dementia are designed to provide health care practitioners with a measure of the severity of illness at a given point in time as well as a means of assessing the progress of mental deterioration over time. Staging methods fall within the global domain because raters must consider a host of factors (e.g., cognition, mood, etc.) during the evaluative process. Two basic assumptions underlie the staging of dementia. First, an individual’s mental abilities deteriorate uniformly. Second, it is assumed that the progressive nature of dementia can be portrayed as a series of nonoverlapping stages (12). Burke et al. (13) claim staging instruments are useful for separating the treatment effect of a medication in a drug trial from the natural, progressive nature of dementia. Staging instruments can also help predict future levels of dependency and institutionalization in the elderly population and, as such, may be used for planning purposes in family support, domestic services and health policy (14). It has also been suggested (15) that staging instruments can help separate Alzheimer sufferers from those with other forms of dementia.
2.2 Clinician Global Impression of Change (CGIC) Scales

2.2.1 Introduction

The first Clinician Global Impression of Change (CGIC) instrument was developed by Guy (16) in 1978 to help study psychiatric diseases. The scale was called the ‘Clinical Global Impressions’ (CGI) and consisted of a severity of illness index, global improvement index and efficacy index. The severity of illness index considered the degree of mental illness exhibited by patients; the global improvement index concerned the extent to which drug treatment was responsible for overall improvements in patient status; the efficacy index was a comparison of the benefits versus side-effects of individual medications. The global improvement index is often the only part of the CGI that is employed in clinical trials because it is believed to best reflect the small treatment effects associated with most antidementia drugs (17). The global improvement index is referred to as the Clinician Interview-Based Impression of Change (CIBIC) when it is singularly employed as an outcome measure.

2.2.2 Description

Constructs Measured: CGIC scales are designed to measure the general physical and mental condition of patients. This includes functional ability with respect to activities of daily living, cognition, behavior and mood.

Administration: The FDA (11) recommends the use of clinician raters who are not personally responsible for the patients being evaluated. This helps eliminate observer bias from other clinical or psychometric assessments and preserves the blinded nature of randomized clinical trials. Since demented patients may not always provide adequate information about their global health status (17), some investigators (17;18) have modified the CIBIC to allow for the inclusion of caregiver input. Meya et al. (17) refer to their modified instrument as the CIBIC-M (‘M’ for ‘modified’); Boothby et al. (18) refer to theirs as the CIBIC-Plus.

Application: Administering physicians conduct interviews when patients are admitted to a trial and then return at one or more later dates (timing is specified in each study’s protocol) to carry out global evaluations. The interview and subsequent evaluation(s) are semi-structured affairs and loose guidelines exist to govern their conduct. For example,
clinicians are not permitted to refer to patient charts or other psychometric test results. A more structured CGIC has been developed by the Alzheimer’s Disease Cooperative Study (ADCS) (19). The ADCS-CGIC requires physicians to use a set of forms as an evaluative guideline. Fifteen areas embracing cognitive, behavioral and social/daily living domains must be assessed to obtain an overall impression of change, which is then scored on a seven-point scale.

Scoring: The CGI severity of illness index uses a seven-point scale (1=normal, 2=borderline mentally ill, 3=mildly ill, 4=moderately ill, 5=markedly ill, 6=severely ill, 7=among the most extremely ill), as does the global improvement index (1=very much improved, 2=much improved, 3=minimally improved, 4=no change, 5=minimally worse, 6=much worse, 7=very much worse). CIBIC instruments generally use the global improvement scale. The CGI efficacy index is a grid containing descriptions of four possible therapeutic drug effects and four levels of severity for drug-induced side-effects. Clinicians choose a description from each set. Both descriptions are then matched on the grid to arrive at a score from 1 to 16.

2.2.3 Evaluation

2.2.3.1 Reliability

Test-retest: Knopman et al. (20) analyzed data from a 30-week, double-blind, parallel-group RCT of tacrine and focused on the 125 patients (mean age=72.1; range=51-89) in the placebo arm who completed the study. The trial employed the CIBIC as an outcome measure; clinicians had been trained to administer the CIBIC expressly for the trial. Test-retest reliability for visits at weeks 16 and 18 was 0.35 (weighted kappa); and 0.51 for visits at weeks 22 and 24.

In a reliability study conducted by Dahlke et al. (21), 37 psychogeriatricians participated. Each clinician selected one patient from his/her practice and was then interviewed by an experienced psychologist. The purpose of the interview was to help the physician assess the chosen patient’s level of dementia on both the CGI severity and global improvement indices. The 37 clinicians provided a second rating for their chosen patients on the CGI severity index eight weeks after the interview. The Pearson correlation coefficient ($r_p$) was 0.65 (95% CI: 0.41-0.81).
Interrater: Dahlke et al. (21) also assessed interrater reliability by having a second physician and a senior nurse rate the patients chosen by the 37 clinicians. Pearson correlation coefficients (r_p) were:

- severity:
  - second physician —> 0.66 (95% CI: 0.42-0.81)
  - senior nurse ——> 0.41 (95% CI: 0.10-0.65)

- global improvement:
  - second physician —> 0.51 (95% CI: 0.22-0.69)
  - senior nurse ——> 0.35 (95% CI: 0.03-0.61)

Boothby et al. (18) videotaped eight subjects (age > 65) with probable AD (NINCDS/ADRDA criteria). Subject caregivers were also videotaped. Eighteen physicians were divided into two groups to score the subjects on the CIBIC-Plus. One group viewed taped caregiver information first, while the other looked at taped subject interviews first. The groups then viewed the tapes again, but in reverse order. Intraclass correlation coefficients (ICCs) of 0.76 and 0.82 resulted when caregiver information was shown first; ICCs of 0.66 and 0.69 were reported when subject information was shown first. The subject-caregiver pairs demonstrating the highest and lowest levels of agreement were rated by seven new physicians for an ICC of 0.70.

2.2.3.2 Validity

Content: The literature does not discuss this area.

Criterion/Concurrent: Knopman et al. (20) also examined concurrent validity using data from a 30-week tacrine trial. They found low correlations between the CGI global impressions index and both the ADAS-cog (r=0.33) and MMSE (r=-0.39). These results were based on ratings taken at the end of the follow-up period. It was not stated whether Pearson or Spearman correlations were reported.

Claus et al. (22) asked clinicians and caregivers to rate the level of “functioning” for 140 memory clinic outpatients with probable or possible AD (NINCDS/ADRDA criteria). Ratings took place over one week and were compared to observations made six months earlier. A three-point scale (worse/no change/improvement) was used; no guidelines were provided to define clinical change. The Cambridge Cognitive Examination (CAMCOG), a short neuropsychological test that helps clinicians diagnose dementia in elderly persons (23), was administered to each patient by a physician.
Percent agreement between the CAMCOG and clinicians’ global assessments of change was 35%; percent agreement between the CAMCOG and caregiver assessments was 21%.

In a study carried out by Korner et al. (24), two psychogeriatricians independently assessed 20 outpatients with memory problems, 16 of whom had probable AD (DSM-III-R criteria). Spearman correlation coefficients between the CGI global improvement index and several other instruments were: MMSE (0.93), CAMCOG (0.94), ADAS-cog (0.93) and ADAS-noncog (0.78).

Construct: Dahlke et al. (21) asked each of the 37 physicians in their study to name six personal assessment criteria for demented patients. A total of 222 such criteria were given and then grouped under seven categories (1 - cognitive/intellectual impairment, 2 - concomitant psychiatric symptoms, 3 - behavioral impairment, 4 - helplessness and dependence, 5 - impairment of language and communication, 6 - social problems, and 7 - somatic disturbances). When specified by a clinician, items in the impairment of language and communication category explained 52% of the variance in the CGI severity index; the cognitive/intellectual impairment category explained 40% of the severity variance. Cognitive/intellectual impairment also explained 24% of the variance in the CGI global impression of change index; the other six categories accounted for less than 24% of the variance in the global index, leaving most of the total variance unexplained.

Dahlke et al. (21) chose the five DSM-III-R criteria (memory impairment, impairment in abstract thinking, impaired judgment, disturbances of higher cortical function and personality change), plus ‘orientation’, as recognized categories for assessing dementia. When physicians indicated that impairment of orientation was the most important assessment criterion, this factor was found to explain 47% of the variance in the severity index and 28% of the variance in the global improvement index. All other criteria accounted for less variance in both parts of the CGI.

Responsiveness to Change: The literature does not discuss this area.

2.2.4 Critical Appraisal

CGIC scales generally demonstrate poor to good test-retest and interrater reliability. Concurrent validity is poor to very good. Some of these results may arise
from the fact that groups who provide global assessments do not necessarily base their ratings on the same domains. Physicians took clinical psychopathology as the basis of determining global improvement, while nurses believed the amount of work needed to care for patients was important (21). Claus et al. (22) found that clinicians tended to focus on patients’ cognitive abilities and caregivers placed an emphasis on patients’ behavioral and functional abilities. Overall, these findings suggest that scores on CGIC instruments are not reflections of an individual’s global degree of health improvement. Rather, scores reflect improvement on whichever domain(s) is/are considered important by the raters.

Turning to construct validity, Dahlke et al. (21) found that substantial elements of the variance in CGI scores could not be explained by the ‘personal’ or ‘recognized’ criteria that were outlined in their study. The authors conclude that this ‘noise’ could reduce the extent to which CGI global improvement ratings reflect real change in patients’ dementia. Change scores would only reflect large therapeutic effects. Indeed, change scores may only be a valid indicator for cognitive and functional decline. Quinn et al. (25) found that the validity and interrater reliability of global measures both diminish when the clinical status of patients improves.

Another results-oriented problem involves CGIC instruments that include caregiver opinion (e.g., CIBIC-plus). Results may differ depending on whether the rater first interviews the patient or caregiver. Schneider et al. (26) reported that the clinical status of patients as measured using the ADCS-CGIC was worse relative to baseline when caregivers, not patients, were interviewed first. These same patterns were also seen in Boothby et al.’s (18) study.

CGIC instruments have problems with their rating scale. Knopman et al. (20) felt that the vagueness of terms used in the seven-point CGI scale (e.g., much improved, minimally improved, etc.) could make the reasons behind a physician’s rating less explicit for outside observers. Rockwood and Morris (15) indicated that the distance between the seven points is not uniform, so the gravity of an individual’s change may not be properly expressed by the ultimate rating. For example, the largest interval (according to Rockwood and Morris) for clinically meaningful change on the improvement scale is between points four (no change) and five (minimal improvement). Clinicians may not
appreciate this fact and could judge the difference between these points to be very slight. Patients rated a ‘five’ on the scale may therefore not have achieved the degree of change implicitly denoted by a comparison to point four (15).

CGIC scales are widely used as outcome measures in AD drug trials. However, the discussion has drawn attention to some substantive problem areas that require more study and/or amelioration before these instruments can be accepted as suitable outcome measures. A first step toward acceptability would be to consider some of the following suggested improvements:

• Rate patients at the last visit before the end of follow-up so that health status at intervening visits will not be confused with health status at baseline (27).
• Develop explicit, qualitative descriptors of clinically important change for at least the middle and endpoints of the seven-point scale to enhance the meaning and uniformity of scoring (20).
• Allow clinicians to consult a caregiver without the patient being present because caregivers can best assess the “ecologic validity of a drug’s benefits.” (20)
• Interview the caregiver first to obtain a more critical opinion of the patient’s condition (18;26).
• Utilize a conservative approach to statistical analysis (employ rank order statistics) because CGI-I scales are ordinal and change is measured individually and not related to outside standards (27).

2.3 Clinical Dementia Rating (CDR)

2.3.1 Introduction

The Clinical Dementia Rating (CDR) was developed in 1977 under the auspices of Washington University’s Memory and Aging Project because a scale was needed to compare AD patients with healthy controls in a longitudinal study (28).

2.3.2 Description

Constructs Measured: Cognitive performance is measured in six areas: memory, orientation, judgment/problem solving, community affairs, home/hobbies and personal care.
Administration: A physician administers the CDR.

Application: The instrument is a semi-structured interview with the patient and an “informed collateral source” (a relative and/or caregiver familiar with the patient’s condition).

Scoring: Each of the six cognitive performance categories is given a score on a five-point scale of impairment (0 = none, 0.5 = questionable, 1 = mild, 2 = moderate and 3 = severe) and an algorithm is then used to derive an overall score (stage) of dementia (28). The overall score can be no greater than the highest level of impairment (CDR 3). The scoring mechanism is occasionally modified by summing the ratings in all six performance categories to obtain a global dementia ranking that can go no higher than 18 points. This scoring method is often referred to as the CDR-Sum of Boxes or CDR-SB. The method is preferred by some because it provides a more detailed and quantitative picture of a subject’s severity of dementia (28;29). In 1996, two additional stages (4 = profound impairment and 5 = terminal impairment) were added to the original scale to account for levels of dementia that are often seen in nursing home patients (30). This modification is called the ‘extended CDR’.

2.3.3 Evaluation

2.3.3.1 Reliability

Test-retest: No literature was found on this subject.

Interrater: Morris et al. (31) selected patients across all five CDR ratings from the (U.S.) Alzheimer Disease Research Center Patient Registry. Eighty-two researchers saw either videotaped or written interviews with these patients. Each researcher was asked to provide a CDR for three patients and some were randomly selected to watch a second set of three taped interviews. Researcher scores were compared to the CDR score given by the principal investigator. Overall ‘agreement’ (no mention of a specific statistic was made) between researcher and principal investigator ratings was 0.79. Agreement for inexperienced raters was 0.74; it was 0.83 for those who had previously used the CDR. Agreement among those who viewed the second set of tapes was 0.83.

Burke et al. (13) videotaped 25 interviews involving patients and caregivers. A clinician scored patients at the time of videotaping. Another clinician blinded to these
first results, saw the tapes later and assigned a second set of ratings. Weighted kappas were used to measure the two physicians’ agreement on CDR ratings, the ratings for each of the six cognitive categories and the score on the Sum of Boxes. These kappas ranged from 0.75-0.94 (indicating excellent agreement).

2.3.3.2 Validity

Content: Morris et al. (31) wrote that the CDR has face validity because it assesses the impact of cognitive deterioration on one’s ability to perform everyday tasks. Rubin et al. (32) and Berg et al. (28) stated that the CDR’s six cognitive performance categories are consistent with the NINCDS/ADRDA diagnostic criteria (33) for AD.

Criterion/Concurrent: Juva et al. (14) took 93 patients from the Helsinki Aging Study and administered the following scales: CDR, Mini-Mental State Examination (MMSE), Index of Activities of Daily Living (Index ADL) and Instrumental Activities of Daily Living (IADL). A general practitioner conducted the CDR, a community nurse carried out the MMSE and “close informants” administered the Index ADL and IADL. Agreement was quantified using the kappa coefficient: CDR-MMSE = 0.56; CDR-Index ADL = 0.46. No kappa was reported for the agreement between the CDR and the IADL.

Hughes et al. (4) looked at the CDR, Blessed Dementia Scale (BDS), Short Portable Mental Status Questionnaire (SPMSQ) and Face-Hand Test (FHT). They recruited 138 community dwelling individuals (age range = 65-80 years): 58 were healthy controls (CDR 0), 16 had questionable dementia (CDR 0.5) and 43 had mild dementia (CDR 1). Pearson correlation coefficients were: CDR-BDS = 0.74; CDR-SPMSQ = 0.84; CDR-FHT = 0.57.

Berg et al. (28) enrolled 43 subjects with mild AD (CDR 1) and 58 healthy controls (CDR 0) over a 19-month period. Validity was assessed between the CDR and the following scales: BDS, Dementia Scale Cognitive (DSC), SPMSQ, Aphasia Battery (AP) and FHT. Concordance between ratings on the CDR and the other scales at 15 and 34 months after entry was estimated using Kendall’s tau B statistic. Excluding dropouts, 31 to 38 patients were available for testing at 15 months and 24 to 33 were available at 34 months. Results were (15 months/34 months): CDR-BDS = 0.53/0.59; CDR-DSC = 0.55/0.70; CDR-SPMSQ = 0.68/0.50; CDR-AP = 0.53/0.68; CDR-FHT = 0.47/0.64.
Dooneief et al. (30) looked at the extended CDR. A total of 956 demented subjects were selected over a six-year period from volunteers participating in a community-based, prospective study of dementia. Measurements on the following scales were taken for each subject at baseline: extended CDR, BDS, ADL and the short Blessed Information, Memory and Concentration (sBIMC) scale. Results using Spearman rank correlation coefficients: between the CDR and increasing functional impairment on the BDS were $r= 0.67$; between the CDR and decreasing independence on the ADL were $r= -0.64$; between the CDR and worsening performance on the sBIMC were $= 0.57$.

**Criterion/Predictive:** Dooneief et al. (30) found that subjects with CDR 4 or 5 had poorer survival, although the results were not statistically significant. Median survival was one year for CDR 5, nearly two years for CDR 4, roughly two-and-a-half years for CDR 3, three years for CDR 2 and three-and-a-half years for CDR 1.

Rubin et al. (32) conducted a small longitudinal study (1979-1981) to see if subjects with questionable dementia (CDR 0.5) progressed to more severe stages of decline. Sixteen subjects with CDR 0.5 from the Washington University Memory and Aging Project were enrolled. Eleven of the 16 progressed to CDR 1 or greater during the follow-up period. Rubin et al. used a homogenous sample to define AD in its “purest form”. The authors call for their study to be replicated on a more heterogeneous sample in order to enhance generalizability.

**Construct:** Juva et al. (14) had a neurologist examine 93 study subjects and then rank dementia severity according to DSM-III-R criteria. The kappa coefficient between the CDR and DSM-III-R criteria was 0.56.

Dooneief et al. (30) found that the extended CDR was only minimally correlated with increasing age ($r = 0.21$) in 956 demented subjects, but they concluded that the profound (CDR 4) and terminal (CDR 5) stages had “prognostic utility with respect to survival”. However, the results for survival did not achieve statistical significance most likely due to the small sample size ($n = 14$).

**Responsiveness to Change:** The literature does not discuss this area.
2.3.4 Critical Appraisal

The CDR is useful because: (i) it provides physicians with a global rating that encompasses a broad range of patient characteristics; (ii) it can be used by neurologists, psychiatrists, psychologists and others who assess cognition in the elderly; and (iii) it focuses on cognition, not on items that may be related to other medical, emotional or social conditions (4). The scale is also amenable to modification. For example, the CDR has been altered to clarify levels of severity in the six cognitive categories (29).

The CDR exhibits good interrater reliability and fair to good concurrent validity. Some findings indicate that the instrument may have predictive value with respect to disease progression and survival. Although no work seems to have been done on test-retest reliability, nothing so far suggests that researchers should avoid this scale when trying to stage AD. The CDR can be used as an eligibility criterion for trial participation or as an outcome measure. Users must note that the instrument tends to measure cognitive aspects of dementia, not a patient’s global health state with respect to AD. The following suggestions should help optimize the CDR’s use in antidementia drug trials:

Develop a more rigid, standardized interview to replace the current semi-structured format, which injects a degree of subjectivity into the CDR rating process and therefore limits the standardization of scores across subjects and studies (31).

Ensure that evaluators rate all six cognitive categories before obtaining an overall score so that the reliability and usefulness of the CDR can be improved (13).

2.4 Global Deterioration Scale (GDS)

2.4.1 Introduction

The Global Deterioration Scale (GDS) was developed in 1982 to stage “the magnitude of cognitive impairment in aging and dementia” (1). Its use is based on the notion that primary degenerative dementia has readily identifiable clinical phases that can be represented on an ordinal scale.

