Received: 7 August 2019 | Revised: 18 November 2019 | Accepted: 3 February 2020 | Published online: 16 May 2020

DOI: 10.1002/dad2.12017

Alzheimer's & Dementia
Diagnosis, Assessment & Disease Monitoring

COGNITIVE & BEHAVIORAL ASSESSMENT

Composite cognitive and functional measures for early stage Alzheimer's disease trials

Lon S. Schneider[1] | **Terry E. Goldberg**[2]

[1]USC Keck School of Medicine, Los Angeles, California

[2]Department of Psychiatry, Columbia University Medical Center, New York, New York

Correspondence
Terry E. Goldberg, PhD, Columbia University Medical Center, 1051 Riverside Drive, Unit 126, New York, NY 10032.
Email: teg2117@cumc.columbia.edu

Abstract

Introduction: Composite scales have been advanced as primary outcomes in early stage Alzheimer's disease trials, and endorsed by the U.S. Food and Drug Administration (FDA) for pivotal trials. They are generally composed of several neurocognitive subscales and may include clinical and functional activity scales.

Methods: We summarized the development of 12 composite scales intended as outcomes for clinical trials and assessed their characteristics.

Results: Composite scales have been constructed from past observational and clinical trial databases by selecting components of individual neuropsychological tests previously used in clinical trials. The atheoretical approaches to combining scales into a composite scale that have often been used risk omitting clinically important measures and so may include redundant, irrelevant, or noncontributory tests. The deliberate combining of neurocognitive scales with functional activity scales provides arbitrary weightings that also may be clinically irrelevant or obscure change in a particular domain. Basic psychometric information is lacking for most of the composites.

Discussion: Although composite scales are desirable for pivotal clinical trials because they, in principle, provide for a single, primary outcome combining neurocognitive and/or functional domains, they have substantial limitations, including their common derivations, inattention to basic psychometric principles, redundancy, absence of alternate forms, and, arguably, the inclusion of functional measures in some. In effect, any currently used composite is undergoing validation through its use in a trial. The assumption that a composite, by its construction alone, is more likely than an individual measure to detect an effect from any particular drug and that the effect is more clinically relevant or valid has not been demonstrated.

KEYWORDS

activities of daily living, Alzheimer's disease, clinical outcomes, clinical trials, cognition, composite outcomes, MCI, mild cognitive impairment, neuropsychological tests, preclinical, prodromal, psychometrics

1 | INTRODUCTION

Recently, the idea of composite scales for primary outcomes for early stage Alzheimer's disease (AD) clinical trials has generated outsized interest. Composites generally include several neurocognitive subscales alone or with functional activity scales that are combined into a single summary score. This approach has been proposed by ad hoc expert panels for preclinical, prodromal, and mild AD trials.[1] This may have been spurred by the concerns of experts in the field and pharmaceutical company sponsors that the established co-primary efficacy outcome criteria for regulatory purposes[2,3] and supported by the U.S. Food and Drug Administration (FDA) were too strict for use in early stage trials where little longitudinal change occurs over the course of the trials.

The FDA offered draft guidance for early stage AD (Guidance for Industry Alzheimer's Disease: Developing Drugs for the Treatment of Early Stage Disease[4,5]) that allowed for the possible use of composite outcomes, stating that they would consider a "composite" outcome, defined as a combined cognitive and functional outcome, as the single primary outcome for efficacy in prodromal AD trials (ie, mild cognitive impairment due to AD), but not in mild AD (usually defined as probable AD[6] with Mini-Mental Status Examination [MMSE] scores from 20 to 26). Moreover, for the concept of preclinical AD (pragmatically defined as little or no cognitive impairment and a positive amyloid biomarker), the FDA would consider a single cognitive outcome (which could be a neuropsychological test or battery of tests) as the sole efficacy criterion for accelerated approval. A more recent FDA non-binding draft document (2018) stages early AD in more detail and is described in Table 1.[7] The European Medicines Agency (EMA) has also suggested use of a composite consisting of cognitive and functional measures in clinical trials of prodromal AD.[8] The EMA did not make further or more-specific recommendations. They did note broadly that other measures of cognition, including executive functions and instrumental activities, could be used as secondary outcomes.

Perhaps as a result of industry perspective, comments from regulators, and the general movement of the field to initiate trials at the earliest identifiable stages of AD, a number of test batteries calling themselves "composite" instruments have been introduced into early stage AD trials. In psychological assessment science, the term composite score has multiple meanings. These include overlap with general intellectual ability, construction of domain-specific sets of tests (eg, a memory composite comprising several different memory tests), specific tests that include a wide variety of individual items from different domains that yield an overall index of ability (eg, MMSE, Alzheimer's Disease Assessment Scale - Cognitive Subscale [ADAS-Cog][11,13,14]) that include such varied measures as word list recall, naming, following commands, constructional praxis, ideational praxis, orientation, word recognition, word finding, test direction set, spoken language, serial subtraction, mazes, number cancellation, and delayed memory, as well as broad multi-domain test batteries that yield an overall summary score. *Here we refer to a composite test as one that assays multiple domains of cognition and function through use of discrete subtests, and then averages the standard score means from these subtests to yield an overall score. We restrict our review to "composites" that meet this criterion.*

Highlights

- Novel aggregations of cognitive and functional measures have been created for use in early stage Alzheimer's disease clinical trials.
- The resulting composites often have redundant tests, lack psychometric data, or may be prone to practice effect confounds, and are generally not validated prior to a trial.
- It is unlikely that they are better than or improve on individual tests or domain-based factor scores.

