Characterizing the effects of sex, *APOE* ε4, and literacy on mid-life cognitive trajectories: Application of Information-Theoretic model-averaging and multi-model inference techniques to the Wisconsin Registry for Alzheimer's Prevention Study

Co-authors: Rebecca L. Koscik[a], Derek L. Norton[b], Samantha L. Allison[a], Erin M. Jonaitis[a], Lindsay R. Clark[c,a,d], Kimberly D. Mueller[a], Bruce P. Hermann[a,e], Corinne D. Engelman[a,f], Carey E., Gleason[c,d], Mark A. Sager[a,d], Richard J. Chappell[b,h], & Sterling C. Johnson[c,a,d]

Affiliations:

[a]Wisconsin Alzheimer's Institute, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA

[b]Department of Biostatistics and Medical Informatics, University of Wisconsin School of Medicine and Public Health, Madison, WI USA

[c]Geriatric Research Education and Clinical Center, William. S. Middleton Memorial Veterans Hospital, Madison WI, USA

[d]Alzheimer's Disease Research Center, University of Wisconsin-Madison School of Medicine and Public Health, Madison, WI, USA

[e]Department of Neurology, University of Wisconsin School of Medicine and Public Health, Madison, WI 53705, USA

[f]Department of Population Health Sciences, University of Wisconsin School of Medicine and Public Health, Madison, WI 53726, USA

[h]Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA

Corresponding author: Rebecca Koscik, 610 Walnut St. (Rm. 944), Madison, WI, 53726

Office ph.: 608-262-6953; fax: 608-265-9122 ; e-mail: rekoscik@wisc.edu

Manuscript body word count: 5266

Abstract word count: 246

Mesh terms:

Kullback-Leibler divergence, model averaging, model likelihoods, model selection, cognitive decline, Alzheimer's disease

Abstract

In this paper we apply Information-Theoretic (IT) model averaging to characterize a set of complex interactions in a longitudinal study on cognitive decline. Prior research has identified numerous genetic (including sex), education, health and lifestyle factors that predict abnormal cognitive decline. Traditional model selection approaches (e.g., backward or stepwise selection), attempt to find models that best fit the observed data; these techniques risk interpretations that only the selected predictors are important. In reality, several models may fit similarly well but result in different conclusions (e.g., about size and significance of parameter estimates), and inference from a single model chosen after a selection procedure can lead to overly simplistic conclusions. Here we utilize longitudinal cognitive data from the Wisconsin Registry for Alzheimer's Prevention to examine the effects of sex and the Apolipoprotein E (APOE) ε4 allele (non-modifiable factors), and reading literacy achievement (modifiable) on cognitive decline. For each outcome, we applied IT model averaging to a model set with combinations of interactions among sex, APOE, literacy, and age. For a list-learning test, model-averaged results showed better performance for women vs men, with faster decline among men; increased literacy was associated with better performance, particularly among men. APOE had less effect in this age range (~40-70). These results illustrate the utility of the IT approach and point to verbal ability as a potential modifier of decline. Whether the protective effect of literacy is due to educational attainment or intrinsic verbal intellectual ability is the topic of ongoing work.

Introduction

Signs of cognitive decline often begin a decade or more before diagnosis of late-onset Alzheimer's disease (AD), a neurodegenerative disease associated with greatly impaired cognition and daily functioning (Price et al., 2009; Price & Morris, 1999). After age, presence of one or more *APOE* ε4 alleles is the strongest predictor of risk of late onset AD (Strittmatter & Roses, 1996; Tang et al., 1996). Several studies report earlier and faster declines in memory or executive function among *APOE* ε4 carriers than non-carriers (i.e., *APOE* by age interactions), with detectable accelerations in decline beginning around age 60 (Caselli et al., 2009; Chang et al., 2014; Wisdom, Callahan, & Hawkins, 2011). Studies also indicate that the rate of ε4-associated cognitive decline and AD risk are moderated by sex with ε4-carriage increasing rate of decline and risk more in women than in men (Altmann, Tian, Henderson, & Greicius, 2014; Beydoun et al., 2012; Koran, Wagener, & Hohman, 2017; Mielke, Vemuri, & Rocca, 2014; Mortensen & Høgh, 2001; Neu et al., 2017; Payami et al., 1996; Riedel, Thompson, & Brinton, 2016). Higher literacy levels (as measured by word reading recognition tasks) have been shown to mitigate age- and/or *APOE*-related cognitive decline in non-demented elders (Kaup et al., 2015; Manly, Touradji, Tang, & Stern, 2003).

No studies, to our knowledge, have investigated the *combined* influences of sex, *APOE* genotype, and literacy on early to late middle-age cognitive trajectories in an integrated analytic framework. Traditional model selection approaches, such as forward, backward, or stepwise selection, attempt to find a model that best fit the observed data. One issue that can arise from such methods is concluding that only the selected predictors are important while assuming those not selected are unimportant (D. R. Anderson & Burnham, 2002). In reality, several models may fit the data similarly well but result in different conclusions, and inference from a single model chosen after a selection procedure can lead to overly optimistic results and conclusions (Claeskens and Hjort 2008). Information-theoretic (IT) modeling techniques offer a way to characterize complex sets of interactions and make multi-model

inference while avoiding the pitfalls of predictor selection methods (D. R. Anderson & Burnham, 2002; Claeskens & Hjort, 2008).

The IT framework focus is on "a small set of science hypotheses, all of which are plausible" (p. 202, (D. R. Anderson & Burnham, 2002)). The IT approach has its roots in biological ecology research (Kenneth P. Burnham, Anderson, & Huyvaert, 2011; Hegyi & Garamszegi, 2011; Richards, 2005; Richards, Whittingham, & Stephens, 2011; Symonds & Moussalli, 2011), and aims to use the relative strength of information among all considered models instead of selecting a single model. The methods start with formulating a reasonably sized collection of models with the same outcome, but different covariate structures (such as different main effects, interactions, etc.). The differences between the models should all be items of interest and scientifically reasonable to include. After fitting all of these models to the data, results are combined across models in proportion to the relative strength of information each model provides. These relative strengths are quantified through the theory of the corrected Akaike Information Criterion (AICc), which estimates the relative differences amongst these models with respect to their Kullback-Leibler divergence, a measure of the distance between a proposed model and the “true” model (Hurvich & Tsai, 1989) . This allows models fitting similarly well to contribute relatively equal amounts of influence on the resulting parameter estimates, while models that fit poorly contribute little or practically nothing to the results.

The aim of this study was to use IT model-averaging to characterize how sex, literacy, and *APOE* genotype influence age trajectories for several neuropsychological tests of cognitive function in a longitudinal sample enriched for risk of developing AD (Wisconsin Registry for Alzheimer’s Prevention (WRAP)). In secondary analyses, we compare the IT model-averaging results with traditional model selection methods.

Methods

WRAP study and participants

WRAP is a longitudinal cohort study enriched for AD-risk via over selection of participants with a parental history of AD (for details, see (Johnson et al., 2017); primary aims of the study include identifying predictors associated with abnormal cognitive decline and estimating their associations. All participants were free of dementia at baseline. At the time of these analyses, there were 1549 WRAP participants in the whole sample (baseline age mean(sd)= 53.7(6.6); parental history of AD n(%)= 1125(72.6%)). To be eligible for inclusion in these analyses, participants had to be free at baseline of MCI and any of four neurological conditions (stroke, Parkinson's disease, multiple sclerosis, epilepsy), have completed at least 2 study visits, and have complete data in the predictors needed for the analyses (n=1256 eligible; n's excluded were as follows: MCI, n=4; neurological disorder, n=59; <2 visits, n=226; incomplete predictors, n=4). Depending upon the outcome, additional subjects were excluded due to missing outcome values resulting in <2 visits with a value (details in Table 1).

[Insert Table 1 approximately here]

Study protocol and outcomes

At each study visit, participants complete comprehensive cognitive assessments, detailed health and lifestyle questionnaires, and provide blood samples for current and future analysis. The first follow-up visit occurs approximately 4 years after baseline, with subsequent visits occurring approximately every 2 years (for details, see (Johnson et al., 2017)). These analyses focus on nine cognitive measures available since baseline of the WRAP study and shown to be sensitive to AD-related cognitive changes. The tests and measures used are: the Rey Auditory Verbal Learning and Memory Test (Schmidt, 1996), sum of learning trials ("AVLT Total") and long delay ("AVLT Delay"); Trail Making Test (Lezak, Howieson, Bigler, & Tranel, 2012) (Reitan, 1958), Trails A and Trails B (seconds to complete); Stroop Color-Word Interference Test (Trenerry, Crosson, DeBoe, & Leber, 1989), number of correct items in two minutes; Controlled Oral Word Association Test (Benton, Hamsher, & Sivan, 1994), total words in 60 seconds for

each letter: C, F, L ("CFL"); the Boston Naming Test (Kaplan, Goodglass, & Weintraub, 2001),"BNT"), and Digit span forward and backward (Wechsler, 1997).