2.4.2 Description

Constructs Measured: Progressive stages of cognitive impairment.
**Administration:** Clinicians are responsible for administering the GDS. No formal training is required beyond simple instructions.

**Application:** The GDS is suitable for adults with symptoms of cognitive decline.

**Scoring:** The printed version of the scale is a grid delineating seven stages: 1=no cognitive decline, 2=very mild cognitive decline, 3=mild cognitive decline, 4=moderate cognitive decline, 5=moderately severe cognitive decline, 6=severe cognitive decline, 7=very severe cognitive decline (1). One or more descriptions of cognitive impairment are given for each stage. Physicians are asked to “rate the subject’s level of cognitive functioning” by selecting one of the stages (34).

### 2.4.3 Evaluation

**2.4.3.1 Reliability**

**Test-retest:** Reisberg et al. (34) report on an earlier study of 38 patients (mean age=69.3; range=54-82), some of whom were demented (either with or without evidence of cerebrovascular disease). The “test-retest correlation coefficient” of the GDS at seven-day to four-month intervals was 0.92.

**Interrater:** Gottlieb et al. (35) assessed 43 ambulatory patients (mean age=72.8; range=55-88) who met the NINCDS/ADRDA criteria for probable AD. The GDS was administered independently by two of the authors, the interval between administrations was a maximum of one week and consultation with one family member was permitted. The intraclass correlation coefficient was 0.82. The two raters were within one point of each other’s scores for 41 of the 43 subjects. In fact, their scores matched for 30 subjects.

Foster et al. (36) selected 40 patients from a municipal psychiatric hospital who had been diagnosed with one of many maladies, including organic mental disorder, personality disorder, adjustment disorder, psychotic and affective disorder and uncomplicated bereavement. Patients were divided into two groups of 20: five attending psychiatrists rated the first group (mean age=63.0; range=30-82) and three research assistants rated the second (mean age=52.6; range=24-81). Each patient was interviewed by one of the raters for 45-60 minutes while the other raters observed. At the conclusion of an interview, all raters independently scored the patient on the GDS. Interrater reliability coefficients were 0.97 (first group) and 0.92 (second group).
2.4.3.2 Validity

Content: The GDS was developed without the use of psychometric and statistical methods. Instead, a more conceptual approach was undertaken and, therefore, some assumptions had to be made regarding disease occurrence, rate of change in severity, interrelationships of different forms of cognitive impairment, behavioral dysfunction, impairment in activities of daily living, and psychiatric symptoms (37). The scale was initially validated against various assessments of brain function and change (34). As a result of this developmental process, the GDS rests on the hypothesis that AD is homogeneous, whereas growing evidence suggests the syndrome may be heterogeneous. Additionally, the instrument can misidentify disease severity (37).

Criterion/Concurrent: Korner et al.’s (24) study (refer to the section on CGIC scales for more information) reported correlation coefficients ($r_s$) from a comparison of the GDS and several other measurement scales. Multiple comparisons were chosen because no true gold standard exists to quantify AD. The coefficients were: GDS-CGI (0.88), GDS-MMSE (0.93), GDS-CAMCOG (0.94), GDS-ADAS-cog (0.92) and GDS-ADAS-noncog (0.66).

Cohen-Mansfield et al. (12) compared the GDS with the Functional Capacity Scale (FCS), a seven-point ordinal staging instrument for dementia. Each point on the FCS represents progressively greater disability. The authors found a strong correlation of 0.97 after neurologists rated 240 normal elderly subjects and 142 cases of AD. The measure of correlation was not clearly specified.

Heun et al. (38) employed receiver operating characteristic (ROC) curve analysis to examine several dementia-related measurement instruments. Their effort was part of a larger epidemiologic study of the general elderly population. Data were reported for 287 individuals who completed more than one of the instruments during a home interview lasting from one to three hours. To achieve 95% sensitivity, scores greater than two were needed to distinguish a demented person from a healthy individual. At 95% specificity, scores greater than three were needed to make the distinction.

Construct: In a retrospective study, Reisberg et al. (1) examined the association between GDS scores and other psychometric assessments such as the Inventory of Psychic and Somatic Complaints in the Elderly (IPSCE). The sample was composed of patients with
very mild to moderately severe dementia. The GDS was found to have statistically significant correlations with 13 out of 19 cognitive elements in the IPSCE and 25 out of 26 miscellaneous psychometric measures (e.g., simple reaction time, perceptual speed, etc.). Actual correlation coefficients were concentrated in the -0.64 to 0.66 range (inverse/negative correlations occur because greater impairment on some scales is signified by a lower score, whereas the GDS quantifies more severe AD with a higher score).

**Responsiveness to Change:** The literature does not discuss this area.

### 2.4.4 Critical Appraisal

The GDS is one of the most frequently used instruments for staging dementia (34). However, clinical trial researchers must be aware that GDS ratings can misstate a patient’s disease severity. Eisdorfer et al. (37) state that severe functional and psychiatric distress are both represented in the later GDS stages. Yet, these problems can occur earlier. Reisberg et al. (34) and Eisdorfer et al. (37) suggest that many of the descriptors of cognitive impairment for the seven GDS stages do not adequately reflect the actual clinical events that occur within those stages.

Another problem may arise when the GDS is used as an inclusion criterion for participation in an RCT. Often, potential subjects must exceed a certain score on the GDS before being allowed into a trial. This helps ensure that participants are severely demented enough to maximize the chance of detecting a treatment effect. The ability to enroll desired patients could be threatened if the GDS misidentifies the stage of dementia. A trial that included patients who were not as demented as desirable could produce results with a large variance. As well, excluding patients who should be included has implications for generalizability and would also require approaching more potential subjects.

These problems suggest that the GDS should not be used to stage dementia in AD drug trials.
2.5  Gottfries-Brâne-Steen (GBS)

2.5.1  Introduction

The Gottfries-Brâne-Steen (GBS) scale was developed by Gottfries et al. (39) in 1982 to provide a means of estimating the different types and degrees of dementia. The developers intended the scale to be an easily administered method of judging (i) physical inactivity, (ii) mental symptoms common to dementia and (iii) intellectual and emotional incapacity. The GBS can be used repeatedly on the same person and it is suitable for evaluating drug treatments, but it is not meant to be a diagnostic tool. The GBS was based on: the Comprehensive Psychopathological Rating Scale, the Sandoz Clinical Assessment Geriatric (SCAG) and the Geriatric Rating Scale of Gottfries and Gottfries (GRSGG).

2.5.2  Description

* Constructs Measured:* Motor function, intellectual function, emotional function and symptoms common to demented patients (e.g., confusion, irritability, anxiety, agony, reduced mood and restlessness).

* Administration:* Staff who are well informed of the patient’s status should administer the scale. Gottfries et al. (39) used a psychiatrist, psychologist and several registered nurses in their original evaluation of the scale’s psychometric properties.

* Application:* Patients with a confirmed diagnosis of dementia.

* Scoring:* Each of the four functional sections of the scale contains several items (e.g., the intellectual section has one item called “impaired recent memory”); three statements commonly follow every item (e.g., one such statement is “patient has some impairment of recent memory but this is evident only in more detailed conversation or testing”). If a particular statement accurately describes a patient, then the rater marks the number corresponding to that statement. These numbers (scores) range from 0 (normal function or absence of symptoms) to 6 (maximal disturbance or presence of symptoms). A total score is obtained by summing the scores of all individual items. However, Gottfries et al. do not provide a qualitative interpretation of different total scores. Potential users must judge what constitutes a ‘good’ or ‘bad’ score.
2.5.3 Evaluation

2.5.3.1 Reliability

Test-retest: Villardita et al. (40) administered six measurement instruments (over a three-day period) to 41 community dwelling patients with probable AD (NINCDS/ADRDA). A second administration was conducted four weeks later. Patients did not undergo any treatment for memory disorders during the four-week period. A trained technician or a neurologist was responsible for evaluating the patients. The ‘test-retest reliability coefficient’ for the GBS was 0.59 (p<0.001).

Interrater: Gottfries et al. (39) tested interrater reliability with 70 nursing home patients (mean age=81; SD=9.0) and 30 psychogeriatric patients (mean age=78; SD=6.2). All subjects had symptoms of dementia. The GBS was administered independently by two nurses in the nursing home and by a nurse, psychologist and physician in the psychogeriatric ward. Each patient was rated a second time within three days. Correlations ($r_s$) ranged from 0.57 to 0.97.

2.5.3.2 Validity

Content: The literature does not report on content validity.

Criterion/Concurrent: Gottfries et al. (39) compared the GBS and the GRSGG using one nurse and the 70 nursing home patients mentioned earlier. The GRSGG measures motor impairment, intellectual impairment and emotional bluntness. It contains simple ‘yes’ or ‘no’ answers and cannot be used for repeated measures. Correlation coefficients (exact type not specified) for the various sections of both scales ranged from 0.42 to 0.92. The authors did not report the interval of time between administrations of both scales.

Jensen et al. (41) compared the GBS with both the MMSE and the Organic Brain Syndrome (OBS) scale. The OBS can be used on demented patients and consists of a patient interview to assess short-term health and a staff/caregiver interview to ascertain long-term health. Twenty-eight demented patients (mean age=76; range=66-89) from either a day care unit or a collective living facility were enrolled in the study. Subjects were diagnosed with AD, vascular dementia, a mix of these two illnesses, bipolar disorder (1 patient) or post-traumatic personality change (1 patient). Correlations ($r_s$) were: GBS-MMSE (-0.80; p<0.001), GBS-OBS (0.82; p<0.001). Aside from mentioning
that a trained interviewer administered the scales, the authors provided no other information on study design (e.g., interval between test administrations).

Villardita et al. (40) also conducted validity testing with the 41 patients in their study. Corrected correlation coefficients (exact type not specified) were (entry/week 4):

GBS-SCAG (0.98/0.98), GBS-MMSE (-0.77/-0.79), GBS-Cognitive Capacity Screening Examination (-0.76/-0.99), GBS-BIMC (0.76/0.78).

Folnegovic-Smalc et al. (42) administered the GBS, MMSE and SCAG to 40 patients (mean age=69; range=46-89) with multi-infarct dementia. Specifics of test administration were not discussed. A high correlation (r=0.73; p<0.0001) was found between the GBS and SCAG. No association was observed between the GBS and MMSE (r=0.02; p=0.9138). The type of correlation coefficient used was not specified.

Construct: Parnetti et al. (43) conducted a principal component analysis on the GBS using 138 outpatients (age>75) fulfilling NINCDS/ADRDA criteria for AD. One hundred sixteen healthy subjects (age>75) served as controls. Three factors, ‘general functioning’, ‘depression’ and ‘restlessness’ accounted for 74% of the variance in GBS scores. Using what is only described as a ‘QROC’ analysis, Parnetti et al. concluded that subjects scoring above 8 on the GBS had an 88% probability of being diagnosed with AD. Patients with a score of less than 8 had a 90% probability of not being diagnosed with AD.

Nyth and Bråne (44) conducted a principal component analysis using 221 patients, the majority of whom were diagnosed with AD, vascular dementia or some unspecified form of dementia. Four components explaining 64.5% of the variance in scores were uncovered. These components were not named. Each component encompassed a mixture of items from the GBS.

Internal Consistency: Nyth and Bråne (44) used the results of the principal component analysis to generate new GBS subscales. These were: impaired orientation and memory, activities of daily living, depression-anxiety and impaired attention and motivation. Two samples were then utilized to calculate the internal consistency of these subscales. The first sample had 98 patients with moderate AD or vascular dementia that participated in a multicentre trial of citalopram versus placebo. The second group had 87 patients with mild to severe AD or vascular dementia who took part in a longitudinal study of
dementia. Cronbach’s $\alpha$ ranged from 0.73 to 0.97 for separate analyses of each subscale and sample.

### 2.5.4 Critical Appraisal

The psychometric properties of the GBS range from fair to very good. The scale appears to be a useful means of quantifying dementia in drug trials. On a cautionary note, the GBS should not be used as a diagnostic tool. Gottfries et al. (39) made this point when they developed the instrument. Olafsson et al. (45) carried out a correlational analysis with 39 patients suffering from multi-infarct dementia and 34 others suffering from AD. They concluded that the GBS was not of diagnostic value in differentiating between these two illnesses.

The GBS appeared in only one of the trials (46) reviewed in this report, most probably because it is not a well-known instrument among North American researchers. Although its creators refer to the GBS as a functional scale, the sections touch upon a number of cognitive, functional and behavioral components. The scale can therefore be considered part of the global domain. Perhaps the future of the GBS in North America is to be a substitute for the confusing number of instruments that fall under the Clinician Global Impression of Change (CGIC) banner. However, concurrent validity between the GBS and CGIC scales must be assessed before any such substitution can be considered.

### 2.6 Mini-Mental State Examination (MMSE)

The Mini-Mental State Examination (MMSE) can be employed as a staging instrument. More will be said about this specific role in the section pertaining to the cognitive domain.

### 2.7 Global Scales: Strengths and Limitations

Global measurement scales (with the possible exception of the GBS) are widely used as primary outcomes in AD drug trials. These scales have shown fair to good reliability and validity (mostly concurrent validity) in many psychometric studies. However, the term ‘global’ appears to be something of a misnomer. Global scales have a tendency to measure specific constructs. Cognition seems to be most popular when
physicians administer the scales. Caregivers and nurses, on the other hand, often rate patients according to behavior, mood or caregiver burden.

If global instruments continue to be used in AD drug trials, then raters should be specifically instructed to consider input from a host of domains. Otherwise, these instruments could become redundant when other scales such as the ADAS-cog are employed in the same trials. Furthermore, FDA guidelines mandating measures of global change in RCTs may not be followed.

There has been little research into responsiveness to change. No standard rule suggests a minimum difference between pre- and post-treatment scores that would have to be observed before one could conclude evidence of a treatment effect. Researchers are left with the common practice of selecting their own clinically meaningful differences. In the case of the AD trials under review, many authors did not explicitly discuss these differences; most either arbitrarily selected a difference or relied on statistically significant differences to indicate an effect.

Research into responsiveness to change must be conducted if global scales are to provide useful indications of treatment efficacy. Until then, the use of these instruments as primary outcome measures should be discouraged. One global scale, the CDR, appears adequate for use in AD drug trials as a baseline measure and/or inclusion criterion.
3  COGNITIVE OUTCOME MEASUREMENT SCALES

3.1 Introduction

The assessment of cognitive function in AD is multidimensional. Examinations typically assess memory, language, praxis, visual perception, attention and abstract reasoning. An overview may be found in Strub and Black (47) and Albert (48;49). According to Zec (50), there are three major objectives for cognitive appraisals, namely (i) early detection, (ii) differential diagnosis and (iii) the measurement of disease severity and progression. Each of these applications necessitates a different combination of neuropsychological tests.

Because successful treatment of dementia usually requires improvement in the primary cognitive symptoms (51), procedures for evaluation of cognition have come to play a key role in clinical trials of AD. Patients can be tested with an array of instruments ranging from composite cognitive scales to individual cognitive tests (52). Composite cognitive scales provide a comprehensive index of cognitive function without emphasizing any particular area of cognition. On the contrary, individual neuropsychological tests target specific areas of cognition (e.g., memory, language), but the individual scores are not meant to be quantitatively combined and cannot therefore provide a comprehensive index of cognition. To indicate their importance as outcome measures in AD drug trials, composite cognitive scales will be reviewed first.

3.2 Alzheimer’s Disease Assessment Scale-cognitive (ADAS-cog)

3.2.1 Introduction

The Alzheimer’s Disease Assessment Scale (ADAS-cog) is the cognitive subscale of the ADAS (53;54). There is also a noncognitive subscale to the ADAS (i.e., the ADAS-noncog) (see section on behavior and mood scales). Originally developed as a global measure of cognitive functioning in AD patients, the ADAS was intended to assess efficacy of drug treatment. However, investigators have realized that the scores from the two subscales are virtually uncorrelated and have concluded that separate reporting of the subscales was more informative (55). In practice, the total score is seldom used.
The ADAS-cog is an 11-item scale that is often selected as a primary outcome measure for efficacy. It is overall one of the most widely used scales in AD drug trials.

3.2.2 Description

Constructs Measured: The ADAS-cog includes 11 items concerned with orientation and memory (immediate recall and recognition), language and praxis (56) (57-59). Memory and orientation items carry the most weight in scoring (50%), followed by language (36%) and praxis (14%).

Administration: Most items are evaluated by tests, although a few (e.g., recall of test instructions, spoken language ability, comprehension of spoken language, and word finding difficulty) are rated by a clinician who grades impairment along a six-point scale.

Application: Intended for use on individuals with cognitive impairment resulting from AD.

Scoring: The scale ranges from 0 to 70. A higher score indicates greater dysfunction. An experienced examiner can administer the scale in 30 to 45 minutes (58).

3.2.3 Evaluation

3.2.3.1 Reliability

Test-retest: Test-retest reliability has been found to be high for the ADAS-cog when cognitively impaired subjects were tested. The Spearman rank-order correlation was 0.92 in a study of 27 well-educated elderly with AD (mean age=65.1; SD=7.4) (54). Similar results were found with AD patients enrolled in the placebo arm of clinical trials (58;59). Test-retest reliability was lower for the ADAS-cog when cognitively intact subjects were tested, but this probably reflected an attenuation of the correlation coefficients that often accompanies restrictions in score range (60). Test-retest reliability, estimated from 440 AD patients enrolled in a German multicentre RCT, was 0.93 (61).

Interrater: The ADAS-cog has high interrater reliability. In a group of 27 well-educated elderly subjects with AD, Rosen et al. (54) reported an intra-class correlation coefficient (ICC) of 0.99. However, Ritchie (62) considered this to be an overestimate because the procedure lacked independent assessments (there was one interviewer and 2 raters). Therefore, true rater variability was likely to have been masked. Standish et al. (63)
reported an ICC of 0.82 for a group of 27 elderly AD subjects. Interviewers were trained medical students. With the same group of subjects, the ICC for intra-rater reliability was 0.86 (63).

**Reliable Individual Change Scores:** From the German multicentre trial’s (61) ICC, it was possible to calculate the standard error of measurement (SEM). The SEM represents that part of the overall variability in score that is due to measurement error and it can be used to derive a confidence interval for reliable individual changes. From the RCT, Weyer et al. (61) estimated that the reliable change score was 7. According to this estimate, a difference of 7 or more would be likely to reflect a real individual change, whereas a smaller difference would be psychometrically irrelevant.

### 3.2.3.2 Validity

**Content:** The ADAS-cog has limitations in terms of content validity. The scale does not assess attention (concentration) as explicitly as it should. Instead, the ADAS-noncog covers this aspect. The ADAS-cog does not assess executive functions, nor does it tap all the components of memory (e.g., delayed recall) (64).

**Construct/Concurrent:** The ADAS-cog showed moderate to good correlation with several cognitive scales. Correlations were: BIMC (-0.78), DRS (0.48) (54), BCRS (0.80), SKT (0.82) (55). Also, the ADAS-cog showed good correlation with the MMSE (range -0.81 to -0.90) (55;65). Doraiswamy et al. (66) re-analyzed data from 1648 RCT patients with AD (DSM-III-R) whose GDS was equal to 4 or 5. Correlations were: ADAS-cog and MMSE (r=0.76); ADAS-cog and GERRI (r=0.40).

**Criterion/Discriminant:** The ADAS-cog discriminated well between normal and AD subjects when a cutoff of 2 SD above the mean of normal subjects was used (50;54).

