Dynamics of Neuropsychological Testing

Michael H. Coen[1,2], Timothy S. Chang[3], Bruce Hermann[4], Asenath La Rue[5], Mark Sager[5]

Department of Biostatistics and Medical Informatics[1]
Department of Computer Sciences[2]
Institute for Clinical and Translational Research[3]
Department of Neurology[4]
Wisconsin Alzheimer's Institute[5]
University of Wisconsin, Madison, WI 53706
mhcoen@biostat.wisc.edu, {tschang3, larue, masager}@wisc.edu, hermann@neurology.wisc.edu

Abstract

How should we analyze repeated trials in neuropsychological testing? It has long been known that experimental subjects display distinct stages of acclimatization and subsequent saturation during cognitive testing (Thurstone, 1927). For example, in list learning tests examining memory, it has been demonstrated that repeated exposure to a fixed enumeration of items can improve recall. However, we think it is equally important to examine acclimatization of the subjects to the test taking procedure itself. In other words, subjects must grow comfortable with the paradigm of the test before we can assume the results correspond with our interpretations of them. In this paper, we examine results of the Rey Auditory-Verbal Learning Test administered to the largest Alzheimer's disease family history cohort. We demonstrate the most informative *signal* in a neuropsychological test may contradict a priori assumptions about the test's interpretation.

Keywords: Neuropsychological testing; statistical analysis; Alzheimer's disease; memory

Introduction

Psychological tests often employ repeated trials of similar or identical tasks. Sometimes, these repetitions are intended to allow subjects to acclimatize to the stimulus and/or decision making paradigms, e.g., as in forced choice experiments (Mitchell & Jolley, 2009). Indeed, in psychoacoustic experiments, subjects may be unable to even distinguish phenomena of interest without substantial prior exposure and early practice rounds are commonly discarded as uninformative.

In neuropsychological tests, it is commonplace to conduct multiple trials of a test, from which summary scores may be derived (Lezak et. al, 2004). Multiple trials can also reveal perseverative errors, which are characteristic of a number of cognitive pathologies. However, in tests focused on evaluating memory, it has been demonstrated that enhanced performance may occur after repeated examinations (Benedict & Zgaljardic, 1998). Thus, subjects are often tested on a smaller number of rounds than might otherwise be desired for acclimatization.

Although examining summary scores averaged over trials is commonplace, the individual trial scores can vary enormously from trial to trial. This may occur even in simple tests such as the Rey Auditory-Verbal Learning Test (AVLT), which is repeated only a few times. In this paper, we provide evidence this is not simply due to naturally occurring variance. Rather, individual trials can be far more *informative* than aggregate summary scores.

We believe that during tests involving a relatively small number of repetitions, subjects are *simultaneously acclimatizing and responding* to the testing procedure itself, which may conflate interpretations of their responses when viewed via aggregate summary measures. In other words, they are still "learning" the test while they are "taking" the test; we believe the shift between these two processes accounts for much variance across trials.

We have previously demonstrated (Coen et al., 2009) that even in non-memory based experiments, such as the Conceptual Set Shifting Task (Milner, 1964) conducted on human and macaque subjects, performance varies substantially but predictably over the course of the trials. Specifically, subjects' performance on the first few rounds is both slow and inaccurate. However, by the third round, it

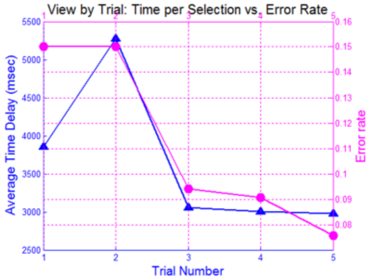

Figure 1 – Viewing the Conceptual Set Shifting Task (CSST) by individual trial. In the CSST, each trial consists of a lengthy procedure of trying to guess a hidden concept correctly 10 times in a row, after which a new secret concept is selected for the next trial. By examining both the times taken per decision and the subjects' error rates, it appears clear that the results become meaningful according to the test's desiderata by the third trial, whereas the first two trials reflect acclimatization. We see this in the dramatic decrease in the average time taken between decisions and the precipitous drop in error rate. This illustrates both that aggregate summary scores combining all five trials are conflating (at least) two different phenomena, and it additionally provides a *signal* that the subject's results after trial 3 more meaningfully reflect performance. (Figure adapted from Coen et al. (2009)).

appears clear that human subjects have *learned* how to take the test, as illustrated in Figure 1. Only after internalizing the rules of the test, do subjects clearly begin to *respond* in ways that meet our expectations.