Given recent results in WRAP, ADNI, and other studies suggesting that higher intraindividual cognitive variability (IICV) at a given visit predicts increased risk of subsequent decline (E. D. Anderson et al., 2016; Gleason et al., 2017; Holtzer, Verghese, Wang, Hall, & Lipton, 2008; Koscik et al., 2016), we also wished to characterize how IICV varied by sex, literacy, *APOE,* and age in our sample. We calculated two versions of IICV. The first represents the version we used in a previous paper (E. D. Anderson et al., 2016; Gleason et al., 2017; Koscik et al., 2016), "4 Test IICV", and was calculated as the standard deviation of z-scores of AVLT Total, Trails A and B, and the Wide Range Achievement Test (3rd ed., "WRAT") reading subtest standard score (Wilkinson, 1993). The WRAT Reading Achievement score when used in middle-aged and older adults is accepted as a proxy for maximal verbal intellectual attainment and quality of education. Higher WRAT Reading scores have been associated with slower memory declines in non-demented elders (Manly et al. 2003). The second IICV version was calculated as the standard deviation of z-scores of the nine tests selected for analysis in this manuscript (9 Test IICV). Before z-scoring occurred, all tests were assessed for normality, and Box-Cox transformed for non-normal items (Trails A and B, BNT, Digit Span Forwards, WRAT, and AVLT Delay plus 1).

Key Predictors and Covariates

Key predictors in this study include age (years), chromosomal sex (M/F), APOE ε4 allele count (i.e., 0, 1, 2 (for details on genotyping methods, see (Darst et al., 2017)), and literacy as measured by WRAT-III Reading test. For subjects without a baseline literacy value, the value at their second visit was used. Additional covariates included include race/ethnicity (non-Hispanic Caucasian vs underrepresented group, URG), if English was the subject's native language (ESL), and socioeconomic status (SES; : 1 = <$20k; 2 = $20k-<$40k; 3= $40k-<$60k; 4=$60k-<$80k, and 5= $80k or more). Missingness in SES was present in 54 (4.3%) participants. To recover these subjects for analysis, their

baseline SES values were imputed through proportional odds regression using baseline values of age, chromosomal sex, race/ethnicity, the Center for Epidemiological Studies of Depression (CES-D) total score (Radloff, 1977), literacy, and years of education.

Data analysis

The steps of the IT-modeling approach are detailed below.

1. Specifying the model set. In preliminary work, we developed an initial set of 12 models based on research indicating non-linear age effects and potential interactions among age, sex, and/or *APOE* genotype among domains represented by our cognitive outcomes (Supplemental Table 1). Based on those results and additional research indicating potential interactions between literacy and sex, *APOE* count or age (on cognitive outcomes), we developed a revised set of 28 research-supported hypothesized models about the relationships between sex, literacy, *APOE*, and age-related cognitive decline (Table 2). We then proceeded with steps 2-7 for each of our cognitive outcomes.

[Insert Table 2 approximately here]

2. Fit each model and check model assumptions. All models used a mixed effects structure, with the fixed effects for each model in the set as specified by Table 2, and subject specific intercepts and age-related slopes as random effects. For outcomes of AVLT Delay and Boston Naming, logistic regression mixed models were used to address the discrete nature and ceiling/floor effects present in the data; the other 9 outcomes used standard linear mixed effects models. For all models, SES was treated as continuous, age and SES were centered to their baseline means, and their associated quadratic terms calculated from these centered values. *APOE* was categorized as ε4 allele count (0, 1, or 2). Each model was fit to the data by maximum likelihood, and model diagnostics were performed on the largest model (model with the most parameters) and the "best fitting" model (lowest AICc value). Diagnostics included checking for homoscedastic and normally distributed residuals (deviance residuals

vs link for logistic regression), outliers, normally distributed random effects, no correlation between random effects and residuals, and overdispersion for the logistic regression. CFL fluency and both IICV measures were square-root transformed to address residuals issues. After diagnostic checks, Stroop Color Word was removed from analysis due to several residuals violations that could not be addressed with reasonable transforms.

Even after reasonable transformations, some issues related to outliers and correlations with random effects persisted. Correlations, up to ~0.3 in magnitude, between the random effects and residuals were noted for all outcomes. Positive correlations, up to ~0.2, between random effects and age / quadratic age were also noted for all outcomes, indicating the complex relationship between age and cognitive tests is not fully captured by these linear mixed effects structures. CFL also had several large positive residual outliers associated with a single subject. A sensitivity analysis for CFL removed this subject and reperformed the entire algorithm; because results did not change in any meaningful way, CFL results presented here include this subject.

3. Extract model statistics for each model. For each model in the set, the extracted model statistics included the number of model parameters (k), Akaike's Information Criterion-corrected (AICc), and the log likelihood statistic. AICc is based on the Kullback-Leibler (K-L) divergence, which is a measure of information loss when model 'g' is used to approximate the true data generating model, model 'f' (Kenneth P. Burnham & Anderson, 2003). For data with n observations and fitted regression model 'g' with k parameters, the formulas for AIC and for AICc are (K. P. Burnham & Anderson, 2002):

$$AIC = -2\ln\left(p\left(y\middle|\hat{\theta}\right)\right) + 2k \qquad AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

where y is the observed data, and $\hat{\theta}$ are the maximum likelihood estimates of the k parameters from model 'g'.

4. Calculate Δ_j's. The minimum AICc across the model set was used to calculate the difference between the best fitting model (i.e., that with the minimum AICc) and model j. For model j:

$$\Delta_j = (AICc\ of\ model\ j) - (minimum\ AICc\ value\ in\ set)\ \ , for\ j = 1, 2, \ldots J\ models\ in\ the\ set$$

5. Calculate model weights. Δ_j's for each model are used to calculate the Akaike weights (w_j's) of all models in the set. Heuristically, w_j represents the likelihood that model j is the K-L best model in the set. The w_j's helps quantify model uncertainty and are used to combine information across the set of models. w_j is calculated as:

$$w_j = \frac{exp(-\Delta_j/2)}{\sum_k exp(-\Delta_k/2)}$$

6. Combine results across models. Results from each model are multiplied by their corresponding weight (w_j), and all weighted results are then summed together for the final model averaged result. When a regression parameter does not appear in a particular model, but does in one or more of the other models in the set, that parameter is set to zero in that particular model. Thus, for the ith regression parameter in one model, the model averaged i^{th} regression parameter is:

$$\bar{\hat{\beta}}_i = \sum_j \omega_j \hat{\beta}_{i(j)}\ \ , where\ \hat{\beta}_{i(j)} = 0\ if\ \beta_i\ does\ not\ appear\ in\ model\ j$$

7. Confidence intervals and inference using model averaged results. To facilitate multi-model inference, model-averaged estimates were evaluated using 95% CI's. CI's for model averaged results were obtained through non-parametric bootstrapping. For each outcome, the data used to fit the model sets was first stratified at the subject level by the total number of visits (2, 3, 4, or 5) each subject had. Within each stratum, subjects were selected, with replacement, back to the number of subjects within that stratum, thus preserving the original number of subjects, observations on subjects, and distribution of follow-up visits. Each bootstrap replicate went through the above process (steps 2-6); 10,000 bootstrap replicates were performed for each outcome. Bootstrap quantiles were used to calculate CI's (using linear interpolation when necessary). Standard CI interpretation methods were used for inference about regression parameters (i.e., CI's that did not overlap with 0 were considered significant at the α=0.05 level).

Bootstrap rationale. While Burnham and Anderson propose analytical methods for model-averaged inference and CI's, these methods are predicated upon assumptions of a limiting normal distribution for the model-averaged estimate(K. P. Burnham & Anderson, 2002). Clauskens and Hjort have shown that, unless one assumes the model weights (w_j's) used are fixed and not random quantities, there is no guarantee of a limiting normal distribution(Claeskens & Hjort, 2008). We were not willing to assume that model weights are fixed quantities. Thus, we utilized bootstrapping methods similar to those proposed by Burnham and Anderson(K. P. Burnham & Anderson, 2002), but chose to, within each bootstrap replicate, reapply the process that generated the original model averaged statistic, instead of performing typical model selection processes within each replicate.

Secondary analyses. ***Comparison of modeling methods.*** To illustrate how results from the above approach differ from some traditional approaches, we compared IT model averaged results for AVLT Total, AVLT Delay, and Trails B, with results obtained via single regression models determined by a best fitting model approach and by a backwards elimination approach. In both methods, the same general mixed effects model structure and use of maximum likelihood fits were used. For the best fitting approach, the model selected was that out of the candidate set which has the minimum AICc value. For backwards elimination, the starting model included all covariates that appeared in any of the models within the candidate set, and the criteria for covariate elimination was which covariate reduced AICc by the largest amount, while preserving the hierarchy of higher order terms. Elimination stopped when removal of any remaining covariate did not reduce AICc. For both methods, CI's and inference from the resulting single models utilized the asymptotic normal properties of regression estimates.

Type I error assessment. No family-wise error rate correction was performed; however, an assessment of the findings in relation to type I error rates was done. For each outcome, 27 unique non-intercept coefficients defined by the model set were examined. Thus, 270 unique coefficients were examined across all 10 outcomes: 100 main effects, 110 two-way interactions, and 60 three-way

interactions. For each of these coefficient groupings, the binomial distribution was used to examine how often one would expect to detect at least the number of significant coefficients found in these analyses (at the 0.05 level), assuming the global null hypothesis that all coefficients are truly zero.