### 3.2.3.3 Responsiveness to change

**Longitudinal Studies:** Data on change in ADAS-cog scores is available from 4 longitudinal studies. The mean decline over one year among AD patients was similar (about 7 units) in 3 of the studies, although the baseline level indicated more impairment in one study (67). By contrast, there was no decline among cognitively intact patients. The progression of disease was assessed in a fourth study (68) where 111 AD patients were evaluated at 6-month intervals; subjects had a mean ADAS-cog of 35 (SD 3; the
variability in scores seems surprisingly small given that scores were said to be in the range from 5 to 69). The mean 1-year change scores plotted against their first measurement displayed an inverse U-shape, which suggests that change scores vary according to baseline measures. For example, according to the curve, the mean yearly change in ADAS-cog would be 10.8 at a baseline level of 25, whereas it would be 13.0 at a baseline level of 35.

Severity: Zec et al. (69) investigated the sensitivity to decline of the ADAS-cog score on subtests in patients with no, mild, moderate or severe AD. (Severity of disease was defined using the MMSE: MMSE≥20, mild AD, n=23; MMSE 10-19, moderate AD, n=33; MMSE 0-9, severe AD, n=5). This analysis indicated that the ADAS-cog score itself and some subtests (e.g., orientation and word recognition) were sensitive at all levels of severity while other subtests showed maximal sensitivity at one level or another of the severity continuum (e.g., early AD: word recall; early and middle AD: word recognition, remembering test instructions, speech items; middle and late AD: commands; late AD: naming objects and fingers, constructional praxis, ideational praxis). According to Zec et al. (69), the differential rate of decline of the ADAS-cog subtests reflects, to some extent, the genuine changing pattern of cognitive impairments with progressing AD. However, these authors reported that some subtests seemed inadequately sensitive to decline in the early stage of disease (e.g., naming and praxis). It follows, they argued, that the ADAS-cog would be inadequately sensitive to detect improvement in these functions when patients with mild to moderate dementia were tested. On the contrary, Ihl et al. (55) found no ceiling or floor effect on the ADAS-cog and Mohs and Cohen (56) found a floor effect. In their reanalysis of 1648 patients using the Geriatric Deterioration Scale, Doraiswamy et al. (66) showed that patients with a GDS=5 had a mean ADAS-cog score that was 50% higher than the mean of patients with a GDS=4.

Age and Education as Confounders: There is limited information concerning age and education effects for the ADAS-cog. There was no age effect found in one study (50) that included only a few subjects over 79 years old. There was no education effect in the same study (50), but the mean values for the ADAS-cog were not reported. An education effect was reported in a study of AD patients enrolled in the placebo arm of a drug trial.
The group with the lowest level of education (no high school degree) had a higher mean ADAS-cog than the group with the highest level of education (post-graduate degree) (23.4 vs. 18.3). Stern et al. (68) showed that age was not associated with the rate of decline on the ADAS-cog. It is still unknown whether education is associated with the rate of decline, an important consideration when measuring change (70).

### 3.2.4 Critical Appraisal

**Clinically Important Change:** The FDA stated that a treatment for AD would be judged clinically important if it was able to reverse the natural history of cognitive decline by at least 6 months (63). One widely quoted study equated a 6-month change on the ADAS-cog with a 4-point change on that subscale (67). A more recent study of the rate of cognitive decline in 111 AD subjects using the ADAS-cog suggests that the rate of decline is not a constant but is dependent on the stage of disease. According to the estimates provided by Stern (68), the mean decline for patients with mild dementia was 8-9 points, while the mean decline for subjects with moderate dementia was 12-13 points. Patients with mild dementia decline on average more slowly than patients with moderate dementia (71).

Standish et al. (63) criticized the guidelines for administering both the ADAS-cog and ADAS-noncog. The guidelines were judged to be inadequate, too brief and too open to interpretation, especially in areas such as cueing, timing of response and wording of questions for the orientation tasks. Inadequate guidelines could provide opportunities for variability in administration and scoring. Standish et al. (63) developed strict guidelines for administration and scoring which insured greater consistency in assessment. The reliability of the ‘standardized’ ADAS was compared to that of the usual version in a randomized, double-blind trial involving a total of 54 elderly AD patients with a broad range of dementia severity levels. While the interrater and intrarater reliability of the standardized ADAS-noncog was substantially better than that of the usual ADAS-noncog, there was no improvement in reliability in the standardized ADAS-cog compared to the regular version.

Despite the lack of improvement, common sense points to the logic of providing testers with one official standardized version that could be used internationally. This
would make comparisons across trials and across cultures more valid and detract users from changing the scale. Already, in the German version of the ADAS-cog, pictures have replaced word lists; this represents a modification of the original scale that concerns the testing of a different cognitive subdomain (63).

Most of the ADAS-cog subtests are limited in their ability to detect change at one end or the other of the severity continuum. For many subtests, detection of improvement appears only possible for a restricted range of severity levels. Differential response to treatment as a function of baseline severity may be due to changes in those subtests that are maximally sensitive to the severity range of interest. This is a limitation as far as use in RCTs is concerned. Jarvik et al. (72) suggested improving the sensitivity of the ADAS-cog by adding delayed recall and delayed recognition, and simplifying the drawing of a cube that does not differentiate between normal and demented elderly.

Mattes (73) scrutinized the ADAS-cog results of 48 AD subjects enrolled in a RCT to identify tasks with floor or ceiling effects. The Word Recall task appeared to be too difficult: subjects remembered on average 1.7, 2.3 and 2.9 words after 3 successive attempts. Some recalled none (floor effect). On the other hand, the Word Recognition task appeared well suited to the subjects and showed neither a floor nor a ceiling effect. The authors suggested eliminating the third Recall attempt (which brought no improvement in performance) and varying the number of words in the Recall task to suit the degree of severity of the patients. These recommendations, although compelling, are based on data from a small number of subjects with unknown levels of dementia severity. In addition, results on the ADAS-cog were incompletely presented. These results need to be replicated before the ADAS-cog is modified.

The limitations of the ADAS-cog should be considered when used as a drug efficacy measure. According to Mohs (71), patients at entry into a long-term trial should be carefully balanced for severity of dementia – the expected natural history of a patient group (e.g., the rate of decline) will be strongly influenced by the severity of dementia at the time the study begins.

The reliable individual change score was estimated from one German trial. Although this estimate provides additional information on the psychometric properties of
the scale, it is not directly applicable to the English-language version. This analysis
should be redone with appropriate data from the English-language version.

3.3 Syndrome Kurtz Test (SKT)

3.3.1 Introduction

The Syndrome Kurtz Test (SKT) is a cognitive scale that was developed and
standardized in German-speaking Europe (74); the initial publication goes back 20 years.
Documentation on the psychometric properties of the SKT comes mostly from studies in
German-speaking patients, although testing of English-speaking patients has been done in
US drug trials (74-76).

3.3.2 Description

Constructs Measured: The SKT is a condensed neuropsychological battery whose
subtests deal with brevity, performance modes and levels of difficulty for mild or
moderate levels of dementia (74). Two factors have consistently been identified by factor
analysis: ‘attention’ (or speed of information processing) and ‘memory’.

Administration: Performance is timed and limited to 60 seconds on the six ‘attention’
subtests (e.g., Naming Objects, Naming Numerals, Arranging Blocks, Replacing Blocks,
Counting Symbols, Reversal Naming). Performance is not timed on the three ‘memory’
subtests (e.g., Immediate Recall, Delayed Recall, and Recognition Memory). Total
administration takes about 15 minutes. A good relation between tester and patient is
necessary for the satisfactory completion of the test (77).

Application: Developed for patients with dementia of mild to moderate degree, the SKT
can be used for patients with psychiatric disease or other conditions that cause cognitive
problems. The main area of application is in follow-up studies and RCTs, but the SKT is
also used in routine clinical testing and basic research. The SKT has been administered
to several thousand elderly persons and has also been adopted as an outcome
measurement in many European RCTs. The availability of 5 parallel forms with
equivalent subtests circumvents any learning effect in drug trials.

Scoring: ‘Attention’ tasks are scored as the number of mistakes plus the number of
seconds to completion (or as 60 seconds if unfinished). ‘Memory’ tasks are scored as the
number of stimuli not recalled or recognized correctly. The summation over the 9 subtests constitutes the raw score. A higher score on the SKT indicates inferior performance.

In German-speaking countries, raw scores are normalized according to the SKT manual (76). SKT norms are stratified on 4 age groups (17-44, 45-54, 55-64 and 65 years and older) and 3 levels of estimated premorbid intelligence (IQ: less than 90, 90-110 and over 110). The normalized scores range from 0 to 27 and identify 5 levels of severity (76):
- 0-4 = no cognitive impairment
- 5-8 = very mild or questionable
- 9-13 = mild dementia
- 14-18 = moderate dementia
- 19-23 = moderate to severe dementia
- 24-27 = severe dementia

Scores on translated versions of the SKT have been treated in different ways: (i) standardized according to the German norms (75), (ii) transformed to T-scores with a mean of 50 and SD of 15 (78) or (iii) used in their raw form (74). Note that the use of mean T-scores produces factor scores for ‘memory’ and ‘attention’ that can be compared directly (78).

3.3.3 Evaluation

3.3.3.1 Reliability

Test-retest: Overall and Schaltenbrand (78) reanalyzed SKT tests for three German antidementia drug trials. A total of 165, 217 and 201 patients with mild or moderate dementia (not restricted to AD or primary degenerative dementia) were involved in each trial; the age distribution was not reported. Patients were assessed twice: 14 days apart in the first two trials and 90 days apart in the third trial. Pearson product-moment correlation coefficients ranged from 0.76 to 0.84 for the ‘memory’ factor; 0.87 to 0.89 for the ‘attention’ factor; and 0.83 to 0.89 for the SKT total score. The correlation coefficients for the subtests were mostly above 0.75. In the three trials, drug treatment
was started soon after the first assessment, although subjects varied from one another
with respect to length of drug exposure. This could have biased the results of the study.

Kim et al. (74) presented a reanalysis of a US multicentre study of linopirdine. The 265 subjects satisfied the criteria for Alzheimer’s disease (NINCDS/ADRDA), had a
MMSE between 14 and 23 and were retested 14 days apart. Correlation was 0.75 for the
‘memory’ factor; 0.93 for the ‘attention’ factor and 0.90 for the SKT total score. The
correlation coefficients ranged between 0.62 to 0.86 for all subtests, except for the
Delayed Recall task (0.52). Age distribution and type of correlation coefficient were not
specified.

3.3.3.2 Validity

Content: Speed vs. Quality: Timed-administration and timed-scoring is a feature unique
to the SKT. This feature is questionable to some psychometricians and
neuropsychologists (74) because it has been traditional in the USA to focus on the quality
(or accuracy) of performance instead of speed. Kim et al. (74) devised an experiment to
examine whether ‘quality’ scoring could add to the information provided by the ‘speed’
scoring of the SKT. They developed a 5-level quality score that was applied to each
‘attention’ subtest of the SKT (recall that the ‘memory’ subtests are not timed);
distinction was made in the scoring between a simple error and a gross error. The sum of
quality scores across the ‘attention’ subtests resulted in the quality ‘attention’ subtest.
Both timed and quality SKT were administered twice at baseline to AD patients (see
above). Test-retest reliability for the ‘attention’ subtest was similar for the timed (0.93)
and the quality (0.86) scores. Correlations between the timed and quality ‘attention’
subtest were also good (0.72 and 0.75); correlations between task items were not as
impressive, ranging from 0.52 to 0.67 on the first testing and 0.49 to 0.77 on the second.
Correlation coefficients for the timed ‘attention’ subtest with the MMSE (-0.57) and the
ADAS-cog (0.63) were similar to coefficients of the quality ‘attention’ subtest with the
same scales - MMSE (-0.55) and ADAS-cog (0.67). Factor analysis provided similar
loadings for the timed and quality SKT. The authors interpreted the striking similarity of
results between the timed and quality scales as highly suggestive of similar underlying
cognitive dimensions being measured in different ways. The experiment appeared to
support the interpretation that time measurements reflected something more cognitively relevant than speed of performance alone.

**Construct/Concurrent:** Weyer et al. (61) compared baseline scores on the SKT and ADAS-cog for 440 patients with AD (NINCDS/ADRDA) who were enrolled in a multi-centre trial. Patients had a mean age of 70 years and MMSE of 10 to 24. The mean of the SKT scores was 18.6 (SD 4.7; range 9-27) and the mean of the ADAS-cog was 34.3 (SD 8.6; range 15-61). The Pearson product-moment correlation coefficient was 0.75. In a previously described reanalysis (74), the correlation coefficient between the SKT and the ADAS-cog was 0.69; between the SKT and the MMSE it was -0.60. When the ‘memory’ and ‘attention’ factors were used instead of the SKT total score, the correlation coefficients were somewhat lower (‘memory’ factor : ADAS-cog 0.47; MMSE -0.38; ‘attention’ factor : ADAS-cog 0.63; MMSE -0.57). Correlation coefficients corrected for random error (disattenuated) were also presented. These were, as expected, higher than the uncorrected ones. The corrected coefficients have theoretical interest only and they cannot be compared to the (uncorrected) correlation coefficients that are usually reported in other studies. It was unclear whether Kim et al. (74) used Pearson’s or Spearman’s correlation coefficients.

In another reanalysis (55), data from 49 clinic patients with probable AD (mean age=70.4 years; range=45-84) showed good correlations ($r_s$) between the SKT and the following cognitive tests: ADAS-cog (0.82); MMSE (0.81); BCRS (0.80).

**Construct/Factor structure:** Factor analysis in the demented elderly has repeatedly shown a two-factor structure interpreted as deficits in ‘memory’ and ‘attention’ (or speed of information processing) (75;76). In cognitively intact elderly, factor analysis showed a three-factor structure, which was interpreted as deficits in ‘memory’, ‘attention’ and ‘language fluency’ (75). Plausible explanations for the differing factor structure between the cognitively intact and impaired elderly were provided by Lehfeld et al. (75) and Kim et al. (74). Cross-cultural stability for the factor structure was shown in one study, where data from Chile, Greece, Russia, England, and Germany were compared (75).
3.3.3.3 Responsiveness to Change

By comparing the distribution of SKT scores from 49 demented subjects with cutoffs on 3 different cognitive scales (BCRS, MMSE and ADAS-cog), Ihl et al. (55) evaluated the SKT for score structure as related to severity of dementia, floor and ceiling effects. They found that the SKT showed sensitivity to levels of severity when contrasted with these three scales. However, the SKT may overestimate severity in comparison to the ADAS-cog and MMSE - a floor effect of this scale seems to happen at a lesser degree of severity. Consequently, the SKT should only be used with subjects with mild or moderate dementia (specific cutoffs for severity are provided vis-à-vis the BCRS, MMSE and ADAS-cog). This conclusion is consistent with the SKT manual, which does not claim to measure the whole range of severity of the disease and recognizes a floor effect on the scale.

3.3.4 Critical Appraisal

Adequate reliability and validity were shown with the SKT in studies involving AD subjects. However, except for one study, the results apply to German-speaking populations and are only valid for these populations (76).

The timed-administration and timed-scoring of the SKT are problematic for some psychometricians. This evaluation does not support such a view. Timed features are common to several individual neuropsychological tests that have proven useful in measuring attention. In addition, one study of the SKT showed that timed performance was as informative as quality scored performance (74).

There is a lack of information on ‘clinically important change’ and ‘reliable individual change scores’, at least in English-language literature and patients. Care should be given to the manner in which results are reported (e.g., raw score, t-scores or standardized scores) when these concepts are interpreted.

Standardization needs to be discussed in the context of translated versions – one suspects the German norms may not apply to all populations.

Factor structure differs in demented versus cognitively intact patients; this aspect should be investigated further.
The SKT features some difficult tasks for demented subjects, such as remembering 12 stimuli (objects). The SKT is limited to the assessment of patients with mild or moderate dementia. A floor effect is to be expected on the scale.

Clearly, there is a need for more evaluative work before the SKT can be adopted as a primary outcome measure in drug trials. The SKT is complementary to the ADAS-cog (which has no attention tasks) and therefore has the potential to fill a unique role in the measurement of cognitive ability in AD patients (78).

3.4 Mini-Mental State Examination (MMSE)

3.4.1 Introduction

The Mini-Mental State Examination (MMSE) (79) was initially developed (i) to assess cognitive aspects of mental function in patients referred for psychiatric evaluation and (ii) to aid in the differential diagnosis of organic (e.g., dementia) vs. functional (e.g., depression) brain disorder. In addition to being the most widely used cognitive screening test, it is also used in longitudinal studies to document the natural history of cognitive impairment and to monitor improvement during therapy, although it was not constructed for such a purpose (55;80). Tombaugh and McIntyre (81) reviewed 26 years of literature on the MMSE and readers are referred to their seminal work for more detail.

3.4.2 Description

Constructs Measured: The MMSE includes 11 questions on orientation, memory (recent and immediate), concentration, language and praxis. All items are evaluated by tests. Variations in the wording, content of some questions, administration and scoring of the MMSE have occurred since its initial development.

Administration: A trained interviewer administers the scale in 5-10 minutes.

Application: Suitable for individuals suspected of having organic or functional brain disorder.

Scoring: The scale ranges from 0 to 30. A higher score indicates less impairment. A cutoff of 23 or less has been adopted for cognitive impairment. A second cutoff of 17 or less is sometimes used for severe cognitive impairment.
3.4.3 Evaluation

3.4.3.1 Reliability

Test-retest: Test-retest reliability has been shown to be high for the MMSE when cognitively impaired subjects were tested (0.80 to 0.95) (81). Similar results were found with AD patients enrolled in the placebo arm of clinical trials (58;59). Lower test-retest reliability was observed when cognitively normal subjects were tested, but this is a likely reflection of the reduction in the correlation coefficient which accompanies a restriction in score range (60;81).

Interrater: Interrater reliability was high for the MMSE (0.50 to 0.75). However, the lack of scoring criteria for some items (e.g., drawing of the pentagons) may have an effect on these correlations (81).

3.4.3.2 Validity

Content: The MMSE is a short screening test that does not assess higher cognitive functions, verbal fluency or all the components of memory. It is, therefore, limited in terms of its ability to assess all the cognitive aspects known to be affected in AD.

Criterion/Concurrent: The MMSE has been compared to a number of cognitive scales. In general, the correlation coefficients were high (0.70 to 0.90) (81). The MMSE has been shown to be highly correlated with the DRS (0.87), SPSMQ (r=0.83), BIMC (r=-0.83 or higher), CAMCOG (r=0.94) (82-84). Two studies examined the correlation between the MMSE and the ADAS-cog and found it to be high (r=-0.81 and -0.90) (55;65).

3.4.3.3 Responsiveness to Change

Longitudinal studies: Decline in MMSE scores over time has been seen in many longitudinal studies of AD patients (81). The rate of decline was generally between 2 and 5 points per year, but displayed great variability both between subjects and across studies (85). This variability was partly explained by the inherently uneven progression of AD and the heterogeneity of patient characteristics.

The contents and psychometric properties of the MMSE are also partly responsible for this variability. Clark et al. (85) provided an illustration of this point. In
their analysis of the Consortium to Establish a Registry for Alzheimer’s Disease (CERAD) data, these authors showed that, under test-retest conditions, the reliability coefficient estimated from the MMSE results of demented patients was high ($r_p=0.87$). However, the corresponding individual change scores ($\text{score}_{T1} - \text{score}_{T2}$) showed a wide distribution that ranged from -8 points to +7 points (mean change score=0.5; SD=2.8). These results indicate the limited utility of the MMSE as a measure of change in individual patients over the short or medium term. Similar conclusions were arrived at by Bowie et al. (86) who showed that the MMSE’s interrater reliability was inadequate for detecting the small changes in cognition observed in current AD drug trials.