In this paper, we reexamine results from the Rey Auditory-Verbal Learning Test obtained from the Wisconsin Registry for Alzheimer's Prevention (Sager, Hermann, & La Rue, 2005). It is the largest family history cohort of its kind, consisting of approximately 1,200 asymptomatic patients. We demonstrate that in using the AVLT to separate familial history from control populations, previously unknown scoring measures found via machine learning approaches provide far more statistically significant results than do *intuitively* designed scoring metrics. This mirrors our previously cited work that the strongest signal – which may not be obvious in advance — is contained after several trials, presumably when subjects have acclimatized to the experiment framework itself. While the data in this paper are drawn from a large cohort study of Alzheimer's disease (AD), there is nothing specific to AD in these results, and we believe these findings, buttressed by our earlier work on the CSST, are of interest in understanding and analyzing the results of neuropsychological testing more generally.

Background

The Rey Auditory Verbal Learning Test (AVLT) (Rey, 1964) is a neuropsychological test consisting of eight trials, of which the first five trials are often scrutinized more carefully for studying Alzheimer's disease (La Rue et al., 2008; Ramakers et al., 2010; Woodard, Dunlosky, & Salthouse, 1999). Briefly, in this test, a psychometrist reads 15 unrelated nouns and the subject repeats as many words as possible in whatever order he or she finds natural.

Summary scores – such as the number of words recalled per trial or the total number of recalled words across all trials – can to varying degrees of confidence differentiate normal persons from those with early-stage AD (Bigler, Rosa, Schultz, Hall, & Harris, 1989; Mitrushina, Satz, & Van Gorp, 1989; Woodard, et al., 1999).

Derivative Performance Measures

Summary scores are often used to create proxy metrics thought to summarize higher-level cognitive functioning. It is often the case, as discussed below, that these proxy measures are averaged across trials to derive aggregate test scores for evaluating patients. It is this process that we deem problematic.

In the AVLT, differences have been noted between persons with mild AD and control groups on *serial position* effects and on *subjective organization* during recall. Persons with AD, even at mild stages, disproportionately recall words from the end of a supraspan list (the "*recency effect*") compared to those at the beginning of the list (the "*primacy effect*") (Capitani, Della Sala, Logie, & Spinnler, 1992; Hermann et al., 1996). The interpretation is that words at the end of the list (i.e., the recent words) are easier for patients with mild AD to remember.

Such derivative learning measures have also been studied in non-demented persons who are at increased risk of developing AD. Ramakers et al. (2010) measured *subjective organization* in the AVLT by examining pairs of words recalled together in subsequent trials and found marginal significant differences between patients diagnosed with mild cognitive impairment that did and did not progress to AD. More recently, La Rue et al. (2008) showed a detectable serial position effect in the Wisconsin Registry for Alzheimer's Prevention; here, asymptomatic persons with a parental family history of AD showed increased reliance on recency in recall compared to controls whose parents did not have AD.

We note that in all of these studies, populations are compared via simple hypothesis testing, where significance is evaluated by a derived *p*-value. However, it is rarely asked what these *p*-values actually mean, whether they can be compared across different tests, or what it means if one does so.

Comparison of *p*-values

It is conceptually and mathematically difficult to compare *p*-values derived from different measures. The common interpretation is a hypothesis test provides the probability that rejection of the null hypothesis is not due to "chance." There is a vast literature on the interpretation of *p*-values (Wasserman, 2004; Ott & Locknecker, 2001); its most simplistic interpretation of p=0.05 is that we believe the detected difference has only a 5% chance of occurring at random. Regardless of interpretation, it is difficult to compare *p*-values. How much "better" is a hypothesis test that provides $p = 0.01$ than one that provides p=0.05? This is exacerbated when different measures are used to obtain these values, all the more so when their stability has not been empirically evaluated.

