Supplementary Material

Executive Function Measures

Hayling sentence completion test. This test is associated with the inhibition domain and consists of two sections (Burgess & Shallice, 1997). Each section has a set of 15 sentences and each sentence has the last word missing. In Section 1, the examiner reads a sentence aloud and the participant completes the sentence as fast as possible. In Section 2, the participant completes the sentence with a word that is completely unrelated or unconnected to the sentence in every way. The task yields two measures of response speed (Section 1 and 2) and an error score (Section 2). Scaled scores range from 1 to 10, with 1 being "impaired" and 10 being "very superior".

Stroop test. This task measures inhibition by requiring participants to inhibit the automatic response of reading a printed word and instead name the color in which the word is printed as fast as possible (Regard, 1981). In Part A, participants name as fast as possible the color of 24 dots printed in blue, red, green, or yellow. In Part C, the colored stimuli are the color names printed in lower case with the color being incongruent to the color name. The performance score is based on the interference index ([Part C – Part A]/Part A).

Brixton test. The Brixton Spatial Anticipation Test (Burgess & Shallice, 1997) is a shifting task consisting of a 56-page booklet, with each page showing 10 circles in two rows of 5, numbered 1-10 for the position, and one position is always colored blue. The position of the colored circle changes as the pages are turned, based on a series of simple rules that vary without warning. Participants are shown one page at a time and are required to decide the position of the next filled circle. Total errors are recorded based on the test manual: maximum of 54 errors and

converted to a scaled score of 10. For the analysis, we used a standard scale score ranging from 1 (impaired) to 10 (very superior).

Color trail test. This test measures shifting as an executive process (D'Elia, Satz, Uchiyama, & White, 1996). In the first part, participants connect encircled numbers 1 through 25 (randomly arranged) in the correct order. Even numbers have a yellow background and odd numbers a pink background. The second part shows encircled numbers from 1 to 25 twice (one sequence with a yellow background, the other pink). Participants are required to connect the numbers from 1 to 25, alternating between pink and yellow circles and choose the circle with the number sequence that is the alternate version of the previous color (D'Elia et al., 1996). Completion time is recorded in seconds and used as the score (standardized). For the analysis, we used the latency score of the second part.

Computational span. This working memory task is used for the updating domain (Salthouse & Babcock, 1991). Participants are asked to solve a series of arithmetic problems while remembering the last digit of each problem they solve in order to be recalled later. There is an increase in the number of problems in a series from one to seven, with three trials at each series length. The measure used was the highest span correctly recalled for two out of three trials.

Reading span. This task is associated with the updating domain and requires participants to answer questions about orally presented sentences while remembering the final word of each sentence for later recall. There is an increase in the number of sentences in the passage from one to seven, with three trials at each series length. The measure used was the highest span correctly recalled for two out of three trials.

Letter series. This task was used for its contribution to the shifting domain. In this test (Thurstone, 1962), participants are required to identify the pattern of a series of letters.

Participants have to decipher the pattern in the target string and then match the letter in the string that is congruent with the pattern presented. The last letter of the series of letters has to be placed in such a way that it would continue the established pattern. For example, a series of letters is arranged in the following order: f g h i j k l. The participant is required to circle the letter that continues the established pattern. The options to circle are: j k l m n. In this case, the correct letter is "m" because it follows the pattern (is the letter that comes after "l"). The outcome measure used was the total number correct out of 20 patterns. We multiplied the scores by 0.5 across the three waves to keep them within range relative to the other variables. We then included these scores to the CFA models.

Letter sets. This task contributed to the shifting domain. Each problem in this test (Ekstrom, French, Harman, & Derman, 1976) has five sets of letters with four letters in each set. Four of the sets of letters are alike in some way. Participants are required to find the rule that makes the four sets alike. One of the set of letters is different from the others and will not fit the rule. Participants circle the set of letters that is different. The outcome measure used was the total number correct out of 15 problems.

Predictors.

We included multiple predictors at baseline (W1):

Genetic. From DNA extraction and genotyping, we included the following genetic polymorphisms: *APOE*, *BDNF*, *IDE*, and *COMT*. For the genetic analyses, a dichotomous genotype categorization was conducted based on the presence or the absence of the risk allele. For *APOE* genotype ε4- (non-risk) composed of ε2ε2, ε2ε3, ε3ε3 combinations, ε4+ (risk) composed of ε4ε4 and ε3ε4 allele combinations. For *BDNF* genotype Met- (non-risk) composed of the Val/Val allele combination, Met+ (risk) composed of the Met/Met and Val/Met allele

combinations. For *IDE* genotype G- (risk) composed of the AA allele combination, G+ (non-risk) composed of the GG and GA allele combinations. For *COMT* genotype Val- (non-risk) composed of the Met/Met allele combination, Val+ (risk) composed of the Val/Val and Val/Met allele combinations. With the exception of the genotype frequency for *IDE* [$\chi^2 = 62.59$ (1), $p < 0.05$], the genotype distribution for *APOE* [$\chi^2 = 0.71$ (1), $p > 0.05$], *BDNF* [$\chi^2 = 1.47$ (1), $p > 0.05$], and *COMT* [$\chi^2 = 2.93$ (1), $p > 0.05$] did not differ significantly from Hardy-Weinberg equilibrium.

