Published in final edited form as:

Perspect Psychol Sci. 2010 October 11; 5(5): 606–621. doi:10.1177/1745691610383510.

Contemporary Modeling of Gene-by-Environment Effects In Randomized Multivariate Longitudinal Studies

John J. McArdle and Carol A. Prescott

Department of Psychology University of Southern California

Abstract

There is a great deal of interest in the analysis of genotype by environment interactions ($G \times E$). There are some limitations in the typical models for the analysis of $G \times E$, including well-known statistical problems in identifying interactions and unobserved heterogeneity of persons across groups. The impact of a treatment may depend on the level of an unobserved variable, and this variation may dampen the estimated impact of treatment. A case has been made that genetic variation may sometimes account for unobserved, and hence unaccounted for, heterogeneity. The statistical power associated with the $G \times E$ design has been studied in many different ways, and most results show that the small effects expected require relatively large or non-representative samples (i.e., extreme groups). In this report, we describe some alternative approaches, such as randomized designs with multiple measures, multiple groups, multiple occasions, and analyses to identify latent (unobserved) classes of people. These are illustrated with data from the HRS/ADAMs study, examining the relations among episodic memory (based on word recall), APOE4 genotype, and educational attainment (as a proxy for an environmental exposure). Randomized clinical trials (RCT) or randomized field trials (RFT) have multiple strengths in the estimation of causal influences, and we discuss how measured genotypes can be incorporated into these designs. Use of these contemporary modeling techniques often requires different kinds of data be collected and encourages the formation of parsimonious models with fewer overall parameters, allowing specific $G \times E$ hypotheses to be investigated with a reasonable statistical foundation.

A simple summary of the role of genetic variation on behavior is provided by the expression

$$Phenotype = f\{Genotype, Environment\} + Error$$

That is, the observed phenotype is some function of the genotype (G) and the environments (E) in which the individual develops, with the further possibility of errors of omission or measurement error (i.e., noise). In theory, this function may take multiple forms but in practice a wide variety of alternatives are difficult to identify.

One complication is the possibility of G by E interaction ($G \times E$) -- whereby gene expression varies depending on the level of the environmental context or, equivalently, the direct effects of the environment on the measured phenotype vary depending on the genotype. Classical examples were based on plant and animal breeding studies (see Tryon, 1940; Cooper & Zubek, 1958). Until recently, testing $G \times E$ in human populations relied on the use of inferred genotypes and observational designs, such as adoption, discordant twin pair, and MZ-DZ twin studies (see Vandenberg & Falkner, 1965; Scarr-Salapatek, 1971; Harden, Turkheimer & Loehlin, 2007; McArdle & Plassman, 2009). More recent studies of $G \times E$ in human behavior have used measured genotypes to help untangle this puzzle (e.g., Caspi et al., 2003). The effect sizes of observed interactions have been very small and these

approaches have been the subject of several important methodological critiques, (e.g., Eaves, 2006; Jooper, Sengupta, & Schmitz, 2007; Monroe & Reid, 2008; Risch et al., 2009).

Another complication is the potential existence of *G and E correlation*. For many behaviors there is a rather obvious correlation between genotypes and environments (e.g., Scarr & McCartney, 1983). That is, persons with specific genotypes are not randomly assigned (or exposed) to environments, and some important correlation of *G* and *E* arises from selection effects. This *G&E* correlation may exist due to evolutionary selection (e.g., skin color and geographical latitude), or mate selection (people have children with partners who have similar traits), or even social selection (e.g., small physical stature leads to being bullied). Of course, on a statistical basis, even if two variables are uncorrelated in the population, they can be correlated in every sub-sample from that population (e.g., Thurstone, 1947).

The purpose of the current paper is not to question whether *Gx E* interactions or *G&E* correlations exist -- We assume that they do and that they are important in some contexts (e.g., Cronbach & Snow, 1977; Wilson, Jones, Coussens & Hanna, 2002; Thomas, 2004; Kendler & Prescott, 2006). Instead we ask, "If a *G* by *E* effect is important, how can we improve our chances of detecting it using current statistical models?" The analyses must be able to deal with *G&E* correlation as well -- either by sampling design or statistical control. To illustrate these issues we present results from analyses examining how variation in a measured gene (APOE4) influences episodic memory (EM) performance in older ages (>60 years). These data do not come from a randomized clinical or field trial, so the *G&E* correlation may exist, but we use high -quality longitudinal data which are publicly available and are useful for presenting key analytic issues (see Shadish, Cook & Campbell, 2002; Rubin, 2006). We illustrate options for fitting variations of *Gx E* models to the data using contemporary techniques from *structural equation modeling* (SEM). We then expand these formal considerations to include some benefits of longitudinal data, and we refit the *Gx E* models using longitudinal data. We then consider some issues of statistical power and the implications of the analytic results for designing *randomized clinical trials* (RCT) or *randomized field trials* (RFT) that include measured genotypes.