The MMSE also displays differential sensitivity to decline - more so than for other longer cognitive scales. The MMSE is sensitive to the profound decline in memory that is caused by mild to moderate dementia; however, as the severity of dementia increases, the MMSE loses its responsiveness to change. In other words, as the MMSE score decreases, the scale becomes less sensitive to the decline in cognitive functioning in AD patients (87) – a feature that has been called the floor effect.

**Severity**: By comparing the distribution of MMSE scores with cutoffs on 3 different cognitive scales (BCRS, SKT and ADAS-cog), Ihl et al. (55) evaluated the MMSE for score structure as related to severity of dementia, floor and ceiling effects. They found that the MMSE displayed a wider than expected range of scores in the moderate stages of disease identified by the cutoffs, which suggested the MMSE had low sensitivity for documenting the course of AD. A ceiling effect was also seen.

### 3.4.3.4 Age and Education As Confounders

Many studies have shown that MMSE scores decrease with increasing age. This age effect was shown in different settings. The age effect, which begins around ages 55 to 60 and accelerates over the ages of 75 or 80, is not a cohort effect as it persists after stratification by education level. As a result, the 23/24 cut-off score for screening on the MMSE may underestimate cognitive impairment in younger subjects, and overestimate it in older ones (88).

Many studies have also shown that MMSE scores are inversely related to years of education (81). The prevalent view is that education introduces a psychometric bias
leading to the misclassification of subjects from different educational backgrounds. However, other hypotheses have been proposed to explain this association. One view is that education is a marker for some other risk factor (e.g., exposure to malnutrition or to noxious substances during life) (89). Another view states that highly educated (or intelligent) people have a higher capacity to adapt to declining brain function; this higher capacity has been called ‘cognitive reserve’ (89). The complete interpretation of the education effect remains unresolved.

Interestingly, education and age have not been found to be associated with the rate of cognitive decline as measured by the MMSE (90). Nevertheless, advanced age and low education of study subjects would be expected to create favourable conditions for the floor effect that plagues the MMSE.

3.4.4 Critical Appraisal

Although the MMSE shows good reliability and validity for its original purpose of screening for dementia, short screening scales were not designed to measure more subtle aspects of cognition. In particular, short scales such as the MMSE may indicate little or no change over time in subjects who would otherwise be shown to have declined substantially if another scale had been used to measure change in status (49) (85). The MMSE is not an ideal outcome measure for AD drug trials, especially if the expected benefits are not large (86).

3.5 MMSE as a Staging Instrument

Kraemer et al. (91) write that the MMSE can be used as a staging instrument for dementia. They propose five stages that relate to MMSE scores: 1 (24-30), 2 (15-23), 3 (8-14), 4 (4-7), 5 (0-3). The authors found good correspondence between the MMSE and GDS after both instruments were administered to 206 patients with probable AD. Correspondence was measured by matching each patient’s MMSE stage (score) and GDS stage. Lower stages on the MMSE (stages 1, 2 and 3) had the best degree of correspondence with GDS staging. Greater divergence between instruments existed when more severely demented patients were rated.

Kraemer et al. (91) discuss the advantages of MMSE staging, namely that the process is easier to employ and less time consuming than the GDS. Also, the MMSE is
familiar to the clinical community. Using the MMSE to stage dementia is preferable when research emphasis is placed on the early phases of dementia and/or cognitive decline. The GDS is more suitable when the focus is on physical, behavioral, neurological or socio-economic changes associated with dementia severity.

The MMSE’s ‘staging’ version (i.e., severity levels defined by cutpoints on the MMSE) was used as selection criteria in many trials. However, the MMSE’s ‘staging’ version was not used as an outcome in any of the AD drug trials. The lack of responsiveness to change of the MMSE suggests that its ‘staging’ version would not be an appropriate outcome measure for AD drug trials.

3.6 Blessed Dementia Scale (BDS)

3.6.1 Introduction

The Blessed Dementia Scale (BDS) was developed in 1968 to examine the association between plaque formation on the brain and psychiatric disorders, especially dementia (92). The scale was primarily intended for psychiatric and geriatric inpatients. Since its development, the BDS has evolved into a tool for evaluating the severity of cognitive dysfunction in AD patients (38). It has also been used to assess the eligibility of AD patients for drug trials.

3.6.2 Description

Constructs Measured: The BDS measures: (i) changes in the performance of everyday activities (e.g., household tasks, recall of recent events, coping with small amounts of money, etc.); (ii) changes in habits related to eating, dressing and continence; and (iii) changes in personality, interests and drive (92).

Administration: An interviewer administers the scale to a relative or friend “in close continual contact with [the] patient.” (92) There are no rules governing the timing or frequency of administrations.

Application: The BDS is suitable for use with psychiatric or demented patients. It is intended to evaluate patient performance in the six-month period immediately prior to test administration.
**Scoring:** The BDS has three sections. The ‘everyday activities’ section has eight questions that are each scored separately according to the following scale: 1=total incompetence; 0.5=partial, variable or intermittent incapacity; 0=fully preserved capacity. The ‘habits’ section is composed of eating, dressing and continence questions that are scored separately. Each question offers four descriptions and the rater selects the one that best describes the patient. Scores range from 0 (clean/proper dressing, unaided dressing, complete sphincter control) to 3 (must be fed, unable to dress, doubly incontinent). The ‘personality, interests and drive’ section consists of a series of statements that are each awarded a score of 1 if they apply to the patient in question. The points from all three sections are simply totaled at the test’s conclusion to obtain a score between 0 (fully preserved capacity) and 28 (extreme incapacity).

### 3.6.3 Evaluation

#### 3.6.3.1 Reliability

**Test-retest:** Test-retest reliability has not been discussed in the literature.

** Interrater:** Cole (93) used 47 inpatients diagnosed with AD or multi-infarct dementia and two trained clinical psychologists to study interrater reliability. Each psychologist independently interviewed one caretaker per patient. The caretaker was required to be a relative who had been caring for the patient before hospitalization. The Pearson correlation coefficient for total scale scores between interviewers was 0.59; the ICC was 0.30.

#### 3.6.3.2 Validity

**Content:** Content validity has not been discussed in the literature.

**Construct/Concurrent:** Morris et al. (94) used 430 AD (NINCDS/ADRDA) patients to study validity. Experienced clinicians administered the BDS and the Short Blessed Test (SBT, which examines orientation, memory and concentration) to caregivers and patients. Experienced psychometricians administered a neuropsychological test battery (including the MMSE) to patients independent of the clinical assessments. Both administrations were conducted at baseline and then annually for four years. Annual change scores
(year\textsubscript{2} - year\textsubscript{1}, etc.) on the BDS were found to be correlated with the SBT (r\textsubscript{p}=0.48) and the MMSE (r\textsubscript{p}=-0.56).

Davis et al. (95) compared the performance of four cognitive instruments, including the BDS, with the CDR. This group hypothesized that the BDS would correlate positively with the CDR because more advanced stages of dementia were thought to produce higher scores on the Blessed test. Twenty-five controls, 25 subjects with questionable AD, 24 with mild AD, 24 with moderate AD and 18 with severe AD were entered into a longitudinal study. Using Kendall’s tau B, the correlation between the BDS and CDR was 0.80. The coefficient between the BDS-cognitive (which omits the section on personality, interests and drive) and the CDR was 0.84.

Criterion/Discriminant: Blessed et al. (92) looked at post-mortem findings for 60 subjects taken from a mix of psychiatric, demented and general inpatients. Pre-mortem BDS scores demonstrated a good correlation (r=0.77; p<0.001) and “broadly linear” relationship with the mean plaque count found upon autopsy. These scores were negatively correlated (r=-0.40; p<0.01) with the actual survival of 40 definitely demented patients (including 26 who were autopsied). Blessed et al. employed unspecified, subjective, clinical judgments to arrive at what they called a “definite” diagnosis.

3.6.3.3 Responsiveness to Change

No substantive work has been published in this area, but Heun et al. (38) did look at the BDS as part of a receiver operating characteristics (ROC) curve analysis. They found a score of greater than 1 would distinguish demented from nondemented patients with 95% sensitivity. The scores at 95% and 99% specificity were greater than 3 and greater than 4.5 respectively. The optimal threshold for differentiation as defined by the ROC curve itself was a score of greater than 1.

Drachman et al. (96) included information on the sensitivity and specificity of the BDS in a study of the Cognitive Assessment Screening Test (CAST). Two groups were employed in this study: (i) 19 mildly to moderately demented subjects (mean age=72.8 years) and 24 age-matched normal controls (mean age=74.9 years) and (ii) 26 nondemented, elderly patients. Neuropsychological test batteries were administered to identify persons with cognitive impairment. The BDS was able to discriminate between
demented and normal subjects with 100% sensitivity and 96% specificity in one experiment where only controls received the neuropsychological batteries. In the second experiment, all subjects received the batteries regardless of health state and the BDS discriminated with 50% sensitivity and 94% specificity.

3.6.4 Critical Appraisal

Reliability testing was limited to one study (93), which found interrater reliability to be low. Results of validity testing were inconclusive. Blessed et al. (92) did show a good correlation between plaque count during autopsy and BDS scores, but Drachman et al. (96) and Heun et al. (38) found questionable results for the sensitivity of the test when the issue was one of discriminating between demented and nondemented patients. Due to these either poor or mixed results, it is not recommended to use the BDS as an outcome measure in AD drug trials. Nor should it be used as an inclusion criterion. Other alternatives such as the ADAS-cog or CDR would appear to be more appropriate choices.

3.7 Individual Neuropsychological Tests

In addition to the previously described scales, there was a mix of individual neuropsychological tests used in the reviewed trials. These tests have been grouped into six categories for purposes of discussion: psychomotor, memory, language, visuo-perceptual function, attention, and abstraction (see Table 2 in Chapter 1).

3.8 Psychomotor

Tests of psychomotor speed are considered useful measures to include in an assessment of treatment effects on cognitive function (97). Performance on many psychological tests, particularly those that are timed, depends on psychomotor speed. Also, normal aging and age-related diseases such as dementia are associated with motor slowing, so it is important to have separate measures of psychomotor speed (97). Two different psychomotor tests were used in the trials, the Finger Oscillation Test (98) and the Choice Reaction Time (99). These tests are ‘control’ tests by which to interpret cognitive results. They will be covered very briefly and a more detailed review can be found in Flicker (97).
3.8.1 **Finger Oscillation Test**

The Finger Oscillation Test requires that the subject depress the key to a tally counter with the index finger as many times as possible within a specified period of time.

3.8.2 **Choice Reaction Time**

This test requires the release of a key or button in response to a signal. If the task requires movement from a resting position to a target position, the overall real time can be broken down into two components, ‘release time’ and ‘travel time’. Release time is partly dependent on cognitive processing speed, whereas travel time is a relatively pure measure of motor speed. Similarly, simple reaction time (one signal – one response) and choice reaction time (the subject must choose the appropriate response for each signal) have different sensitivities to aging and dementia.

3.8.3 **Evaluation**

According to Flicker (97), both of these crude tests of motor speed cover a wide range of difficulty, have adequate reliability (which can be improved with practice trials), adequate construct validity and sensitivity to aging and dementia. Each test takes 5 minutes or less to apply. It is not advisable to use more complex tasks beyond these tests to assess motor speed. Greater complexity usually makes larger demands on cognitive function and any test results in the realm of pure motor speed would be difficult to interpret.

3.9 **Memory**

3.9.1 **Buschke Selective Reminding Test (BSRT)**

3.9.1.1 **Description**

The Buschke Selective Reminding Test (BSRT) Boston Naming Test (BNT) (100) measures verbal learning and memory.  

*Constructs Measured:* The domains examined are immediate recall, cued recall, multiple-choice recognition recall and delayed recall. 

*Administration:* The test is administered by an examiner who reads or shows subjects a list of 12 words. The subject is then asked to recall as many words as possible. Each
subsequent trial involves the selective presentation of only those items not recalled on the immediately preceding trials. The examiner can distinguish between short- and long-term components of memory by measuring recall of items not presented during a given trial. Maximum performance is sought for immediate recall by allowing up to 12 possible consecutive trials. Both adult and child versions are available. The adult version requires 30 minutes.

Application: The BSRT has become the most widely used test for assessing memory following head injury. The BSRT can also differentiate between normal and mildly demented elderly persons.

Scoring: An explicit and complex scoring scheme takes into account the number of mistakes and repetitions by the subject as well as reminders from the interviewer. Different scores are calculated for the various components of memory: short-term recall (STR), long-term storage (LTS), long-term retrieval (LTR) and consistent long-term retrieval (CLTR). A higher score indicates better performance. Normative data exist for different age groups (up to 91 years of age), as well as for men and women. Education appears to have little effect on the results of the BSRT.

3.9.1.2 Evaluation

According to Flicker (97), the reliability and construct validity of the BSRT is adequate. The BSRT is also responsive to change, being sensitive to aging and dementia. However, the BSRT, like other word list tests, is prone to floor effects (i.e. the test may be too demanding for the cognitive capacities of the subject). While on theoretical grounds, the BSRT is devised to measure recall from short-term and long-term memory, the construct validity of these items has yet to be established.

3.9.2 Wechsler Memory Scale–Revised (WMS-R)

3.9.2.1 Description

The Wechsler Memory Scale–Revised (WMS-R) Boston Naming Test (BNT) (100) represents a major revision to the Wechsler Memory Scale (WMS) battery, the
most frequently used clinical measure of memory in the US. The purpose of both batteries is to provide measures of various aspects of memory function.


*Administration:* An interviewer administers the 13 subtests of the WMS-R. The first subtest is an information orientation question that is used for screening purposes and is not scored. The other subtests are scored. The delayed tasks are undertaken after a 30-minute delay. The time required to complete the whole test (not including delay) is about 45 minutes.

*Application:* Clinical investigations using the WMS-R showed that the test is sensitive to memory disturbances and may characterize the learning and memory disorders in a number of different patients groups including those with AD.

*Scoring:* The test yields five age-corrected summary scores: one for general immediate memory, one for general delayed memory, one for attention/concentration and two separate indices for immediate visual and immediate verbal memory. The test manual provides explicit guidelines for scoring and weighting. The manual also provides norms for individuals from 16 to 74 years of age and information about significant differences between any two scores.

3.9.2.2 *Evaluation*

Test-retest reliability data from the developers was judged to be adequate. Concurrent validity with a verbal learning test was also adequate. Factor analysis of the WSM-R showed evidence of two major factors: ‘general memory’ (which does not appear to correlate with IQ) and ‘attention/concentration’ (which correlates highly with intelligence).

The WMS-R is considered to be an improvement over the original version. However, it has several limitations: administration time is longer relative to the WMS; no parallel forms exist; normative data for the very old (75+) is lacking; and there is a recognized floor effect. Other problems: verbal memory performance continues to contribute more heavily to the score than other components of memory; the absence of
recognition tasks limits the ability of the scale to differentiate between patient groups. Level of education is correlated with each of the five scores and should be taken into account when interpreting individual scores.

3.10 Language

According to Flicker (97), language tests have been underused in research for antidementia treatment. Language function is often perceived as resistant to conditions that compromise brain function, at least to the extent that vocabulary test scores are sometimes treated as an index of premorbid intelligence. Although some AD patients can retain intact syntax into advanced stages of the disease, word finding difficulty is often present early on.

3.10.1 Controlled Oral Word Association (COWA)

3.10.1.1 Description

The Controlled Oral Word Association (COWA) Boston Naming Test (BNT) (100) is a language test which requires the spontaneous production of words beginning with a given letter or arising from a given class (e.g., animals) within a limited amount of time. Other names for this test include Word Fluency and FAS (F, A, S are the most commonly used letters for this popular test). COWA is the most frequently administered language test in clinical trials of treatment effects in the elderly.

*Construct Measured:* Verbal association fluency is the construct of interest. The test has a time limit, so performance is dependent on speech production speed, which is in turn dependent on cognitive processing speed.

*Administration:* The subject is asked to produce as many words as possible in a limited period. Inadmissible words (e.g., variations, repetitions, proper nouns, wrong words) are explained. The test takes about 5 minutes.

*Application:* Adults or children.

*Scoring:* The score is the sum of all admissible words. Inadmissible words are not counted as correct.
3.10.1.2 Evaluation

Interater reliability is very high. Test-retest reliability in adults after 19-42 days is also high. The test is weakly correlated with age. In adults, factor analysis showed factor loading mainly to a ‘verbal knowledge’ factor. Scores are sensitive to specific brain pathologies.

According to Spreen and Straus (100), the designation ‘Word Fluency’ is misleading “since verbal productivity in conversation or in continuous sentences is not measured. Instead the test measures production of individual words under restricting search conditions” (p. 221). Several authors have failed to find reduced word fluency in AD patients, possibly because of ceiling effects (97). Nevertheless, AD patients are reported to make more errors by intrusion, repeat or variation (100).

3.10.2 Boston Naming Test (BNT)

3.10.2.1 Description

The Boston Naming Test (BNT) (100) assesses the ability to name pictured objects. Constructs Measured: Accuracy and speed of naming come into play in this test. Sixty drawings ranging in an ordered fashion from simple high-frequency vocabulary to rare words are presented one at a time on cards. When a response is not elicited within 20 seconds, the subject is offered a verbal cue and, if necessary, a phonemic cue.

Administration: Adults begin the test with item 30. If any of the next eight items are failed, the interviewer is instructed to proceed backward from item 29 until eight consecutive words are passed. The interviewer then proceeds forward until a failure point is reached. Credit is given if the word is named in no more than 20 seconds. Numbers of correct and incorrect answers are recorded. The administration time is variable.

Application: The test is widely used for identifying aphasia in adults and children.

Scoring: A total naming score is derived from summing the number of correct answers between the baseline item and the ceiling item while adding the number of items below baseline. A higher score means better performance.
3.10.2.2 Evaluation
Test-retest reliability data are not available. Correlation between forms was high in healthy controls and in Alzheimer’s patients (two equivalent forms were obtained by dividing the original scale). Information on construct validity appears to be available in children only. BNT scores are related to an index of aphasia severity. The number of errors on the test can discriminate between cognitively intact subjects, AD subjects with mild impairment, and AD subjects with moderate impairment. Visual-perceptual integrity should be verified if errors occur on the test. The norms accompanying the test are based on small groups of adults; normative data for the elderly are scantier. An age or education effect becomes apparent when subjects reach 70+ years.

3.11 Visuo-Perceptual Function

Tests of visuo-perceptual function have been shown to be sensitive to damage of the parietal cortex. Since the parietal cortex is a brain area showing signs of damage early in the course of AD, this type of tests would seem very appropriate for detecting drug-induced improvements. Unfortunately, these tests have not been routinely employed in AD drug trials.

3.11.1 Hooper Visual Organization Test (VOT)

3.11.1.1 Description
Assesses the ability to rearrange pictures that have been disarranged. This test is similar to other fragmented figure tests.

*Constructs Measured:* The test places demands on perceptual differentiation and conceptual reorganization.

*Administration:* The subject is shown 30 drawings of common objects on 4” by 4” cards placed in a binder. Each object is cut into two or more pieces that are illogically arranged on the card. The subject is asked to name the object. Testing takes from 10 to 15 minutes. The correct naming of each object is required. Quality of the responses may be important.