Software used. All analyses were performed using R version 3.4.0. Proportional hazard models were fit using 'polr' in the MASS package; mixed effect regression were fit using 'lmer' and 'glmer' in the lme4 package; AICc-based model statistics were calculated using 'aictab' in the AICcmodavg package; baseline characteristics and tests were computed using the tableone package; plots were created using the ggplot2 package; bootstrapping was performed utilizing HTCondor version 8.6.3.

Results

Sample Characteristics

Sample characteristics are shown in Table 1 overall and by sex. Men and women did not differ in terms of age, follow-up, proportion URG or ESL, and IICV. The sexes differed on *APOE* ε4 count, SES, and all cognitive tests at baseline. Women performed better than men at baseline on all tests except BNT and digit span (both forwards and backwards).

Model-averaged Results

Model averaged coefficients and corresponding 95% CI's for all non-intercept terms are shown in Tables 3 (AVLT Total, AVLT Delay, $\log_{10}$Trails A and $\log_{10}$Trails B, and CFL) and 4 (Digit Span forward and backward, BNT, 4 Test and 9 Test IICV). In each table, 95% CI's that exclude 0 are identified by bold-face; gray shading denotes CI's containing 0. AICc's and weights for all models in the set are presented in Supplemental Table 2 for all outcomes. Results are summarized below; for each outcome, results are supported by a two-panel figure (left-panel depicts 95% CIs that exclude 0 and right-panel depicts predicted outcome scores by age and selected predictors using model-averaged estimates).

[Insert Tables 3-4 approximately here]

Memory. For AVLT Total, Model 21 was best-fitting, contributing a weight of .562 to model averaging; Model 26 contributed .183 and Model 25 contributed .139 (weights in supplemental Table 2); all other model weights were under .05. Significant interactions included ε4 count 1*quadratic age, sex*age, and sex*literacy (left-panel, Figure 1). Figure 1 (right-hand panel) depicts AVLT Total model-averaged predicted performance for the latter two interactions since the estimated beta for the interaction with age and ε4 count 1 was ostensibly 0. There's a larger gap in AVLT Total scores for high vs low literacy in men than in women and men show faster AVLT Total age-related decline than women.

[Insert Figure 1 approximately here]

The best fitting model for AVLT Delay was also Model 21 (weight=.531), followed by Model 15 (.123) and Model 25 (.119). Variability in AVLT Delay was explained largely by SES, sex, literacy and age, with age-related AVLT Delay decline steeper for men than women (see Supp. Fig. 2 for CI's for and predicted trajectories).

Executive function. The best fitting models for Trails A were Models 19 (weight=.574), 21 (.248), and 25 (.064). The left panel of Supplemental Figure 3 depicts significant beta estimates and their CI's while the right-hand panel depicts $\log_{10}$(Trails A) predicted values vs age, stratifying by sex and literacy (high vs low). Predicted values indicate that women of high literacy perform worse on Trails A then women of low literacy while the opposite is true for men. Although the interactions of sex*quadratic age, and sex*age* *APOE ε4* counts of 1 or 2 were statistically significant, all had estimated betas extremely close to 0 (Table 4; Supp. Fig. 3).

The best fitting models for Trails B were Models 6 (weight=.330), 26 (.300), 12 (.124), and 19 (.109). Variability in Trails B was explained by SES, URG status, sex, literacy, age (linear and quadratic), and a 3-way interaction with sex**APOE ε4* count 1*age (left-hand panel of Supp. Figure 4). Though significant, the estimated beta for the sex*age**APOE* interaction was again close to 0. Predicted age-related trajectories show slower times for men than women and lower vs higher literacy (right-hand

panel, Supp. Fig. 4; no predicted values are shown for men and *APOE* ε4 count=2 due to the small cell size (n=4)).

Working memory. The best fitting models for Digit-span Forwards were Models 6 (weight=.307), 26 (.234), 19 (.163), and 12 (.124). Variability in Digit-span Forwards was explained by sex, age (linear and quadratic) and literacy; significant interactions included sex*literacy and literacy*quadratic age. The latter beta estimate was near zero; sex differences in the outcome are negligible at low literacy levels while men outperform women at high literacy levels (Supp. Figure 5). The best fitting models for Digit-span Backwards were Models 13 (weight=.583), 15(.168), 20(.117), and 15(.104). Variability in Digit-span Backwards was explained by quadratic age, literacy, and sex with greater sex differences at higher literacy levels (Supp. Figure 6).

Language. The best fitting models for CFL were Models 13(weight=.416) and 14 (.251). Variability in CFL was explained by age, sex, literacy, and interactions sex*_APOE_ count 1, sex*literacy, and sex*_APOE_ count 1*quadratic age, although the three-way beta estimate was essentially 0 (Supp. Figure 7). Women did better overall and sex differences were smaller in ε4 count=0 than count=1. All improved with age. The best fitting models for BNT were Models 6 (weight=.630), 12 (.239), and 8 (.128). Variability in BNT was explained by SES, URG status, sex, age, ESL status, and literacy level with higher literacy, SES, male sex and older age associated with better performance, while URG and ESL status were associated with lower BNT scores (Supp. Figure 8).

Intraindividual variability. The best fitting model for 4 Test IICV was Model 26 (weight=0.959); all other models contributed <0.05 to the parameter weights. Variability in 4 Test IICV was accounted for by sex, age, and literacy, with significant sex*quadratic age, sex*literacy, and literacy*age interactions (Figure 2). At lower levels of literacy, women show higher 4 Test IICV than men and IICV declines with age; in contrast, IICV increases with age at higher literacy levels and is at times higher among men than women (Figure 9). The best fitting model for 9 Test IICV was also Model 26 (weight=0.890; Supp. Figure

9); significant parameters indicate that 9 Test IICV patterns differed most between high and low literacy men.

[Insert Figure 2 approximately here]

Model comparisons. In secondary analyses of 3 outcomes, we compared estimated betas and 95% CI's between the IT, best fit, and backward selection approaches for coefficients that were significant in any of the methods, per outcome. Graphs of point estimates and CI's for the three approaches are provided in Supplementary Figures 10, 11 and 12, for AVLT Total, AVLT Delay, and Trails B, respectively.

Type I error assessment. Out of the 100 unique main effects examined across all 10 outcomes, 42(42%) were significant. Under the global null hypothesis for all coefficients, the probability of detecting at least 42 coefficients at the 0.05 level is 2.50×10^{-29}. For two-way interactions, 15(13.6%) of 110 unique coefficients were significant, corresponding to a probability of 0.00012 under the global null. For three-way interactions, 4(6.67%) of 60 unique coefficients were significant, corresponding to a 0.180 probability under the global null.

Discussion

In this study, we used information-theoretic (IT) model averaging techniques to characterize how sex, *APOE,* and literacy modify age-related cognitive and IICV trajectories in a sample that was middle-aged and free of clinical impairment at baseline assessment(mean(sd) age at baseline and last visit: 53.7(6.6) and 62.8(6.8), respectively). In the WRAP data, we observed age-related declines for all cognitive outcomes except the two language-related measures (CFL and BNT). Age-related declines in IICV were associated with lower literacy levels while IICV tended to increase with age among participants with higher literacy. Significant but small quadratic age effects were observed for a few

outcomes. *APOE* ε4 count showed significant modifying effects on age-related trajectories on four outcomes, although the corresponding estimated interaction betas were ostensibly zero. Sex and literacy were consistently significant predictors of measures of memory, executive function, working memory, language and intra-individual cognitive variability including significant main effects, and sex*age and/or sex*literacy interactions for a subset of outcomes.

Compared to those with no *APOE* ε4 alleles, carriage of one or two ε4 alleles is associated with greater risk of AD (Neu et al., 2017) and faster or earlier cognitive decline in certain domains. For example, in a sample of cognitively normal adults (mean baseline age ~60 years and followed an average of ~5 years), Caselli et al (Caselli et al., 2009) reported accelerated age-related decline on AVLT Delay among *APOE* ε4 carriers (vs non-carriers) beginning prior to age 60. Predicted annual rate of AVLT Delay change, however, was very small for both carriers and non-carriers in age-ranges similar to our sample: e.g., 50-59, 0.07 vs 0.08 and 60-69, 0.04 vs -0.03 for non-carriers vs carriers, respectively). In a meta-analyses, Wisdom and colleagues (Wisdom et al., 2011) also reported increased age-related decline among *APOE* ε*4* carriers relative to non-carriers particularly at older ages with Cohen's d estimated effect sizes in the small range (i.e., |max|<.20; (Wisdom et al., 2011)). Results in our sample also showed small effects. Specifically, significant age-modifying effects of one *APOE* ε*4* allele (vs 0) were noted for AVLT Total, Trails A and B, and CFL, although beta estimates for these interactions were essentially zero (and not clinically meaningful)..

Previous research has also shown sex differences in risk of MCI (Roberts et al., 2012) and AD (e.g., (Altmann et al., 2014)), and rates of decline (e.g., (McCarrey, An, Kitner-Triolo, Ferrucci, & Resnick, 2016; Mortensen & Høgh, 2001)), with other studies showing evidence of potential sex*ε4 interactions (Payami et al., 1996) . In the WRAP sample, men performed significantly worse than women on all cognitive measures except digit span, BNT, and CFL. Sex*age interactions showed faster decline in men for both AVLT measures. These results are consistent with those reported for the Mayo Clinic Study of

Aging in which memory, memory decline, and hippocampal volume were worse among men compared to women (Jack CR, Jr, Wiste HJ, Weigand SD, & et al, 2015) A small number of significant sex**APOE**age interactions were observed, but beta estimates were essentially 0. As we follow the WRAP cohort into their 70's, larger effects may emerge for these complex interactions.

