ELSEVIER

Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 1 (2015) 103–111

Alzheimer's & Dementia

Practice effects due to serial cognitive assessment: Implications for preclinical Alzheimer's disease randomized controlled trials

Terry E. Goldberg[a,*], Philip D. Harvey[b,c], Keith A. Wesnes[d], Peter J. Snyder[e], Lon S. Schneider[f]

[a]*Litwin Zucker Center for the Study of Alzheimer's Disease, Feinstein Institute, Hofstra North Shore LIJ School of Medicine, Manhasset, NY, USA*
[b]*Department of Psychiatry and Behavioral Sciences, University of Miami Miller School of Medicine, Miami, FL, USA*
[c]*Research Service, Miami VA Healthcare System, Miami, FL, USA*
[d]*Wesnes Cognition, Streatley on Thames, UK*
[e]*Department of Neurology, Alpert Medical School of Brown University & Rhode Island Hospital, Providence, RI, USA*
[f]*Departments of Psychiatry, Neurology, and Gerontology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA*

Abstract

Introduction: Practice effects are characteristic of nearly all standard cognitive tasks when repeated during serial assessments and are frequently important confounders in clinical trials.

Methods: We summarize evidence that gains in neuropsychological test performance scores associated with practice effects occur as artifactual changes associated with serial testing within clinical trials. We identify and emphasize such gains in older, non–cognitively impaired individuals and estimate an effect size of 0.25 for composite cognitive measures in older populations assessed three times in a 6- to 12-month period.

Results: We identified three complementary approaches that can be used to attenuate practice effects: (1) massed practice in a prebaseline period to reduce task familiarity effects; (2) tests designed to reduce practice-related gains so that item-specific driven improvements are minimized by using tasks that minimize strategy and/or maximize interitem interference; and (3) well-matched alternate forms.

Discussion: We have drawn attention to and increased awareness of practice effect–related gains that could result in type 1 or type 2 errors in trials. Successfully managing practice effects will eliminate a large source of error and reduce the likelihood of misinterpretation of clinical trials outcomes.

Keywords: Alzheimer's disease; Practice effects; Cognition; Clinical trials; Serial assessment; Preclinical Alzheimer's disease; Neuropsychology

1. Introduction

Practice effects are characteristic of serial neurocognitive assessments, including those used in clinical trials. They refer to changes in test performance attributed to increasing familiarity with and exposure to test instruments, paradigms, and items. Nevertheless, these effects are often underappreciated. Our own work in this area [1–3] has identified them as important in the interpretation of both outcomes in clinical trials and in longitudinal studies of patients with schizophrenia. Here, we discuss the relevance of these findings to clinical trials for Alzheimer's disease (AD) and mild cognitive impairment (MCI), a stage often thought to be transitional between cognitive health and AD, and, notably, preclinical AD [4]. Preclinical AD at stages 1 and 2 refers to those individuals who have cerebrospinal fluid or positron emission tomography evidence of amyloid-β abnormalities and/or "downstream" neurodegeneration but do not demonstrate cognitive changes; at stage 3, individuals additionally suffer from subtle cognitive changes. For preclinical AD, the assessment of cognition has been suggested

*Corresponding author. Tel.: +1-516-562-0410; Fax: +1-516-562-0401.
E-mail address: tgoldber@nshs.edu

http://dx.doi.org/10.1016/j.dadm.2014.11.003

by the Food and Drug Administration (FDA) as a suitable and sole primary end point for the accelerated approval of a pharmaceutical treatment (FDA Draft Guidelines for Early Stage AD) [5]. For recent clinical trials in AD and MCI, studies typically used designs comparing cognition between the drug and placebo groups, assessed on several occasions but within a relatively short period of 18 months to 2 years, and with the end point or final assessment used as the outcome. That end point, however, may be strongly influenced by previous testing as we show in Sections 3 and 7 below. Thus, the serial testing used in these clinical trials may result in unappreciated but artifactual gains across a range of neuropsychological measures, including speed of processing, episodic memory, executive function, and working memory.

Practice effects may result from several different factors and in our view can be divided into two components. The first can be termed task familiarity and occurs early in serial assessment with given cognitive tasks. It involves the subject gaining full comprehension of the directions for the task necessary for context memory (e.g., that letters and numbers alternate in Trail-Making Test B), some knowledge of the sequence of a task (e.g., that multiple trials of a word list will be administered), and stimulus response mapping (e.g., use of a response pad in an N back test). Some task familiarity effects may be due to procedural learning, an aspect of cognition that remains relatively uncompromised in AD [6]. Even if the active treatment outperforms the placebo when both arms show practice effects, this effect may be due to an enhancement of procedural memory, which will not generate substantial benefit to the everyday cognitive function of patients with AD [7]. The second component can be termed practice-related effects. These include gains made over multiple exposures to the test because of familiarity with specific items (e.g., words on a list, a story to be recalled). Developing strategies over time that alter performance (e.g., clustering words semantically on a verbal list-learning test) might occur either as a task familiarity phenomenon or as a practice-related phenomenon. The distinction between these two components is important beyond nomenclature because it directly suggests different trial design and test construction strategies for their reduction (see Sections 3 and 7 below). If not managed, these practice effects could result in improvements that are unrelated to valid drug-placebo differences in a clinical trial.