METHODS

The data used in this paper come from the publicly available *Aging, Demographics, and Memory Study* (ADAMS), a part of the *Health and Retirement Study* (HRS; see Langa et al., 2005; Plassman et al., 2008; McArdle, Fisher & Kadlec, 2007). The ADAMS/HRS sample initially included a sub-population of 1,700 individuals selected from the HRS with the ultimate goal of a detailed in-person neurological evaluation to assess dementia status. After several initial screenings, *N*=856 were evaluated, and three diagnostic groups were formed -- Non-Cases (35.6%), Demented cases (40.1%), and Cognitively-Impaired not Demented (CIND; 20.6%) individuals. In the analyses presented here, our outcome is Episodic Memory score (EM), based on number of words recalled following a few minutes delay. Recent analyses of these ADAMS/HRS data by McArdle, Koontz, Langa and Plassman (2009) focused on organizing the longitudinal trajectories of memory performance among ADAMS participants whose Word Recall scores were obtained on multiple occasions as part of the HRS data collection. Figure 1 shows the age-related results of longitudinal mixed-effects models of Word Recall for three different diagnostic groups (from McArdle et al., 2009). The pattern of changes overage show declines for all groups, but steeper declines for the CIND and the Demented groups.

Among many available variables possibly associated with memory, we selected for use in these analyses: APOE ε4 genotype, Education, and Age. APOE was selected because of its possible role in dementia (see Bäckman, Jones, Berger, Laukka, & Small, 2005; Bertram &

Tanzi, 2008) and because it was available on a large portion of the ADAMS sample. There was no attempt at a randomized manipulation in the ADAM study, and there are few purely environmental measures. So, to illustrate these methods, we use Educational Attainment as a proxy for an environmental exposure. Of course, we realize that Educational Attainment is not purely an environmental variable and has genetic roots (see Baker et al., 1996; McArdle & Plassman, 2010). But years of education is reliably measured, and for nearly all participants it was completed prior to the memory assessments (and, as a coincidence, it begins with “E”).

Table 1 is a list of sample means, variances, and correlations for the five variables used here: Episodic Memory score at the first HRS assessment (EM[1]), Age at time 1, APOE $\epsilon 4$ genotype, Education at time 1, and the numerical product of Education and APOE. In the subsequent analyses, Age is rescaled to be in decades and centered around 60 years, Genotype is coded as 0 or 1 to represent the absence or presence of an $\epsilon 4$ allele, Education is centered around 12 years, and $G \times E$ is the product of the coded education and genotype variables. (Only 14 individuals in the sample had two copies of the $\epsilon 4$ allele, so we combine them with those with one copy and code the score APOE4 as 0 or 1). In this sample, the correlation of Memory and Age is negative (-0.37), Memory and Education is positive ($+0.41$), Age and Education are negative (-0.14), but APOE4 is largely uncorrelated with all others (i.e., the Point-Biserial APOE4 with Education $r = -0.002$). As an aside, this lack of a correlation between the genotype and the putative environment is exactly what we would try to achieve with a typical RCT design because this independence allows us to separate observed effects of the environment and gene-environment interactions.

These values are based on using all available data. APOE genotype information is available on $N=842$ of the ADAMS/HRS sample, but other variables (Education) are only available on $n=772$ of the same persons. We address the incomplete data by using an SEM-EM-type algorithm (Little & Rubin, 1987; McArdle, 1994; Cnaan, Laird & Slasor, 1997) that allows us to examine these basic summary statistics “as if all persons were measured on all variables.” These statistics shown in Table 1 are fairly close to the pair-wise estimates and this typically indicates these data meet the minimal conditions of “missing at random” (MAR, Little & Rubin, 1987; Rubin, 2006). Most importantly, these estimated statistics do not suffer from some common statistical problems (local linear dependency), and this also means that we can routinely use whatever information is available from every person. It is possible that dealing with incomplete data as *unobserved but important scores* can alter inferences about impacts. This becomes much more of an issue when we consider the longitudinal data at every age, including considerations of the loss of participants due to selective attrition (see McArdle, Small, Bäckman & Fratiglioni, 2005). Thus, this estimate of incomplete data statistics is an essential part of all models used here. Of course, a relatively complex expression is required when we wish to include participants with partially incomplete data. To carry out subsequent analyses we use a variety of new computer programs which allow us to estimate parameters using all available data (e.g., computer scripts available upon request; see McArdle, 1996).

Using this approach, we conducted new analyses of these data using contemporary SEM based on regression and latent curve models. In the first analysis we estimate a simple regression of the Memory performance only at the initial testing (EM[1]), regressed on the APOE4 genotype (G), the Educational Level (E), and their potential interaction ($G \times E$). Once again, although educational attainment is not typically thought of as an environmental manipulation, especially since it is influenced by the genetic background, we can use it as a convenient measured proxy for multiple environmental exposures. In a second set of analyses we use a multiple group regression model to study whether sub-groups defined by genotype differ in their EM[1] scores. A third set of analyses explores the possibility of

latent subgroups using regression mixture models and tests the relationship of these classes to $G \times E$ interaction.