Literacy level as measured by word reading tasks is considered a proxy for verbal intellectual ability and quality of education; lower literacy has been associated with faster memory decline (Manly et al., 2003). Higher literacy levels have also been shown to be associated with resilience to *APOE* 4-related cognitive decline (Kaup et al., 2015; Vemuri P, Lesnick TG, Przybelski SA, & et al, 2014). Our results also indicated that better literacy was associated with better performance across cognitive outcomes. Although literacy did not modify age-related trajectories, benefits were stronger for men than women for AVLT Total, Trails A, CFL and the two digit span measures.

IICV has been shown to vary some over the life span (Salthouse & Soubelet, 2014). In our data, lower levels of literacy were associated with steady age-related decline in 4 Test IICV with higher IICV consistently observed among women; at higher literacy levels, IICV was generally higher among men and increased with age for both sexes. Our 4 Test IICV is the same as used in previous WRAP analysis and similar to what's been used in other analyses (Holtzer et al 2008; Anderson et al, Gleason et al, in press) indicating that IICV predicts MCI and AD. As IICV is developed as a potential biomarker for AD risk, researchers may examine whether risk-indicating value is constant across underlying demographics such as age, sex, and literacy.

Traditional model selection methods such as stepwise regression are prone to overfitting the data, producing overconfident estimates whose standard errors do not account for the degrees of freedom in the search process (Hastie, Trevor, Tibshirani, Robert, & Friedman, Jerome, 2009). Shrinkage methods, such as Lasso, can help select important predictors with respect to the outcome, but parsimonious models and predictive accuracy are the typical goal, and statistical inference can be

difficult. Bayesian model averaging combines information from posterior distributions of parameters of interest across several models, weighting each by its posterior model probability (Hoeting, Madigan, Raftery, & Volinsky, 1999). However, it is important is important in Bayesian methods to formulate reasonable prior distributions for all parameters and model probabilities, and this can be prohibitive when the set of models under consideration is large (Claeskens & Hjort, 2008). By using the IT approach in this paper, we obtained the robustness benefits of model averaging without the overhead of Bayesian methods, while still yielding familiar statistical outputs that support inferences (i.e. point estimates, CI's). IT methods also allow examination of several scientific questions of interest, with each question corresponding to its own model, but where all results are combined into a single set of standard outputs. When only one or a few model structures are necessary to handle the scientific question(s) of interest, such methods may be of little value considering the extra complexity and time required. However, since age-related decline is influenced by many factors, the IT approach offers an attractive framework for evaluating competing hypotheses about modifiers of decline.

Both aspects of our secondary analyses indicate that the IT results do not over-identify significant findings and may guard against overestimating effects. First, in our comparison of the IT approach with best fit and backwards elimination approaches, main effects estimates between the three methods were generally very similar, though the IT method tended to have the widest confidence intervals. Relationships between the three methods were more complex for quadratic effects and interactions. Backwards elimination commonly "found" interactions that best fit did not, while IT tended to attenuate the estimated interaction coefficient ostensibly to zero. In addition, the IT method rarely found such coefficients significant if the best fit method did not include it. Second, the numbers of significant main and two-way interaction effects detected in the IT method were well above those expected from random change under the global null resulting in a low estimated probability of type I error. This low probability lends confidence to conclusions about significant IT model-averaged effects,

especially for specific interactions that were detected across multiple outcomes. For all 4 significant three-way interactions, the IT method estimates have very tight CI's that, while not containing zero, are extremely close to zero, especially compared to estimates that were found by backwards elimination. This suggests that IT methods might further guard against overconfident results by estimating some coefficients to be ostensibly zero, even when the finding itself is significant.

The generalizability of our results is limited by cohort characteristics, including that our sample is highly educated, enriched for AD risk, has limited follow-up on participants from URG, and has few males homozygous for ε4. In addition, use of *APOE ε4* count is just one of many possible ways of parameterizing *APOE*-associated risk.

Conclusions

While carriage of an *APOE ε4* allele is known to be the greatest modifier of age-related risk for late onset AD, many factors may hasten or slow the rate of cognitive decline during mid-life. When several predictors and their interactions are considered simultaneously, certain model selection procedures can result in overestimation of significance of certain parameters and loss of information about others that are accounting for similar amounts of variability. The IT model averaging approach offers a framework that allows results from multiple plausible hypotheses to provide weighted model-averaged parameter estimates and CI's, with CI's excluding zero interpreted as significant. The application of this methodology to the WRAP sample using a set of models designed to characterize mid-life relationships among sex, *APOE*, literacy and age suggest preclinical age-related trajectories are modified more by sex and literacy levels than by *APOE* ε4 allele count in this age range. Future applications of the IT methodology will examine the interplay of sex and literacy with other potential cognitive trajectory-modifying variables such as polygenic risk, AD biomarkers, or lifestyle factors (e.g., exercise or diet).

Acknowledgments

This research was supported by the National Institutes of Health awards R01 AG027161, UL1 TR000427 and by donor funds including the Wisconsin Alzheimer's Institute Lou Holland Fund. Portions of this research were supported by resources at the Wisconsin Alzheimer's Institute, the Wisconsin Alzheimer's Disease Research Center and the Geriatric Research Education and Clinical Center of the William S. Middleton Memorial Veterans Hospital, Madison, WI. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the NIH or the Veterans Administration. We gratefully acknowledge the WRAP study team who have carefully acquired the longitudinal data, and the WRAP participants who make this research possible.

References

.Altmann, A., Tian, L., Henderson, V. W., & Greicius, M. D. (2014). Sex modifies the APOE-related risk of developing Alzheimer disease. *Annals of Neurology*, *75*(4), 563–573.

Anderson, D. R., & Burnham, K. P. (2002). Avoiding pitfalls when using information-theoretic methods. *The Journal of Wildlife Management*, 912–918.

Anderson, E. D., Wahoske, M., Huber, M., Norton, D., Li, Z., Koscik, R. L., … Asthana, S. (2016). Cognitive variability—A marker for incident MCI and AD: An analysis for the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *4*, 47–55.

Benton, A. L., Hamsher, K., & Sivan, A. B. (1994). *Multilingual Aphasia Examination: Manual of instructions . Iowa City, IA: AJA Associates*. Inc.

Beydoun, M. A., Boueiz, A., Abougergi, M. S., Kitner-Triolo, M. H., Beydoun, H. A., Resnick, S. M., … Zonderman, A. B. (2012). Sex differences in the association of the apolipoprotein E epsilon 4 allele with incidence of dementia, cognitive impairment, and decline. *Neurobiology of Aging*, *33*(4), 720–731. e4.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference* (2nd ed.). New York, NY: Springer.

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*(1), 23–35.

Caselli, R. J., Dueck, A. C., Osborne, D., Sabbagh, M. N., Connor, D. J., Ahern, G. L., … Woodruff, B. K. (2009). Longitudinal modeling of age-related memory decline and the APOE ε4 effect. *New England Journal of Medicine*, *361*(3), 255–263.

Chang, Y.-L., Fennema-Notestine, C., Holland, D., McEvoy, L. K., Stricker, N. H., Salmon, D. P., … Initiative, A. D. N. (2014). APOE interacts with age to modify rate of decline in cognitive and brain changes in Alzheimer's disease. *Alzheimer's & Dementia*, *10*(3), 336–348.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging* (Vol. 330). Cambridge University Press Cambridge.

Darst, B. F., Koscik, R. L., Racine, A. M., Oh, J. M., Krause, R. A., Carlsson, C. M., … Bendlin, B. B. (2017). Pathway-Specific Polygenic Risk Scores as Predictors of Amyloid-β Deposition and Cognitive Function in a Sample at Increased Risk for Alzheimer's Disease. *Journal of Alzheimer's Disease*, *55*(2), 473–484.

Gleason, C. E., Norton, D., Anderson, E. D., Wahoske, M., Washington, D. T., Umucu, E., … Carlsson, C. M. (2017). Cognitive Variability Predicts Incident Alzheimer's Disease and Mild Cognitive Impairment Comparable to a Cerebrospinal Fluid Biomarker. *Journal of Alzheimer's Disease: JAD*.

Hastie, Trevor, Tibshirani, Robert, & Friedman, Jerome. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York, NY: Springer.

Hegyi, G., & Garamszegi, L. Z. (2011). Using information theory as a substitute for stepwise regression in ecology and behavior. *Behavioral Ecology and Sociobiology*, *65*(1), 69–76.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, *14*(4), 382–401.

Holtzer, R., Verghese, J., Wang, C., Hall, C. B., & Lipton, R. B. (2008). Within-person across-neuropsychological test variability and incident dementia. *Jama*, *300*(7), 823–830.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307. https://doi.org/10.1093/biomet/76.2.297

Jack CR, Jr, Wiste HJ, Weigand SD, & et al. (2015). Age, sex, and apoe ε4 effects on memory, brain structure, and β-amyloid across the adult life span. *JAMA Neurology*, *72*(5), 511–519. https://doi.org/10.1001/jamaneurol.2014.4821

Johnson, S. C., Koscik, R. L., Jonaitis, E. M., Clark, L. R., Mueller, K. D., Berman, S. E., … Sager, M. A. (2017). The Wisconsin Registry for Alzheimer's Prevention: A Review of findings and current directions. *BioRxiv*.