*Application:* The test was originally developed to screen for brain damage. More recent studies have attempted to better delimit its use. The test is used in adults of all ages.
Scoring: The score is the total of correct responses – partial credit is given for some partly correct answers.

3.11.1.2 Evaluation

Test-retest reliability and split-half reliability are high. The validity of this test has been debated. An age decline on this test was observed and interpreted as right hemisphere functional decline with age. Spreen and Strauss (100) consider this test to be useful to explore individual difficulties in perceptual organization, but not to confirm the presence of brain damage. Results for patients with language disorders must be interpreted with caution, since the test requires naming. That is, language function must be somehow controlled for in order to use the test to assess visuo-perceptual ability. The test does not appear to have been validated in AD patients.

3.11.2 Rey-Osterrieth Complex Figure Test (RF)

3.11.2.1 Description

The purpose of this test is to measure visuo-spatial constructional ability and visual memory.

Constructs Measured: planning and organizational skills, problem solving, as well as perceptual, motor and memory functions.

Administration: The subject copies the Rey-Osterrieth figure, which is a complex two-dimensional geometric figure. Then, with a delay varying from a few minutes to 45 minutes, the subject is asked without warning to reproduce it from memory. Some investigators also measure immediate recall. The utility of immediate recall is unclear and delayed recall is favoured. The test requires about 10 minutes (excluding the delay).

Application: The test is used as a diagnostic aid with respect to brain lesions.

Scoring: Strict and comprehensive scoring guidelines are provided. Two measures of performance are usually derived: a score measuring visuo-constructional ability and a score reflecting the amount of information retained over time. The accuracy of the original copy should be carefully verified because slight mistakes can be indicative of certain types of brain damage. Interpretation of the RF should consider the qualitative aspects of performance, such as copying strategy. Copying strategy is related to recall performance. A disorganized piecemeal approach to the copy of the figure may result in
an accurate production – but recall will tend to be poor. A higher score is indicative of better performance. The highest possible score is 36.

3.11.2.2 Evaluation

Adherence to strict scoring criteria yields high inter-rater reliability. Practice effects occur in normal adults (improvement of about 10% can be expected). The test is useful for distinguishing frontal lobe, right and left hemisphere lesions. Information from the test may also be useful in differentiating between different memory disorders. Normative data exists for healthy adults with above average education aged 15 to 85. Both age and intellectual level contribute to performance on the RF. The test does not appear to have been used much in AD patients. Given the complexity of the figure, a floor effect on this test is likely in AD patients (the test would be too demanding).

3.12 Attention

3.12.1 Trail Making Test

3.12.1.1 Description

The purpose of the Trail Making Test (100) is to assess speed of visual search, attention, mental flexibility and motor function. There is a part A and a part B to the test.

* Constructs Measured: Factor analysis indicates loading on two factors designated as rapid visual search and visuo-spatial sequencing.

* Administration: The subject is presented with a paper and pencil and is asked to draw a line between circles on the page; inscribed in the circles are numbers (part A) or letters (part B). The objective is to link one point to another point in an ordered fashion (e.g., 1,2, ... to 25 or A, B, ... to Z) until the last point is reached. The time for administration is 5 to 10 minutes.

* Application: The test is used in subjects with brain damage (including traumatic head injury, Korsakof and other types of damage) and in neuropsychiatric populations. The test is used for initial assessment and again in follow-up during recovery.

* Scoring: If the subject makes an error, the interviewer calls this to his attention immediately. The subject is asked to proceed from the point the mistake occurred. The
score is the time in seconds taken to complete the task. Errors count only in the increased time of performance. Normative data are available.

3.12.1.2 Evaluation

Reliability was reported to be higher for part A than for part B. It has been noted that the use of time scores instead of both error counts and time scores reduces reliability because error correction may take a variable amount of time, depending on both the interviewer and the subject’s ability to understand. The test is highly sensitive to brain damage, including head injury and alcoholism. Part B is more sensitive than part A to brain damage; part B requires more information-processing ability than part A. Performance on this test does not seem to be influenced by the presence of aphasia. One can gain useful information by observing performance in part B, including the ability to shift course during an ongoing activity and the ability to deal with more than one stimulus at a time. Scores are strongly affected by the education level of the subject. The effect of intelligence (IQ) is less pronounced but still noticeable in part B.

3.12.2 Digit Symbol –Subtest from the WAIS-R

3.12.2.1 Description

The Digit Symbol test (101) is a symbol substitution task. It consists of four rows containing in all 100 small blank squares, each paired with a randomly assigned number from 1 to 9. Above these squares a key is printed that pairs each number with a different nonsense symbol. The subject is asked to fill each blank square as quickly as possible with the symbol corresponding to the number. Timing is stopped after 90 seconds and the score is the number of correct answers. At a disadvantage on this test are elderly whose vision or visuomotor coordination is impaired or who have difficulty understanding the instructions, as well as elderly with pronounced motor slowing or low educational levels. The Digit symbol subtest is considered more sensitive to brain damage than other WAIS-R subtests. Despite its sensitivity, the test is of little use in predicting the laterality of a lesion. Performance on the test is affected by many different components, which probably explains the lack of specificity of the test.
3.12.3 Symbol Digit Modalities Test

The Symbol Digit Modalities Test (SDMT) reverses the presentation of material of the WAIS-R Digit Symbol subtest described above so that the symbols are printed above the blank squares and the numbers are used as key to the symbols. This allows the patient to respond verbally as well as in writing. The test features 110 items (rather than 90) and is scored over 90 seconds. A written and oral administration is encouraged, whenever possible. Education is positively correlated with the score on both the written and oral administration and therefore needs to be taken into account. Significant impairment on both administrations reflects visual perceptual, visual scanning, or oculomotor defects, or general mental or motor slowing. As with the Digit Symbol subtest, this test is sensitive to normal effects of aging as well as brain dysfunction. Normative data exists up to age 74.

3.12.4 Digit Span subtest of the WAIS-R

3.12.4.1 Description

The Digit Span subtest comprises two different tests, the Digit Forward and the Digit Backward. Both tests are similar in that they consist of seven pairs of random number sequences that the tester reads aloud to the patient at the rate of one per second. The Digit Span Forward: The subject’s task is to repeat each sequence exactly as it is given to him. When a sequence is repeated exactly, the tester moves on to a longer sequence until patient failure or correct repeat of a 9-digit sequence. Digit Forward measures efficiency of attention. It is a relatively stable capacity that is resistant to the effects of dementing diseases.

The Digit Span Backward: The subject’s task is to repeat each sequence in an exactly reversed order. This test is more of a memory test than the Digit Forward; it involves ‘double-tracking’ in the simultaneous operations of memorization and reversal of information. This test is very vulnerable to the diffuse brain damage that occurs in dementing illnesses.
3.12.4.2 Evaluation

These tests involve different mental activities and are affected differently by brain damage. The Digit Forward is stable with advancing age, whereas the Digit Backward span typically shrinks. The averaging of the Forward and Backward scores recommended by the WAIS-R manual is discouraged by Lezak (101) who observes that a high score on the Digit Forward can conceal a low score on the Digit Backward. Lezak recommends keeping the Forward and Backward scores separate for maximum validity.

*Mental Control, subtest of the Wechsler Memory Scale-Revised:* The Mental Control subtest checks automatisms (the alphabet) and simple conceptual tracking (counting by fours from 1 to 53). It is a crude test of attention.

3.13 Abstraction

3.13.1 Category Retrieval Test

We were not able to locate the Category Retrieval Test purportedly used by Tariot and al. in their selegiline trial (99). The reference provided by Tariot et al. (99) cited Lezak’s textbook (101) and mentioned the COWA test (a test of language). In her textbook, Lezak (101) did not disclose any information on any Category Retrieval Test; she however discussed the Category Test.

3.14 Strengths and Weaknesses for Individual Neuropsychological Tests

3.14.1 Strengths

Adequate reliability and validity were shown for the tests we reviewed. When used in a battery, these tests have proven useful in providing a cognitive profile that describes relative strengths and weaknesses as well identifies problem areas. Individual neuropsychological tests also help with the early detection of dementia.

3.14.2 Weaknesses

There is a lack of adequate normative neuropsychological data on the elderly – this is considered an important and challenging area for development (50;69). Educational norms that are equally lacking are also considered necessary, as performance on many tests is influenced by premorbid (predementia) intellectual achievement.
Testing may be long and cumbersome to the patients, depending on the number of tests administered.

### 3.14.3 Responsiveness to Change

Responsiveness to change was rarely discussed. This is probably not surprising since neuropsychological testing focuses on detection and identification of specific cognitive deficits, rather than on change over time.

An outstanding and unique study in the area of responsiveness to change is the study by Loscascio et al. (102), wherein the differential responsiveness to change of individual neuropsychologic tests was demonstrated. These investigators followed 60 normal and 123 AD patients over a 4-year period. They monitored changes in cognition using 11 neuropsychological tests. The following observations were recorded: (i) tests of recent information recall were best for discriminating intact from demented subjects or detecting mild from more severe dementia; (ii) all tests used had a relatively poor ability to separate stages of AD; and (iii) tests that showed linear decline were best at following the course of disease. Loscascio et al. (102) concluded that memory tests would not be useful in longitudinal studies of cognition in AD because of floor effects early in the course of disease. Also, cognitive tests were not efficient ways of denoting different stages of disease. This supports the current usage of global tests for staging. Tests of word fluency and visual recall should be included as outcome measures in any therapeutic trial claiming to alter or slow the course of cognitive deterioration. Information on measurement properties with regard to change is crucial, but still limited.

### 3.15 Cognitive Scales: Strengths and Limitations

Two approaches to cognitive evaluation were seen in the AD trials. Investigators using the first approach relied solely on cognitive scales; investigators using the second approach relied on individual neuropsychometric tests in addition to one or more cognitive scales.

The MMSE and BDS are viewed by some (103) as being (i) valid scales for measuring the severity of cognitive impairment and (ii) valuable aids for screening potential trial subjects. The use of either instrument is recommended by the NINCDS/ADRDA criteria (33) to document a patient’s cognitive profile during AD
diagnosis. However, the ceiling effect known to plague the MMSE (103), and the mixed validity results/dearth of reliability results for the BDS, should preclude either scale from being adopted as a main outcome measure in AD drug trials. The MMSE can be used to document the progression of drug trials and obtain a universally meaningful index of dementia severity (104), while the BDS should be avoided until more is known about its reliability and validity.

The ADAS-cog and SKT both demonstrated adequate reliability and validity for use in AD drug trials. However, more work must be done in the area of responsiveness to change. A failure to develop this area might lead to uncertainty regarding the scores that would be needed to demonstrate a treatment effect. Beyond this concern, both scales have shown themselves to be good outcome measurement instruments.

Neuropsychological tests should be secondary/exploratory outcome measures in antidementia drug trials because of their limited scope (each test measures a single part of cognition rather than cognition as a whole). Also, testing becomes long and strenuous for the patient when these batteries are utilized as outcome measures. Due to the existence of many cognitive subdomains and neuropsychological tests, care must be taken to select the most optimal instrument for measuring change in the subdomain of interest. The use of different tests across trials can reduce the comparability of RCTs that are evaluating the same drug.

Collaboration between epidemiologists, neuropsychologists and geriatricians, which has been lacking, is absolutely necessary to develop and refine cognitive scales that have good measurement properties, relevance to the disease of interest and practical use in AD drug trial.
4 FUNCTIONAL AND QUALITY OF LIFE OUTCOME MEASUREMENT SCALES

4.1 Introduction: Functional Disability Scales

In the framework of the WHO Classification of Impairment, Disability, and Handicap (CIDIH), functional disability refers to any restriction or inability to perform an activity in the manner, or within the range, considered normal for a human being (105;106). Since deterioration in functional performance is an integral part of the course of AD, the importance of measuring level of disability in clinical trials is increasingly recognized (107-109). Such an outcome is considered meaningful primarily because it is presumed that if an antideementia drug improves cognitive function, this effect should extrapolate to a change in functional performance (107;108). Secondly, it is important because disability has a major impact on quality of life of the AD patients and their caregivers (110). Thirdly, functional deterioration is a predictor of institutionalization and can thus assist health care providers in making placement and management decision with respect to the nature of the resources that need to be allocated (110;111). Finally, functional performance information is helpful in determining the level of assistance that the caregiver must provide for the elderly living in the community or in institutions (110).

Although there is a consensus on the importance of assessing function, there are a number of methodological and theoretical issues that limit its practical usefulness in the context of pharmacological trials for AD. Indeed, in the majority of trials reviewed there was no statistically significant impact on function. According to Beck and Frank (110), the problems that affect AD research can be attributed to: (i) variability in conceptual and operational definitions of functional outcomes; (ii) lack of sensitivity inherent to the use of broad task categories and one-dimensional rating scales; (iii) lack of contextual perspective, and (iv) focus on physical disability rather than on a cognitive impairment framework.

Numerous functional instruments are available for assessing AD patients. In 1991, Kluger and Ferris (112) estimated that 40 such instruments were available and, since then, several more have been developed (113) (108). The lack of a universally
accepted measure of functional disability is reflected in the multiplicity of scales that were used in this review.

Twelve of the 25 trials reviewed in Report I (A Comparative Analysis of Clinical Trials) assessed functional problems and of these four relied on more than one measurement scale. There were 11 different outcome scales among these 12 trials combined. With the exception of 2 scales, functional scales were generally identified as secondary outcomes. All referenced scales are briefly described, although some of them were cited only once or twice. This approach was beneficial in broadening our understanding of the benefits and limitations involved in assessing functional disability and quality of life in AD drug trials.

4.2 The Geriatric Evaluation by Relative’s Rating Instrument (GERRI)

4.2.1 Description

The Geriatric Evaluation by Relative’s Rating Instrument (GERRI) (114) is a scale in which significant others judge elderly outpatients who show symptoms of mental decline. It consists of 49 short sentence items designed to assess the frequency of typical behavioral disturbances and complaints in the areas of cognitive functioning (20 items), social functioning (18 items), and mood (11 items). The instrument was developed to provide information useful in geriatric pharmacology.

Scoring: Items are rated on a 5-point frequency scale that ranges from “almost all the time” to “almost never”. An average score (from 1.00 to 5.00) based on all applicable items is computed and the higher the score, the greater the impairment. The scale is designed to be completed every two weeks, with the items to be judged on the basis of behavior demonstrated in the previous 2-week period.

4.2.2 Evaluation

4.2.2.1 Reliability

Interrater: Inter-rater reliability was high for the GERRI total scale score when 45 subjects of varying levels of severity of Primary Degenerative Dementia were evaluated by two significant others (114). The intraclass correlation coefficient (ICC) was
estimated as 0.94. For the cognitive functioning, the social functioning and the mood sub-scales, the coefficients were 0.96, 0.94 and 0.66 respectively. For individual items, the ICC ranged from 0.02 to 0.98 (median of 0.72).

**Internal Consistency:** For the GERRI total scale score and its cognitive functioning sub-scale, the Cronbach’s alpha was 0.96 and 0.95 (114). These values are considered very high. For values of alpha greater than 0.90, DeVellis (115) recommends reconsideration of the choice of items as one must be concerned with redundancy. This has not been done for the GERRI Scale. For scales with heterogeneous items, the Cronbach’s alpha is known to underestimate true reliability. However this does not seem to be the case for the mood sub-scale (alpha of 0.66) that includes aspects such as “reports he/she feels sad”, “appears restless and fidgety”, “reports feeling worthless”, “appears to be easily annoyed or angered”, etc. This indicates that the mood sub-scale is not reliable.

### 4.2.2.2 Validity

**Content Validity:** The content validity of the GERRI, with respect to item coverage and relevance, has not been reported in the reviewed articles.

**Concurrent Validity:** Rozenbilds and his colleagues (116) found that the GERRI cognitive function and social function sub-scales showed moderate to good correlation with (i) the London Psychogeriatric Rating Scale (nurse as informant; assesses physical, behavioral, and mental functioning), (ii) the MMSE, and (iii) the Activities of Daily Living Test. Correlation coefficients ranged from 0.35 to 0.59. However, there was little correlation between the GERRI mood sub-scale and a Visual Analogue Scale for Depression as well as with all the previously listed scales (116).

**Construct Validity:** An analysis of variance was performed by Rozenbilds and his colleagues (116) to assess the discriminant validity of the GERRI Scale in terms of differentiating among three groups of subjects having low, moderate and high cognitive impairment, based on the GDS (34). Both the GERRI total score and the cluster scores differentiated the three severity groups (116). It is important to note that the three-dimensional structure of the scale has not been ascertained through factorial analysis. The clusters of items are however supported by high face validity and strength of inter-item correlations (114).
4.2.2.3 Responsiveness to Change

The sensitivity to severity of dementia has been demonstrated for three broad levels of severity (116). From the available information, it is not clear if small functional changes can be detected using the GERRI Scale. This is a potential limitation in RCTs. Clinically important changes have not been discussed. Because the total score is obtained by averaging the individual items (rather than adding them up), the use of two decimal digits is necessary to prevent loss of sensitivity.

A ceiling effect appears to be avoided with the GERRI Scale. The mean scores ranged from 1.4 to 3.8 (potential of 1.00 to 5.00), for subjects classified from 2 to 7 on the GDS (116). However, a floor effect is likely to be found with more severely impaired subjects.

4.2.3 Critical Appraisal

Further assessment of construct validity appears desirable, since there is very little theoretical basis to the scale. Test-retest reliability has not been studied. The validity and reliability of the mood sub-scale appear to be perceptibly weaker than for the other sub-scales that comprise the instrument.

The GERRI Scale emphasizes different aspects of functioning of the elderly that appear to have a lot of impact on the person’s quality of life and burden of care. Moreover, an essential characteristic of this scale is the fact that it relies on untrained informants. As such, it may be a very relevant additional outcome measure in RCTs.

4.3 Physical Self-Maintenance Scale (PSMS)

4.3.1 Description

The Physical Self-Maintenance Scale (PSMS) developed by Lawton & Brody in 1969 (117) is an observer rating scale that addresses activities of daily living in geriatric patients. It was adapted from a scale developed in psychiatry by Lowenthal (cited in (117)).

*Construct Measured:* The construct of physical self-maintenance is measured through competence in six behaviors: toileting, feeding, dressing, grooming, locomotion, and
bathing. From a theoretical perspective, the construct is well grounded and viewed as a specific component in the hierarchy of human behavioral competence (117) (118).

Administration: The PSMS scale can be completed by untrained personnel staff, based on information provided by one or several informants (subject, caregiver, friend, etc.). The scale focuses on concrete observable behaviors.

Scoring: Each of the behavioral areas is given a score of 1 or 0, leading to an overall score that ranges from 0 to 6. Five descriptive scale points have been developed for each behavior, based on a Guttman scaling technique. Accordingly, items are worded to increase in strength so that once respondents agree with one statement, they should agree with all statements that express less complex functional behavior and disagree with statements that express more complex functional behavior (7).

4.3.2 Evaluation

4.3.2.1 Reliability

Data obtained from 265 subjects from a variety of sources showed that the Guttman scaling criteria was adequately met. Two studies were conducted on interrater reproducibility. Firstly, two nurses were asked to independently rate 36 patients with a variety of self-care deficits ($r_p=0.87$). Secondly, research assistants independently rated 14 other impaired and nonimpaired patients ($r_p=0.91$). There are no data on the internal homogeneity and test-retest stability of the PSMS.