The remaining analyses use the longitudinal Memory data – In a longitudinal mixed-effects model we estimate latent levels and slopes based on up to seven occasions of data on Episodic Memory (EM[t], where $t=1$ to 7). As part of this model we estimate the regression of the measured APOE4 and Education and their interaction on the EM[t] latent levels and slope scores. The main goal is to eliminate some of the time-independent random noise and obtain a more accurate picture of the long-term effects of the $G \times E$ interaction. We then use the multiple groups model and a latent growth mixture model to explore the possibility of multiple trajectories for memory scores. The relationship of these trajectories to the concept of $G \times E$ interaction will be developed as we proceed.

RESULTS

Analysis 1: Regression Modeling in the Analysis of $G \times E$ Interactions

Our initial model presented above assumes we have at least one observed phenotype (Y), at least one measured environment or exposure (E), and at least one measured genotype (G). However, the functional relationships among these predictive elements were not fully specified. The most typical model fit to such data is based on the simple linear additive regression model. This standard regression model is depicted as a path diagram in Figure 2a, and is widely used because it takes into account the potential correlation across levels of E and G —i.e., estimating $G \times E$ in the context of $G \times E$ correlation. The regression weights are estimates of the independent (i.e., conditional mean) differences in the phenotypic scores that are expected from independent differences in the G and the E predictors.

Figure 2b includes a potential genotype -environment interaction by including the new product variable $G \times E$. Note the other model parameters take on different interpretations. For example, the regression weight for E is now the expected change in the phenotype for a one unit change in E when the G effect=0 and the $G \times E$ effect=0. The basic ANOVA/Regression principles are well known, so we will not elaborate further (e.g., see Tukey, 1949; Freeman, 1973; Aguinis, Beaty, Boik, & Pierce, 2005). However, it is noteworthy that the model accounts for correlation of *measured* environment with *measured* genotype, but we do not deal with unmeasured genotypes or environments. Any positive correlation between the environmental exposure and background genetic variation will appear as an increase in the main effect of the environment. Conversely, positive correlations between unmeasured environments and measured genetic variation will appear as an increase in the main effect of measured genotype.

The results of a multiple linear regression analysis are presented in Table 2. Following the typical statistical conventions, we consider “significant” effects to be those with a t value >1.96 (i.e., parameter estimate/standard error). The initial model (2a) uses only Age as a predictor and yields an overall $R^2=13\%$, with a significant intercept for EM at age 60 (6.29 words correct), a significant prediction of Age (-1.23 words lost per decade), and an improvement in fit ($\Delta L^2=57$) compared to a baseline model with no predictors (results not shown). The second model (2b) uses both Age and APOE as predictors, and this results in little improvement in fit ($\Delta L^2=2$ compared to 2a). The overall $R^2=14\%$, so the change in explained variance is small ($\Delta R^2=1\%$), and the APOE4 effect is not significant ($t=1.5$). The next model (2c) adds Education as an additional predictor, and yields an overall $R^2=25\%$, with significant predictors of Age (-1.11 words lost per decade), and Education ($+0.21$ words gained per year of schooling), and this substantially improves the model fit ($\Delta L^2=63$ compared to 2b). The final model (2d) adds the product (APOE $\times$ Education) and yields an overall $R^2=29\%$, with significant predictors of Age (-1.10 words lost per decade),

and Education (+0.24 words gained per year of schooling). Of some importance, the linear effect of APOE4 and the interaction with Education are not significant at the conventional test level (i.e., $t=-1.93$ for $G \times E$). Another standard way to test the interaction hypothesis is to drop the interaction term from the model (2c vs 2d) but this yields only a small loss of fit ($\Delta L^2=1$ on $df=1$).

Analysis 2: Multiple Group Regression Modeling for GxE Interactions

A serious limitation of the standard regression model approach is it assumes the residual terms are independent of the predictors and have the same size at all points on the regression line – i.e., *homogeneity of regression*. A simple way to restate this assumption is to say that the model parameters apply to all observations equally well. This is rarely considered a reasonable assumption in applied research, because there are likely to be sub-groups of individuals for whom different forms of these models apply.

One contemporary approach to the analysis of heterogeneous sub-groups can be found in SEM work on *multiple group* models (e.g., Sörbom, 1979; Horn & McArdle, 1992). Here, we defined independent groups based on APOE4 genotype and allowed the parameters representing the impact of any predictor on the outcome to vary over groups. The model can also include a difference in the intercepts to index the main effect of genotype. Instead of modeling GxE as a variable in the model, differences between the groups in the slope serve as an index of the $G \times E$ interaction. Each group also has an unobserved residual with possibly different variances, and this is also a testable hypothesis. This multiple group approach seems very reasonable when genotypes are nominal categories, such as when there are multiple markers or the genotypes do not have a clear ordering with respect to their impact on the phenotype.