Functional. We used baseline (a) pulse pressure (PP; equals systolic blood pressure (BP)–diastolic BP, in mmHG) based on average of eight BP readings, (b) body mass index (BMI; equals weight/height^2, in kilograms/meters^2), (c) peak expiratory flow (PEF; largest volume of air expired over three attempts, in litres/minute), and (d) grip strength (average hand strength, in kilograms/force).

Lifestyle. Lifestyle factors were (a) everyday physical activity (based on n = 4 self-report questions), and (b) everyday novel cognitive activity (n = 27). These lifestyle variables are part of the VLS-Activities Lifestyle Questionnaire (i.e., Hultsch, Hertzog, Small, & Dixon, 1999; Runge, Small, McFall, & Dixon, 2014) and are based on a nine-point scale (i.e., never = 0, daily = 8) that rates frequency of participation.

Mobility. Mobility markers included (a) balance or timed turn (360-degree turn, in seconds) and (b) gait or timed walk (20 feet, in seconds).

Demographic. We included baseline participants' age (in years), education (total years), and sex (male or female).

DNA extraction and genotyping.

Saliva samples were collected according to Oragene DNA Genotek technology protocol, including preparation and stabilization (see McFall et al., 2013). Genotyping was carried out by using a Polymerase Chain Reaction Restriction Fragment Length Polymorphism strategy to analyze the allele status for *APOE* (determined by the combination of the SNPs rs429358 and rs7412), *BDNF* (rs6265), *IDE* (rs6583817), and *COMT* (rs4680).

Statistical Analyses

Two sets of preliminary statistical analyses were performed. Using confirmatory factor analysis (CFA) and eight EF indicators, we established the best fitting EF latent variable by testing one-, two-, and three-factor models. For example, we specified a one-factor model that included all eight indicators (standardized prior to analysis) loading on the EF factor. Factor scaling was established by fixing the factor loading of one indicator (Hayling) from each factor (i.e., from wave, wave 2, and wave 3) to a value of 1.0. The factor loadings are the correlation coefficients between the EF indicators and the EF factor and indicate the variance explained by the EF indicators on the EF factor. All factor loadings were significant ($p < .001$). The second analysis involved testing the measurement invariance of the EF latent variable model across three waves. Model fit for CFA and invariance testing was determined by using standard indices: (a) χ^2 (good fit producing a non-significant test, $p > .05$), (b) comparative fit index (CFI; value of $\geq .95$ is good and $\geq .90$ is adequate fit), (c) root mean square error of approximation (RMSEA; value of $\leq .05$ is good and $\leq .08$ is adequate fit), and (d) standardized root mean square residual (SRMR; good fit determined by a value of $\leq .08$), (Kline, 2011; Little 2013). We used factor scores from the best fitting EF latent variable to produce the individualized trajectories (level and slope) across a broad band of aging. Factor scores are linear combinations of an observed variable (i.e., EF) and take into account what is shared between a measure and a factor (i.e.,

variance). In the present study, we derived factor scores from the latent variable made up of eight EF measures. Each EF measure contributes to the factor score. The purpose factor scores serve is to provide information about an individual's placement or ranking on the factor (s) (DiStefano, Zhu, & Mindrila, 2009). CFA uses standardized information to create factor scores. Consequently, factor scores are standardized scores that have a metric similar to a z-score, with values ranging approximately between -3 and 3.

In all subsequent growth models, age as a continuous variable was used as the metric of longitudinal change. In this study, we centered age at 75 years for two reasons: (1) this is the rough mid-point of the age distribution and (2) previous VLS studies have showed this age to represent a typical inflection point for aging-related change in EF and other cognitive performance domains (i.e., McFall et al., 2013, 2014; Small, Dixon, McArdle, & Grimm, 2012). Therefore, the centering age was used to interpret intercept differences that represent the performance level attribute. All trajectories were comprised of the individual EF performance and age at the first, second, and third wave of testing. When fully assembled, the EF trajectory distribution covered a 40-year band of aging (53-95 years). Specifically, each participant contributed EF and age data at each of the three data collection points thus accounting for chronological age directly. Although each participant contributed to the three data collection points, not all participants had two or more waves of data. An important advantage of the present structural equation modeling (SEM) approach is that it allows for the inclusion of data from all participants, even those with just one wave of data. Using maximum likelihood estimation, SEM estimates values for all waves (Little, 2013). Specifically, we used robust maximum likelihood estimation based on all available information from every EF variable included in the covariance matrix. Notably, within statistical growth curve models, attrition does not translate to subjects

lost at all waves; there is no need for "listwise deletion" in these models, as all participants providing data at one, two, or three waves are included in the analyses.