In the context of learning and memory, practice effects would not be valid indices of specific cognitive enhancement if they do not generalize or transfer readily to other tasks or real-world activities that draw on ostensibly similar cognitive operations [8]. This is often referred to as the "transfer of training" problem. Thus, practice effects that do not relate to concurrent improvements in broad domains of cognition may be viewed as item or paradigm specific. They may also engage different cognitive operations and neural systems (e.g., procedural learning) than those thought to be treated in the intervention [9]. Also, some studies have shown that improvements in performance with repeated exposure can be used as prognostic indicators, including those related to MCI to AD conversion [10] and survival [11,12]. However, detailed discussion of these is outside the scope of the present article, which focuses on the adverse impacts of practice effects on clinical trial outcomes. Rather, in the context of a clinical trial we will cover in detail the interpretative and statistical problems associated with practice effects (see especially Sections 5.4 and 7).

We begin with a selective review of the literature on practice effects in AD, MCI, and older healthy controls as they relate to trials. We then present an example of how practice effects were confounded with treatment effects from the schizophrenia literature. Based on the literature and the schizophrenia studies, which strongly suggest that practice effects are present and large enough to obscure or be mistaken for a treatment signal, we first discuss an array of possible solutions. Next, we make recommendations for managing practice effects in preclinical AD trials based both on our review and experience in the psychometrics of test construction. It is important from the outset to recognize that our purpose is not to review the practice effect literature comprehensively. This has already been done [10,13]. Rather, our purpose is to draw out the confounding implications of practice effects in clinical trials in non–cognitively impaired older populations and suggest concrete remedies.

2. Methods

We first selectively review the literature in MCI and AD with the intention of demonstrating that even in presumptively amnestic subjects, practice effects can be identified in some cohorts. Our review in the AD and MCI groups is not meant to be exhaustive or comprehensive but rather to suggest that such effects are plausible occurrences. We then shift our focus to older, cognitively healthy individuals to demonstrate that such effects are common and measureable in serial assessment paradigms and to determine the approximate magnitude of practice effects on cognitive tests in this group. This latter group will be the focus of intense interest as the AD field moves toward secondary prevention trials in the preclinical AD spectrum.

3. Results

3.1. AD and MCI samples

Practice effects in AD have not been discussed often. Perhaps, this is the result of an expectation that many patients are substantially amnestic and unable to learn and consolidate item-level information over repeated testing. However, memory impairments are dependent on individual differences in premorbid ability and disease stage, thus creating some variability in training. Furthermore, impairments in

other domains may not be as severe as those in neural systems associated with the primary amnestic symptoms (hippocampal and medial temporal lobe pathology) and so may also increase the likelihood of practice effects (e.g., some learning may engage striatal procedural learning systems that may be relatively intact in AD). However, we acknowledge that practice effects in AD and MCI are smaller than those in healthier populations, and their significance perhaps arguable. At the same time, it should at least be considered that practice effects could serve to obscure ongoing cognitive deficits in progressive degenerative dementia.

Even with these caveats, practice effects were observed and commented on in early clinical trials of tacrine in AD [14]. Practice effects on the Mini-Mental State Examination were discernible using repeated-measures statistical techniques in AD [15]. Indirect evidence for practice effects in AD patients comes from the large number of cholinesterase inhibitor clinical trials in which both the drug and placebo groups demonstrated improvements in performance-based outcomes measures early in the trial (in the 3- to 6-month period) [16–18] with maximum effect sizes (ESs) in the 0.10 to 0.15 range. In contrast, practice effects in Alzheimer's Disease Neuroimaging Initiative's (ADNI) AD subjects, tested at 6 monthly intervals over the first 2 years, were negligible.

In ADNI's MCI group, practice effects were evident for logical memory, immediate and delayed recall, at 12 months, but were small for most other measures (T.E.G., unpublished data, 2014). In a clinical trial of MCI subjects that examined the effect of two treatments, both generally deemed ineffective (placebo and vitamin E), gains in multiple nonmemory domains (e.g., language [semantic fluency and naming], executive [digits backward, digit symbol, number cancellation], and visual processing [clock drawing]) were observed at 6 and 12 months [19]. ES gains ranged from about 0.06 to 0.30 in these domains. It is possible that these practice effects reduced the ability of the measures to detect disease-related declines in placebo groups and masked treatment effects.

In sum, practice effects can be identified in studies involving MCI and AD subjects. The magnitude of practice effects can vary across cohorts, and the literature is unclear in identifying those tests that might be most sensitive or insensitive to practice effects. Additionally, the magnitude of such practice effects (when present) diminishes as the trial progresses, and by the end of the trial, disease progression often eliminates these practice effects and decline can be observed. However, even when decline is observed, this may be an underestimate of the true decline in cognitive abilities because of practice effects. Furthermore, performance on the repeated measures at early time points beyond baseline may be inflated in both treatment and placebo groups.