RESEARCH IN CONTEXT

1. Systematic review: We identified 12 composites based on literature searches in clinical trials.gov and PubMed as of June 2019. We also utilized Vellas et al. (2015),[1] Mortamais et al. (2016),[11] and our own specialized knowledge of the literature to identify composites.
2. Interpretation: Some recent early stage Alzheimer's disease (AD) trials have relied on cognitive or cognitive/functional aggregates of individual tests and subtests. These have been implemented without attention to psychometrics, redundancy of measures, validity, and practice effects. The implications of such measures have not been considered in terms of empirical comparison to individual or domain-specific factors, or cognitive architecture more generally. Use of the measures appears to reflect a "validation by fire" approach.
3. Future directions: We suggest more careful attention to test selection and psychometrics in the construction of composites and testing of composites in several samples prior to use in a pivotal clinical trial.

In addition, relevant to this review are several older measures. The Repeatable Battery for the Assessment of Neuropsychological Syndromes (RBANS) was also designed and co-normed as a single test. In the RBANS, the "total scale" composite index score is derived from five domain scores, each contributing equally to the total score (immediate memory, delayed memory, language, attention, visual spatial/construction). It includes detailed psychometric information, four alternate forms (2012 manual), and is used in the European Prevention of Alzheimer's Dementia Consortium registry.[9]

The Neuropsychological Test Battery (NTB) has been promoted in various forms for use in AD.[10] It originally comprised multiple tests of verbal and visual memory and executive function, used a composite z-score derived from nine individual test measures (Wechsler Memory

TABLE 1 FDA Draft Guidelines (2018) for Staging Early Alzheimer's Disease and Measuring Change

Stage 2: Patients with characteristic pathophysiologic changes of AD and subtle detectable abnormalities on sensitive neuropsychological measures, but no functional impairment. Use of measures suggested for Stage 3 but in a trial of sufficient duration to observe decline, may be a necessity. The emergence of subtle functional impairment signals a transition to Stage 3.
Stage 3: Patients with characteristic pathophysiologic changes of AD, subtle or more apparent detectable abnormalities on sensitive neuropsychological measures, and mild but detectable functional impairment. The functional impairment in this stage is not severe enough to warrant a diagnosis of overt dementia. The FDA emphasizes the sensitivity of both cognitive and functional measures, or an integrated instrument, for this stage.
Stage 4: Patients with overt dementia. This diagnosis is made as functional impairment worsens from that seen in Stage 3. Of relevance to this manuscript are Stages 2 and 3 with the FDA focus on sensitive cognitive and functional instruments. As will be seen below, this is the motivation for composite instruments.

Scale visual immediate, Wechsler Memory Scale verbal immediate, Rey Auditory Verbal Learning Test [RAVLT] immediate, Wechsler Memory Digit Span, Controlled Word Association Test, Category Fluency Test, Wechsler Memory Scale visual delayed, Wechsler Memory Scale verbal delayed, and RAVLT delayed—comprising summed delayed recall and recognition performance components[10]). It was validated in data from a single AD clinical trial (N = 372 subjects). It had high test-retest reliability and an interpretable factor structure (memory/executive). The tests themselves were established and existing clinical neuropsychological measures. It should be noted that there are multiple versions of the NTB that differ from each other in test selection, apparently on an ad hoc basis.

Another level to the composite definition has been added by the FDA. As above, the FDA's draft guidelines consider tests that combine measures of cognitive function and everyday function as composites for use in prodromal AD. Unfortunately, they provide only one example for such a composite, namely the Clinical Dementia Rating Scale Sum of Boxes (CDR-SB) that can be used for regulatory purposes as a single primary efficacy measure for pivotal clinical trials in patients with prodromal AD. The FDA has been silent on the use of composites for preclinical AD except to state that they would consider allowing a cognitive assessment alone (ie, "isolated cognitive measure") as the efficacy test for accelerated marketing approval of a drug for preclinical AD.

In succeeding sections we discuss some of the issues associated with this approach, including the conceptual and scientific basis of composites, their reported psychometric properties (if any), and what exactly a composite may mean in terms of its implications for understanding a treatment response.

2 | MATERIALS AND METHODS (SEARCH STRATEGY)

We identified 12 newer composites using literature searches in clinical trials.gov and PubMed as of September 1, 2017: composite AND preclinical AD; composite AND MCI; composite AND prodromal AD. In clinicaltrials.gov we generated a list of possibly relevant studies by a search conditioned by the following term: Alzheimer's and outcome and composite. We also utilized Vellas et al. (2015),[1] Mortamais et al. (2016),[11] and our own specialized knowledge of the literature to identify composites.

We operationally defined composites as instruments that combined several clinical tests of multiple domains and a derived single outcome score (eg, z-score), and wherein the instrument was proposed for or used in a clinical trial for preclinical, prodromal, or mild AD populations. We do not discuss efforts to improve sensitivity of individual scales through item-response theory or Rasch statistics, and we do not comment on composites based on a single cognitive domain.