4.3.2.2 Validity

In order to demonstrate the construct validity of the PSMS, the authors determined its correlation ($r_p$) with four measures believed to tap the concept of functional competence. In total, 180 subjects were involved in this study but few received all five evaluations. The correlations were PSMS-IADL (0.61) ($n=77$), PSMS-Physical Classification (6-point rating of physical health) (0.62) ($n=130$), PSMS-Mental Status Questionnaire (0.38) ($n=152$), and PSMS-Behavior and Adjustment rating scales (0.38) ($n=98$).
4.3.2.3 Responsiveness to Change

No data are available on responsiveness to change. However, use of the Guttman scaling technique properties imply a loss of redundancy which can be problematic when sensitivity is desirable and the error rate can be high. In addition, every item is scored as a dichotomy, forcing a loss of discriminating ability (118).

4.3.3 Critical Appraisal

The PSMS is a brief objective assessment of activities of daily living for the geriatric patients. It is theoretically well grounded and has proven to be useful for the evaluation of the institutionalized elderly. For testing psychopharmacological treatment effects, the versions of the scale that use additive ratings could be preferable (118). Moreover, in the context of assessing community living elderly persons, it is likely to demonstrate a strong ceiling effect. Testing of its psychometric properties is incomplete.

4.4 Instrumental Activities of Daily Living (IADL)

4.4.1 Description

The Instrumental Activities of Daily Living (IADL) scale was developed Lawton and Brody in 1969 (117) to measure elderly persons’ competence in instrumental self-maintenance, that is, the complex activities required in everyday functioning. This scale was developed and tested concurrently with the Physical Self-Maintenance Scale (PSMS), which addresses more basic self-care abilities. Several self-rated versions of the IADL scale were later developed (105).

Construct Measured: The construct of instrumental competence is operationalized differently for women and men. For women, the set of behaviors assessed include telephoning, shopping, food preparation, housekeeping, laundering, use of transportation, use of medicine and ability to handle money. For men, the areas of food preparation, housekeeping, and laundering are excluded.

Administration: The rating of the IADL scale can be done by a variety of personnel, using one or several informants. The scale focuses on concrete behaviors and fosters objective judgments.
Scoring: Each of the behavioral areas is given a score of 1 or 0, leading to an overall score that ranges from 0 to 8 for women and from 0 to 5 for men. A higher score indicates better performance. Between three and five descriptive scale points have been developed for each behavior, based on the Guttman scaling technique.

4.4.2 Evaluation

4.4.2.1 Reliability

There are no data on the internal homogeneity and test-retest stability of the IADL. The authors report a study in which 12 subjects were assessed by one interviewer with another rater present but not participating. The reproducibility of the IADL scale was estimated with $r_p=0.85$.

4.4.2.2 Validity

In order to demonstrate construct validity of the IADL, Lawton and Brody (117) determined its correlation ($r_p$) with four measures believed to tap the concept of functional competence. In total, 180 subjects were involved in this study but few received all five evaluations. The correlations were IADL-PSMS (0.61) (n=77), IADL-Physical Classification (6-point rating of physical health) (0.40) (n=50), IADL-Mental Status Questionnaire (0.48) (n=74), and IADL-Behavior and Adjustment rating scales (0.36) (n=44).

4.4.2.3 Responsiveness to Change

Loss of sensitivity due to the Guttman scaling technique and the dichotomized scale is likely to occur as for the PSMS scale.

4.4.3 Critical Appraisal

The IADL is a very frequently used and often cited instrument for assessing the instrumental competence of elderly patients. The scale is well anchored from a theoretical point of view (118) and the behaviors that are included are likely to be affected in the first stages of dementia. For its usage in the 1990s, the distinction that is imposed on male and female activities is, however, questionable. Extensive testing of the
reliability of the IADL has not been done. For testing psychopharmacological treatment effects, the versions of the scale that use additive ratings are preferable (118).

4.5 Blessed Dementia Scale, Part 1 (BDS Part 1)

4.5.1 Introduction

The Blessed Dementia Scale was originally developed in 1968 by Blessed, Tomlinson and Roth (92) to demonstrate the relationship between severe mental decline in the elderly and the presence of changes in brain tissue. The scale contains two parts: (1) an evaluation of activities of daily living and changes in personality, interests and drives and (2) an assessment of orientation, memory, and concentration. It is important to note that several alternate versions of the BDS have been developed over time (119). In this study, two papers refer to the original version of the BDS (99;120) and one to a modified version (121). The following paragraphs present the BDS Part 1 and succinctly report on the modified Blessed Dementia Scale (mBDS).

4.5.2 Description of the BDS Part 1

Construct Measured: The BDS was designed to measure the degree of intellectual and personality deterioration shown by the elderly. To operationalize this dementia construct, Part 1 of the scale quantifies the patient's ability to deal with the practical tasks of everyday life. The scale comprises 22 items related to competence in personal, domestic and social activities, such as ability to perform household tasks, to cope with small sums of money or to find their way in familiar surroundings.

Administration: The BDS is rated after questioning a close relative or friend about the patient's performance in the tasks during the preceding six months.

Scoring: The BDS Part 1 dementia score is computed by adding the scores allocated for each item. Scoring varies according to three groupings of items. For performance of everyday activities (8 items), total incompetence is allocated a score of 1, partial, variable, or intermittent incapacity is given a score of 0.5 and fully preserved capacity is given a score of 0. For scoring of habits (3 items), descriptive response options are presented and respectively attributed a score of 3, 2 or 0. For changes in personality, interests and drives (10 items), ratings are either 1 or 0, based on specific observable
behaviors. Thus, the total score could lie between 0 (fully reserved capacity) to +28 (reflecting extreme incapacity). Based on their experience, Zerssen and his colleagues (122) have proposed that a total score of the BDS Part 1 of 15 to be indicative of moderate dementia in AD.

4.5.3 Description of the mBDS

The mBDS consists of items measuring both activities of daily living (ADL) and instrumental ADL functions (Erkinjuntti, cited in (123)). Three ADL functions (eating, dressing and continence) are scored from 0 to 3 (formerly the 'habits' items) and eight items assessing changes in everyday activities, such as performing household tasks, coping with money or finding one's way, are scored 0, 0.5 or 1 (formerly the 'everyday activities' items). The total score ranges thus from 0 (independent) to 17 (dependent). In a study involving 795 elderly subjects (123), the scale was shown to discriminate well between demented and non-demented subjects (estimated AUC=0.90 to 0.96). A cut-off point at 1.00 was suggested for screening for dementia.

4.5.4 Psychometric Properties of the BDS Part 1

4.5.4.1 Reliability

No reliability testing of the BDS Part 1 has been reported (119).

4.5.4.2 Validity

The authors of the scale (92) studied 264 patients classified into various diagnostic groups and found that patients with senile dementia registered poorer mean scores on the BDS scales than the other groups (Student's \( t \) tests). In 76 postmortem investigations, they found a negative association between senile plaque counts and scores on the BDS scales (\( r=0.77 \)). Mayeux, Stern and Spanton (124) found a moderate correlation (\( r=0.50 \)) between the BDS part 1 and the modified Mini-Mental State Examination (mMMSE). Together, these findings support the construct validity of the scales.

Stern, Hesdorffer and colleagues (125) have submitted the items of the BDS Part 1 to a Principal Component Analysis using data from 187 subjects diagnosed with probable AD. This produced 4 independent factors: Cognitive (7 items, score range 0-7),
Apathy/Withdrawal (3 items, score range 0-3), Personality change (6 items, score range 0-6) and Basic self-care (3 items, score range 0-3) that were considered to be more descriptive of specific behavior changes. These findings do not support the original use of the scale.

*Responsiveness to Change:* No data are available on this measurement property.

### 4.5.5 Critical Appraisal

The BDS Part 1 specifically addresses a population of demented elderly, which is relevant for its use within AD clinical drug trials. The scale has good potential for screening individuals for dementia. However there is a lack of evidence as to its sensitivity in detecting small changes that may have occurred following specific interventions. Study of the BDS's psychometric properties is incomplete. The use of altered versions also complicates comparisons across sites (119).

### 4.6 The Nüremberg Geriatric Observation Scale (NAB)

#### 4.6.1 Description

The Nüremberg Geriatric Observation scale (NAB) is comprised of 15 items assessing activities of daily living. It was developed by Oswald and Schaltenbrand for German-speaking patients (cited in (46)) as part of the Nüremberg Gerontopsychological Inventory (NAI) (126;127). The NAI is a battery of measurements to evaluate intervention-induced changes in old age. Very little English language literature is available on the NAB and its psychometric properties.

*Constructs Measured:* Need for care/ability with regards to activities of daily living. Domains examined include: appearance, constructive occupation, help in financial affairs, everyday activities, speech comprehension, speech expression, behavior outside the house, help with washing, arrangements for visiting public areas, hearing, seeing, help with dressing, walking, diet and help with eating.

*Administration:* The patient's performance is scored by relatives or nurses based on their observations.
Scoring: Each item is measured on a 3-point rating scale. Response options vary according to the item targeted. The total score ranges from 15 to 45 points; a score decrease denotes an improvement.

4.6.2 Evaluation and Critical Appraisal

There is no English text on the reliability of the NAB scale. With regards to validity for assessing demented subjects, the items hearing and seeing are less relevant and more closely related to deficiencies than to functional disabilities. In one study (126) involving 44 subjects, moderate correlation coefficients (unknown type) were found with other tests included in the NAI, namely the cognitive performance tests (0.42), the NAS (Gerontopsychological self-rating) (0.43) and the NAR (Behavior rating) (0.53). The available information is insufficient to demonstrate the validity and sensitivity of the NAB.

4.7 The Nüremberg-Alters-Alltags-Skala (NAA)

The Nüremberg-Alters-Alltags-Skala (NAA) is comprised of 20 items assessing activities of daily living from the patient's point of view. Each item is measured by a 3-step scoring. The total score ranges from 20 to 60 points; a score decrease denotes an improvement. The NAA was developed by Oswald and Schaltenbrand for German-speaking patients (given in (46)) as part of the Nüremberg Gerontopsychological Inventory (NAI). The NAI is a battery of measurements to evaluate intervention-induced changes in old age. No English language literature is available on the NAA and its psychometric properties.

4.8 Activities of Daily Living Checklist (ADLC)

The Activities of Daily Living Checklist (ADLC) was used in two metrifonate drug trials, both conducted by Becker and his colleagues (128;129). It is important to note that the provided references on the ADLC are inadequate and falsely suggest that this scale has been published, which is not the case. In one of the drug trials (128), the scale is succinctly described as a 50-item list of common daily activities. The caregiver rates each item with the same 7-point continuum ranging from very much improved to very much deteriorated. The ADLC compares current behaviors to behaviors prior to
study entry. No testing of this scale has been done and there is, as yet, no justification for its use.

4.9 The Interview for Deterioration in Daily Living in Dementia (IDDD)

4.9.1 Description

The Interview for Deterioration in Daily Living in Dementia (IDDD) (130) is a 33-item questionnaire designed to assist clinicians in detecting functional loss in community living demented patients. The original development and testing of the IDDD was done and published in Dutch.

**Constructs Measured:** The IDDD measures functional disability in self-care (16 items such as washing, dressing and eating) and complex activities (17 items such as shopping, writing and answering the telephone).

**Administration:** The IDDD is completed by caregivers in the context of a structured interview. To prevent their interpretation of the severity of the disability, the caregivers are asked to give concrete descriptions of the functioning of the patient.

**Scoring:** For all items, initiative to perform activities and performance of activities are evaluated. According to Burns and his colleagues (131), the severity of the impairment is rated on a 7-point scale, where 1-2=no or slight impairment, 3-4=mild impairment, 5-6=moderate impairment and 7=severe impairment, giving a total range score of 22-231.

After the initial development of the scale, alternative response options were used by Teunisse and her colleagues (130) to allow comparison with baseline function. There were three options: 1=same frequency of assistance needed, 2=more often assistance needed, and 3=nearly always assistance needed. In this case, the calculated sum score range from 33 (no deterioration) to 99 (severe deterioration).

4.9.2 Evaluation

**Reliability:** High internal consistency is reported with a $\alpha=.94$ (cited in (132)). No other reliability data are available.

**Validity:** Some evidence of construct validity can be inferred from a study conducted by Teunisse and her colleagues (130) on the overall severity of dementia. In this study (n=30), the IDDD was found to correlate with other related constructs, such as cognitive
impairment (r=0.77), behavioral disturbances (r=0.67), burden experienced by the caregiver (r=0.47) and with duration of illness (r=0.45).

4.9.3 Critical Appraisal

This functional rating scale appears to be appropriate to assess community-living patients with mild and moderate levels of dementia. It assesses a substantial proportion of complex activities likely to be affected during the first stages of the AD. Moreover, the number of non-redundant items in the scale is viewed positively since it may increase the sensitivity of the tool. However, although some construct validity could be inferred from one study, empirical information on the testing of the IDDD and its measurement properties is seriously lacking.

4.10 Unified Activities of Daily Living Form (Unified ADL)

The Unified Activities of Daily Living Form (Unified ADL) was devised in 1973 by Donaldson and his colleagues (133) to contain all self-care and mobility variables commonly used to assess patients’ functional status between 1950 and 1970. A 20-item scale was produced, out of which the Barthel, Katz and Kenny indexes standard scores could be derived. The need for assistance is scored for every item, on a 10-point rating scale. The psychometric properties of the scale have not been studied.

4.11 Dependence Scale (DS)

4.11.1 Description

The Dependence Scale (DS) was developed in 1994 by Stern and his collaborators to serve as a consistent milestone in the progression of dementia. This scale stages the dependence level of the patients, in terms of their need for help in performing occupational functions. It is designed to be administered to an informant who is well informed about the patient's day-to-day activities. In total, 13 items are included and coded either on a 3-point scale (no need, occasional need, frequent need) (2 items) or on a 2-point scale (no, yes) (11 items). A dependence level (Level 0 to Level 5) is subsequently derived, according to precise criteria. A higher level indicates more severe
dementia. Also included in the scale is an Equivalent Institutional Care (EIC) chart, to score the level of care the patient is receiving.

### 4.11.2 Evaluation

#### 4.11.2.1 Reliability

To assess inter-rater reliability, Stern and his team (134) studied 20 informants who were independently interviewed by two raters. The results revealed 100% agreement on dependence levels and ICCs of 0.90 for the composite scores and of 0.73 for the EIC scores. Based on a sample comprising 233 subjects, the scale's internal consistency was estimated with a Cronbach's alpha of 0.66. Subjecting responses to Principal Components Analysis revealed three subscales which had increased alpha values: Cognitive support (4 items, alpha=0.93), Assistance-elder active (5 items, alpha=0.87), and Assistance-elder passive (4 items, alpha=0.78).

#### 4.11.2.2 Validity

With regards to validity, significant correlation coefficients (p<0.01) with cognitive and functional impairment measures were demonstrated in the same study involving 233 patients (134). The scales used included the mMMS, the Clinical Dementia Rating Score, the BDRS (four subscales) and the EIC rating. The BDRS scores and the dependence level were both independently related to the mMMS scores, suggesting that these scales do not overlap but tap different concepts. The relationship between dependence level and patient's living situation was also demonstrated (p<0.01) as well as a time trend for increased dependency (Hotelling's F=29.68, p<0.001; n=150). Finally, the hierarchical structure of scale items was assessed using a Guttman scalogram. The items were found to meet this requirement, as indicated by a coefficient of reproducibility of 0.97 and a coefficient of scalability of 0.66.

#### 4.11.3 Critical Appraisal

The Dependence Scale shows adequate reliability in terms of inter-rater reproducibility and internal consistency of subscales. Validity has been studied from several perspectives, building up justification for its use. Indeed, results demonstrated
the relevance of the dependence concept for studying the demented elderly and the sensitivity of the scale to disease progression. In conclusion, the Dependence Scale is a potentially valuable outcome measure in epidemiological research on AD.

4.12 Overall Conclusion on the Functional Disability Scales

The functional disability scales that were used in this review exhibit several limitations with respect to their psychometric properties. Lack of evidence for reliability and validity has been noted in several cases, even with classic scales developed decades ago, such as the IADL, the PSMS and the BDS. The choice of these older scales over the more recently developed ones is also questionable, since updated theoretical and methodological knowledge cannot be integrated and exploited. However, the most serious problem encountered in this review is that many of the scales show little responsiveness to symptomatic antidementia drugs, even in the presence of measurable cognitive improvement. Given the nature of the global behaviors these instruments were developed to assess, it comes as no surprise that they may be sensitive only to substantial changes. Such changes are unlikely to occur with less severe levels of dementia. To explain low responsiveness, it can also be argued that there are many factors outside the cognitive ability that influence the functional capacity of the elderly, such as physical and sensory impairments or psychosocial issues. Therefore, it may be appropriate to adapt the theoretical framework of assessing function from a performance to a cognitive focus. In the area of clinical trials for AD, there is thus a need to develop specific tools for the assessment of functional outcomes. One of the most recent instrument reviewed, the Dependence Scale, confirms that such an endeavor is valuable.

4.13 Introduction: Quality of Life Scales

The assessment of quality of life is becoming increasingly important in AD drug trials (135-138) (139) and AD research (140) (141). According to the WHO (cited in (117)), quality of life can be defined as the integration of cognitive functioning, activities of daily living, social interactions and psychological well-being.
Five trials identified quality-of-life scales as secondary outcome measures. The Quality-of-Life (QoL) (142) and the Progressive Deterioration Scale (PDS) (143) were used in four trials (135-138) and one trial respectively (144). The two scales are presented below.

4.14 The Quality of Life Scale (QoL)

4.14.1 Description

The Quality of Life (QoL) assessment (142) is a patient-rated, 10-item scale that evaluates a person's feeling of well-being. It was developed to assess effective living behaviors or positive conditions of living through the use of several measurable social indicators. The QoL is intended for mental health in or outpatients.

**Constructs Measured:** Social indicators encompassing working, leisure, eating, sleeping, social contact, earning, parenting, loving, environment and self-acceptance.

**Administration:** Can be self-administered or filled out by external judges. Internal and external assessments are likely to differ (142).

**Application:** The QoL requires minimal abilities in reading and writing. The total score can reach a maximum of 500. However, the item related to parenting is not applicable for individuals who have no children. In this case, the total raw score has a maximum of 450.

**Scoring:** The QoL items are assessed along a 6-point rating scale and assigned 0, 10, 20, 30, 40 or 50 points. The lowest and highest anchor points have been defined for each social indicator but it is not clear, in the original publication, whether or not the respondents are presented with these definitions. Cutoffs scores are as follows: 350 or more means fairly successful conditions of living and quality of life; 250-349 suggests painful but adequate coping; 100-249 indicates that persons are both suffering a great deal and seeking immediate help; institutionalized mental patients score less than 100. These cutoffs are based upon QoL assessment pilot studies carried out on patients undergoing psychotherapy (142).

4.14.2 Evaluation

Blau (142) did not assess the QoL's reliability or validity. However, some evidence of content validity (content relevance and coverage) is provided with the
description of the instrument's construction. In the published paper (142), the author acknowledges the need for revision-oriented field testing and standardization.

4.14.3 Critical Appraisal

The QoL scale focuses on the presence of effective social behaviors rather than their absence. Moreover, the instrument's content is consistent with the definition of quality of life given by the WHO (139). However, the scale appears to be in an early stage of development and data on reliability, validity and responsiveness to change are lacking. This may, in part, explain why inconsistent or non-significant results were found and reported in the context of the donepezil drug trials (135-138).