Research Goal 1: Examining Variability in EF Longitudinal Trajectories

Latent growth modeling was used to establish an EF latent growth curve. Using age as the metric of longitudinal change, the equation for a linear growth curve for a single individual is represented as follows:

$$Y_{io} = \gamma_{0i} + \gamma_{li} * Age + e_{i0}$$

In this equation, i refers to the individual; o refers to the occasions of measurement; Y_{io} represents the scores for each individual i at occasion o; γoi is the intercept for each individual i; γ_{li} is the regression slope for each individual i as a function of age; e_{io} is the error in prediction for each individual i at occasion o (Little, 2013). Using this design, we can see developmental changes that are related specifically to age. Furthermore, each individual's EF would be affected by changes that occur at the particular age of EF measurement. The model stipulates age at each time point as the metric of change. It also takes age into consideration as a covariate and in effect controls for age across all time points, including baseline age. Model fit was determined by conventional indices: (a) -2 log likelihood (-2LL), (b) Akaike Information Criterion (AIC), (c) Bayesian Information Criterion (BIC), and (d) deviance statistics (D). When comparing models, the best fitting model was significantly better than a previous model based on D and contained the lowest values for -2LL, AIC, and BIC.

Research Goal 2: Establishing Latent Classes of EF Trajectories

Latent class growth analysis (LCGA) was used to analyze individualized EF latent variable trajectory data. LCGA is a data-driven technique that identifies subgroups of individuals with similar patterns of change over time on a latent variable (see Andruff, Carraro, Thompson,

Gaudreau, & Louvet, 2009). Each individual has a unique developmental course; however, the distribution of individual differences is specified by a finite set of polynomial functions corresponding to a discrete trajectory. Intercept and slope parameters are estimated for each trajectory to account for the magnitude and direction of change, which can be distinct across trajectories (Nagin, 2005). These parameters are fixed to be equal for each individual within trajectories and individual differences are captured by the trajectories in the model. Essentially, the latent class growth model is defined as the following: each trajectory is described as the latent variable ($y*_{it}$) that represents the scores on a dependent variable (Y) for a given trajectory (j) at a specific time (t). This is defined by the following function:

$$(y*_{it}) = \beta^j_0 + \beta^j_1 X_{it} + \varepsilon_{it}$$

In the above equation, X_{it} represents the independent variable (age) introduced in a regular (linear) term. ε_{it} is the disturbance term with an assumed normal distribution and a mean of zero and constant standard deviation. Lastly, β^j_0 and β^j_1 are the parameters that define the intercept and slope of the trajectory for a specific subgroup (j; Andruff et al., 2009).

Research Goal 3: Biomarker Risk Predictors Discriminating Classes of EF Trajectories

Random Forest Analysis (RFA; Kuhn & Johnson, 2013) was used to determine the most important predictors discriminating classes of EF trajectories. RFA is a classifier that evolves from predictor trees. In order to classify a new instance, each predictor tree provides a classification for the data. These classifications are collected by random forest (RF), which chooses the most voted prediction as the result (Mao & Wang, 2012). RFs are a combination of tree predictors, with each tree depending on the value of the random vector independently sampled and with equal distribution as all the trees in the forest (Breiman, 2001). To assess relative level of importance, we used the mean decrease in accuracy (MDA). MDA is an

importance measure that ranks and selects variables. It is used to quantify the importance of a variable by measuring prediction changes in accuracy, when the values of each variable are permuted in a random order to the original observations (Calle & Urrea, 2010). We used the out of bag (OOB) error rate to get an unbiased estimate of the classification error as trees were added to the forest. This unbiased estimate is what cross-validation or separate test phases aim to accomplish. Therefore, since RFA produces this internally, such methods are not needed. The uncorrelated predictors were generated from a bootstrap sample, which omitted about 37% of the data. The OOB then estimated each predictor's vote over the omitted data from its bootstrap sample (Bylander, 2002). The OOB error rate produces two results: (a) normalized root mean square error (NRMSE; Oba et al., 2003) and (b) proportion of falsely classified entries (PFC). Good performance is indicated by values closer to 0. The RFA model produced the following results: NRMSE = 0.233; PFC = 0.07. The forest error rate depends on the correlation between any two trees in the forest; the higher the correlation, the higher the error rate (Breiman, 2001). Therefore, with the OOB results we assumed that the trees were uncorrelated. There was also no multicollinearity between predictors.