A standard statistical answer to this question is that comparing *p*-values is useful only when it provides additional insight into the problem at hand. In other words, comparing $p = 0.05$ and $p = 0.0005$ may have little meaning unless the process by which p was lowered is informative. Thus, a smaller p value may not be inherently better unless we have some understanding of how it was obtained. (The most straightforward example of this would be a lookup table, which can provide arbitrarily low *p*-values. However, if we realize that an approach, for example, simply overfits the data, it is no longer of any interest.)

Our Approach

In this paper, we construct derivative performance measures for evaluating the results of neuropsychological testing. By observing the effects of combining different metrics on test results, we can derive confidence that incorporating particular data (or "signals") does indeed help in hypothesis testing, namely, in separating test populations. As such, this is a valid domain for comparing *p*-values and one where doing so makes sense. Namely, it tells us whether including additional factors in a given hypothesis test makes it more or less powerful. It simultaneously allows us to include features that capture the test's internal dynamics, even when these are unknown in advance. For example, we may not know (or even be able to well-define) the transition between acclimatization, test taking, and saturation. However, we demonstrate that these can be learned reliably.

Experimental Methodology

Participants

The study methods for the Wisconsin Registry for Alzheimer's Prevention (WRAP) began enrollment in 2001; a detailed summary is in Sager et al. (2005). Briefly, WRAP participants are English speaking adults between the ages of 40 and 65 years with at least one parent with autopsy-confirmed or probable AD (McKhann et al., 1984). Control participants had mothers surviving to at least 75 years and fathers to at least 70 years without Alzheimer's disease, other dementia, or significant memory deficits.

Procedures

A wide assortment of data were collected, including clinical measures, health history, extensive neuropsychological testing, including AVLT responses, and chemical panel data. This included data corresponding to the Apolipoprotein ε4 (APOE) gene, a biomarker widely suspected to be implicated in onset of AD.

Derivative Measures

Subjective organization as explained in (Ramakers et al., 2008) was measured for a patient between subsequent trials for the first five trials (trial 1 and trial 2, trial 2 and trial 3, trial 3 and trial 4, trial 4 and trial 5). Subjective organization is calculated for trial i to i+1 as $j - \frac{2c(c-1)}{hk}$, where j is the number of pairs of items recalled on trial i and i+1 in adjacent positions, c is the number of common items recalled on both trial, h is the number of items recalled on trial i and k is the number of items recalled on trial i+1. Serial position primacy was calculated as described in (La Rue, et al., 2008) for the first five trials, where primacy was the percentage of the first four words from the AVLT that were recalled.

Fine grained AVLT analysis

Reflecting on derivative measure such as primacy and subjective organization, we noticed they did not capture the low-level differences in recall or our intuitions of what they represented.

For example, Figure 2 illustrates the insensitivity of these measures to seemingly gross differences in performance. This is largely due to the effects of partitioning responses as equivalent based on histograms – rather than their actual recall order. On inspection, it appears that Figures 2b and 2c are much more similar than Figures 2a and 2b. Therefore, 2a and 2b should not have the same subjective organization score; nonetheless, they do, as recall order is ignored entirely. These examples highlight that very different recall strategies are not being captured by these measures. In the present work, we use microstructure in the test results to find signals that are otherwise lost in analyses that examine binned regions of recall regardless of their precise order.

We developed a new derivative measure of AVLT to investigate details in subject recall using the Euclidean distance between trials i and i+1. For two recall trials with a different total number of words recalled, zeros were filled in at the end of the shorter recall trials. As a concrete example, we calculate the Euclidean measure for trials [1,2,3,4] and [8,7,4,3,1,2,6] by calculating the Euclidean distance between [1,2,3,4,0,0,0] and [8,7,4,3,1,2,6]. We calculated the measure between sequential trials from the first five learning trials. We note the signal between trials 3 and 4 was so strong when viewed this way that many other point-wise distance metrics worked similarly, as described below. Euclidean distance was selected for its simplicity, although several more esoteric distance metrics (Deza & Deza, 2009) provided slightly increased performance in hypothesis testing.