3.2. Older healthy individuals

A number of studies involving repeated cognitive testing in healthy elderly groups in longitudinal studies or non–central nervous system–related clinical trials have identified practice effects both at short and longer between-test intervals [20–23]. Several of these studies indicated that such effects were large enough to reduce or eliminate age-related decline over several years [24,25]. In ADNI, practice effects were noted for older controls (mean age = 75 years) on several tests including those involving speed, logical memory, and naming. The ESs of these increases listed in Table 1 were in the low to medium range (0.20–0.30). Findings from ADNI may be especially important because they use several of the tests likely to be used in preclinical AD trials, and the subjects are likely to be representative of potential participants in preclinical AD trials.

Table 1
Maximal practice effects in the ADNI healthy control group with test-retest intervals of 6 months or 1 year

	Years	Effect size
Trails A	0.5, 1, 2	0.22–0.31
DSST	1, 2	0.19–0.20
Boston Naming Test	0.5, 1, 2, 3, 5	0.31–0.48
Rey AVLT delayed	1, 2	0.20–0.31
Logical immediate	1, 2, 3, 4, 5, 6	0.43–0.72
Logical delayed	1, 2, 3, 4, 5, 6	0.25–0.53

Abbreviations: ADNI, Alzheimer's Disease Neuroimaging Initiative; AVLT, Auditory Verbal Learning Test; DSST, Digit Symbol Substitution Test.

We next focus on two very recent studies that provide further quantitative support for this view. The first is a meta-analysis of practice effects and the second is a recent and large study of practice effects (not included in the meta-analysis). With respect to the former, Calamia et al. [13] reported a comprehensive meta-analysis in healthy controls and various neuropsychiatric groups that examined domain-specific effects as well as composite effects and accounted for various modifying factors, including age, test-retest interval, and test domain. For any given test, the number of studies ranged from 8 to 143 and from 12 subjects to 186. Calamia et al. determined a composite practice effect of z (regression weight) = 0.24 in a modal middle-aged subject retested at a 1-year interval. In a well-conducted study using the Mayo Clinic neurocognitive battery, Machulda et al. [26] found an ES of 0.24 for their global composite in a sample of 947 cognitively stable individuals (mean age, 78 years) who were tested three times over a 30-month period (and five times over 60 months). Improvements were largest in the learning and memory and smallest in language tests. Attention and speed measure gains fell between the two aforementioned domains. Based on the moderator regression weights for age and retest interval provided in the report by Calamia et al, as well as our own experience [1,2] and that of others [26] with multiple reassessments (in which magnitude of improvement with a third assessment is approximately half that of the initial T0–T1 reassessment), we estimate a performance gain of approximately 0.25 in healthy individuals in the 70- to 75-year age range undergoing three assessments within a 6- to 12-month period.

Although the sum total of improvement seems quite small, in cases who are at risk for development of MCI or dementia, this amount of improvement could mask several years' subtle decline in cognition.

Practice effects may not be restricted to cognitive measures. Harvey et al. [27] demonstrated substantial practice effects on the part of older (mean age = 68 years) healthy controls in the ability to perform tests of functional skills when assessed at 18-month retest intervals. These individuals were tested up to three times with a single form of a performance-based functional capacity measure.

3.3. Schizophrenia clinical trials

Before discussing how practice effects may make interpretation of clinical trial results challenging, we address germane findings from schizophrenia antipsychotic trials to anchor our interpretation of AD spectrum–related practice effect findings to a concrete, nonhypothetical example of the problems of separating practice effects from treatment effects. We first identified practice effects as a largely unrecognized but pervasive problem in a clinical trial involving putative cognitive-enhancing antipsychotic drugs in first-episode subjects with schizophrenia and a healthy control group [1]. Both groups were serially assessed at three time points in a 16-week period with a comprehensive set of neurocognitive measures. Both groups improved over time and to a similar degree with an ES of 0.35 from T0 to T2. As the healthy control group's improvement could only be attributed to practice (based on item familiarity, given that alternate forms were not used in the testing battery), schizophrenia-related improvement of the same magnitude could most parsimoniously be attributed to practice as well. Furthermore, it is important to appreciate that although schizophrenia subjects have widespread cognitive impairments in the mild to moderate range, these did not preclude the group from demonstrating a practice effect. Such practice effects have since been observed in multiple studies, and it is now clear that practice effects are the best explanation for improvement in several large trials in which patients were randomized to second-generation antipsychotic drugs or to a first-generation antipsychotic comparator and assessed at multiple time points (e.g., [28,29]). Thus, there is substantial evidence that the influences of these practice effects had unfortunate consequences, as a confusing state evolved with claims and counterclaims made for the "benefits" of one antipsychotic or another.

4. Discussion

It is our hope that by drawing attention to this issue, the field will begin to develop novel strategies that may overcome changes in cognitive performance that are solely due to practice effects. An organized attempt to reduce practice effects will not only eliminate a large source of noise in AD trials but also reduce the likelihood of misinterpreting outcomes. As such, we believe that this issue should be dealt with proactively in the design of clinical trials.