3 | RESULTS

3.1 | Overview of trials using composites in clinicaltrials.gov

Of the 64 trials our searches yielded and listed in Supplement 1, a total of 23 were not relevant because they did not use composites of cognitive measures, did not pertain to the appropriate diagnostic group, or did not include an identifiable composite of any sort (Items 1, 13, 22, 26, 27, 28, 29, 30, 32, 42, 46, 48, 52, 53, 54, 56, 57, 58, 59, 60, 62, 63, and 64). Of the remaining trials, three used the Alzheimer's Prevention Initiative (API) composite, two the RBANS, two the Preclinical Alzheimer Cognitive Composite (PACC), one the Dominantly Inherited Alzheimer Network (DIAN), and one the Alzheimer's Disease Composite Score (ADCOMS) (see previous and subsequent text for detailed descriptions of these composites). The majority of trials did not list the tests in the composite being used as an outcome. Multiple other trials (Items 6, 7, 8, 9, 16, 19, 31, 34, 39, 41, and 63) used a "composite" of a single domain (eg, a memory composite, an executive composite). Although these did not conform to our definition of a composite, their relatively wide use may be of interest to the reader.

3.2 | Characterization of individual composite measures: descriptions and comments

We selected 11 composites for discussion based on published work that included descriptions of the measures being used and at least some information on selection criteria for individual tests and psychometric properties and validation procedures of the composite. These descriptions of select composites may have utility for the reader in understanding generally potential strengths and limitations as the field moves forward in implementing these tests in trials. The final list of

TABLE 2 List of older and more recent composites and their test construction parameters

Test	Derivation	Orientation	CDR	MMSE	Verbal list Mem	Story Mem	Coding (Digit Sym)	Psychometric description	Alter. form	Function
RBANS[12]	Rational				✓	✓	✓	✓	✓	
NTB[10,11]	Rational				✓			✓		
PACC[13]	Sensitivity	✓		✓	✓	✓	✓			
IADRS[14]	Rational	✓			✓					✓
ADCOMS[15]	Sensitivity	✓	✓		✓					✓
TOMM[16]	Rational				✓					
DIAN TU[17]	Rational	✓		✓	✓	✓	✓			
API APOE[18]	Sensitivity	✓			✓	✓	✓			
API ADAD[19]	Sensitivity	✓			✓					
ZAVEN[20]	Rational				✓	✓	✓	✓		
CCS 3D[21]	Rational	✓			✓		✓			
AIBL[22]	Sensitivity	✓	✓	✓						
GuidAge[23]	Rational	✓			✓				✓	
MAPT[24]	Rational	✓			✓		✓		✓	

RBANS[12], Repeatable Battery for the Assessment of Neuropsychological Syndromes; NTB[10,11], Neuropsychological Test Battery; PACC[13], Preclinical Alzheimer Cognitive Composite; IADRS[14], Integrated Alzheimer's Disease Rating Scale; ADCOMS[15], Alzheimer's Disease Composite Score; TOMM[16], The TOMMORROW Study; DIAN TU[17], Dominantly Inherited Alzheimer Network Trial; API APOE[18], Alzheimer's Prevention Initiative Apolipoprotein E; API ADAD[19], Alzheimer's Prevention Initiative Autosomal Dominant Alzheimer's Disease; ZAVEN[20], Z-scores of Attention, Verbal fluency, and Episodic memory for Nondemented older adults composite; CCS 3D[21], Composite Cognition Score-3 Domain; AIBL[22], Australian Imaging Biomarkers and Lifestyle; GuidAge[23], Long-term use of standardized Ginkgo biloba extract for the prevention of Alzheimer's disease; MAPT[24], Multidomain Alzheimer Prevention Trial.

relevant articles is in Table 2,[11–25] where we describe characteristics of the composites including methods of their derivation, source, tests included, and availability of psychometric information.

3.3 | Critical commentary about the composites

3.3.1 | API[19]

The API composite used an effect size type change score of decline in PS1 mutation carriers (N = 78) in a large Colombia kindred who were unimpaired at baseline. Various combinations of sensitivity of individual variables to decline were examined, although it is unclear how this was done. Those combinations that were robust across the 2-year and 5-year follow-up periods were included in the final composite, and a weighting procedure was then employed. The set of tests selected was the following: MMSE orientation, CERAD naming, CERAD word recall, CERAD constructional praxis, and Ravens Progressive Matrices. No information about psychometrics (eg, test-retest reliability, normality of distribution, ceiling, and floor effects), correlations with global change measures, or function was provided. The composite does not have alternate forms. No cross-validation was conducted.

3.3.2 | API[18]

A composite was constructed based on longitudinal data from nearly 1100 individuals in various Rush studies including Religious Orders, Aging, and Minority. Multiple cognitive tests were examined for sensitivity to decline using a mean change to SD (of the change score) ratio, and various combinations were constructed by systematic computerized search to identify the best "composite" in identifying progression from normal cognition to clinically diagnosed AD. The final selection was based not only on combinations that performed well; they were then evaluated for representation of relevant cognitive domains, robustness across individual years prior to diagnosis, and occurrence of selected items within top performing combinations. The optimal composite cognitive test score (unweighted) comprised of seven cognitive tests/sub-tests (MMSE time orientation, Ravens, category fluency, symbol digit coding, naming, naming recall, and logical memory delayed) that combined showed a standardized change over2- and 5-year periods, defined as mean/SD of the change (MSDR) = 0.96. By comparison, the most sensitive individual test score after covariate adjustment was Logical Memory – Delayed Recall, MSDR = 0.64, although unadjusted category fluency had an MSDR of .83. Weighting did not improve sensitivity. No other psychometric data were reported. (Of note, a different composite was alluded to on the clinicaltrials.gov website for the API APOE4 trial; it consisted of the RBANS, MMSE, and Ravens with subtests unspecified).