Table 3 is a summary of the results of models fit to the same data as before, but using a two-group SEM regression in which each group is defined by level of APOE4 genotype (i.e., group 0 has no $\epsilon 4$ alleles, comprising 74% of the sample; group 1 has one or two $\epsilon 4$ s, comprising 26%). The parameter estimates from the first model (3a) assume invariance of all parameters. This yields an age regression where the starting point for EM at age 60 is high (6.16 words), the age decline is large (-1.10 words per decade), the education impact is positive (+0.24 per year of schooling), and the prediction of EM[1] from age and education within each group is fairly reasonable ($R^2=27\%$). The misfit is large ($L^2=1737$) indicating the groups are not exactly equal in all parameters. The second model (3b), allows each group to follow a different regression – we allow the regression parameters of EM[1] on Age to differ over groups, and this is equivalent to allowing both age-related main effects and interactions. The change in fit is not large ($\Delta L^2=7$ vs 3a) but the heuristic principle is useful. In the first group (APOE4=0), the starting point for EM at age 60 is high (6.36), the age decline is large (1.2), and the age prediction is relatively high ($R^2=31\%$). The second group ($\epsilon 4=1$) has a lower starting point (5.5 at age 60), but a less steep decline (0.7) and less prediction from age ($R^2=15\%$). The third model (3c) also allows the residual variance to differ over groups, and this does not improve the fit very much ($\Delta L^2=1$). The separation into these two groups based only the simple regression model does not dramatically improve the fit relative to the cost in fitting additional parameters, indicating little evidence for $G \times E$ interactions using this approach. However, this can be a useful approach for exploring genotypic differences that are less dependent upon model assumptions and the specific form of the interaction.

Analysis 3: Latent Class Regression Modeling for GxE Interactions

It is worth considering whether we have unmeasured sub-groupings of people in our data set. One approach to the analysis of unobserved but heterogeneous sub-groups can be found

in the work on *latent mixture* models (e.g., McLachlan & Peel, 2000; Muthén & Muthén, 2002; Bauer & Curran, 2003). The defining feature of this model is that a regression model is appropriate for each subgroup of persons, but there may be completely different regressions for different unobserved subgroups. To understand what people are likely to be in which classes, we can also write a prediction model (based on logistic regression) where the probability of latent class membership is influenced by the environments, genotypes, and their interactions. The use of multiple parameters and a combined model is relatively new in the context of measured genotypes, but it is potentially a reasonable and convenient way to deal with the problem of heterogeneity in a single group data collection design.

Table 4 is a summary of analyses using the same raw data as used previously, but here we fit a two-class latent mixture, allowing parts of the regression of EM[1] on Age and Education to differ over classes. The first model (4a) allows only the intercepts to differ over groups. The result suggests that about 85% of individuals are largely identified with class 1 and the other 15% are in class 2. There was a minor possibility of misclassification to groups when the groups were observed, but when the groups are not based on observed scores this is a major issue. Of some importance here is that the potential for misclassification is higher in the first group (31% vs 10%). As a description of these results, the first latent class follows an age curve where the intercept for memory at age 60 is lower (5.7 words) while the second class is higher (6.2). The model requires both groups to have the same age decline (-1.02 words per decade), and education effect ($+0.21$ per year of schooling), and the prediction of EM from age is fairly reasonable ($R^2=30\%$). As part of the same SEM, we fit a logistic regression equation describing this group separation from APOE4 status. If the two latent classes identified in this analysis were to be believed, then APOE4 genotype is not a significant predictor ($t=1.1$), and we would conclude the subgroups are not associated with measured genotype.

In a second model (4b), we allow all regression coefficients to vary between groups, and this results in the latent classes shifting quite a bit. A first latent class (comprising 61% of individuals) has a relatively low score at age 60 (5.51), a small age-related decline (-0.53), and low predictability ($R^2=24\%$). In contrast, the second latent class (comprising 39% of individuals), has a high score at age 60 (8.27) but strong age-related decline (-1.99) and relatively high predictability ($R^2=52\%$). The impact of education does not change much over classes (0.21 vs 0.24), even though these were estimated separately. This model seems to improve the fit to the data slightly ($\Delta L^2=12$ vs 4a), but this separation into classes is not predictable from APOE4 status ($t=0.4$), and the misclassification is relatively large (28% and 27%). The final model (4c) allows the residual variance to differ over classes and this improves the fit only slightly ($\Delta L^2=2$ vs 4b) and the resulting parameter values have similar interpretations. Although this separation into two classes does not dramatically improve the fit (relative to the cost in fitting additional parameters), these analyses are useful for understanding the evidence for heterogeneity that emerge in subsequent analyses using the longitudinal data.