It is important to note that some theoretical work has been conducted on assessing quality of life in AD and objective measures of quality of life have been described in the literature (141;145). Examples of currently used quality of life instruments include: the EuroQoL, the SEIQoL, the SF-36, the NHP, etc. (141). Methodological issues in measuring quality of life, such as responsiveness and use of self-reports by persons with cognitive-communicative problems, have been addressed in review papers (141;146). In light of the more recently published work on quality of life theory and the other available instruments, the choice of the QoL scale as an outcome measure is questionable.

4.15 The Progressive Deterioration Scale (PDS)

4.15.1 Description

The PDS was developed in 1989 by DeJong and colleagues (143) to measure the quality of life of AD patients. It consists of 27 factors designed to provide a sensitive indicator of differences and changes in the quality of life.

Constructs Measured: PDS examines activities of daily living and instrumental activities of daily living. Losing such abilities is believed to profoundly affect the patient's quality of life. Examples of content areas (11 total in PDS) include: extent to which a patient can leave the immediate neighborhood; use of familiar household implements; involvement in family finances, budgeting, etc; self-care and routine tasks.
Administration/Application: The scale was designed to be self-administered by the caregiver using a bipolar visual analog scale. The scale requires 10 to 15 minutes to complete is based on observations about the patient’s daily behavior.

Scoring: Each question is scored by measuring the distance along the line on a scale from 0 to 100. A composite score is derived from averaging across the items for a maximal score of 100. A higher score indicates better quality of life.

4.15.2 Evaluation

4.15.2.1 Reliability

Internal consistency was reported through split-half reliability with Kuder-Richardson coefficients ranging from 0.92 to 0.95 (143). The PDS appeared to be reliable on a test-retest basis with one week between administrations. With respect to three stages of AD (based on GDS stages 2-3, 4-5 and 6-7) and for the entire sample, \( r_p \) are respectively of 0.89 (n=14), 0.78 (n=44), 0.78 (n=65) and 0.90 (n=123).

4.15.2.2 Validity

Good content validity is derived from the description of the instrument development procedures. The latter involved in-depth interviews with caregivers and subsequent step-by-step data reduction. With respect to construct validity, the PDS was found to discriminate between non-demented elderly and AD patients (n=80) with a 95% accuracy level in classification (143).

4.15.2.3 Responsiveness to change

In a study involving n=141 subjects, DeJong et al. (143) demonstrated that the PDS can differentiate between three stages of AD as measured by the GDS (t ratio<0.01). However in their study of 62 subjects, Knopman and Gracon (147) found high levels of inter-subject variability. Consequently, it may be difficult to define a reliable increment of change over a short period of time.

4.15.3 Critical Appraisal

The PDS was specifically designed to assess differences and changes in the quality of life of patients with AD as the disease progresses. It has been shown to be
sensitive to three severity stages of dementia and to have good reliability and validity. However, small changes may remain undetected because of large intra-subject variance (147). Furthermore, it is important to note the apparent overlap with domains assessed by functional outcome measures. Nonetheless, the scale may provide a useful means of clinically evaluating AD medications.

4.16 Overall Conclusion to the Quality of Life Scales

Several authors (139;145;148) confirm that the study of quality of life with dementia is still in its infancy, both from conceptual and operational viewpoints. Quality of life assessment tools tend to blur the distinctions between the biological aspects of the disease and their social and functional consequences. At this point in the process of global harmonization of dementia guidelines, it is considered reasonable to collect quality of life data (during Phase III drug trials), but not require an improvement in measured quality of life to be a condition for demonstration of efficacy and subsequent approval (139). The review of the QoL and PDS highlight several limitations in psychometric properties that tend to confirm this point of view.
5 BEHAVIOR AND MOOD OUTCOME MEASUREMENT SCALES

5.1 Introduction

Behavior and mood problems are well-recognized and pervasive components of AD (149). There is, however, lack of consensus in the scientific community on behavior and mood problems. Various terms are used to refer to these problems, including behavioral disorders, psychiatric symptoms, noncognitive symptoms and neuropsychiatric symptoms. In addition, some controversy exists as to the specific problems considered characteristic of AD and their prevalence at each stage of the disease (149-151).

The relationship among behavioral problems, cognitive status and functional status in AD is complex and, as knowledge accrues, the interaction of these three spheres is taking on more importance in drug development (150). According to experts (152;153) (154) (155), some assessment of behavior and mood is necessary in antidementia drug trials to monitor symptoms which may exacerbate cognitive or functional disability and to evaluate the possible effects of antidementia drugs on behavioral symptoms. In practice, behavioral and mood problems in AD are of great importance, as their presence bears on the level of care required for the clinical management of patients (150;156), on caregiver burden, as well as on the cost of care (157).

5.2 Selection for Review

Thirteen of the reviewed trials assessed behavior and mood problems while the remaining 12 did not. It is interesting to note that, in many of the trials which did not assess behavior and mood, functional or quality of life assessments were carried that included items compatible with the identification of behavior and mood disturbances. There were ten different scales among these 12 trials – a remarkable degree of diversity. Behavior and mood disturbances appeared to take on particular importance in six trials that relied on more than one scale for assessment. We chose for evaluation the Brief Psychiatric Rating Scale (identified as a primary outcome in (158)), the ADAS-noncog (the single most popular scale), as well as the Behavior Rating Scale for Dementia and the NeuroPsychiatric Inventory (both dedicated to dementia patients). Two scales were
not reviewed: the BfS (122) (German references were cited) and the IPSC-E (159) (160) (time constraints). The other scales (i.e., RAGS, DMAS, CSSD, DBD), which were used in only one or two trials, are briefly described. There were no significant differences between placebo and treatment groups reported on any of the behavior and mood scales not or briefly reviewed.

5.3 Brief Psychiatric Rating Scale (BPRS)

5.3.1 Introduction

The Brief Psychiatric Rating Scale (BPRS) (161) was originally developed to evaluate treatment effects in psychopharmacology research in young or middle-aged adults and was imported directly from psychiatric studies into the assessment of dementia (162).

5.3.2 Description

Constructs Measured: The BPRS contains 16 items emphasizing psychiatric problems such as somatic concern, anxiety, emotional withdrawal, conceptual disorganization, guilt feelings, tension, mannerisms and posturing, grandiosity, depressive mood, hostility, suspiciousness, hallucinatory behavior, motor retardation, uncooperativeness, unusual thought content and blunted affect. The items were originally chosen on the basis of their usefulness in measuring the response of psychiatric patients to therapy (163).

Administration: An interviewer administers the BPRS through direct observation of, and conversation with, the subject. No caregiver is used. The interview is directed (each item is introduced with a short description of the target symptom), but unstructured. The interviewer rates each item on a 7-point scale of severity, ranging from 0 ('not present') to 6 ('extremely severe'); the interviewer is asked to rate the severity in comparison with the population of (psychiatric) patients who do have the symptom in question.

Scoring: The total score is obtained by summing across item scores. A higher score indicates more problems. The range of scores is 0 to 96.
5.3.3 Evaluation

5.3.3.1 Reliability

Interrater: Mack and Patterson (163) evaluated 61 community patients with AD and 20 cognitively intact spouses of these patients. Information was sought on behavior occurring in the week prior to the interview. Fifty-five of these 81 subjects were rated by two interviewers during a session attended by both, but conducted by one interviewer. The estimated prevalence of symptom-specific disturbances ranged from 0 to 50% and was low for most items. However, most patients and controls had at least one symptom rated as present. The kappa statistic ranged from 0.134 to 1.00 (median 0.448). Agreement on presence of symptoms was only 33% for all symptoms combined. The low prevalence of symptoms limited the assessment of reliability.

5.3.3.2 Validity

Content: The factor structure of the BPRS was investigated in 117 patients with AD. Ownby et al. (164) deleted six symptoms, which were found not to contribute meaningful information to the analysis. While commenting on the remarkable stability of the BPRS, the investigators considered that the interpretation of the factor structure required an interpretation specific to AD patients (i.e., one that differed from that of psychiatric patients).

Construct/Concurrent: A study conducted by Mack and Patterson (163) compared the frequency of ratings on the BPRS to those on the BEHAVE-AD and the Cornell Scale for Depression in Dementia (CSDD). The analysis showed that some important content areas were not represented in the BPRS (e.g., fears and phobias), while other areas were specific to cognitively intact subjects but not to AD patients (e.g., somatic concern). Overall, the BPRS was found to be less sensitive to symptoms than the Behave-AD and CSSD. The BPRS discriminated less well between the two groups of subjects than the BEHAVE-AD.

5.3.4 Critical Appraisal

The BPRS is of limited relevance for assessing elderly patients suffering from dementia (163;164). The scale has important shortcomings in terms of content as it does
not cover the whole range of behavior and mood problems seen in AD. Direct patient interview, which is featured in the BPRS, may lead to inadequate behavior assessment since the patient is often unable to provide an accurate report; caregiver interview is preferable. Clearly the BPRS, initially developed to evaluate pharmacologic treatment of psychiatric conditions at large, is ill-suited as an outcome measure in antidementia clinical trials.

5.4 Alzheimer's Disease Assessment Scale-noncognitive (ADAS-noncog)

5.4.1 Introduction

The Alzheimer's Disease Assessment Scale-noncognitive (ADAS-noncog) (53;54;56) is the subscale of the ADAS that focuses on behavior and mood problems. As mentioned earlier, the ADAS was developed to measure change in AD patients following pharmacologic treatment. The ADAS-noncog has 10 items grouped into 6 categories (affective disturbance, hallucinations, delusions, activity disturbance, attention and appetite).

Constructs Measured: The items of the ADAS-noncog are depression, tearfulness, delusions, hallucinations, pacing, increased motor activity, tremors, concentration/distractibility, uncooperativeness during testing and decreased/increased appetite. Some researchers have simplified the ADAS-noncog to a 9-item scale by omitting the item on hallucinations.

Administration: Items are rated by a clinician, based on observation of the patient's behavior and on an interview with the patient's caregiver. The period of interest is the week preceding the interview. The administration of the scale is directed, but unstructured.

Scoring: Noncognitive items are scored on a 6-point scale from 0 (no impairment) to 5 (most impairment). Descriptors of severity specific to each item are provided (56). The maximum score is 50 (or 45 for the 9-item scale); a higher score indicates a higher level of problems.
5.4.2 Evaluation

5.4.2.1 Reliability

Test-retest: Test-retest reliability after one month in 27 patients with AD was reported to be 0.59 by the scale developers (56). Standish et al. (63) obtained a higher test-retest coefficient (ICC=0.70) in a different group of 27 elderly with AD. Raters were trained medical students. Weyer et al. (61) reported a test-retest reliability ($r_p$) of 0.98 for a group of 440 elderly with AD over a three to four week period. Inter-rater reliability in the same initial group of 27 patients with AD was 0.95 (ICC) (56). On the other hand, Standish et al. (63) reported ICCs of 0.42 and 0.55 for his group of 27 elderly.

Using data from their test-retest assessment, Weyer et al. (61) estimated that the reliable individual change score on the ADAS-noncog was 3 points.

5.4.2.2 Validity

Content: An internal consistency coefficient of $\alpha=0.83$ was obtained by Weyer et al. (61) on the ADAS-noncog, indicating a high degree of homogeneity of item contents. Whether or not the coefficient was boosted by a ‘halo effect’ (i.e., tendency to rate symptoms of unknown severity in the same manner as previous symptoms) could not be ascertained.

Construct/Concurrent: Weyer et al. (61) correlated scores of 440 patients with AD, comparing the ADAS-noncog with the MMSE (-0.49), SKT (0.50), ADAS-cog (0.67) and the Nurses’ Observation Scale for Geriatric Patients (NOSGER) (0.79). The NOSGER, not previously described, assesses observable behaviors in six areas relevant to daily functioning (memory, instrumental activities of daily living, activities of daily living, mood, social behavior and disturbing behavior).

5.4.2.3 Responsiveness to Change

Time: The fact that the ADAS-noncog is used in many drug trials to monitor change is sometimes presented as an indication that the scale is responsive to change. It should be noted, however, that scale endorsement and scale responsiveness are two different things.
Severity: In the group of AD patients evaluated by Weyer et al. (61), the 166 subjects with mild dementia had a group mean value for the ADAS-noncog equal to 9.4 (SD 4.9), while the 274 subjects with moderate dementia had a group mean value equal to 15.3 (SD 6.8). These values are consistent with the increase in behavior and mood problems with increasing dementia severity.

5.4.3 Critical Appraisal

The ADAS-noncog is a well-evaluated scale that shows adequate reliability and validity. However, there is evidence that the scale is less reliable when used by nonclinicians. Standish et al. (63) developed a standardized version of the ADAS-noncog (SADAS-noncog) and showed that the scale benefits from a more structured administration.

With 10 items, the ADAS-noncog is considered to be a “brief” scale. It assesses a combination of psychopathology and other abnormalities (162). For example, three of its items are outside the domain of behavior and mood (tremors, concentration/distraction and decreased/increased appetite). Other relatively frequent problems specific to AD are absent from the scale’s content such as aggressiveness, anxiety and phobia (165). There is reason to believe the scale may be insensitive to change. For reasons related to contents, the ADAS-noncog does not appear to be the scale of choice to monitor behavior and mood problems in AD. As expressed by Ferris et al., “[b]ehavioral assessment in clinical trials should focus on the behavioral domain and exclude any cognitive functions, because these are covered by their own specific instruments. In addition, these behavioral assessments should be able to isolate drug side effects without bias against the overall efficacy of the drug” (166).

5.5 CERAD Behavior Rating Scale (C-BRSD)

5.5.1 Introduction

The CERAD Behavior Rating Scale (C-BRSD) was developed by the Consortium to Establish a Registry for Alzheimer’s Disease (CERAD) to assess behavior in AD patients through interviews with primary caregivers (167). Scale items were selected
from a literature review and consultation with experts to represent a broad range of behavioral symptoms (167).

5.5.2 Description

*Constructs Measured:* The C-BRSD contains 48 items that can be grouped into 8 factors including depressive symptoms, psychotic symptoms, defective self-regulation, irritability/agitation, vegetative features (i.e., sleep, tiredness, appetite), apathy, aggression and affective lability (167) (168). Frequency of behaviors is quantified, but severity (judged to be more difficult to anchor) is not. Of the 48 items, 40 are rated on a 4-point scale from ‘1’ (present 1-2 days in the last month) to ‘4’ (>16 days in the last month); the remaining items are rated as a yes/no response (167) (168).

*Administration:* A trained interviewer completes the C-BRSD during a structured interview with the patient’s primary caregiver. The scale assesses behavior over the preceding month, but also takes note of behavior more than one month ago and since the beginning of the disease (167) (168). The interviewer is expected to make an overall judgement about the validity of the caregiver’s response. The scale is administered in 20 to 45 minutes.

*Scoring:* There have been no recommendations made on scoring the C-BRSD in the development article of Tariot et al. (168). Two options for scoring have been reported in the literature. These are (i) summing the frequency ratings across items (169) and (ii) counting the number of behavioral problems endorsed (170). In either case, a higher score indicates more behavior problems.

5.5.3 Evaluation

5.5.3.1 Reliability

*Test-retest:* Test-retest reliability was assessed in the context of the Alzheimer’s Disease Cooperative Study (ADCS) Instrument Project (169). Repeat interviews were performed one month apart with caregivers of cognitively intact subjects (n=64) and caregivers of AD subjects with different levels of disease severity (n=241). Disease severity (5 levels) was established according to cutoffs on the MMSE and each severity level had equal
representation. Reliability, measured using Spearman’s rank correlation coefficient, was 0.62 for intact subjects and ranged from 0.70 to 0.89 for demented subjects.

*Interrater:* Interrater reliability was established in 104 subjects from a convenience sample of 303 AD subjects by having two raters rate the same interview (167). Interrater reliability was assessed as overall agreement and weighted kappa. Overall agreement was high, ranging from 91.3% to 100%. Weighted kappas ranged from 0.77 to 1.00 with most items having a weighted kappa between 0.90 and 1.00. Although they concluded that interrater reliability is high on the C-BRSD, the authors (167) were careful to point out that, had the ratings been done by two independent interviewers, the agreement might had been lower. Also, since the sample as a whole did not include subjects with extensive psychopathology, the rate of agreement was probably inflated by the high agreement on absence of symptoms.

### 5.5.3.2 Validity

*Construct:* Construct validity has been supported by factor analysis, which indicated loading on 8 factors (see above) (168).

*Construct/Concurrent:* The concurrent validity of the C-BRSD was evaluated for depressive symptomatology. Jacobs et al. (170) classified AD patients as depressed and nondepressed according to a semi-structured interview for behavior symptoms that enables diagnosis of clinical depression (both minor and major). Of the 69 patients selected, 29 were diagnosed as depressed. On the basis of their responses to the C-BRSD anxiety and depression items, 70% of patients were correctly classified as depressed or nondepressed.

### 5.5.3.3 Responsiveness to Change

Data on responsiveness to change were obtained from a single study, namely the ADCS Instrument Project (169). In this study, the C-BRSD was scored by summing the frequency ratings across items, while deleting the last item for ‘occurrence of any other behavior symptoms’; scores ranged from 0 to 164. Cognitively intact and AD subjects were evaluated at regular intervals within a one-year period.

*Longitudinal study:* The change score on C-BRSD between the 1-month and 12-month assessment was investigated for responsiveness to change (169). Change scores ranged
from \(-2.5\) (SD 20.8) in the most severely affected group to \(3.7\) (SD 17.5) in the group of intermediate severity. The magnitude of mean change was similar in the cognitively intact subjects (i.e., \(-1.5\)), but the variation of results was much smaller (i.e., SD 4.5). The authors concluded that the C-BRSD indicated little change in behavior over the 12-month period.

**Severity:** In the ADCS Instrument Project, the C-BRSD discriminated between AD patients and cognitively intact controls (169). The mean (baseline) C-BRSD score for intact patients was 4.7 (SD 8.3), while the mean (baseline) score of AD patients ranged from 18.9 (SD 12.8) in the lowest severity level to 40.4 (SD 26.8) in the highest severity level.

### 5.5.4 Critical Appraisal

The C-BRSD was shown to have good reliability. However, the assessment of interrater agreement was probably inflated by the inclusion of a large number of subjects with mild symptomatology. In this context, additional information on the reliability of the scale that includes more disturbed subjects is warranted. Adequate concurrent validity for the depression and anxiety symptoms was shown.

One expects to observe an increase in the frequency of behavioral symptoms with disease progression (166). This was not demonstrated with the C-BRSD, possibly owing to a 12-month period not being adequate to observe change. However, increase in behavior and mood problems was also not shown with decreasing MMSE scores. Patterson et al. (169) observed that the C-BRSD represents an additive combination of diverse items. Although some items may deteriorate as dementia progresses, other items may actually improve. There seems to be no good reason to assume that all items follow the same progression. Thus, Patterson et al. (169) envision the development of homogeneous factor-based subscales that would make the tracking of behavior problems more specific and meaningful (169).
5.6 Neuropsychiatric Inventory (NPI)

5.6.1 Introduction

The Neuropsychiatric Inventory (NPI) is a scale developed by Cummings et al. in 1994 (171). The NPI assesses a wide range of neuropsychiatric disturbances and was developed to help distinguish among different dementias. The inclusion of symptoms known to be rare in AD but common in other types of dementia (e.g., euphoria and disinhibition; compulsive and repetitive behaviors) increases the diagnostic utility of the scale. The NPI (157;171) is based on a structured interview with a caregiver who is a reasonably good observer of the patient's behavior. Screening questions, subquestions and rating on frequency and severity are the main features of the scale.