Metric Combination

We constructed a new aggregate measure using the aggregate function:

$$M_{Aggregate}(\theta) = \alpha\bar{M}_{Prim(i)} + \beta\bar{M}_{SO(j,k)} + \gamma\bar{M}_{Euc(l.m)}$$

where $\bar{M}_{Prim(i)}$ is the (normalized) primacy score on trial i,

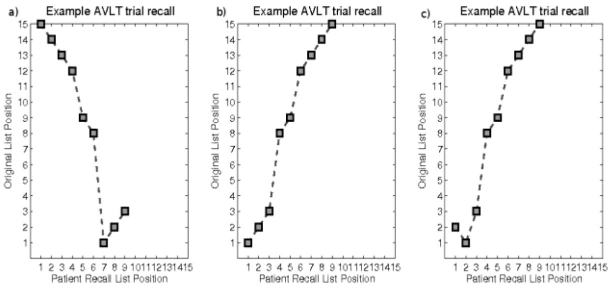

Figure 2: We represented the recalled words of a trial by the position from the original list. This figure showed the recall position on the x-axis and the original list position on the y-axis. Each subfigure had example trials that recall the same words but in different orders. 2a) and 2b) resulted in the same serial position scoring because order is not taken into consideration. Yet the recall strategies seem almost opposite. The subjective organization score when comparing 2b) and 2a) was identical to comparing 2b) and 2c). However, it seemed that 2b) and 2c) were much more similar than 2a) and 2b) and should not have the same subjective organization score if order were considered. These examples highlight that different recall strategies are not captured by these measures. Our goal is to use microstructure in the test results to find signals that are otherwise lost in analyses that examine binned regions of recall.

***P*-values separating by trial**	
Trial	**Primacy**
1	0.00599
2	0.0744
3	0.994
4	0.852
5	0.361
1-5	0.187

***P*-values separating by comparing pairs of successive trials**		
Trial	**Subjective Organization**	**Euclidean Distance**
1-2	0.0268	0.604
2-3	0.449	0.237
3-4	0.0206	5.07×10^{-04}
4-5	0.134	0.536
1-5	0.199	.0212

Table 1. ANOVA *p*-values for family history using individual trials with primacy and consecutive trials with subjective organization and Euclidean distance. ANOVA was performed while controlling for a genetic biomarker, age, sex and education level. For primacy, only trial 1 family history was significant (*p*=0.0059). For subjective organization, family history was significant for trials 1-2 (*p*=0.0268) and trials 3-4 (*p*=0.0206). For the Euclidean measure, family history was significant for trial 3-4 (*p*=0.000507). SO=subjective organization. The bottom row of each table shows the summary score for the measure across all trials, which is typically employed in the literature.

$\bar{M}_{SO(j,k)}$ is the (normalized) score of subjective organization between trials j and k, and $\bar{M}_{Euc(l.m)}$ is the (normalized) Euclidean distance between trials l and m. We normalized each measure to have a domain between [0, 1] to eliminate arbitrary scaling differences in their scoring methodology. For example, the maximum distance for primacy = 1, whereas the maximum Euclidean distance is approximately 35.21. Thus, we did not want one measure to arbitrarily dominate the scoring because of variability in its output.

To find parameters $\theta = \langle \alpha, \beta, \gamma, i, j, k, l, m \rangle$, we employed stochastic gradient descent (Bertsekas & Nedic, 2003), using $\theta_{i+1} = \theta_i - \varphi \nabla M_{Aggregate}(\theta_i)$, where the objective minimization was over the *p*-value derived from an unpaired t-test employing $M_{Aggregate}(\theta_i)$. However, we noticed the following interesting result. Namely, the function appeared weakly convex over a wide range of values for parameters $\alpha, \beta,$ and γ, all of which provided extremely similar results. This was confirmed via an extensive uniform grid search over this parameter space, alleviating concerns of over-fitting. For simplicity, we therefore set α=β=γ=1, yielding a final measure of:

$$M_{Aggregate} = \bar{M}_{Prim(1)} + \bar{M}_{SO(1,2)} + \bar{M}_{Euc(3,4)}$$

We conducted intensive ANOVA-based permutation tests to validate this measure. The effects of the combinations of the measures are shown in Table 2.