Critically, given our review of the literature, it is likely that individuals who are at high risk for AD (symptomatic or asymptomatic) and who are enrolled in trials and have intact cognition or subtly impaired cognition will demonstrate robust practice effects on many of the tests used. If not controlled, these practice effects may mask subtle decline in a placebo group, reducing the ability to detect improvements, if any, in the active treatment group. In this latter group, conservatively, an ES of 0.25 units would likely be larger than decline over a 6- to12-month period. Alternatively, effects might be misinterpreted as active drug effects in trials in which cohort differences in learning were present and positive drug effects would have to be substantial to be detected (see below).

5. Solutions

Several solutions to the problem of practice effects can be considered (Table 2).

5.1. Use of alternate forms

On the face of it, the use of alternate forms would seem to be a straightforward approach [30]. However, some tests with problem-solving components (e.g., the Wisconsin Card Sorting Test and similar procedures) and that require discovery of an overarching and discrete set of rules are

Table 2
Rules of thumb for addressing practice effects in clinical trials

Approach	Advantages	Pitfalls
Use of a control group	Necessary in placebo-controlled randomized clinical trial	Prone to confounding a practice effect with a treatment effect
Massed practice (multiple prebaseline testing)	Repeated testing during a prebaseline period may result in a task familiarity–based asymptote and can reduce interindividual and intraindividual variance due to subjects not fully understanding task demands.	Differential asymptotes between tests; possible ceiling effects; occlusion of treatment effects
Reliable change index	Is rigorous	Applicable to cases, not group means
Alternate forms	Clearly reduce practice effects	Forms may not be equivalent in difficulty level; influence of strategy formation
Practice-insensitive tests	Interpretation of improvement or lack of decline is straightforward	Relevance of cognitive operations; sensitivity to treatment

not easily amenable to this solution. Second, in other problem-solving tasks, generation of various strategies can result in improved performance across testing, even with alternate forms. Third, alternate forms may not be equivalent: They are often quite different in difficulty. This was recently demonstrated on the Auditory Verbal Learning Test (AVLT) in the non–cognitively impaired sample from the ADNI study in which two parallel versions, forms A and B, were administered over 5 years. Form B was administered at 6 months and 3 years, and form A at 1, 2, 4, and 5 years. Compared with performance at the start of the study, the two forms showed either declines or improvements [24]. Fig. 1 presents the data from which it can be seen that compared with performance at the start of the study, version B of this verbal list-learning test was associated with significantly reduced performance at 6 months and 3 years, whereas version A showed improvements, which were significant for both immediate and delayed recall at 2 years. Studies in schizophrenia have also revealed significant discrepancies across forms of common tests in repeated-measures designs [31]. Thus, alternate forms pose a significant problem in that the more testing points that there are, the more alternate forms are required. When forms of different difficulty are administered over time, the true course of functioning becomes very challenging to discern. In many ways, practice effect variance across single forms is a more easily managed problem than alternate forms of variable difficulty.

5.2. Reliable change index

The reliable change index (RCI) is a rigorous approach that yields information on the number of individual subjects who demonstrate gains above and beyond practice. A confidence interval identifies the extent to which an individual subject would have to improve to demonstrate progress beyond a practice effect beyond reasonable doubt. The statistic is dependent on not only differences in means between time points but also the variance of the difference and the practice effect for untreated cases [2,32]. This statistic is critical for treatment development because it allows for estimation of the magnitude of change that exceeds the practice effect on a placebo-corrected basis that excludes all non–treatment-related differences. Nevertheless, this is a conservative statistic that does not consider the fact that for nearly every treatment, not all treated cases respond. Thus, the number needed to treat must be crossed with the RCI for the outcome measures to understand the magnitude of a group response that would suggest a truly responsive subgroup. A similar approach uses regression models that take into account individual demographics and initial level of performance to define an expected score and a confidence interval. This approach may be more flexible than the RCI approach [32]. Additionally, the information output is based on the case count, not group mean differences. Nevertheless, cases can be combined and contrasted across conditions, although we are not aware of any phase 3 trial that has used either of these statistics.

5.3. Prebaseline massed practice

In this approach, an attempt is made to reach an asymptote in task familiarity–driven gains during a lead-in or prebaseline period of the trial by administering tests multiple times in a short period (e.g., two to three administrations within a day). This approach has been advocated for many years [33,34]. Although several studies [35,36] have indicated that two to three practice trials before baseline result in asymptotic performance, others [37] indicate a mixed picture, with different tests demonstrating asymptotes over different number of repetitions. The intent in using massed practice is to minimize issues of comprehension of instructions, strategy formation, and inefficient stimulus