3.3.3 | ADCOMS[15]

Selection of tests was based on an effect-size type measure over the course of an amnestic MCI trial. After this initial phase the variables were used to predict progression, that is, the length of the trial in a

partial least-squares regression. The outcome is arguably appropriate but results in some restriction of range. Use of the regression is unusual because it is generally applied in situations when multiple dependent measures are examined. In addition, it does not eliminate redundant predictor variables as might a stepwise linear regression. Partial least-squares coefficients served as weights in the validation sample for the following variables: ADAS-Cog memory, word finding, and orientation; MMSE orientation and drawing; and CDR-SB. Thus, this composite focuses on orientation and memory. It is unclear how differentially weighting the three nearly identical orientation measures affects scoring. The discovery sample consisted of 963 subjects from multiple amnestic MCI studies including Alzheimer's Disease Neuroimaging Initiative (ADNI) (N = 405), as well as selected Pfizer/Eisai trials. The validation groups consisted of a mild AD placebo sample from Alzheimer Disease Cooperative Study (ADCS) (N = 264) and a donepezil treatment group of N = 469 (Pfizer trial). Two subgroups within the discovery sample (APOE4 carriers and cerebrospinal fluid [CSF] amyloid beta [Aβ]–positive individuals) served as (pseudo-) validation groups. No information about psychometrics (test-retest reliability, normality of distribution, ceiling, and floor effects), correlations with global change measures, or function was provided. The composite does not appear to have alternate forms. A partial least-squares approach was used to find subsets of tests to group together and to then quantify sensitivity by seeking the largest standardized change (ie, MSDR). It is unclear whether other approaches beyond partial least squares might have been explored. New results from a phase 2 trial of BAN2401 show that it differed only trivially in sensitivity to decline compared to the CDR-SB or the ADAS-Cog.[25]

3.3.4 | PACC A4[13]

The PACC was based on rational determination that sensitivity to change in preclinical AD would involve episodic memory, executive function, and orientation. Data were derived from three samples: ADCS, ADNI, and Australian Imaging Biomarkers and Lifestyle (AIBL). The ADCS PACC (N = 505 healthy controls) included free and cued selective reminding, NYU paragraph delayed recall, digit symbol substitution, and MMSE (total). The AIBL PACC used a different list learning test (CVLT). ADNI (N = 97) required use of a delayed recall from ADAS-Cog. No psychometric data were reported but alternate forms were available for some of the tests (namely verbal memory). No information was provided on these alternate forms as to equivalence. The composite was validated in the cohorts that included amyloid positive/negative subjects (from ADNI and AIBL). Decline was steeper in the positive group (N = 87) than in the negative group (N = 274). Weighting was tested but did not improve fit in the test sample. Since reports on the PACC were published there have been several criticisms as well as a replication in the Harvard Aging study.[26] The first has to do with the memory measures used. These have different semantic encoding demands and use either delay or total recall. The second has to do with whether measures should be weighted unequally or weighted equally after z-scoring. Kryscio[27] in examining the PACC also noted several other problematic features. First, in the discovery samples, cohorts differed in the exact test used, so substitutions were made, making it difficult to fully establish validity. Second, practice effects might change slopes differentially between or simply add noise to placebo and treatment. Thus, the PACC[13] that is being applied in the A4 trial is based on expert opinion and clinical judgment followed by test data validation, as opposed to the training-data-driven assessments and model building that some others have applied; and substitutes the Wechsler Memory Scale-Revised Logical Memory – Delayed Recall (WMS-R LM II) for the NYU paragraph delayed recall.

3.3.5 | DIAN TU[17]

The measures included in this composite were the International Shopping List, WMS-R LM II, digit symbol coding, and MMSE (complete). They were chosen based on reduced ceiling and floor effects, low variability, and "sensitivity to subtle decline" before clinical diagnosis in mutation carriers. No data have been provided as yet on details about psychometrics, cross validation, and so on.

3.3.6 | AIBL prodromal MCI[22]

Burnham studied the rate of decline in 37 MCI subjects who were amyloid positive in Australian Imaging Biomarkers and Lifestyle (AIBL). After examining change scores in 27 measures, two measures were combined for maximum effect: CDR-SB and MMSE (complete). These were z-scored and weighted equally. Other than demonstrating that the composite measure was sensitive to decline (in the discovery sample) it was shown that it could reduce sample sizes in a clinical trial through increased power. No other psychometrics were provided.

3.3.7 | GuidAge[23] Composite

The Long-term use of standardized Ginkgo biloba extract for the prevention of Alzheimer's disease (GuidAge) composite was developed from 1414 older subjects in a placebo group over 5 years. It included trails, free and cued selective reminding (alternate forms), and the MMSE orientation item. These were weighted equally. It was unclear how they were chosen. The composite predicted change on the CDR from 0 to 0.5 after 2 or 3 years and progression to AD dementia over 5 years. No other psychometrics were reported.

3.3.8 | MAPT[24] Composite

A cognitive composite was used in the Multidomain Alzheimer Prevention Trial (MAPT) trial on the effects of omega-3 fatty acids and "multidomain" intervention on cognitive decline in 1680 non-demented older

subjects. It was similar to the GuidAge composite.[23] The composite consisted of free and cued selective reminding (with two alternate forms), MMSE orientation, symbol digit coding, and semantic fluency. It was deemed valid because it was sensitive to *APOE* status and amyloid, although no results were shown. No psychometric information was provided. Inspection of a results table indicated that tests of speed and orientation may have been marginally more sensitive to treatment than the composite as a whole.