Constructs Measured: The original scale evaluated 10 items: delusions, hallucinations, dysphoria (depressed mood), anxiety, agitation, euphoria, apathy, irritability, disinhibition and aberrant motor behavior (pacing and rummaging). Two items - nighttime behavior and changes in appetite and eating behaviors - were added after completion to expand the number of syndromes assessed. The psychometric properties of the NPI were established with the original 10-item scale (171).

Administration: The NPI is administered by an interviewer to a knowledgeable caregiver. Screening questions are asked first to minimize administration time while providing an overview of each specific behavioral domain. If the screening question suggests the presence of problems in the domain being investigated, seven or eight additional subquestions are asked to the caregiver who then rates the frequency of the behavioral disturbance on a 4-point scale (1= < 1 per week; 4=once a day or more) and the severity on a 3-point scale (1=mild, 2=moderate, 3=severe).

Scoring: The total score for each domain is calculated by multiplying the frequency rating by the severity rating. A total score is calculated by adding the scores across domains. A higher score represents more problems. The maximum total score is equal to 120 when 10 domains are assessed and to 144 when 12 domains are assessed.
5.6.2 Evaluation

5.6.2.1 Reliability

Test-retest: Test-retest reliability was determined by conducting 2 interviews 3 weeks apart with 20 caregivers; half of the interviews were done by telephone (157). Correlation coefficients were 0.79 for overall frequency and 0.86 for overall severity. Correlation coefficients were somewhat lower for some specific domains, such as severity of agitation (0.51), severity of irritability (0.53), frequency of anxiety (0.51), and frequency of irritability (0.51).

Interrater: Two raters scored the NPI responses of interviews conducted with one of the raters (157). Raters were blind to each other's ratings. Forty-five subjects with a wide range of MMSE scores (mean 17.4) were assessed. Results, which were expressed as percent crude agreement between raters, ranged from 90% to 100% for different behaviors. Inter-rater reliability was similar for frequency and severity ratings.

5.6.2.2 Validity

Content: The scale was submitted to a 'Delphi panel' composed of 10 experts in the field of behavior and geriatrics (157) (171). Each panel member rated the screening and subquestions for content validity. Only one item was identified as needing reformulation (e.g., 'troublesome behavior' was reworded as 'aberrant motor behavior').

Construct/Concurrent: The NPI was compared to the BEHAVE-AD by correlating questions measuring similar behaviors on both scales (157) (171). Forty patients and their caregivers participated in this study. Twenty patients had AD, 9 had vascular dementia and 11 had some other type of dementia (mean MMSE = 19.2). Most item pairs showed moderate to good correlations ($r_S = 0.54$ to 0.80). Comparison of the NPI 'dysphoria' item with the Hamilton Rating Scale for Depression (HAM-D) showed moderate correlation ($r_S = 0.62$).

5.6.2.3 Responsiveness to Change

Time: It is recognized that behavioral and mood disturbances are not present in all patients and do not necessarily progress in tandem with other aspects of dementia (156;157). Lack of steady progression can reduce the opportunities for measuring change.
and thereby constrain the assessment of responsiveness to change for behavioral scales. It was shown in an open-label study of tacrine that NPI scores improved in treated patients; dose-related improvements were concerned with apathy and aberrant motor behavior (157). For lack of better data, these results were invoked as evidence of the scale’s responsiveness to change.

**Severity:** Cognitively intact elderly differ from demented elderly on NPI profile, except for dysphoria, disinhibition and irritability. NPI scores showed an absence of problems in 40 cognitively intact elderly. By contrast, data from 50 subjects with AD show that most subjects have some form of psychopathology identified by the NPI. The disturbances whose frequency increased with dementia severity were agitation, dysphoria, anxiety, apathy and aberrant motor behavior (157).

### 5.6.3 Critical Appraisal

The content validity was well established by a panel of international experts. Reliance on a caregiver insures that a comprehensive assessment of behavior is possible. Reliability and validity are satisfactory. However, we noted the following inconsistencies or limitations. The time period over which behavioral symptoms were appraised appeared to be the preceding 4 weeks, but this is not clearly specified in the description of the scale in different settings. Variations in the time period of recall are likely to affect the reliability and validity of the scale. The authors offered no justification for their scoring system based on the product of frequency and severity. Inter-rater reliability was expressed as overall agreement, a measure not consistent with the comparison of continuous ratings. There was little information to substantiate the reported low false negative rate of the screening questions (less than 5%). We did not secure a copy of the scale and could therefore not appraise the relevance of the screening questions and subquestions. The results provided in the tacrine open-label study do not constitute valid evidence of responsiveness to change.

### 5.7 Relative’s Assessment of Global Symptomatology (RAGS)

The Relative’s Assessment of Global Symptomatology (RAGS) (172) is intended to assess patient’s behavior in the community and is completed by a close relative or friend. The 21 items in this self-administered scale sample psychiatric symptoms and
behavior rather than social or interpersonal adjustment. The presence of each symptom is rated on a 5-point scale, from ‘1=not at all’ to ‘5=extremely’. Symptoms and behavior within the week prior to test administration are to be considered by the rater. The authors suggest that each item should be scored separately. However, if needed, a mean score can be calculated by summing the individual items and dividing by 21 - the mean score range is 1 to 5, an increase in score indicating an increase in symptoms.

The RAGS was able to distinguish demented patients from other patient groups and the scale’s depression items showed evidence of concurrent validity when compared to the Inventory of Psychic and Somatic Complaints-Elderly (159). However, several items refer to problems experienced by demented elderly irrespective of their behavior and mood state and reflect cognition or global health rather than mood (ex: need for assistance in ADLs, fatigue, lack of motor coordination, forgetfulness, disorientation, speech difficulty, variability in mental functioning). Clearly, this scale, not developed for patients with AD, has limited content validity for assessing behavior and mood problems specific to AD.

5.8 The National Institute of Mental Health Dementia Mood Assessment Scale (DMAS)

The National Institute of Mental Health Dementia Mood Assessment Scale (DMAS) (173) is intended as an objective measure of mood in cognitively impaired subjects. The scale is administered by gathering information over the preceding week from all sources including the family or nursing staff. The scale has 24 items: the first 17 items are designed to measure depression severity, while the last 7 items are intended to provide an indication of dementia severity. Indicators of severity (i.e., descriptor sentences) are provided for even-numbered steps (i.e., 0, 2, 4, and 6), but are not provided for odd-numbered ones (i.e., 1, 3, 5). Symptoms and behavior within the week prior to test administration are to be considered by the rater. The score range for the 17-item subscale is 0 to 102 and 0 to 42 for the 7-item subscale; a higher score indicates more problems. Factor analysis of the 17 items identified four factors: depression, social interaction, anxiety, and vegetative symptoms. The 7-item subscale was shown to correlate well with some measures of cognition (173).
The scale developers indicate that the DMAS may not be appropriate for severely demented or uncooperative patients. Irrespective of the degree of dementia severity, the DMAS has limited usefulness for the purpose of assessing behavior and mood. On the one hand, the 17-item subscale is mainly concerned with depression and is thus restricted in the scope of problems it evaluates. On the other hand, the 7-item scale does not address behavior and mood problems at all.

5.9 Dementia Behavior Disturbance Scale (DBD)

The Dementia Behavior Disturbance Scale (DBD) (174) is a 28-item scale developed to quantify behavior disturbances among patients with dementia. Behavior disturbance was defined as 'the outward manifestation of some underlying cognitive, psychological, or physiological deficit - regardless of etiology - likely to cause stress to those caring for the patient' (p.221). Items were sampled from the domains usually associated with dementia: passivity, agitation, eating disturbances, aggressiveness, diurnal rhythm disturbances, and sexual misdemeanor. To increase conceptual clarity, the items selected for inclusion refer to observable behavior (e.g., accusations) and not to ideation associated with the behavior (e.g., paranoia). The scale was designed to be used outside the clinical setting with the caregiver as informant, either as an interview or as a self-administered questionnaire. Each behavior is rated on a Likert scale for frequency of occurrence during the past week (0=never; 4=all the time). The score range is 0 to 112; a higher score indicates more problems. No time frame was recommended, but the validation study was done using a 1- and 2-week time frame.

The scale developers report preliminary evaluation results showing adequate test-retest reliability and construct validity. However, they note that some behaviors are very common among elderly with dementia (e.g., repetitive questions; losing things; unwarranted accusations), while some others are so rare that they could for all intents purposes be excluded from the scale (e.g., indecent exposure, screaming, throwing food). Although they can be stressful to the caregiver, some scale items refer to daily activities or performance (e.g., excessive sleep during the day; incontinence of urine or stool) and are clearly outside of the realm of behavior and mood problems.
5.10 Cornell Scale for Depression in Dementia

The Cornell Scale for Depression in Dementia (CSDD) (175) is a 19-item scale derived from current concepts of the phenomenology of depression (including mood-related changes; behavioral disturbance; physical signs; cyclic functions; ideational disturbances). Rated primarily on the basis of observation, this scale does not include items, such as phobias or obsessions, which require reliable patient self-reporting. The interview is carried out in two steps. First the clinician interviews the patient's caregiver on each item of the CSDD. Next, the clinician briefly interviews the patient on the basis of CSDD items and observation. If there is a large discrepancy between the caregiver and patient interviews, then the clinician interviews the caregiver again to clarify the reason for disagreement. Scoring is done on the basis of the clinician's best judgement: 0=absent, 1=mild or intermittent, 2=severe. Total score ranges from 0 to 38; a higher score indicates more severe depression. Total administration time is about 30 minutes (20 minutes with the caregiver; 10 minutes with the patient). The CSDD is unique in being a validated depression scale for patients with dementia (176).

5.11 Overall Conclusion to the Behavior and Mood Scales

The measurement of behavior and mood disturbances is difficult because shared terminology and a body of theoretical concepts akin to cognition are both lacking (177). The behavior and mood domain often includes indicators of function, changes in mood, socially inappropriate actions, and psychiatric symptoms of hallucinations and delusions, as though all share the same biological substrate. By including widely varying items in their definition, with a seeming lack of boundaries, the different scales are not easily comparable. The BPRS, for example, does not adequately cover an appropriate range of problems in AD. The ADAS-noncog (53;54;56) includes items outside the range of the domain, but does not include others important to AD. Further, in a cognitive paradigm, stimuli are well controlled, but the stimuli of behavior and mood are often unidentified, leading to additional components of measurement error.

The use of behavior and mood scales in clinical trials is based on the belief that an intervention should over the course of time ameliorate (or prevent) symptoms without inducing other behavior and mood problems. While the course of deterioration for
behavior and mood is still uncertain in AD, some indirect evidence (e.g., correlation with severity of dementia) suggests a possible mode of decline (150), although not over the short term(149;151). Many AD drug trials exclude patients with behavior and mood problems and include the domain as a secondary outcome. These RCTs are well suited to assess the side effects (or mixed effects) of an intervention on behavior and mood, provided the scales used are sensitive to change.

A number of qualities must be considered in the interpretation of behavior and mood scales. The interval of observation (or period of recall) is often left unspecified in these scales, and without a clearly and consistently defined interval, the measurements lack coherence and comparability (177). The NPI, for example, has an uncertain time frame (157;171).

The source of information for raters will introduce error that cannot entirely be remedied. For example, scales using self-reported data, which is partly the case with the BPRS, may be unreliable for measuring anything other than immediate states (162). Also, recall of transitory past events by subjects with even mild dementia is suspect, as was suggested by a validity study of the Geriatric Depression Scale (178). Family caregivers tend to lack the skills needed for sophisticated observation and there may also be a reliance on the previous relationship with the subject. Other informants have the disadvantage of fewer and shorter periods of observation. In particular, direct observation by clinician, while sophisticated, is extremely limited and only of use with acute change of behaviors and fast-acting interventions.

There are different strategies for assessment. Some scales are cursory in their assessment (such as the ADAS-noncog), some are detailed (such as the BRSD), and others have a mixed strategy of initial screening and subsequent probing (such as the NPI). The strategy should be considered for choosing an appropriate scale for the needs of the clinical trial, with the limitations of time and thoroughness kept in mind.

Finally, instrument scoring is often ad hoc, without any real evidence of validity. Scoring by summation of the presence of items is logical, provided that the items being summed are truly one-dimensional in AD (177;179) (172) (174). A scale’s responsiveness to change is of the utmost importance and should, over the course of the trial, separate spontaneous remission from the emergence of new behavioral symptoms
(particularly with additive scores) (162). For this reason, global improvement may be too lofty a goal. When the range of behavior problems is comprehensive, the scale’s utility will be enhanced by examining factor-based subscales targeting more focused behavioral constellations (155). As in other domains, many of the behavior and mood scales also lack an empirically driven understanding of clinically meaningful changes.
6 GENERAL DISCUSSION AND CONCLUSION

We critically reviewed the outcome scales, categorized by domain, that were used in the AD drug trials under review (I: A Comparative Analysis of Clinical Trials). The domains subjected to assessment were akin to those proposed by Berg and Montgomery (140) as metrics of effectiveness in AD research. These domains were: (i) the global domain, assessed by single overall ratings; (ii) the cognitive domain, containing both cognitive scales and individual neuropsychological tests; (iii) the functional domain, which included functional disability scales and quality-of-life measures; and (iv) the behavior and mood domain. Other miscellaneous measures that were used rarely, such as time to occurrence of death (121), were not considered.

Knowledge of the psychometric properties of these outcome measures is crucial because they help determine the extent to which the scales are useful for assessing the efficacy of treatments in AD drug trials. Use of reliable and valid measures ensures that drug efficacy results are well substantiated and can be correctly interpreted. Use of instruments with little reliability, validity and responsiveness to change can seriously undermine the credibility of trial findings.

One important observation that can be drawn from this review is the wide diversity of instruments in use for each domain (see Tables 1 and 2 in Chapter 1). From a methodological point of view, the number of scales affected the depth of information that could be obtained for each individual instrument. It was virtually impossible to retrieve all the documentation pertaining to the array of scales that were reviewed, particularly in domains where scale diversity was greatest (e.g., cognition and behavior and mood). However, since thorough database searches and reviews of retrieved papers’ lists of references were conducted, we are confident that the most important references published in English were examined. It is also important to note that the available literature was unequal among scales. Consequently, some instruments were reviewed in more detail than others, and little regard was paid to the frequency of use.

From a theoretical point of view, the diversity of scales can impact on any comparisons that can be made across trials (14;107;108;139). Scale diversity was clearly a feature of the reviewed trials and consensus on scale use across trials was limited. In
the global domain, the CDR was recommended but was only used once (121) (99) as a
definite primary outcome. For the cognitive domain, the ADAS-cog and the SKT have
shown themselves to be good outcome measurements. However, only the ADAS-cog was
used in several trials (128;135-138;154;180-184) as a primary outcome measure. In the
functional domain, no scale was used in more than three [e.g., IADL (99;154;182), PSMS
(99;154;182), GERRI (154;158;183)] or four trials [e.g., QoL(135-138)]. In the behavior
and mood domain, there was somewhat greater consistency of use for the ADAS-noncog
(128;158;180-182;184;185) and the BPRS (98;99;158;186).

Some promising instruments were reviewed, but there was no clear evidence in
favor of any one particular scale. Within each domain, the reasons for this were diverse.
In the global domain, the construct that was assessed did not always appear to be 'global'
in nature and sometimes lacked a strong theoretical background. In the cognitive domain,
even the most accepted scales were lacking in items of adequate difficulty levels for
mildly or moderately demented elderly. For functional scales and quality of life
measures, the need for a more specific tool for the demented elderly was recognized. The
lack of consensus in the medical community on the definition of behavior and mood
problems specific to AD subjects was highlighted in our review. This can partly explain
the inappropriate use of behavior and mood scales with respect to the subjects included in
AD drug trials.

From the perspective of the entire group of scales, one of the most common
problems was the absence of reporting on reliable individual change. As is evident from
this review, correlation coefficients were used almost exclusively by authors to report on
the reliability of the reviewed scales. We consider the reporting incomplete. Indeed,
correlation coefficients can be spuriously high when the range of patient values is large
and, being dimensionless, they cannot be related directly to the scale’s measurement
error. Reliability assessment should go beyond reliability coefficients and provide the
reader with the scale’s measurement error expressed in the units of the scale. This
information could then be used to appropriately identify individual subjects who respond
to treatment.

Another common problem concerning the measurement properties of the
reviewed scales was the lack of data on responsiveness to change. Data on responsiveness
to change were not available for the majority of the scales we reviewed, although these scales were selected for the purpose of showing change. This explains why, in general, investigators did not disclose a priori information on the amount of change to be considered clinically important, but rather relied on the statistical significance of their results to identify change. A closely related problem of measurement properties is the ceiling and floor effects that were observed or suspected with several scales (e.g., CIBIC, SKT, MMSE, GERRI, PSMS, IADL) and would limit the ability to measure change. Taken globally, the impact of these limitations is considerable because they can work to either overestimate or underestimate the true efficacy of the drugs that were investigated in the trials.

A less pervasive yet important issue encountered in this review is the use of scales that are adapted from tests developed in countries other than the ones where the RCTs were conducted. In some cases, the validity of the scale (e.g., SKT) was established in the country of origin, but this information was not available in the country of adoption; in others, the cross-cultural stability of the translated scale was not established (e.g., NAA, NAB), thereby casting doubts on the interpretation of study results. The advent of multicentre drug trials, which enroll patients from different cultures and education levels, and the globalization of health concerns suggest that more research is needed in this area.

When considering the value of scales results, administration procedures also need to be considered. Standardized instructions must be followed whenever they exist. For example, it is unacceptable to substitute a proxy respondent for a patient in a scale that focuses on patient perception, unless the authors suggest otherwise. Other common mistakes include changing items, devising new ones and breaking down the scale in order to use specific parts and omitting others. These modifications jeopardize the psychometric properties of the scale and impede the interpretation of the results obtained. Training of raters also strongly impacts on the reliability of scale scores. Inappropriate training can lead to attenuation of effect. For this reason, it is particularly important to insure homogeneous training to raters involved in multicentre trials. Learning or practice effects arise from repeated administration of a scale to the same subjects. Although it may not be a major issue in Alzheimer's disease, too short intervals between repeated measurements can potentially affect scoring. Particularly in cognitive scales, special
precautions must be used to alleviate these effects, for example, by using various sets of stimuli. Moreover, scales must be used with population with appropriate levels of symptom severity if misleading floor and ceiling effects are to be avoided.

Several recommendations for scale improvement can be listed. The most important one is to focus on the development of scales specific to AD drug trials. Particular emphasis must be put on responsiveness to change in order to detect clinically significant improvements in AD subjects with respect to global status, cognition, functional status, quality of life, behavior and mood. The Alzheimer Disease Cooperative Study-Instrument Development Project (187;188) is an example of how a plan for scale improvement can be implemented. In view of current developments and the growing availability of instruments, RCT investigators remain ever more responsible for optimal choice of outcome measurement scales.
7 REFERENCE LIST


(86) Bowie P, Branton T, Holmes J. Should the Mini Mental State Examination be used to monitor dementia treatments? Lancet 1999; 354:1527-1528.


(142) Blau TH. Quality of life, social indicators and criteria of change. Prof Psychol 1977; 8:464-473.


8  BIBLIOGRAPHY (BY AUTHOR)


Bowie P, Branton T, Holmes J. Should the Mini Mental State Examination be used to monitor dementia treatments? Lancet 1999; 354:1527-1528.


Kanowski S, Herrmann WM, Stephan K et al. Proof of efficacy of the ginkgo biloba special extract EGb 761 in outpatients suffering from mild to moderate primary degenerative dementia of the Alzheimer type or multi-infarct dementia. Pharmacopsychiatry 1996; 29(2):47-56.