Analysis 4: Longitudinal Latent Curve Models in the Analysis of G x E Interactions

The use of longitudinal data permits analyses that could not be considered otherwise, and here we use principles derived from *latent curve models* (LCM; McArdle, 1986, 2008; McArdle, Fisher, & Kadlec, 2007). As we use LCMs here, these models are standard mixed effects growth equations based on age-at-testing (see Albert, Blacker, Moss, Tanzi & McArdle, 2007; McArdle et al, 2009). We fit functional forms in which the phenotype at any occasion is the sum of three latent scores: (1) an intercept score (f_0), representing the reliable score at age 60, (2) a slope score (f_1), representing the systematic changes over each decade of age, and (3) a unique disturbance score ($d[t]$), representing the age-independent

fluctuations in the observations. The basis function $A[t]$ used here is a fixed linear function of age-at-testing and the latent variables. A path diagram of this model is shown in Figure 3.

Although there are many ways to consider the use of environmental and genetic measured variables, we will follow the regression principles used above. It is straightforward to examine these as influences on the newly estimated common latent variables. That is, we now estimate latent variable regressions where the impacts of environments, genotypes, and their interaction are based on the longitudinal phenotypes -- the parameters predicting the latent intercepts (α) and the age slopes (β) are estimated simultaneously along with the levels and slopes. A key feature of this analysis is that we can estimate the G , E , and $G \times E$ effects after conditioning out the occasion -specific latent scores ($u[t]$), which may represent noise or other variation not common across times of measurement.

Longitudinal data from the HRS Episodic Memory scores ($EM[t]$) includes one to seven occasions of measurement per person (as in Figure 1). Four models for these data are presented in Table 5. The first model fitted (5a) is a standard linear growth/decline model, and this provides a large improvement in fit ($L^2=2242$ on $df=3$) over a baseline or no-change model (results not shown). The fixed (mean) parameters indicate a predicted recall of about 6 words ($\alpha_0=5.62$) at age 60 with a decline of about half a word ($\beta_0=-0.59$) per decade. We also estimate a large variance at age 60 ($\sigma_0^2=3.66$), an important variation in the slope ($\sigma_1^2=0.68$), a negative covariance between the age-60 level and subsequent slope ($\sigma_{01}=-1.74$), and a substantial residual component ($\sigma_u^2=1.94$). To place these variances in context, we can calculate the intra-class ratio of variances at age 60, which suggests about 35% of the total variance ($\eta^2[60]=1.94/(3.66+1.94)*100$) is due to random noise. In another calculation we find the latent variance gets smaller over age (due to large negative covariances), so this constant sized noise ($\sigma_u^2=1.94$) represents a larger proportion of the score variation in older ages.

The second model fitted (5b) adds to the previous model by including APOE4 status as a predictor of both level and slopes. The improvement in fit is very small ($\Delta L^2=3$ vs 5a), indicating no main effects of APOE4 and no interaction of APOE4 with age. The third model (5c) adds Education and the fit is also improved slightly ($\Delta L^2=3$ vs 5b). Here the Education parameters are positively related to level (+0.27 per year of Education) but negatively related to the latent slope (-0.05). This kind of results is all too often interpreted (i.e., as indicating a greater decline for the higher educated) but this pattern of correlations is not actually a test of any substantive statement. Most importantly, the APOE4 parameters are still not significant. The fourth model (5d) allows for an interaction between Education and APOE4. The model fit is dramatically improved ($\Delta L^2=92$ vs 5c), providing statistical evidence for a significant $G \times E$ interaction. The parameters for the $G \times E$ interaction on both the $EM[t]$ levels and slopes are negative, indicating the presence of the $\epsilon 4$ allele is associated with a lower starting point and steeper decline and these effects are magnified with higher levels of Education.

The results of model 5d include a predicted initial level of 5.7 words at age 60 ($\sigma_0^2=2.7$), and initial level is significantly predicted by Education ($\alpha_e=0.31$ words/year; $t=7.1$) but this is offset by the Education -by-APOE4 interaction ($\alpha_i=-0.19$; $t=-2.3$). Together these effects account for about 47% of the initial level variance. This model also has significant slopes over age ($\beta_0=-0.59$, $\sigma_0^2=0.6$), and significant predictors of these slopes include both Education ($\beta_e=-0.08/\text{year}$; $t=3.8$) and the Education-by-APOE interaction ($\alpha_i=+0.12$; $t=-2.9$), which together account for about 69% of the slope variance.

One way to understand the broad results is to plot the *expected trajectories over age* for selected subgroups. Figure 4 shows the parameters from model 5d plotted over age for six

groups defined by three educational levels (8, 12, or 16 years) and two APOE4 groups (0 vs 1 or 2 $\epsilon 4$ alleles). The plot shows: (1) Age-related declines in Episodic Memory in all groups; (2) clear Education-related differences in their starting points (both groups with 16 years of education are higher than those with 12 years and these are higher than those with 8 years); (3) the effects of having an $\epsilon 4$ allele are not large, especially for the middle levels of education, but we do find some crossing of the curves, suggesting that the $\epsilon 4$ allele is associated with more decline in the presence of low education, but in the presence of high education is associated with a lower starting point but less decline. The complete story of this small interaction is fairly complicated.