Statistical Analysis

We examined each of these terms in isolation and in combination on the AVLT results. Type III sum of squares analysis of variance (ANOVA) was performed accounting for family history, genetic biomarkers, age, sex and education level as predictors and the measures as the response variable.

prim t1	**Euc t3-4**	**SO t1-2**	**SO t3-4**	***p*-value**
✓	✓			2.83×10^{-05}
✓		✓		0.0016
✓			✓	0.2844
	✓	✓		3.69×10^{-05}
	✓		✓	0.2415
✓	✓	✓		7.00×10^{-06}
✓	✓		✓	0.00658

Table 2. Measure combination *p*-value from permutation test. Three combinations have family history *p*-values lower than the lowest individual measure which is Euclidean measure trial 3-4 (*p*=5.07 × 10^{-4}). These include primacy trial 1 and Euclidean trial 3-4 (*p*=2.83 × 10^{-5}), Euclidean measure trial 3-4 and subjective organization trial 1-2 (*p*=3.69 × 10^{-5}), and primacy trial 1, Euclidean trial 3-4 and subjective organization trial 1-2 (*p* =7.00 × 10^{-6}). prim = primacy, Euc = Euclidean, SO = subjective organization, t = trial, t*x*-*y* = comparing two trials.

We used a permutation test to compute *p*-values as proposed by Fischer, employing 10^7 permutations (Cox & Hinkley, 1979; Fisher, 1935). Namely, we permuted the labels of the given predictor and repeatedly derived the *p*-values as the test-statistic using ANOVA. We calculated the percentage of permutations where a *p*-value was returned with a lower value than our original ANOVA *p*-value. This is known as the *Fisher p-value* and its iterated computation provides a far more meaningful rejection of the null-hypothesis than a single use of an unpaired t-test. It strongly demonstrates that the predictors and the labels are not independent of one another. Additionally, we derived the pairwise Pearson correlation coefficients of the constituent measures to confirm that they are capturing different cognitive phenomena.

Results

One major outcome of this effort is that we achieved a highly reliable Fisher *p*-value of 7.00×10^{-6} for our aggregate measure, $M_{Aggregate}$. However, while separating family history from control populations has been the primary interest of prior work concerning AD, our concern is focused on the contribution of each term in this aggregate towards separating these populations.

Specifically, by using this framework, we can measure the information provided by each term towards the result of the hypothesis test. To this end, we determined their Fisher *p*-values in isolation, as shown in Table 1, which summarizes results of ANOVA for primacy, subjective organization and the Euclidean measure individually. It is clear that these measures are differentially informative across the trials, whereas their aggregate, summary scores are far less so. Surprisingly, primacy is only informative in the first trial, while subjects are still acclimatizing to the experiment.

	trial 1-2 SO	trial 3-4 SO	trial 3-4 Euclidean
trial 1 primacy	0.188 (0.135, 0.242)	0.162 (0.107, 0.216)	0.0277 $(-2.81\times10^{-2}, 8.34\times10^{-2})$
trial 1-2 SO		0.254 (0.201, 0.306)	-0.0134 $(-6.92\times10^{-2}, 4.24\times10^{-2})$
trial 3-4 SO			-0.26 (-0.312, -0.207)

Table 3. Pearson Correlation between measures (95% confidence intervals). The correlation between the significant measures and trial were between -0.226 and 0.301. These low Pearson correlation coefficients show weak correlation between any of these two measures. SO = subjective organization.

This leads us to question what precisely is being measure here. Similarly, the disconnect between informative trials for subjective organization seems to indicate that it is a proxy for some yet unknown measure. On the other hand, the extremely low *p*-value for Euclidean distance (and many other measures, as discussed below) leads us to believe that something happening here is so significant that one can almost not help but notice it. Clearly, a significant cognitive transition is occurring at this point, but it would be premature to attribute a cause to it.