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston naming test*. Pro-ed.

Kaup, A. R., Nettiksimmons, J., Harris, T. B., Sink, K. M., Satterfield, S., Metti, A. L., … Yaffe, K. (2015). Cognitive resilience to apolipoprotein E ε4: contributing factors in black and white older adults. *JAMA Neurology*, *72*(3), 340–348.

Koran, M. E. I., Wagener, M., & Hohman, T. J. (2017). Sex differences in the association between AD biomarkers and cognitive decline. *Brain Imaging and Behavior*, *11*(1), 205–213.

Koscik, R. L., Berman, S. E., Clark, L. R., Mueller, K. D., Okonkwo, O. C., Gleason, C. E., … Johnson, S. C. (2016). Intraindividual Cognitive Variability in Middle Age Predicts Cognitive Impairment 8–10 Years Later: Results from the Wisconsin Registry for Alzheimer's Prevention. *Journal of the International Neuropsychological Society*, *22*(10), 1016–1025.

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment* (5th ed.). New York, NY: Oxford University Press.

Manly, J. J., Touradji, P., Tang, M.-X., & Stern, Y. (2003). Literacy and memory decline among ethnically diverse elders. *Journal of Clinical and Experimental Neuropsychology*, *25*(5), 680–690.

McCarrey, A. C., An, Y., Kitner-Triolo, M. H., Ferrucci, L., & Resnick, S. M. (2016). Sex differences in cognitive trajectories in clinically normal older adults. *Psychology and Aging*, *31*(2), 166–175. https://doi.org/10.1037/pag0000070

Mielke, M. M., Vemuri, P., & Rocca, W. A. (2014). Clinical epidemiology of Alzheimer's disease: assessing sex and gender differences. *Clinical Epidemiology*, *6*, 37.

Mortensen, E. L., & Høgh, P. (2001). A gender difference in the association between APOE genotype and age-related cognitive decline. *Neurology*, *57*(1), 89–95.

Neu, S. C., Pa, J., Kukull, W., Beekly, D., Kuzma, A., Gangadharan, P., … Redolfi, A. (2017). Apolipoprotein E Genotype and Sex Risk Factors for Alzheimer Disease: A Meta-analysis. *Jama Neurology*, *74*(10), 1178–1189.

Payami, H., Zareparsi, S., Montee, K. R., Sexton, G. J., Kaye, J. A., Bird, T. D., … Litt, M. (1996). Gender difference in apolipoprotein E-associated risk for familial Alzheimer disease: a possible clue to the higher incidence of Alzheimer disease in women. *American Journal of Human Genetics*, *58*(4), 803.

Price, J. L., McKeel, D. W., Buckles, V. D., Roe, C. M., Xiong, C., Grundman, M., … Dickson, D. W. (2009). Neuropathology of nondemented aging: presumptive evidence for preclinical Alzheimer disease. *Neurobiology of Aging*, *30*(7), 1026–1036.

Price, J. L., & Morris, J. C. (1999). Tangles and plaques in nondemented aging and" preclinical" Alzheimer's disease. *Annals of Neurology*, *45*(3), 358–368.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*(3), 385–401.

Reitan, R. M. (1958). Validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and Motor Skills*, *8*(3), 271–276.

Richards, S. A. (2005). Testing ecological theory using the information-theoretic approach: examples and cautionary results. *Ecology*, *86*(10), 2805–2814.

Richards, S. A., Whittingham, M. J., & Stephens, P. A. (2011). Model selection and model averaging in behavioural ecology: the utility of the IT-AIC framework. *Behavioral Ecology and Sociobiology*, *65*(1), 77–89.

Riedel, B. C., Thompson, P. M., & Brinton, R. D. (2016). Age, APOE and sex: triad of risk of Alzheimer's disease. *The Journal of Steroid Biochemistry and Molecular Biology*, *160*, 134–147.

Roberts, R. O., Geda, Y. E., Knopman, D. S., Cha, R. H., Pankratz, V. S., Boeve, B. F., … Petersen, R. C. (2012). The incidence of MCI differs by subtype and is higher in men: the Mayo Clinic Study of Aging. *Neurology*, *78*(5), 342–351. https://doi.org/10.1212/WNL.0b013e3182452862

Salthouse, T. A., & Soubelet, A. (2014). Heterogeneous ability profiles may be a unique indicator of impending cognitive decline. *Neuropsychology*, *28*(5), 812.

Schmidt, M. (1996). *Rey auditory verbal learning test: A handbook*. Western Psychological Services Los Angeles, CA.

Strittmatter, W. J., & Roses, A. D. (1996). Apolipoprotein E and Alzheimer's disease. *Annual Review of Neuroscience*, *19*(1), 53–77.

Symonds, M. R., & Moussalli, A. (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology*, *65*(1), 13–21.

Tang, M.-X., Maestre, G., Tsai, W.-Y., Liu, X.-H., Feng, L., Chung, W.-Y., … Tycko, B. (1996). Relative risk of Alzheimer disease and age-at-onset distributions, based on APOE genotypes among elderly African Americans, Caucasians, and Hispanics in New York City. *American Journal of Human Genetics*, *58*(3), 574.

Trenerry, M. R., Crosson, B., DeBoe, J., & Leber, W. R. (1989). Stroop neuropsychological screening test. *Odessa, FL: Psychological Assessment Resources*.

Vemuri P, Lesnick TG, Przybelski SA, & et al. (2014). Association of lifetime intellectual enrichment with cognitive decline in the older population. *JAMA Neurology*, *71*(8), 1017–1024. https://doi.org/10.1001/jamaneurol.2014.963

Wechsler, D. (1997). *WAIS-III: Wechsler adult intelligence scale*. Psychological Corporation.

Wilkinson, G. S. (1993). *WRAT-3: Wide range achievement test administration manual*. Wide Range, Incorporated.

Wisdom, N. M., Callahan, J. L., & Hawkins, K. A. (2011). The effects of apolipoprotein E on non-impaired cognitive functioning: a meta-analysis. *Neurobiology of Aging*, *32*(1), 63–74.

Tables and Figures

Table 1: Sample characteristics

	Overall	Female	Male	p-value*
N	1256	873 (69.5%)	383 (30.5%)	
Age (mean (sd))	53.67 (6.55)	53.56 (6.57)	53.92 (6.51)	0.238
URG (N (%))	85 (6.8)	61 (7.0)	24 (6.3)	0.715+
ESL (N (%))	21 (1.7)	16 (1.8)	5 (1.3)	0.636+
Follow-up years (mean (sd))	9.10 (2.60)	9.04 (2.63)	9.22 (2.53)	0.34
Number of visits (N (%))				
2	135 (10.7)	102 (11.7)	33 (8.6)	0.124+
3	265 (21.1)	180 (20.6)	85 (22.2)	
4	432 (34.4)	310 (35.5)	122 (31.9)	
5	424 (33.8)	281 (32.2)	143 (37.3)	
APOE ε4 count (N (%))				
0	767 (61.1)	523 (59.9)	244 (63.7)	<0.001+
1	438 (34.9)	303 (34.7)	135 (35.2)	
2	51 (4.1)	47 (5.4)	4 (1.0)	
SES (N (%))				
1	39 (3.1)	26 (3.0)	13 (3.4)	<0.001^
2	122 (9.7)	99 (11.3)	23 (6.0)	
3	267 (21.3)	202 (23.1)	65 (17.0)	
4	241 (19.2)	171 (19.6)	70 (18.3)	
5	587 (46.7)	375 (43.0)	212 (55.4)	
RAVLT total (mean (sd))	51.10 (7.95)	52.61 (7.30)	47.64 (8.30)	<0.001
RAVLT delayed recall (mean (sd))	10.46 (2.82)	10.96 (2.60)	9.33 (2.99)	<0.001
Trails A (mean (sd))	26.81 (8.55)	26.28 (8.24)	28.02 (9.11)	0.001
Trails B (mean (sd))	62.06 (24.61)	60.21 (22.16)	66.28 (29.03)	<0.001
Boston Naming 60 (mean (sd))	57.02 (3.09)	56.87 (3.19)	57.35 (2.84)	0.002
CFL fluency (mean (sd))	43.37 (11.07)	43.99 (10.75)	41.94 (11.65)	0.002

Digit span backwards (mean (sd))	7.09 (2.22)	7.00 (2.17)	7.31 (2.32)	0.024
Digit span forwards (mean (sd))	10.51 (2.18)	10.40 (2.11)	10.76 (2.31)	0.006
Stroop color word, (mean (sd))	108.12 (20.63)	109.86 (19.58)	104.09 (22.42)	<0.001
9 Test IICV (mean (sd))	0.81 (0.23)	0.80 (0.22)	0.82 (0.24)	0.069
4 Test IICV (mean (sd))	0.73 (0.33)	0.73 (0.33)	0.73 (0.33)	0.803

*Note: Comparisons made with Mann-Whitney, unless noted with + (Fisher's exact) or ^ (Chi-squared); # of participants omitted from model averaging for that outcome due to <2 visits with that outcome's data: AVLT Total and Delay, 7; Trails A and B, 6; Digit span forward and backward, 7; Stroop CW, 21; BNT, 10; CFL, 10; 9 Test IICV, 26; 4 Test IICV, 9.