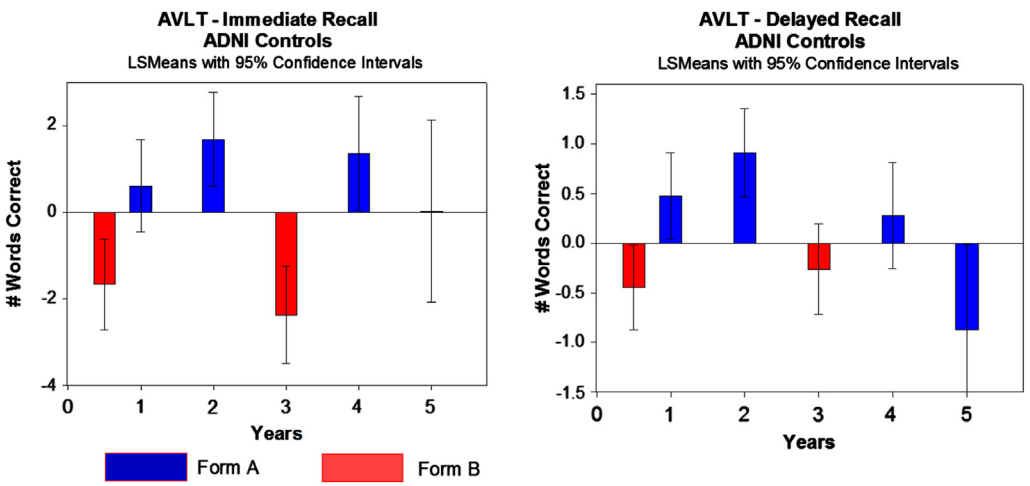

Fig. 1. Change from year 0 data in the ADNI studies for the Rey Auditory Verbal Learning Test. Ascending values reflect improvements and descending impairments. The two parallel forms (A and B) are clearly not of equivalent difficulty. Abbreviations: ADNI, Alzheimer's Disease Neuroimaging Initiative; AVLT, Auditory Verbal Learning Test; LS, Least Square.

response mapping due to lack of familiarity, as opposed to practice-related effects such as item exposure–related gains.

One aspect of this approach that may be problematic if performance is relatively high at baseline is that a ceiling effect associated with practice could prevent further change due to treatment. Another potential problem with this approach is that, after baseline, an artifactual decline can result because of the use of an alternate form of the test not used in the prebaseline period or possible loss of retention of task familiarity–based knowledge of instructions, sequence, strategies, and so on. Additionally, other work has suggested that practice may recruit, engage, or otherwise occlude the same or opposing neural systems as those targeted by treatment and interfere with connectivity, neurochemical, or plasticity-related alterations specific to active treatment [38].

5.4. Control groups

We do not agree with the frequently made argument that the use of a control group obviates practice effect issues because the essential comparison in a clinical trial is between groups at end point and not change over time. First, it is easier to detect a change signal against a "flat," no practice effect background than against a noisier, practice effect-plus-treatment effect background. In keeping with this view, it has been demonstrated that a procognitive drug effect (e.g., amphetamine in schizophrenia and donepezil in AD patients) was much more likely to be identified using tests that did not show practice effects, rather than using domain-similar tests that were prone to practice effects [3,39]. These findings may have been the result of reductions in practice-associated variance. Alternatively, repeated exposure to the same stimulus reduced potential plasticity in the neural system targeted by the drug, thereby obscuring any benefit of the drug on cognitive function. Second, if time is explicitly taken into account in repeated measures in the comparison of groups in a trial [40], an ES of 0.20 to 0.30, corresponding to a barely noticeable between-group effect for treatment, would be added to a practice-effect ES of 0.25 in more intact populations. Thus, the gain in the active treatment group would have to be as high as 0.50 for a difference between treatment groups to be detected (as a group × time interaction). This is large and would require a seemingly improbably efficacious compound, particularly in cases with MCI or dementia. Nevertheless, we acknowledge that this is a statistically driven model, and there is neither positive nor negative data that directly bear on it.

5.5. A priori development of tests that are not prone to practice effects

Sensitivity to the issue of practice effects in the a priori design of a variety of tests that assay multiple cognitive domains offers several advantages: use of cognitive science paradigms that minimize individual item recall and strategy shifts that could differentially impact performance, development of alternate forms from large item pools, and comprehensive co-normed data for the battery versions. Thus, a novel battery of tests was recently constructed that used specific principles from the cognitive science literature to substantially reduce practice effects: (1) multiple items, a restricted set of stimuli that serve to induce interference, and alternative and equivalent forms with different items and sequences in tests of attention, working memory, and executive function and (2) for episodic memory, obligatory common encoding of items (to reduce strategy changes). Preliminary results in 29 healthy controls (age range, 20–50 years) who were tested three times in 16 weeks suggested reduced practice effects (for all tests below an ES = 0.15), robust psychometrics, and lack of ceiling and floor effects (T.E.G., unpublished data). Several tests have also directly addressed practice effects (via alternate forms) and have been used in AD-related clinical trials. These include, but are not restricted, to the following: Repeatable Battery for the Assessment of Neuropsychological Status [41], CogState Battery (including the Groton Maze Learning Test) [42–44], and the Cognitive Drug Research (CDR) System [45–48], but their factor structure, the computational operations demanded by particular tests, and their relationship to everyday function have not always been fully delineated.

6. Recommendations

Three approaches to attenuating practice effects involve (1) massed practice in a prebaseline period so that a task familiarity having to do with comprehension of instructions, development of simple strategies, stimulus response mapping, and testing sequence is increased; (2) use of tests with multiple similar items, standardized encoding, and so on that capitalize on cognitive science principles that reduce recall of individual items or protect against large strategy shifts that might influence recall; and (3) well-matched alternate forms to minimize item exposure. Each has different strengths, pragmatic implications, and economic costs. Our list of suggestions is included in Table 2.