3.3.9 | ZAVEN[20]

The Z-scores of Attention, Verbal fluency, and Episodic memory for Nondemented older adults composite (ZAVEN) was derived from a group of older controls in AIBL (N = 423) using digit symbol, letter fluency (considered as a test of executive function, though speed of processing may be more accurate), CVLT total recall, and logical memory delayed recall followed up to 6 years. The composite demonstrated faster decline in amyloid-positive subjects. It also performed better than the PACC. The ZAVEN demonstrated high test-retest reliability, but no other psychometric information was provided. No alternate forms were constructed, nor was a cross-validation group utilized.

3.3.10 | CCS-3D Merck[21]

The Composite Cognition Score-3 Domain (CCS-3D) is calculated as the mean of three domain z-scores (episodic memory, executive function, and attention processing). Each of these three-domain z-scores is calculated as the mean of domain-specific tests (after transformation to z-scores), as follows:

Episodic Memory: Immediate Word Recall, Delayed Word Recall, Word Recognition, and Orientation (all from ADAS-Cog); Executive Function: Digit Span Test (Backwards), Trails B Test, COWAT letter fluency, and CERAD Verbal Fluency Test; Attention/Processing Speed: Trails A Test, Digit Span Test (Forwards), and digit symbol coding. It was "designed to cover aspects of cognition not well assessed by the ADAS-Cog," but nevertheless one domain consists entirely of ADAS-Cog measures. No other information was provided about psychometric characteristics, alternate forms, or sensitivity.

3.3.11 | TOMMORROW[16]

The following tests were chosen based on "experience in international studies" of MCI and AD: memory (CVLT-2, Brief Visual Memory Test), executive function (Trails B, digit span backwards), attention (digit span forward, Trails A), visual spatial function (Clock, Brief Visual Memory), language (naming, semantic, and letter fluency). No published information is available on psychometric characteristics of the composite. The individual memory scales were used to assess the onset of MCI in a clinical trial.

3.3.12 | iADRS[14]

Outcomes from four studies (including the late and early MCI cohorts of ADNI, and the mild and moderate AD patients included in the two Expedition trials of solanezumab sponsored by Eli Lilly) were examined to determine the largest standardized change over 80 weeks to 36 months, depending on the sample, for multiple cognitive and functional measures. ADAS-Cog 14 and ADCS ADL were selected. These were weighted unequally based on the number of items (90 for cognition and 56 for function); this results in cognition having primacy. Psychometric analyses were confined to a principal components analysis, which unsurprisingly identified two factors, one cognitive and one functional. The resulting measure was sensitive to tracking disease progression and donepezil benefits in an ADCS MCI study. However, it was unclear if the integrated composite performed better than the ADAS-Cog alone.

3.4 | Recent and older composites: summary

Two older (NTB and RBANS) and 12 newer scales can be compared (Table 2).[10–24] Only the two older composites (NTB and RBANS) and one of the new composites (ZAVEN) included psychometric information. Two included measures of everyday function. The majority of the new scales do not have alternate forms; alternate forms are important for reducing practice effects. Those that utilize alternate forms, do not report on their equivalence or lack thereof. Nearly all composites included a test of verbal list learning, although the exact type differed (eg, Selective Reminding, CVLT, AVLT, Shopping List). WMS Logical Memory for stories was the most widely used episodic memory test. Most scales used measures of both list learning and memory for stories. This is a reasonable approach as it has been shown that memory for stories may be as specific, but less sensitive than, list learning in identifying MCI subjects[28] and their correlation is only moderate (r = .40, ADNI data, Goldberg unpublished data). In addition, speed of processing was assessed with digit symbol coding in many of the newer scales. Nearly all composites used an orientation scale/item that was obtained from an existing scale, namely, the CDR, ADAS-Cog, or MMSE. (Indeed, many composites required the extraction of items from test batteries that were often administered by two or more clinicians.) To a surprising degree, many the selection criteria for individuals tests of many composites appear to rely on a process similar to ratiocination. By this we mean, deliberate, rational reasoning or argument, perhaps based on old judgment, but uninformed by data. Very few of the composites used an independent validation sample after their initial derivation in a "discovery" sample.

4 | DISCUSSION

As can be seen by the publication dates in Table 2,[10–24] composites appear to be in a proliferative phase. It remains unclear, however,

if composites will perform better than assays of individual cognitive domains (eg, episodic memory, orientation, executive function). It could also be said that the CDR, ADAS-Cog, and Mini-Mental State used in some of the composites above, were themselves constructed as composites. That is, they include items from multiple cognitive domains (eg, memory, orientation, language, visual-motor function) combined into a single test and score. Although they do not meet our criteria of a composite (combining separate tests statistically to yield a single performance value), they function as composites. In a sense this a "back to the future" moment for the field, as they appear to be reprises of mental status examinations.

Despite the seeming enthusiasm for these instruments, several lines of potential criticism also need to be considered. First, in batteries based on sensitivity, use of mean change/SD change followed by subjective selection of tests based on domains and using partial least-squares regression, the redundancy of the measures are not taken into account. Rather, surprisingly, none of the composite development strategies used models such as stepwise regression, a statistical procedure that attempts to identify largely independent predictors (ie, non-overlapping and hence not prone to collinearity). For instance, the ADCOMS has three measures of orientation that are combined with other tests for its total score. In our own work we found that orientation, as tested for example, in mental status examinations, the ADAS-Cog, and the CDR, was a surrogate memory measure that was highly correlated with neurocognitive measures of episodic memory.[24] We also found it associated with the degree of integrity of temporal and medial temporal lobe regions.[29] It is unclear, however, if orientation is "better" or different from other memory measures. Critically, we also demonstrated that from a psychometric perspective, orientation is particularly ill-suited as a measure in healthy or relatively healthy controls who are similar to subjects in preclinical samples because of extreme ceiling effects. Thus, in healthy controls approximately 80% of subjects score at ceiling and another 17% score 9/10 on the MMSE orientation items. Even in amnestic MCI, approximately 42% of subjects score at ceiling.