Table 2: Model set fixed effects structure

Model #	Model hierarchy	SES and SES^2	ESL	URG	Sex	Age	WRAT	APOE	Age^2	Sex * Age	Sex * Age^2	WRAT * Age	WRAT * Age^2	APOE * Age	APOE * Age^2	Sex * WRAT	Sex * APOE	Sex * WRAT * Age	Sex * WRAT * Age^2	Sex * APOE * Age	Sex* APOE * Age^2
1	baseline model	x	x	x	x	x															
2	M1 + WRAT	x	x	x	x	x	x														
3	M1 + APOE	x	x	x	x	x		x													
4	M1 + Age^2	x	x	x	x	x			x												
5	M1 + WRAT + APOE	x	x	x	x	x	x	x													
6	M1 + WRAT + Age^2	x	x	x	x	x	x		x												
7	M1 + APOE + Age^2	x	x	x	x	x		x	x												
8	M1 + WRAT + APOE + Age^2	x	x	x	x	x	x	x	x												
9	M1 + Sex * Age	x	x	x	x	x				x											
10	M9 + Sex * Age^2	x	x	x	x	x			x	x	x										
11	M2 + WRAT * Age	x	x	x	x	x	x					x									
12	M11 + WRAT * Age^2	x	x	x	x	x	x		x			x	x								
13	M3 + APOE * Age	x	x	x	x	x		x						x							
14	M13 + APOE * Age^2	x	x	x	x	x		x	x					x	x						
15	M9 + WRAT * Age	x	x	x	x	x	x			x		x									
16	M9 + APOE * Age	x	x	x	x	x		x		x				x							
17	M11 + APOE * Age	x	x	x	x	x	x	x				x		x							
18	M15 + APOE * Age	x	x	x	x	x	x	x		x		x		x							
19	M2 + Sex * WRAT	x	x	x	x	x	x									x					
20	M3 + Sex * APOE	x	x	x	x	x		x									x				
21	M9 + Sex * WRAT	x	x	x	x	x	x			x						x					
22	M9 + Sex * APOE	x	x	x	x	x		x		x							x				
23	M19 + Sex * APOE	x	x	x	x	x	x	x								x	x				
24	M21 + Sex * APOE	x	x	x	x	x	x	x		x						x	x				
25	M1 + Sex * WRAT * Age	x	x	x	x	x	x			x		x				x		x			
26	M1 + Sex * WRAT * Age^2	x	x	x	x	x	x		x	x	x	x	x			x		x	x		
27	M1 + Sex * APOE * Age	x	x	x	x	x		x		x				x			x			x	
28	M1 + Sex * APOE * Age^2	x	x	x	x	x		x	x	x	x			x	x		x			x	x

Table 3: Model averaged parameters for all model terms AVLT Total, AVLT Delay, $\log_{10}$Trails A, $\log_{10}$Trails B, and Square-root CFL

Model term	log10 Trails A	log10 Trails B	AVLT Total	AVLT Delayed	Sqrt CFL
SES	**-0.008 (-0.015, -0.002)**	**-0.013 (-0.02, -0.006)**	**0.442 (0.018, 0.86)**	**0.056 (0.001, 0.114)**	0.047 (-0.001, 0.095)
SES^2	**0.005 (2.78E-05, 0.01)**	0.005 (-0.001, 0.012)	-0.261 (-0.537, 0.013)	-0.012 (-0.048, 0.024)	-0.003 (-0.039, 0.032)
URG	**0.063 (0.035, 0.091)**	**0.099 (0.067, 0.131)**	**-2.301 (-3.973, -0.677)**	-0.177 (-0.386, 0.035)	-0.003 (-0.194, 0.183)
Male	**0.034 (0.022, 0.053)**	**0.037 (0.021, 0.051)**	**-5.799 (-6.709, -4.908)**	**-0.64 (-0.773, -0.537)**	**-0.173 (-0.277, -0.024)**
Age	2.67E-04 (-3.72E-04, 0.001)	**0.003 (0.002, 0.004)**	**-0.096 (-0.148, -0.057)**	-0.006 (-0.011, 0.001)	**0.014 (0.01, 0.017)**
ESL	-0.002 (-0.06, 0.054)	-0.009 (-0.064, 0.049)	-0.988 (-3.959, 1.905)	-0.176 (-0.539, 0.254)	-0.014 (-0.396, 0.379)
Male x Age	-1.51E-04 (-0.002, 0.001)	1.01E-04 (-0.001, 0.002)	**-0.104 (-0.178, -0.005)**	**-0.01 (-0.021, -0.001)**	-0.001 (-0.01, 0.003)
one APOE ε4	2.69E-05 (-0.007, 0.009)	0.001 (-0.001, 0.014)	-0.033 (-1.186, 0.052)	-0.002 (-0.134, 0.046)	0.009 (-0.023, 0.146)
two APOE ε4	-2.03E-05 (-0.028, 0.007)	4.36E-04 (-0.008, 0.022)	-0.081 (-2.456, 0.015)	-0.014 (-0.401, 0.003)	-0.012 (-0.23, 0.043)
Male x one APOE ε4	2.41E-06 (-0.011, 0.017)	4.39E-05 (-0.001, 0.007)	3.41E-04 (-0.996, 0.925)	0.003 (-0.029, 0.222)	**-0.029 (-0.368, -1.25E-06)**
Male x two APOE ε4	1.28E-04 (-0.01, 0.075)	1.89E-05 (-0.01, 0.008)	0.118 (-2.439, 9.44)	0.005 (-0.368, 1.23)	-0.004 (-0.845, 0.84)
Age x one APOE ε4	2.11E-07 (-5.73E-05, 0.002)	5.76E-08 (-1.82E-06, 9.14E-05)	-7.57E-06 (-0.004, 0.001)	-2.25E-05 (-0.006, 0.001)	2.05E-07 (-3.72E-04, 4.92E-04)
Age x two APOE ε4	-1.12E-06 (-0.003, 7.38E-06)	1.89E-07 (-2.48E-06, 1.95E-04)	-2.79E-05 (-0.013, 0.002)	-1.42E-04 (-0.031, 0.001)	-9.33E-06 (-0.007, 7.55E-06)
Age^2	4.96E-06 (-1.48E-06, 1.32E-04)	**6.74E-05 (1.16E-06, 1.43E-04)**	**-0.001 (-0.007, -5.22E-06)**	1.06E-06 (-5.95E-05, 3.71E-04)	-2.32E-05 (-0.001, 1.03E-06)
Male x Age^2	**-8.76E-06 (-2.29E-04, -1.03E-08)**	1.12E-05 (-5.02E-05, 1.38E-04)	1.04E-04 (-0.004, 0.005)	-2.37E-06 (-0.001, 1.29E-05)	2.96E-05 (-3.24E-06, 0.001)
Age^2 x one APOE ε4	3.62E-09 (-7.51E-06, 8.92E-05)	5.41E-14 (-1.65E-20, 1.50E-07)	**-3.80E-21 (-6.02E-12, -6.18E-32)**	-6.12E-19 (-2.74E-08, 4.51E-11)	-1.66E-38 (-4.99E-27, 6.95E-28)
Age^2 x two APOE ε4	8.71E-09 (-1.30E-05, 2.01E-04)	-1.02E-13 (-2.00E-07, 2.03E-13)	-6.14E-21 (-4.74E-12, 4.94E-17)	-2.54E-18 (-2.36E-07, 3.84E-12)	-2.93E-38 (-1.90E-26, 6.40E-28)
Male x Age x one APOE ε4	**-7.52E-07 (-0.005, -1.89E-11)**	**-1.01E-14 (-7.53E-07, -3.29E-28)**	1.53E-21 (-2.83E-13, 3.13E-12)	9.58E-18 (-1.22E-09, 6.91E-07)	6.76E-37 (-6.22E-29, 1.24E-24)
Male x Age x two APOE ε4	**2.77E-06 (3.53E-10, 0.014)**	1.16E-14 (-4.73E-08, 1.13E-07)	1.94E-20 (-2.38E-14, 2.35E-10)	2.95E-17 (-2.89E-08, 1.75E-06)	4.95E-36 (-1.74E-29, 1.79E-23)
WRAT reading	**0.001 (2.76E-06, 0.002)**	**-0.002 (-0.003, -0.001)**	**0.142 (0.091, 0.197)**	**0.018 (0.01, 0.025)**	**0.024 (0.018, 0.03)**
Male x WRAT reading	**-0.002 (-0.004, -7.27E-06)**	-0.001 (-0.003, 3.36E-05)	**0.113 (0.005, 0.213)**	0.007 (-1.94E-04, 0.02)	**0.012 (0.001, 0.023)**
Age x WRAT reading	-3.70E-06 (-1.12E-04, 2.54E-05)	-6.60E-06 (-1.07E-04, 7.96E-05)	-5.13E-05 (-0.004, 0.004)	-4.21E-05 (-0.001, 2.04E-04)	4.55E-05 (-4.26E-06, 0.001)
Age^2 x WRAT reading	-2.63E-08 (-6.09E-06, 5.25E-06)	2.74E-06 (-2.96E-07, 1.61E-05)	-2.34E-06 (-2.92E-04, 2.64E-04)	-1.66E-07 (-3.77E-05, 1.99E-06)	2.12E-06 (-5.98E-08, 6.13E-05)
Male x Age x WRAT reading	-1.85E-06 (-1.24E-04, 1.03E-04)	6.46E-07 (-1.31E-04, 1.31E-04)	0.001 (-0.003, 0.009)	6.11E-05 (-5.62E-05, 0.001)	-3.18E-05 (-0.001, 1.03E-04)
Male x Age^2 x WRAT reading	2.27E-08 (-7.72E-06, 8.58E-06)	-3.08E-06 (-2.16E-05, 1.22E-07)	-2.70E-05 (-0.001, 3.34E-04)	1.28E-07 (-6.11E-06, 3.61E-05)	-2.32E-06 (-7.99E-05, 1.35E-06)
Male x Age^2 x one APOE ε4	-1.06E-08 (-2.30E-04, 6.12E-07)	-1.10E-16 (-6.63E-09, 1.58E-10)	3.71E-22 (-8.99E-16, 1.21E-12)	-3.42E-19 (-7.80E-08, 2.43E-11)	**-1.19E-37 (-5.49E-25, -1.24E-49)**
Male x Age^2 x two APOE ε4	-4.06E-08 (-0.001, 3.08E-05)	1.26E-15 (-9.45E-13, 8.32E-08)	-2.84E-21 (-4.52E-11, 6.42E-20)	-6.96E-18 (-4.33E-06, 2.47E-13)	2.66E-37 (-2.49E-28, 3.36E-25)