Table 2 summarizes ANOVA results using the combination measures, including the one corresponding to $M_{Aggregate}$. This table allows us to examine how incorporating various measures increases separability differentially. The fine-grained Euclidean distance between trials 3 and 4 dominates clearly here; it provides the strongest signal for distinguishing these populations. This is the case even though there was no prior basis for expecting the difference between trials 3 and 4 was the single most important factor in distinguishing these populations. Thus, a simple machine learning approach applied to this problem, accompanied by a rigorous statistical analysis, revealed a far more nuanced cognitive transition than has ever been previously apparent in this test. We discuss further consequences of this below.

Finally, we note that Table 3 presents Pearson correlation coefficients between these terms using a 95% confidence interval. This demonstrates they are largely uncorrelated (i.e., they are measuring different effects). The largest absolute correlation is 0.312, which is considered small for the Pearson coefficient in cognitive test (Cohen, 1988).

Conclusions

This paper has made three primary claims:

1) Using aggregate scores in repeated neuropsychological testing can be highly misleading. Rather, examining individual trials and the differences between them can be far more informative than summary measures.
2) In tests with a relatively short number of repetitions, we believe acclimatization effects will be conflated with expected test results, particularly in early trials. This reinforces the point in (1) and stresses the need to look for "signals" in the results that may reflect a transition from reliance on working memory to engagement of secondary memory processes or are indicative of other cognitive phenomena. It is clear that different, independent measures were sensitive to different aspects of the learning and recall process. Note that we do not claim to understand why the transition from trials three to four is so significant. Clearly, further investigation is called for.
3) Postulating "expected" cognitive phenomena, such as Subjective Organization, may not be the most profitable avenue for analyzing neuropsychological testing results. Rather, there is value in "listening" to the data. Namely, by looking for signals that demonstrate a significant event has occurred, we may arrive at new understandings for cognitive phenomena underpinning the test that could not have been expected a priori.

A contribution of this paper is the demonstration that a simple machine learning framework, along with a rigorous statistical treatment, can reveal previously unknown cognitive phenomena. We note that the a variety of measures more exotic than Euclidean distance, such as Smith-Waterman alignment (Durbin, Eddy, Krogh, & Mitchison, 1998), were highly sensitive to the transition between trials 3 and 4, sometimes decreasing *p* by orders of magnitude. Thus, there appears to be something highly significant happening at this point in the test. It is interesting that a similar strong "transition" signal in a later trial has been shown to be highly significant in another neuropsychological test (Coen et al., 2009). This transition may be indicative of a shift in how subjects are approaching the test; reflecting a transition from reliance on working memory to engagement of secondary memory processes; or demonstrating cognitive adaptation or other effects.

More speculatively, because we can demonstrate that each additional measure contributes something new, we are constructing more informative methods for separating populations in neuropsychological tests. Our goal is to explore minimizing the Bayes error between the groups to the point where we can tentatively classify individuals, rather than distinguish populations.

While this paper has demonstrated clear benefits with respect to AVLT evaluation, we believe its approach is quite general and can be applied to a variety of conventional neuropsychological tests. As such, it supports the view that performance measures should not be viewed as competing with one another. Rather, each evaluation method can tell a different story about a patient's performance during the dynamic and complex cognitive processes involved in neuropsychological testing.

Acknowledgements

This research was supported by NIH grants 1UL1RR025011 (University of Wisconsin Clinical and Translational Science Award) and M01RR03186 (University of Wisconsin General Clinical Research Center), Helen Bader Foundation, Northwestern Mutual Foundation, Extendicare

Foundation. Additional support included the Wisconsin Alumni Research Foundation, the School of Medicine and Public Health, the Department of Biostatistics and Medical Informatics, and the Department of Computer Sciences at the University of Wisconsin-Madison. We gratefully acknowledge the assistance of Rebecca Koscik and Erin Jonaitis. The study protocol was approved by the institutional review board at the University of Wisconsin Hospital and Clinics and written informed consent was obtained from each patient (HSC #H-2006-0202 on 3/2009 and HSC #2001-329 on 11/2009).

References

Benedict, R., & Zgaljardic, D. (1998). Practice effects during repeated administrations of memory tests with and without alternate forms. *Journal of Clinical and Experimental Neuropsychology, 20*(3), 339-352.