A clinical trial using tests that are designed to avoid the impact of practice-related effects would look very much like current trials in which baseline assessment at T0 by version A of the test was followed by three more assessments (T1-version B, T2-version C, and T3-version D) over an 18-month period. Administration of test versions would be counterbalanced. For trials in which prebaseline test administration was used (i.e., "massed practice") to reduce task familiarity effects, two to three assessments before baseline might reduce these artifacts [9,10].

For example, for a task of paired associate learning for which there is benefit in understanding the paradigm, two prebaseline assessments (version X) could serve to increase task familiarity, with another version (A) serving as baseline. This would be followed by three more assessments (T1-version B, T2-version C, and T3-version D). Of course, ceiling effects would have to be considered and minimized if

necessary. Strategy changes in digit span (e.g., "chunking" items or covert rehearsal) or in visual search during Trail-Making Part B (appreciating that a number or letter may be "under" one's hand) and using semantic clustering in memory tests might also be stabilized in prebaseline administrations. It might not be possible, however, to estimate in advance how many massed practice sessions are required to eliminate the possibility of subsequent improvements. This solution might be especially useful in populations with more substantial baseline impairments such as MCI or AD, where improvement to ceiling is less likely.

7. Final thoughts

Consider the following thought experiment. An outcome of a prevention trial in preclinical AD suggests that a neurodegenerative cascade has been arrested. This effective disease-modifying treatment would result in stability in cognitive function or small improvements, insofar as the neurodegenerative effects are reversed. The untreated group would decline, albeit subtly. A sensitive set of tests assaying important cognitive domains and resistant to practice effects would accurately monitor this scenario. In contrast, using tests subject to practice effects, both groups would improve, inaccurately representing the drug's efficacy, resulting in the strong possibility that differential effects would be masked as the within-group change would be much greater than the between-group change, and resulting in a serious type 2 error. In other words, the cognitive signal would have been misaligned with underlying neurobiological changes associated with neurodegeneration (neural system compromises) and rectifications thereof. We think that interpretation would be parsimonious and accurate if a treatment-related signal could be identified in a group that would otherwise demonstrate measurable subtle decline across time points.

Acknowledgments

The authors thank Dr. Richard Keefe for comments on an earlier draft of this work.

T.E.G. has received funding for the writing of this article from the NIA (3R01 AG038734) and DOD (W81XWH-12-1-0084). He receives royalties from NeuroCog Trials for use of the BACS in clinical trials.

P.D.H. is a consultant for the following companies: AbbVie, Boehringer-Ingelheim, En Vivo, FORUM Pharma, Genentech, Lundbeck, Otsuka-America, Roche, Sunovion, and Takeda. He conducted contract research for Genentech.

L.S.S. has received grants from the NIH P50 AG05142, R01 AG033288, R01 AG037561, and UF1 AG046148, the State of California, the Alzheimer's Association for a registry for dementia and cognitive impairment trials and grants or research support from the Alzheimer's Disease Cooperative Study (NIA, UCSD), Baxter, Genentech, Johnson & Johnson, Eli Lilly, Lundbeck, Novartis, Pfizer, and Tau Rx. He has served as a consultant for and received consulting fees from AbbVie, AC Immune, Allon, AstraZeneca, Baxter, Biogen Idec, Biotie, Bristol-Myers Squibb, CereSpir, Chiesi, Elan, Eli Lilly, En Vivo, GlaxoSmithKline, Johnson & Johnson, Lundbeck, MedAvante, Merck, Novartis, Piramal, Pfizer, Roche, Servier, Takeda, Tau Rx, Toyama, and Zinfandel.

P.J.S. is a consultant for CogState, Ltd. (Melbourne, Australia).

K.A.W. was until recently an employee and shareholder in Bracket, who provide the CDR System to the clinical trial industry.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.dadm.2014.11.003.

RESEARCH IN CONTEXT

1. Systematic review: In this review, we selectively examined literature on practice effects in serial cognitive assessment in healthy elderly, preclinical Alzheimer's disease (AD), and AD.
2. Interpretation: Practice effects are frequently found in preclinical AD and reliably observed in healthy elderly individuals. Their effect size is estimated to be moderate (Cohen's d=.25). We further suggest that individuals with preclinical AD, who by definition are normal or near-normal cognitively, will demonstrate practice effects that are similar in magnitude to those in the healthy elderly. We provide scenarios by which such effects could easily cloud interpretation of results in clinical trials, such that a drug effect is confounded with a practice effect and/or does not align with neurobiological processes, including aging and neurodegeneration.
3. Future directions: Critically, we also offer a set of concrete recommendations on how to manage practice effects in clinical trials utilizing multiple and distinct approaches.

References

[1] Goldberg TE, Goldman RS, Burdick KE, Malhotra AK, Lencz T, Patel RC, et al. Cognitive improvement after treatment with second-generation antipsychotic medications in first-episode schizophrenia: Is it a practice effect? Arch Gen Psychiatry 2007;64:1115–22.

[2] Goldberg TE, Keefe RS, Goldman RS, Robinson DG, Harvey PD. Circumstances under which practice does not make perfect: a review of the practice effect literature in schizophrenia and its relevance to clinical treatment studies. Neuropsychopharmacology 2010; 35:1053–62.