Note: "E" represents "x10^"; e.g. 1.23E-04 = 1.23x10^-4

95% CI's that do not overlap 0 (are significant) are identified by bold-face font; gray shading denotes CIs that overlap 0.

Table 4: Model averaged parameters and 95% CI's for all model terms and CFL, BNT, Digit Span Forwards, Digit Span Backwards, 4 Test IICV and 9 Test IICV

Model term	Boston Naming	Digit Span Forwards	Digit Span Backwards	Sqrt 4 test IICV	Sqrt 9 test IICV
SES	0.012 (-0.052, 0.075)	-0.001 (-0.116, 0.111)	0.031 (-0.08, 0.145)	-0.003 (-0.013, 0.006)	-0.006 (-0.012, 0.001)
SES^2	**-0.053 (-0.095, -0.001)**	-0.05 (-0.121, 0.024)	-0.032 (-0.105, 0.048)	0.001 (-0.006, 0.008)	-0.001 (-0.006, 0.003)
URG	**-0.695 (-0.888, -0.487)**	-0.346 (-0.794, 0.1)	-0.26 (-0.719, 0.194)	-0.021 (-0.056, 0.016)	0.006 (-0.02, 0.032)
Male	**0.225 (0.122, 0.353)**	**0.344 (0.102, 0.628)**	0.22 (-0.041, 0.425)	**0.022 (1.97E-04, 0.043)**	**0.029 (0.013, 0.043)**
Age	**0.014 (0.009, 0.02)**	**-0.014 (-0.025, -0.006)**	0.004 (-0.004, 0.018)	**0.001 (2.79E-04, 0.002)**	0.001 (-1.35E-04, 0.002)
ESL	**-0.672 (-0.983, -0.351)**	-0.236 (-1.068, 0.604)	0.012 (-0.796, 0.849)	0.007 (-0.057, 0.073)	0.039 (-0.002, 0.083)
Male x Age	-1.29E-05 (-0.011, 0.001)	-1.59E-04 (-0.016, 0.016)	-0.001 (-0.03, 0.004)	2.10E-04 (-0.002, 0.002)	0.001 (-1.01E-04, 0.003)
one APOE ε4	-8.14E-08 (-0.209, 0.005)	0.001 (-0.042, 0.21)	-0.007 (-0.192, 0.046)	-3.21E-14 (-4.89E-08, 3.58E-11)	-4.58E-05 (-0.021, 4.17E-05)
two APOE ε4	0.024 (-0.004, 0.437)	-0.002 (-0.268, 0.201)	0.011 (-0.125, 0.399)	9.65E-15 (-5.56E-09, 2.23E-08)	5.72E-05 (-0.004, 0.021)
Male x one APOE ε4	-1.06E-07 (-0.008, 8.98E-08)	-0.005 (-0.597, 3.04E-04)	-4.96E-05 (-0.058, 0.003)	2.64E-18 (-7.61E-14, 8.71E-12)	1.15E-06 (-0.004, 0.021)
Male x two APOE ε4	-1.16E-07 (-0.007, 2.82E-04)	-0.005 (-1.054, 0.292)	-3.87E-04 (-0.243, 9.44E-05)	-6.43E-18 (-1.29E-11, 1.19E-12)	8.33E-06 (-0.004, 0.168)
Age x one APOE ε4	1.39E-10 (-1.23E-07, 5.44E-06)	4.99E-07 (-8.04E-05, 1.57E-04)	-2.01E-05 (-0.019, 1.01E-07)	2.79E-15 (-4.45E-12, 6.67E-09)	1.72E-07 (-4.47E-05, 0.002)
Age x two APOE ε4	-2.89E-11 (-4.74E-06, 2.76E-06)	-4.76E-07 (-2.38E-04, 2.25E-04)	1.72E-05 (-0.001, 0.016)	1.07E-14 (-1.20E-13, 2.41E-08)	4.61E-07 (-3.63E-05, 0.003)
Age^2	**-0.001 (-0.002, -0.001)**	**-0.001 (-0.002, -2.19E-05)**	**-0.001 (-0.003, -5.34E-05)**	**1.69E-04 (1.39E-05, 2.73E-04)**	**1.50E-04 (2.84E-05, 2.15E-04)**
Male x Age^2	1.68E-06 (-5.65E-07, 0.001)	3.14E-04 (-1.75E-05, 0.003)	1.74E-04 (-4.14E-06, 0.003)	**-2.03E-04 (-3.78E-04, -1.23E-05)**	-1.21E-04 (-2.43E-04, 1.48E-05)
Age^2 x one APOE ε4	-4.22E-53 (-1.58E-40, 3.89E-45)	3.68E-48 (-2.02E-39, 2.70E-37)	-4.83E-47 (-1.31E-36, 9.17E-40)	1.36E-27 (-5.71E-27, 3.12E-18)	2.63E-08 (-1.05E-08, 1.98E-04)
Age^2 x two APOE ε4	-2.89E-52 (-2.67E-40, 6.78E-44)	-3.50E-48 (-8.37E-38, 5.21E-38)	-8.24E-47 (-1.58E-36, 1.02E-38)	1.30E-27 (-9.95E-21, 1.21E-18)	-2.33E-08 (-1.63E-04, 3.83E-05)
Male x Age x one APOE ε4	9.19E-52 (-6.53E-53, 8.76E-39)	8.40E-48 (-5.94E-42, 2.10E-35)	-6.06E-48 (-3.05E-36, 2.26E-39)	-3.26E-27 (-2.33E-17, 9.20E-25)	-2.78E-07 (-0.004, 1.22E-06)
Male x Age x two APOE ε4	-2.59E-51 (-7.83E-39, 8.47E-42)	-2.61E-47 (-1.42E-35, 3.11E-38)	4.18E-48 (-2.58E-38, 7.81E-36)	-1.06E-26 (-8.36E-17, 2.84E-23)	-8.26E-07 (-0.009, 0.002)
WRAT reading	**0.046 (0.038, 0.051)**	**0.073 (0.053, 0.083)**	**0.079 (0.059, 0.087)**	**-0.003 (-0.004, -0.002)**	-7.39E-05 (-0.001, 0.001)
Male x WRAT reading	2.60E-05 (-1.06E-05, 0.018)	**0.014 (3.42E-04, 0.051)**	**0.003 (2.05E-05, 0.043)**	**0.007 (0.005, 0.009)**	**0.002 (4.09E-06, 0.003)**
Age x WRAT reading	9.30E-05 (-2.52E-05, 0.001)	1.62E-04 (-3.61E-04, 0.002)	-1.03E-04 (-0.001, 0.001)	**3.28E-04 (2.05E-04, 4.61E-04)**	4.41E-05 (-2.53E-05, 1.32E-04)
Age^2 x WRAT reading	-1.48E-06 (-2.56E-05, 2.95E-05)	**2.75E-05 (1.70E-07, 1.60E-04)**	6.62E-06 (-4.67E-05, 1.05E-04)	4.44E-06 (-5.00E-06, 1.50E-05)	4.43E-07 (-6.87E-06, 8.11E-06)
Male x Age x WRAT reading	2.61E-07 (-4.02E-04, 0.001)	-3.42E-04 (-0.003, 1.45E-04)	-1.35E-04 (-0.003, 2.30E-04)	1.43E-04 (-6.86E-05, 3.42E-04)	9.19E-05 (-1.99E-05, 2.33E-04)
Male x Age^2 x WRAT reading	-5.76E-08 (-6.12E-05, 1.94E-05)	-1.95E-05 (-2.06E-04, 4.47E-05)	-4.03E-06 (-1.41E-04, 7.72E-05)	-6.68E-06 (-2.05E-05, 7.43E-06)	-2.10E-06 (-1.34E-05, 6.90E-06)
Male x Age^2 x one APOE ε4	8.12E-53 (-1.04E-53, 7.22E-40)	1.19E-49 (-8.64E-40, 9.43E-38)	-2.76E-49 (-9.82E-38, 1.01E-39)	-2.78E-28 (-2.07E-18, 1.01E-25)	-2.37E-08 (-3.35E-04, 9.57E-08)
Male x Age^2 x two APOE ε4	1.10E-52 (-4.00E-44, 5.31E-40)	-4.20E-49 (-2.62E-37, 7.14E-39)	-1.92E-48 (-1.38E-36, 1.55E-40)	-1.64E-29 (-1.13E-20, 1.08E-18)	-7.37E-09 (-0.001, 2.60E-04)