Second, several of the sensitivity-derived composites utilize weights. Although weights certainly increase "fit" within the discovery group and in some cases within the validation sample, they might not increase sensitivity in other cohorts. Weights might also change naturalistically with increasing disease severity. The cohorts used in these cases were not clinical trials cohorts and perhaps are not likely to be those in future trials. Third, and potentially impacting on generalizability, several of the composites were derived from the same databases, namely ADNI and AIBL, and also were not clinical trials samples.

Most reports on composite scales have not been overly disciplined in presentation of important psychometric data. The majority of the newer composites did not include alternate forms to reduce practice effects. This is somewhat surprising in that it is now clearly established that even older healthy subjects can generate significant practice effects of an effect size of Cohens d = 0.25 over two to three assessments.[30] Remarkably, this is an effect size observed with marketed cholinesterase inhibitors and planned for in most recent disease-modifying trials. As we discussed elsewhere, such practice effects add to the variance, reduce power, and may prevent or otherwise mask cognitive scores from aligning with neurodegenerative or other biomarker changes. It is also striking that even basic psychometric information such as test-retest reliability and ceiling and floor effects was often not reported for the newer instruments.

Furthermore, in our view, what is said by various authors about test-domain relations is sometimes mischaracterized. Thus, digit symbol coding is frequently considered to be a test of speed of processing based on factor analytic loadings, not an executive test, or as in one composite, a test of visual spatial ability. Similarly, letter fluency has been considered to be an executive test in one composite and a language measure in another. Thus, claims that one test or another reflect executive function, or speed, or language (eg, trails, digit symbol, and fluency) should be backed by a factor analysis or structural equation modeling, not by simple assertion.

A more theory-driven interpretation of composites may rest on perspectives on the structure of cognition, that is, how different elements of cognition are integrated or modularized to proceed with intelligent thought, focused attention, and goal-directed action. Composites imply that multiple tests can be meaningfully aggregated to form an overall score of cognition. Indeed, over the past 75 years, factor analytic studies of groups of cognitive tests have generally yielded a robust general intelligence factor (sometimes called "g") with a high eigenvalue that accounts for much of the common variance shared by seemingly disparate tests. In other words, this result suggests that all tests are correlated with each other to some degree. As such, a composite measure is in a position to capture basic cognitive architecture as well as damage to that architecture, for example, in the case of neurodegeneration, which presumably would result in reduction of g. In this view even the Wechsler IQ test is a composite. It assesses multiple ability areas utilizing various subtests (eg, language, visual perception, speed, working memory) that when averaged yield a Full Scale IQ score. Thus, for proponents of the view that there is a general or g factor that explains much of the variance in intelligence, composites are a rational way to measure this construct.

A second reason a composite may be advantageous is empirical. Several new composites are based on those tests that demonstrate the steepest decline (whether due to psychometric sensitivity such as difficulty level or sensitivity to neurobiological change). That is, they are assumed to be most sensitive over time to AD progression. Third, composites, in principle, can hurdle psychometric deficiencies of an individual test. They might reduce the ceiling and floor effects of individual tests. They may also improve test-retest reliability insofar as they stabilize variance by including a larger item pool. Finally, they may reduce statistical problems of multiplicity.

There are several lines of argument against this view, however. Some factor analytic and structural equation modeling studies suggest that human cognitive architecture comprises multiple factors or domains that often consist of spatial, language, memory, working memory, and speed factors that are relatively dissociable from one another. At the neural systems level of explanation, identification of specialized brain regions that support different abilities (eg, Broca's and Wernicke's regions for speech and language comprehension; the so-called

fusiform face area, medial temporal lobe for episodic memory processing, prefrontal cortex in cognitive control and executive function) have also been identified. Thus, any metric that combines different cognitive domains might dilute a specific impairment or a treatment response to said cognitive impairment. To the extent that AD initially and preferentially manifests itself in the medial temporal lobe episodic memory system, and to the extent that memory is the most robust predictor of MCI to AD progression,[31,32] combining memory tests with tests from other domains might dilute sensitivity.

With respect to memory, we appreciate that impairment in this domain is a diagnostic criterion for amnestic and multi-domain MCI, subtypes most likely to progress to AD, and to AD itself in older diagnostic systems (eg, the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition [DSM-IV]). Furthermore, post mortem histopathology and amyloid and tau PET ligand studies have frequently indicated substantial abnormalities in temporal lobe regions in MCI and preclinical AD. Similarly, morphometric magnetic resonance (MR) studies in MCI and *APOE* ε4 carriers in mild AD have indicated prominent atrophy in the medial temporal regions that support memory. Moreover, because memory failures can be associated directly with loss of functional competence[33] and are also associated with symptoms that burden the caregiver, such as repetitive questioning,[34] one might reasonably focus on this domain. In addition, a treatment, be it drug or cognitive training, might target a specific cognitive domain. Indeed, for most experimental drugs for AD there is no logical reason that they specifically affect those tests in a composite that show the steepest decline with AD. These criticisms again point to the possibility that a composite might dilute a valid or "real" treatment effect.