[3] Pietrzak RH, Snyder PJ, Maruff P. Amphetamine-related improvement in executive function in patients with chronic schizophrenia is modulated by practice effects. Schizophr Res 2010;124:176–82.
[4] Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, et al. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7:280–92.
[5] Kozauer N, Katz R. Regulatory innovation and drug development for early-stage Alzheimer's disease. N Engl J Med 2013;368:1169–71.
[6] Budson AE, Price BH. Memory dysfunction. N Engl J Med 2005; 352:692–9.
[7] Wesnes K, Pincock C. Practice effects on cognitive tasks: a major problem? Lancet Neurol 2002;1:473.
[8] Thorndike EL, Woodworth RS. The influence of movement in one mental function upon the efficiency of other functions. Psychol Rev 1901;8:247–61.
[9] Kelly AM, Garavan H. Human functional neuroimaging of brain changes associated with practice. Cereb Cortex 2005;15: 1089–102.
[10] Duff K, Lyketsos CG, Beglinger LJ, Chelune G, Moser DJ, Arndt S, et al. Practice effects predict cognitive outcome in amnestic mild cognitive impairment. Am J Geriatr Psychiatry 2011;19:932–9.
[11] Dodge HH, Wang CN, Chang CC, Ganguli M. Terminal decline and practice effects in older adults without dementia: the MoVIES project. Neurology 2011;77:722–30.
[12] Meinert CL, Breitner JC. Chronic disease long-term drug prevention trials: lessons from the Alzheimer's Disease Anti-inflammatory Prevention Trial (ADAPT). Alzheimers Dement 2008;4(1 Suppl 1):S7–14.
[13] Calamia M, Markon K, Tranel D. Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. Clin Neuropsychol 2012;26:543–70.
[14] Eagger S, Morant N, Levy R, Sahakian B. Tacrine in Alzheimer's disease. Time course of changes in cognitive function and practice effects. Br J Psychiatry 1992;160:36–40.
[15] Galasko D, Abramson I, Corey-Bloom J, Thal LJ. Repeated exposure to the Mini-Mental State Examination and the Information-Memory-Concentration Test results in a practice effect in Alzheimer's disease. Neurology 1993;43:1559–63.
[16] Rogers SL, Farlow MR, Doody RS, Mohs R, Friedhoff LT. A 24-week, double-blind, placebo-controlled trial of donepezil in patients with Alzheimer's disease. Donepezil Study Group. Neurology 1998; 50:136–45.
[17] Birks J, Harvey RJ. Donepezil for dementia due to Alzheimer's disease. Cochrane Database Syst Rev 2006;CD001190.
[18] Rogers SL, Doody RS, Mohs RC, Friedhoff LT. Donepezil improves cognition and global function in Alzheimer disease: a 15-week, double-blind, placebo-controlled study. Donepezil Study Group. Arch Intern Med 1998;158:1021–31.
[19] Petersen RC, Thomas RG, Grundman M, Bennett D, Doody R, Ferris S, et al. Vitamin E and donepezil for the treatment of mild cognitive impairment. N Engl J Med 2005;352:2379–88.
[20] Bartels C, Wegrzyn M, Wiedl A, Ackermann V, Ehrenreich H. Practice effects in healthy adults: a longitudinal study on frequent repetitive cognitive testing. BMC Neurosci 2010;11:118.
[21] Rabbitt P, Diggle P, Smith D, Holland F, Mc Innes L. Identifying and separating the effects of practice and of cognitive ageing during a large longitudinal study of elderly community residents. Neuropsychologia 2001;39:532–43.
[22] Krenk L, Rasmussen LS, Siersma VD, Kehlet H. Short-term practice effects and variability in cognitive testing in a healthy elderly population. Exp Gerontol 2012;47:432–6.
[23] Frank R, Wiederholt WC, Kritz-Silverstein DK, Salmon DP, Barrett-Connor E. Effects of sequential neuropsychological testing of an elderly community-based sample. Neuroepidemiology 1996; 15:257–68.
[24] Wesnes KA, Schneider LS. Are neuropsychological tests such as those used in ADNI suitable for long-term trials of cognition enhancers for preclinical Alzheimer's disease? J Nutr Health Aging 2012;16:810.
[25] Salthouse TA, Tucker-Drob EM. Implications of short-term retest effects for the interpretation of longitudinal change. Neuropsychology 2008;22:800–11.
[26] Machulda MM, Pankratz VS, Christianson TJ, Ivnik RJ, Mielke MM, Roberts RO, et al. [Formula: see text]Practice effects and longitudinal cognitive change in normal aging vs. incident mild cognitive impairment and dementia in the Mayo Clinic Study of Aging. Clin Neuropsychol 2013;27:1247–64.
[27] Harvey PD, Reichenberg A, Bowie CR, Patterson TL, Heaton RK. The course of neuropsychological performance and functional capacity in older patients with schizophrenia: Influences of previous history of long-term institutional stay. Biol Psychiatry 2010;67:933–9.
[28] Keefe RS, Bilder RM, Davis SM, Harvey PD, Palmer BW, Gold JM, et al. Neurocognitive effects of antipsychotic medications in patients with chronic schizophrenia in the CATIE trial. Arch Gen Psychiatry 2007;64:633–47.
[29] Davidson M, Galderisi S, Weiser M, Werbeloff N, Fleischhacker WW, Keefe RS, et al. Cognitive effects of antipsychotic drugs in first-episode schizophrenia and schizophreniform disorder: a randomized, open-label clinical trial (EUFEST). Am J Psychiatry 2009; 166:675–82.
[30] Beglinger LJ, Gaydos B, Tangphao-Daniels O, Duff K, Kareken DA, Crawford J, et al. Practice effects and the use of alternate forms in serial neuropsychological testing. Arch Clin Neuropsychol 2005; 20:517–29.
[31] Harvey PD, Palmer BW, Heaton RK, Mohamed S, Kennedy J, Brickman A. Stability of cognitive performance in older patients with schizophrenia: an 8-week test-retest study. Am J Psychiatry 2005;162:110–7.
[32] Temkin NR, Heaton RK, Grant I, Dikmen SS. Detecting significant change in neuropsychological test performance: a comparison of four models. J Int Neuropsychol Soc 1999;5:357–69.
[33] McClelland GR. The effects of practice on measures of performance. Hum Psychopharmacol 1987;2:109–18.
[34] Duff K, Westervelt HJ, McCaffrey RJ, Haase RF. Practice effects, test-retest stability, and dual baseline assessments with the California Verbal Learning Test in an HIV sample. Arch Clin Neuropsychol 2001; 16:461–76.
[35] Collie A, Maruff P, Darby DG, McStephen M. The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test-retest intervals. J Int Neuropsychol Soc 2003; 9:419–28.
[36] Falleti MG, Maruff P, Collie A, Darby DG. Practice effects associated with the repeated assessment of cognitive function using the CogState battery at 10-minute, one week and one month test-retest intervals. J Clin Exp Neuropsychol 2006;28:1095–112.
[37] Beglinger LJ, Ahmed S, Derby MA, Siemers E, Fastenau PS, Crawford-Miller J, et al. Neuropsychological practice effects and change detection in people with schizophrenia. Schizophr Res 2003;62:191–4.
[38] Vinogradov S, Fisher M, de Villers-Sidani E. Cognitive training for impaired neural systems in neuropsychiatric illness. Neuropsychopharmacology 2012;37:43–76.
[39] Pietrzak RH, Maruff P, Snyder PJ. Methodological improvements in quantifying cognitive change in clinical trials: an example with single-dose administration of donepezil. J Nutr Health Aging 2009;13:268–73.
[40] Olanow CW, Rascol O, Hauser R, Feigin PD, Jankovic J, Lang A, et al. A double-blind, delayed-start trial of rasagiline in Parkinson's disease. N Engl J Med 2009;361:1268–78.
[41] Karantzoulis S, Novitski J, Gold M, Randolph C. The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): Utility in detection and characterization of mild cognitive impairment due to Alzheimer's disease. Arch Clin Neuropsychol 2013;28:837–44.
[42] Pietrzak RH, Snyder PJ, Jackson CE, Olver J, Norman T, Piskulic D, et al. Stability of cognitive impairment in chronic schizophrenia