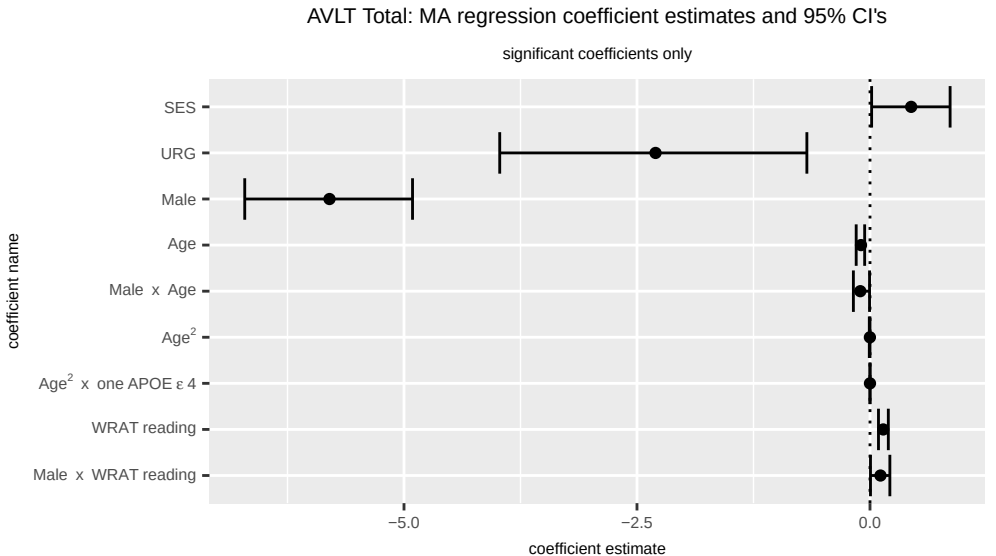
Note: "E" represents "x10^"; e.g. 1.23E-04 = 1.23x10^-4

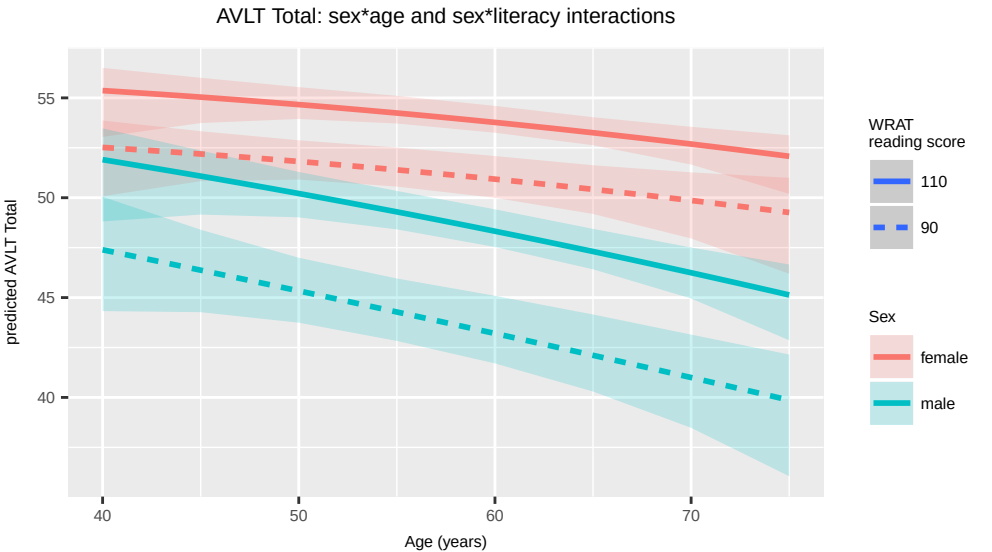
95% CI's that do not overlap 0 (are significant) are identified by bold-face font; gray shading denotes CIs that overlap 0.

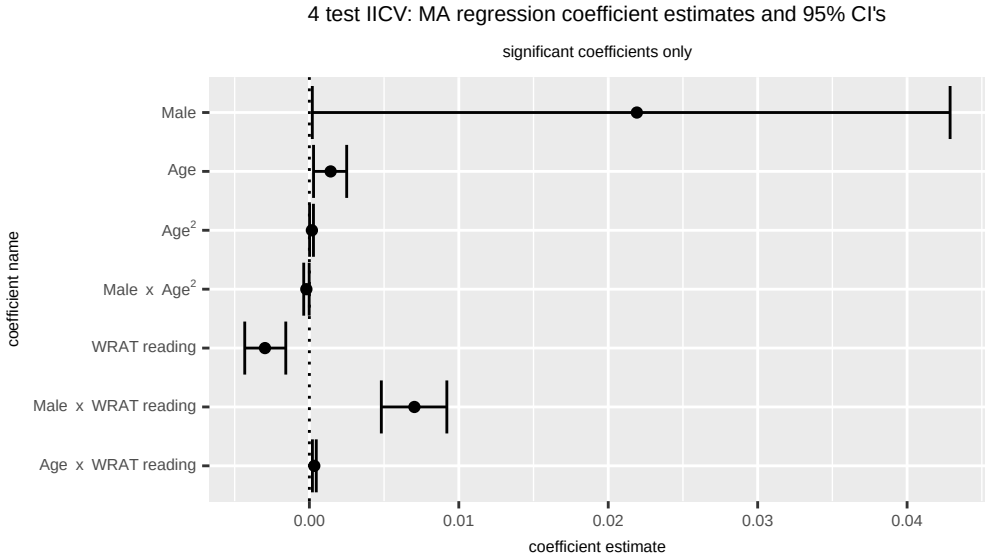
Figures Captions

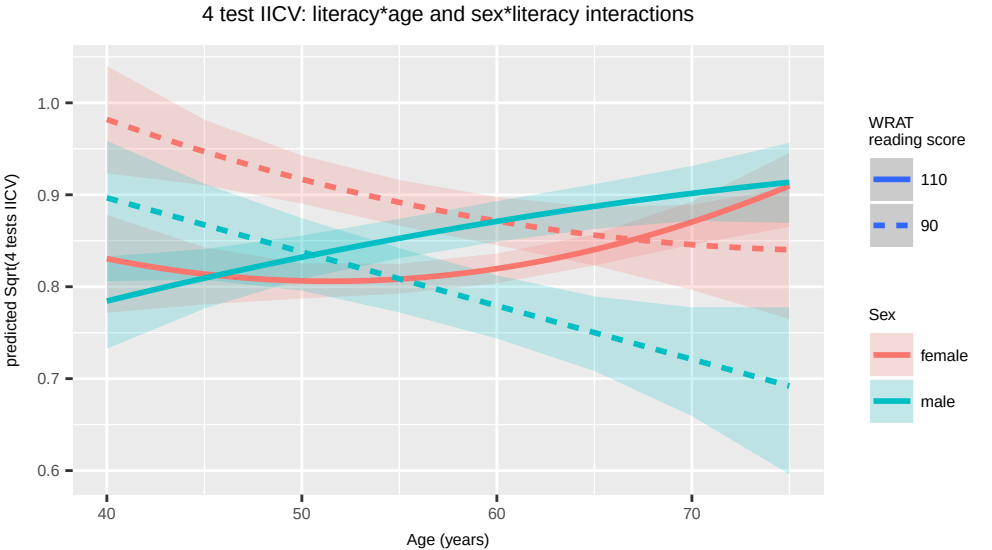
Figure 1. AVLT Total: 95% CI's for Significant MA Parameters (left panel) and Predicted Values (right)

Figure 2. 4 Test IICV: 95% CI's for Significant Model-Averaged Parameters (left) and Predicted Values (right)


AVLT Total: MA regression coefficient estimates and 95% CI's
significant coefficients only
SES
URG
Male
Age
Male x Age
Age^2
Age^2 x one APOE ε 4
WRAT reading
Male x WRAT reading
coefficient name
−5.0
−2.5
0.0
coefficient estimate


AVLT Total: sex*age and sex*literacy interactions
predicted AVLT Total
55
50
45
40
40
50
60
70
Age (years)
WRAT reading score
110
90
Sex
female
male


4 test IICV: MA regression coefficient estimates and 95% CI's
significant coefficients only
coefficient name
Male
Age
Age^2
Male x Age^2
WRAT reading
Male x WRAT reading
Age x WRAT reading
0.00
0.01
0.02
0.03
0.04
coefficient estimate


4 test IICV: literacy*age and sex*literacy interactions
predicted Sqrt(4 tests IICV)
1.0
0.9
0.8
0.7
0.6
40
50
60
70
Age (years)
WRAT reading score
110
90
Sex
female
male