It is also unclear what advantage a combined cognitive and function measure would have compared to separate cognitive and function co-primaries. Thus, the majority of composites do not include a measure of function, although the FDA and European Medicines Agency (EMA) have provided non-binding guidance on this.[4] Perhaps the cognitive and function tests could be weighted in an optimized manner, but this has not been examined explicitly.

Finally, we note that in the majority of composites that we described, memory and orientation comprised a disproportionate number of items/tests. Although it could be argued that multiple memory measures are necessary, given the importance of this domain in diagnosis and the fact that correlations between verbal list learning and memory for stories is rather low (r = .40, T. Goldberg, unpublished, ADNI data) weighing orientation so heavily is more problematic.

Of interest, the FDA has provided additional information on use of composites. In the context of a composite based on subscales of an agitation inventory, they suggested that test construction should include psychometric analyses (eg, reliability), construct validity, ability to detect change, and use of anchor scales to define responders. (They also noted that the studies done to obtain these data should be conducted prior to confirmatory trials.) In our view, these criteria could just as easily be applied to cognitive composites.

Beyond "validation by fire" for composites within registered, large, phase 2 and 3 pivotal clinical trials that are personnel intensive, time-consuming, and particularly expensive, there are better ways to gain an understanding of how a given composite is likely to behave in a trial. Elsewhere we suggested use of a clinical trial armature that involves randomizing subjects to test-type (new composite vs established measure) followed by serial assessments in subjects that are similar to AD spectrum trial patients. No treatment would be provided. Psychometric properties could easily be deduced and contrasted with older measures without confounding effects of interference between measures.

In sum, this review addresses the increasing and somewhat uncritical use of composite cognitive/functional measures in early stage AD trials, and highlights the psychometric properties of these measures or the lack of reported properties. We also note the implicit assumptions of the composites from the standpoint of cognitive architecture; and that these may not fit the realities of the clinical phenotypes or neurobiology of AD. Finally, we note the "back to the future" moment wherein some composites rather resemble a reassembly of mental status examinations focusing on combining brief assessments of orientation and memory, and with the similar strengths and limitations of such bedside examinations.

CONFLICT OF INTEREST

L.S.S. reports grants and personal fees from Eli Lilly, personal fees from Avraham, Ltd, personal fees from Boehringer Ingelheim, grants and personal fees from Merck, personal fees from Neurim, Ltd, personal fees from Neuronix, Ltd, personal fees from Cognition, personal fees from Eisai, personal fees from Takeda, personal fees from vTv, grants and personal fees from Roche/Genentech, grants from Biogen, grants from Novartis, personal fees from Abbott, outside the submitted work. T.E.G. reports royalties from VeraSci for use of the BACS cognitive screening instrument in clinical trials.

REFERENCES

1. Vellas B, Bateman R, Blennow K, et al. Endpoints for pre-dementia AD trials: a report from the EU/US/CTAD Task Force. *J Prev Alzheimers Dis*. 2015;2(2):128-135.
2. Leber P. Guidelines for the clinical evaluation of anti-dementia drugs, 1st draft. Rockville, MD: United States Food and Drug Administration 1990. (Unpublished)
3. Leber P. Criteria used by Drug Regulatory Authorities. In: Qizilbash N, Schneider L, Chui H, et al., eds. Evidence-based Dementia Practice. Oxford: Blackwell Science Ltd; 2002:376-387.
4. FDA (2013). United States Food and Drug Administration. Guidance for industry Alzheimer's disease: Developing drugs for the treatment of early stage disease (FDA-2013-D-0077) DRAFT (US Food and Drug Administration, Center for Drug Evaluation and Research).
5. Kozauer N, Katz R. Regulatory innovation and drug development for early-stage Alzheimer's disease. *N Engl J Med*. 2013;368:1169-1171.
6. McKhann GM, Knopman DS, Chertkow H, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*. 2011;7(3):263-269.
7. FDA. Early Alzheimer's disease: Developing drugs for treatment guidance for industry. Available at: https://www.fda.gov/…fda-guidance…/alzheimers-disease-developing-drugs-treatment. Published February 2018.