It is interesting to note that, when employing all the longitudinal data to estimate these common and more reliable level and slope components, no main effect of APOE4 was found, but the small effect of the APOE4 $\times$ Education interaction is now significant. We note this same parameter was not significant in regressions based on the cross-sectional data (Analysis 1). These new interaction effects may have emerged for several reasons: (1) the number of data points (and hence the *dfs*) has increased, (2) precision is gained by eliminating the random noise, and (3) a latent slope has been formed as an individual-level variable. It is nearly impossible to isolate the meaning of level and slope impacts when latent level-slope correlation is present (here, $\rho_{01} = -0.6$), although this is often tried (see Reynolds et al., 2005). In the current example, the negative level -slope correlation reflects a tendency for those who start with better memories to show more decline over time. Perhaps more importantly, this negative correlation implies that the latent scores seem to have less variance over age. We return to this point in later discussion.

Analysis 5. Multiple Group Trajectory Models in the Analysis of G $\times$ E Interactions

A potential limitation of the LCMs used in Analysis 4 is the fundamental assumption that a single basis curve is best for all individuals (i.e., homogeneity of trajectories). It can be difficult, if not impossible, from fitting a standard LCM to know if this is a reasonable assumption. As noted previously, any multiple group SEM, where all the model parameters may differ over groups, may be useful. Of special interest here is the potential difference in latent basis $A[t]$ over groups, because this defines the shape of the curve.

The results of Table 6 are based on the same longitudinal data but we fit a two-group model based on APOE4, allowing us to test if EM[t] trajectories over Age differ by genotype. In the first model (6a), all the regressions are allowed to differ over groups. The overall change in the likelihood is not large ($\Delta L^2 = 6$), and many of the parameters are similar. This model suggests that the APOE4=0 group (including 74% of individuals) follow an age curve where the average score at age 60 is high ($\alpha_0 = 5.7$) and the age decline is large ($\beta_0 = -0.6$). As part of the same model, the APOE4 = 1 group follow a slightly different latent curve where the average score at age 60 is lower ($\alpha_0 = 5.4$) and the age decline is smaller ($\beta_0 = -0.5$). These differences are not large, but within each group the level and slope variance are fairly small ($\sigma_0^2 = 2.7$ and $\sigma_1^2 = 0.6$), suggesting these groups are fairly homogeneous. The final model (6b) allows all covariances of the latent variables to be free to vary across groups and there are only small differences in the results compared to model 6a.

Analysis 6: Latent Curve Mixture Models in the Analysis of G $\times$ E Interactions

As considered in the cross-sectional models in Analysis 3, heterogeneity over groups may not be directly measured by the measured predictors. We consider the possibility that the participants are members of latent classes characterized by different trajectories of memory over time. Once again, the defining feature is that the latent curve model is appropriate for each subgroup of persons, but there may be completely different latent curves for different people. Also, as before, it follows that we can write a logistic expression where the

probability of group trajectory membership is influenced by the environments, genotypes, and their interactions. The use of this combined model is usually not applied to measured genotypes, but it could be a reasonable and convenient way to deal with heterogeneity in a single group longitudinal data design.

Table 7 is a summary of results based on the same longitudinal data but fitting a two -class latent mixture model. Model 7a allows the intercepts of the model of EM[t] on Age to differ over classes, and the change in the likelihood is not large ($\Delta L^2=12$ relative to a baseline model of no differences). The results suggest that about 84% of the individuals follow an age curve where the average score at age 60 is high ($\alpha_0=5.9$) and the age decline is large ($\beta_0=-0.5$). The second class, with 16% of the sample, follows a different latent curve where the average score at age 60 is slightly lower ($\alpha_0=5.0$) but the age decline is the same ($\beta_0=-0.53$). The prediction of the classes is not related to APOE4 ($t<1.96$).

The second latent class model (7b) seems very different. Here we allow all latent components, means and variances, to differ, and the gain in fit is substantial ($\Delta L^2=77$ vs 7a). Of note here is that the first large group (93%) starts high ($\alpha_0=5.6$) and has a steep decline ($\beta_0=-0.53$), whereas the smaller group (7%) starts very low ($\alpha_0=1.4$) and changes little ($\beta_0=0.08$). Also, the regression of Education on initial memory is positive within class one ($\alpha_2=0.31$) but negative within class 2 ($\alpha_2=-0.26$). Within each group the level and slope variance are fairly small ($\sigma_0^2=2.6$ to 2.7 and $\sigma_1^2=0.4$ to 0.6). Here, genotype is associated with the group separation, with the $\varepsilon_4>0$ being much more likely in class 2. If this outcome of two latent curve classes were correct, then a substantial part of the differences in the group trajectories can be attributed to genotype.