References

Andruff, H., Carraro, N., Thompson, A., Gaudreau, P., & Louvet, B. (2009). Latent class growth modelling: A tutorial. *Tutorials in Quantitative Methods for Psychology*, *5*(1), 11-24. doi:10.20982/tqmp.05.1.p011

Bylander, T. (2002). Estimating generalization error on two-class datasets using out-of-bag estimates. *Machine Learning*, *48*(1-3), 287-297. doi:10.1023/A:1013964023376

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. doi: 10.1023/A:1010933404324

Burgess, P. W., & Shallice, T. (1997). *The Hayling and Brixton tests*. Thurston, Suffolk, England: Thames Valley Test Company.

Calle, M. L., & Urrea, V. (2010). Letter to the editor: Stability of random forest importance measures. *Briefings in Bioinformatics*, *12*(1), 86-89. doi:10.1093/bib/bbq011

D'Elia, L. F., Satz, P., Uchiyama, C. L., & White, T. (1996). *Color trails test: Professional manual*. Odessa, FL: Psychological Assessment Resources.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*(20), 1-11.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.

Hultsch, D. F., Hertzog, C., Small, B. J., & Dixon, R. A. (1999). Use it or lose it: Engaged lifestyle as a buffer of cognitive decline in aging? *Psychology and Aging, 14*(2), 245-263. doi:10.1037/0882-7974.14.2.245

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd Edition). New York, NY: Guilford Press.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York, NY: Springer-Verlag.

Little, T. D. (2013). *Longitudinal structural equation modeling.* New York, NY: Guilford Press.

Mao, W., & Wang, F. (2012). *New advances in intelligence and security informatics*. Oxford: Zhejiang University Press.

McFall, G. P., Wiebe, S. A., Vergote, D., Jhamandas, J., Westaway, D., & Dixon, R. A. (2014). IDE (rs6583817) polymorphism and pulse pressure are independently and interactively associated with level and change in executive function in older adults. *Psychology and Aging*, *29*(2), 418-430. doi:10.1037/a0034656

McFall, G. P., Wiebe, S. A., Vergote, D., Westaway, D., Jhamandas, J., & Dixon, R. A. (2013). IDE (rs6583817) polymorphism and type 2 diabetes differentially modify executive function in older adults. *Neurobiology of Aging, 34*(9), 2208-2216. doi:10.1016/j.neurobiolaging.2013.03.010

Nagin, D.S. (2005). *Group-based modelling of development.* Cambridge, MA: Harvard University Press.

Oba, S., Sato, M. A., Takemasa, I., Monden, M., Matsubara, K. I., & Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, *19*(16), 2088-2096. doi:10.1093/bioinformatics/btg287

Regard, M. (1981). *Cognitive rigidity and flexibility: A neuropsychological study.* Unpublished doctoral dissertation, University of Victoria, Victoria, British Columbia, Canada.

Runge, S. K., Small, B. J., McFall, G. P., & Dixon, R. A. (2014). APOE moderates the association between lifestyle activities and cognitive performance: Evidence of genetic

plasticity in aging. *Journal of the International Neuropsychological Society*, *20*(5), 478-486. doi:10.1017/S1355617714000356

Salthouse, T. A., & Babcock, R. L. (1991). Decomposing adult age differences in working memory. *Developmental Psychology, 27*(5), 763-776. doi: 10.1037/0012-1649.27.5.763

Small, B. J., Dixon, R. A., McArdle, J. J. & Grimm, K. J. (2012). Do changes in lifestyle engagement moderate cognitive decline in normal aging? Evidence from the Victoria Longitudinal Study. *Neuropsychology, 26*(2), 144-155. doi:10.1037/a0026579

Thurstone, T. G. (1962). *Primary mental abilities: Grades 9-12*, *1962 revision*. Chicago: Science Research Associates.

Table 1

Class comparison for RF follow-up analyses

Class	Predictors	C-Statistic
a vs. b	Novel cognitive activity, pulse pressure, education, physical activity	0.66; 95% *CI* (.59 – .72)
a vs. c	Age, education, novel cognitive activity, gait	0.77; 95% *CI* (.71 – .82)
a vs. d	Education, novel cognitive activity, pulse pressure, age, gait, body mass index, balance	0.84; 95% *CI* (.77 – .91)
b vs. c	Novel cognitive activity, education, age, physical activity, gait	0.63; 95% *CI* (.58 – .68)
b vs. d	Education, novel cognitive activity, *BDNF*, gait	0.72; 95% *CI* (.63 – .80)
c vs. d	Education, novel cognitive activity, gait, *BDNF*	0.55; 95% *CI* (.46 – .64)

Note. Class a = very high level and shallow declining slope; Class b = moderate level and notably declining slope; Class c = low level and substantially declining slope; Class d = very low level and steepest declining slope; *BDNF* = Brain Derived Neurotrophic Factor; *CI* = Confidence Interval.