8. European Medicines Agency. Draft guideline on the clinical investigation of medicines for the treatment of Alzheimer's disease and other dementias. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2016/02/WC500200830.pdf. Accessed April 21, 2016.
9. Ritchie K, Ropacki M, Albala B, et al. Recommended cognitive outcomes in preclinical Alzheimer's disease: consensus statement from the European Prevention of Alzheimer's Dementia project. *Alzheimers Dement*. 2017;13(2):186-195.
10. Harrison J, Rentz DM, McLaughlin T, et al, ELN-AIP-901 Study Investigator Group. Cognition in MCI and Alzheimer's disease: baseline data from a longitudinal study of the NTB. *Clin Neuropsychol*. 2014;28(2):252-268.
11. Mortamais M, Ash JA, Harrison J, et al. Detecting cognitive changes in preclinical Alzheimer's disease: a review of its feasibility. *Alzheimers Dement*. 2016;13(4):468-492.
12. Randolph C, Tierney MC, Mohr E, Chase TN. The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): preliminary clinical validity. *J Clin Exp Neuropsychol*. 1998;20(3):310-319.
13. Donohue MC, Sperling RA, Salmon DP, et al. The preclinical Alzheimer cognitive composite: measuring amyloid-related decline. *JAMA Neurol*. 2014;71(8):961-970.
14. Wessels AM, Siemers ER, Yu P, et al. A combined measure of cognition and function for clinical trials: the Integrated Alzheimer's Disease Rating Scale (iADRS). *J Prev Alzheimers Dis*. 2015;2(4): 227-241.
15. Wang J, Logovinsky V, Hendrix SB, et al. ADCOMS: a composite clinical outcome for prodromal Alzheimer's disease trials. *J Neurol Neurosurg Psychiatry*. 2016;87(9):993-999.
16. Romero HR, Monsch AU, Hayden KM, et al. TOMMORROW neuropsychological battery: German language validation and normative study. *Alzheimers Dement*. 2018;4:314-323.
17. Bateman RJ, Benzinger TL, Berry S, et al. The DIAN-TU Next Generation Alzheimer's prevention trial: adaptive design and disease progression model. *Alzheimers Dement*. 2017;13(1):8-19.
18. Langbaum JBS, Hendrix SB, Ayutyanont N, et al. An empirically derived composite cognitive test score with improved power to track and evaluate treatments for preclinical Alzheimer's disease. *Alzheimers Dement*. 2014;10(6):666-674.
19. Ayutyanont N, Langbaum JB, Hendrix SB, et al. The Alzheimer's Prevention Initiative composite cognitive test score: sample size estimates for the evaluation of preclinical Alzheimer's disease treatments in presenilin 1 E280A mutation carriers. *J Clin Psychiatry*. 2014;75(6): 652-660.
20. Lim YY, Snyder PJ, Pietrzak RH, et al. Sensitivity of composite scores to amyloid burden in preclinical Alzheimer's disease: introducing the Z-scores of attention, verbal fluency, and episodic memory for nondemented older adults composite score. *Alzheimers Dement*. 2016;2:19-26.
21. Voss T, Li J, Cummings J, et al. Randomized, controlled, proof-of-concept trial of MK-7622 in Alzheimer's disease. *Alzheimers Dement*. 2018;4:173-181.
22. Burnham SC, Raghavan N, Wilson W, et al. Novel statistically-derived composite measures for assessing the efficacy of disease-modifying therapies in prodromal Alzheimer's disease trials: an AIBL study. *J Alzheimers Dis*. 2015;46(4):1079-1089.
23. Coley N, Gallini A, Ousset PJ, et al. Evaluating the clinical relevance of a cognitive composite outcome measure: an analysis of 1414 participants from the 5-year GuidAge Alzheimer's prevention trial. *Alzheimers Dement*. 2016;12(12):1216-1225.
24. Andrieu S, Guyonnet S, Coley N, et al. MAPT Study Group. Effect of long-term omega 3 polyunsaturated fatty acid supplementation with or without multidomain intervention on cognitive function in elderly adults with memory complaints (MAPT): a randomised, placebo-controlled trial. *Lancet Neurol*. 2017;16(5):377-389.
25. Satlin A, Wang J, Logovinsky V, et al. Design of a Bayesian adaptive phase 2 proof-of-concept trial for BAN2401, a putative disease-modifying monoclonal antibody for the treatment of Alzheimer's disease. *Alzheimers Dement*. 2016;2:1-12.
26. Buckley RF, Mormino EC, Amariglio RE, et al. Sex, amyloid, and APOE $\varepsilon 4$ and risk of cognitive decline in preclinical Alzheimer's disease: Findings from three well-characterized cohorts. *Alzheimers Dement*. 2018;14(9):1193-1203.
27. Kryscio RJ. Secondary prevention trials in Alzheimer's disease: the challenge of identifying a meaningful endpoint. *JAMA Neurol*. 2014;71(8):947-949.
28. Weissberger GH, Strong JV, Stefanidis KB, Summers MJ, Bondi MW, Stricker NH. Diagnostic accuracy of memory measures in Alzheimer's dementia and mild cognitive impairment: a systematic review and meta-analysis. *Neuropsychol Rev*. 2017;27(4):354-388.
29. Sousa A, Gomar JJ, Goldberg TE, for ADNI. Neural and behavioral substrates of disorientation in mild cognitive impairment and Alzheimer's disease. *Alzheimers Dement*. 2015;1(1):37-45. https://doi.org/10.1016/j.trci.2015.04.002.
30. Goldberg TE, Harvey PD, Wesnes KA, Snyder PJ, Schneider LS. Practice effects due to serial cognitive assessment: implications for preclinical Alzheimer's disease randomized controlled trials. *Alzheimers Dement*. 2015;1(1):103-111.
31. Gomar JJ, Bobes-Bascaran MT, Conejero-Goldberg C, et al. Utility of combinations of biomarkers, cognitive markers, and risk factors to predict conversion from mild cognitive impairment to Alzheimer disease in patients in the Alzheimer's disease neuroimaging initiative. *Arch Gen Psychiatry*. 2011;68(9):961-969.
32. Schmand B, Eikelenboom P, van Gool WA; Alzheimer's Disease Neuroimaging Initiative. Value of neuropsychological tests, neuroimaging, and biomarkers for diagnosing Alzheimer's disease in younger and older age cohorts. *J Am Geriatr Soc*. 2011;59(9):1705-1710.
33. Goldberg TE, Koppel J, Keehlisen L, et al. Performance-based measures of everyday function in mild cognitive impairment. *Am J Psychiatry*. 2010;167(7):845-853.
34. Rockwood K, Fay S, Jarrett P, Asp E. Effect of galantamine on verbal repetition in AD: a secondary analysis of the VISTA trial. *Neurology*. 2007;68(14):1116-1121.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Schneider LS, Goldberg TE. Composite cognitive and functional measures for early stage Alzheimer's disease trials. *Alzheimer's Dement*. 2020;12:e12017. https://doi.org/10.1002/dad2.12017