DISCUSSION

Summary of Results for Episodic Memory Over Time

We repeat that Educational Attainment was used here for illustrative purposes and is not purely an environmental variable. Indeed, recent twin study results suggest the heritability of educational level is about 40% (e.g., Baker et al., 1996; Silventoinen, Kaprio & Lahelma, 2000; McArdle & Plassman, 2010). This means that any “*GxE*” finding may be indicative of a *GxG* interaction. Of course, this can happen with any variable assigned the artificial label of an “environment.” If our measured genotype and environmental exposure had been correlated, the interpretation of these modeling results would not be as clear and we would need to consider alternative mechanisms (e.g., mediation of APOE effects on education by early memory). The role of educational attainment in later life cognition is certainly complex (e.g., Gatz et al., 2006).

A summary of the results from our new modeling analyses follows:

1. Regression analysis of episodic memory as measured at the first occasion did not suggest significant effects of APOE4 genotype or genotype-by-education. Higher educational attainment was associated with higher initial levels of memory and steeper decline in memory over age.
2. Splitting the data into multiple cross-sectional groups did not alter this result.
3. Exploratory searches for multiple latent classes did not lead to evidence for *GxE* based on cross-sectional results.
4. When the longitudinal memory data were analyzed, there was a significant interaction of APOE4 genotype and educational status on memory decline. This occurred even though there was not a significant main effect of genotype on memory.

5. Splitting the data into multiple longitudinal groups did not alter this result.
6. Latent growth mixture models were considered in these longitudinal models, and the results suggest these may be some further differences between groups.

These basic results for Episodic Memory are consistent with findings from a recent study applying latent variable twin models to memory data from the NAS-NRC twin sample (McArdle & Plassman, 2009, 2010). Using typical twin assumptions, the impact of additive genetics was about 50%. Thus, there appear to be important genetic influences on memory, but the current analyses suggest that, in this sample, little of the variation in memory is attributable to APOE4 genotype (but see Caselli et al. 2009).

Given the associations reported between the APOE ϵ 4 allele and lower educational attainment in other samples (see Baxter, Caselli, Johnson, Reiman & Osborne, 2003; Caselli et al., 2004) the absence of an association in this sample may be surprising. A possible explanation is that samples of older adults differ in the degree to which educational level is an indicator of socioeconomic status vs a more direct measure of early life cognitive abilities, which would be expected to have a closer relation to genetic variation.

Implications for Including Genotype Information in a Randomized Trial

It is well known that *random assignment to treatment groups* permits key statistical assumptions about unobserved variables to be assured. That is, group assignment is assumed to be uncorrelated with all residual terms (in large samples) and this leads to unambiguous attribution of the treatment impacts in size and direction (e.g., Fisher, 1928; Mosteller & Boruch, 2002; Rubin, 2006). For the purposes of this discussion, *randomization to groups eliminates the correlation (in the long run) between the genotypes and environments*. In practice, this lack of a correlation leads to both ease of inference, and increased power to detect the main effects of *E* and *G*, as well as any *G*×*E* interaction. However, this can only be assured in large sample studies.

Randomized clinical trials (RCT) or *randomized field trials* (RFT) are widely used to isolate and estimate the specific effects of a well-defined treatment (Mosteller & Boruch, 2002). The RCT/RFT designs have multiple strengths in the estimation of causal influences, and there is no barrier to including the effects of measured genotypes as a part of this design. In randomized experiments investigators are often uncertain about the optimal exposure to training so they randomly assign a specific “dose” of training to individuals, and then examine the outcome or “response to training” in terms of increased performances on all types of measures (e.g., see Baltes, Dittmann-Kohli & Kliegl, 1986). The same analytic principles apply to this design, but the dose-response curve is often examined using regression techniques, and the curves can be nonlinear and, hopefully, identify the optimal dosage to achieve the peak outcome. If the training dose is randomized it is possible to make an inference about the impact of any level or dose, including the transfer of training. This randomization also allows estimation of the independent effects of the genotype with greater precision. The RCT design also reduces the possibility of correlation of treatment with unmeasured background genetic variation.

More precision is also gained from using a pre-post training design. This design allows the direct measurement of change, as well as estimating the differential impacts of the training on changes (Baltes & Nesselroade, 1979). It has become well known that, due to individual differences, the goals of randomization are not always achieved in randomized trials (e.g., Rubin, 2006). The possibility that training may be effective for only some individuals is a reasonable confound (e.g., Shadish, Cook & Campbell, 2002). Analyses based on mean differences and dose-response impacts within and between groups may need to be combined to deal with these issues, and the impacts of genotype on training may be evaluated as well.