over brief and intermediate re-test intervals. Hum Psychopharmacol 2009;24:113–21.

[43] Pietrzak RH, Olver J, Norman T, Piskulic D, Maruff P, Snyder PJ. A comparison of the CogState Schizophrenia Battery and the Measurement and Treatment Research to Improve Cognition in Schizophrenia (MATRICS) Battery in assessing cognitive impairment in chronic schizophrenia. J Clin Exp Neuropsychol 2009; 31:848–59.

[44] Maruff P, Thomas E, Cysique L, Brew B, Collie A, Snyder P, et al. Validity of the CogState brief battery: relationship to standardized tests and sensitivity to cognitive impairment in mild traumatic brain injury, schizophrenia, and AIDS dementia complex. Arch Clin Neuropsychol 2009;24:165–78.

[45] Allain H, Neuman E, Malbezin M, Salzman V, Guez D, Wesnes K, et al. Bridging study of S12024 in 53 in-patients with Alzheimer's disease. J Am Geriatr Soc 1997;45:125–6.

[46] Vellas B, Cunha L, Gertz HJ, De Deyn PP, Wesnes K, Hammond G, et al. Early onset effects of galantamine treatment on attention in patients with Alzheimer's disease. Curr Med Res Opin 2005;21:1423–9.

[47] Galvin JE, Cornblatt B, Newhouse P, Ancoli-Israel S, Wesnes K, Williamson D, et al. Effects of galantamine on measures of attention: results from 2 clinical trials in Alzheimer disease patients with comparisons to donepezil. Alzheimer Dis Assoc Disord 2008;22:30–8.

[48] Wesnes K, Edgar C, Andreasen N, Annas P, Basun H, Lannfelt L, et al. Computerized cognition assessment during acetylcholinesterase inhibitor treatment in Alzheimer's disease. Acta Neurol Scand 2010;122:270–7.