Bertsekas, D., & Nedic, A. (2003). *Convex Analysis and Optimization*. Nashua, NH: Athena Scientific.

Bigler, E. D., Rosa, L., Schultz, F., Hall, S., & Harris, J. (1989). Rey-Auditory Verbal Learning and Rey-Osterrieth Complex Figure Design performance in Alzheimer's disease and closed head injury. *Journal of clinical psychology, 45*(2), 277.

Capitani, E., Della Sala, S., Logie, R. H., & Spinnler, H. (1992). Recency, primacy, and memory: reappraising and standardising the serial position curve. *Cortex; a journal devoted to the study of the nervous system and behavior, 28*(3), 315-315.

Coen, M. H., Selvaprakash, V., Dassow, A. M., Prudom, S., Colman, R., & Kemnitz, J. (2009). *Modeling the Role of Memory Function in Primate Game Play*. Paper presented at Thirty-First Annual Conference of the Cognitive Science Society, Amsterdam, Netherlands.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*: Lawrence Erlbaum.

Cox, D. R., & Hinkley, D. V. (1979). *Theoretical statistics*: Chapman & Hall/CRC.

Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (1998). *Biological sequence analysis*: Cambridge Univ. Press.

Fisher, R. A. (1935). The design of experiments. 1935. *Oliver and Boyd, Edinburgh*.

Hermann, B. P., Seidenberg, M., Wyler, A., Davies, K., Christeson, J., Moran, M., & Stroup, E. (1996). The effects of human hippocampal resection on the serial position curve. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 32*(2), 323-334.

La Rue, A., Hermann, B., Jones, J. E., Johnson, S., Asthana, S., & Sager, M. A. (2008). Effect of parental family history of Alzheimer’s disease on serial position profiles. *Alzheimer's & dementia: the journal of the Alzheimer's Association, 4*(4), 285.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology, 34*(7), 939-944.

Milner, B. (1964). Some effects of frontal lobectomy in man. *The frontal granular cortex and behavior*, 313-334.

Mitchell, M. L., & Jolley, J. M. (2009). *Research Design Explained*: Wadsworth Publishing.

Mitrushina, M., Satz, P., & Van Gorp, W. (1989). Some putative cognitive precursors in subjects hypothesized to be at-risk for dementia. *Archives of Clinical Neuropsychology, 4*(4), 323-333.

Ott, R.L., & Longnecker, M.T. (2001). An Introduction to Statistical Methods and Data Analysis. DuxburyPress.

Ramakers, I. H., Visser, P. J., Aalten, P., Bekers, O., Sleegers, K., van Broeckhoven, C. L., . . . Verhey, F. R. (2008). The association between APOE genotype and memory dysfunction in subjects with mild cognitive impairment is related to age and Alzheimer pathology. *Dement Geriatr Cogn Disord, 26*, 101-108.

Ramakers, I. H., Visser, P. J., Aalten, P., Maes, H. L., Lansdaal, H. G., Meijs, C. J., . . . Verhey, F. R. (2010). The predictive value of memory strategies for Alzheimer's disease in subjects with mild cognitive impairment. *Archives of clinical neuropsychology : the official journal of the National Academy of Neuropsychologists, 25*(1), 71-77.

Rey, A. (1964). L'examen clinique en psychologie [The clinical examination in psychology]. *Paris: Presses Universitaires de France*.

Sager, M. A., Hermann, B., & La Rue, A. (2005). Middle-aged children of persons with Alzheimer's disease: APOE genotypes and cognitive function in the Wisconsin Registry for Alzheimer's Prevention. *Journal of Geriatric Psychiatry and Neurology, 18*(4), 245-249.

Thurstone, L. L. (1927). Psychophysical analysis. *The American journal of psychology, 38*(3), 368-389.

Wasserman, L. (2004). *All of statistics: a concise course in statistical inference*. New York, NY: Springer Verlag.

Woodard, J. L., Dunlosky, J. A., & Salthouse, T. A. (1999). Task decomposition analysis of intertrial free recall performance on the Rey Auditory Verbal Learning Test in normal aging and Alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology, 21*(5), 666-676.