The goals of almost any training are typically thought to be broad in scope – the goal is not to train the specific task but to impact the construct underlying the tasks (e.g., Horn, 1972; Baltes, Dittmann-Kohli, & Kliegl, 1986). For many reasons, it is quite typical to measure multiple variables in training experiments. These multivariate approaches merge naturally with the measured genotype models presented earlier – either by including genotype as a predictor of the latent factor, or by using genotype as a stratifying variable in multiple group versions of these multivariate models.

The multiple variable structural modeling approach used here can be generalized to consider more general possibilities for experimental impacts and for evaluating non-random confounds and subsequent causal inference (see Blalock, 1985; Shadish, Cook, & Campbell, 2002; Rubin, 2006). One recent example of SEM analysis of randomized trial data comes from the recent *Advanced Cognitive Training for Independent and Vital Elderly* (ACTIVE) study (Jobe et al., 2001; Ball et al., 2002; Willis et al., 2006). The published analyses of the ACTIVE data have used more advanced ANOVA methods than those used by others in the classical literature (e.g., Baltes et al., 1986). The statistical models used by McArdle & Prindle (2008) apply contemporary dynamic SEM to examine the ACTIVE training impacts. These analyses focus on the concept of the latent change score as the basis of the sequential influences of one variable upon another over time (McArdle, 2008). This dynamic SEM approach allows us to: (1) examine hypotheses about both Near and Far variables together, (2) examine mean differences between training groups, dose-response functions of training both between and within groups, (3) fit these hypotheses as models for latent constructs representing the multiple outcomes, and (4) explore further differences between and within groups. The distribution of any key latent parameters were thought to can come from a “mixture” of two or more overlapping distributions or “latent classes,” and the techniques described earlier were used to explore heterogeneity within the trained group (see McArdle & Prindle, 2009). In this way we analyzed the impacts of experimental manipulation on the dynamic trajectories, both within and between variables.

Confounds in Statistical Modeling

The linear models used here are an extremely simple functional form and should be considered a useful first step. The calculation of this model by regression methods is designed to explicitly deal with $G \times E$ correlation, as long as this correlation is not too high (i.e., co-linear). However, to identify the regression parameters we need to assume the residual is completely uncorrelated with the predictors, and this is never completely true. To obtain effective test statistics we need to assume the residuals follow a normal distribution and while this may be true for random noise, it is unlikely to be true for omitted predictors, either relevant genotypes or environments. As noted previously, there may be important covariation between E and unmeasured genotypes or between G and unmeasured environmental factors that will complicate the simple interpretation of G , E , and $G \times E$ estimates. Thus, any model is best considered as only a first approximation.

The inclusion of a product term in a model (3b) deals with $G \times E$ Interaction in a specific way. Initially, this test of interaction used above was reported by Tukey (1949) as a “single degree of freedom for no-additivity.” We have not fully considered the limitations of significance tests in this class of modeling, but control over Type-I errors (α -level) is of great concern when there are many significance tests examined, as is frequent in analysis of measured genotyping data (Eaves, 2006). It is clear that multiple non-independent tests will lead to inflated α -levels, and replication is essential. The control over Type-II errors is also a big issue, largely because the effects of any single gene can be very small, and statistical power ($1-\beta$) is compromised.

It is well known that statistical biases are created by the selection of the range of genotypes and environments being studied. In prior methodological research on the interaction model, some practitioners have argued that the inclusion of nonlinear terms such as $G \times E$ should only be acceptable if (a) significant main effects are present, and (b) a significant increase in precision (usually judged by incremental R^2) is obtained. In the current framework, it clearly makes sense to include the main effects in a regression model such as equation (3). The main effects allow the product term to depart from a pure parabola, a form that is unlikely to represent any real data collection. That is, although we might like to design pure measures, there is no sense in thinking that our sampling techniques have allowed for an exact centering of the interaction effects (i.e., a perfect cross). On the other hand, the main effects are simply control variables, so considerations about the significance of the effects, or defining increased fit, are not relevant to a standard statistical interpretation of these equations.

Our use of multiple group and latent class models allowed us to explore alternative forms of genotype main effects and interactions. Such approaches may uncover important heterogeneity between genotypes (in predictor variances or covariances; in outcome predictability) that would not be detected using simple linear models. A multiple group or latent class approach can also be used to investigate whether evidence of $G \times E$ is sensitive to variable scaling and model assumptions.

Any formal statistical model has limitations. For instance, in using the regression we assume both the outcome (phenotype) and the predictors (genotypes, environments) are measured perfectly, and this is hardly ever a reasonable assumption (see Wong, Day, Luan, Chan & Wareham, 2003). When any variable in the model is measured with random error, bias occurs through an inflation of the residual variance, with a corresponding decrease in the precision of the parameters (i.e., inflated standard errors). This applies to the phenotype, or environments, or even the genotypes. (Although automated genotyping is highly accurate, many markers do not index functional genetic variation, but are located nearby and are thus imperfectly correlated with the genotypes of interest). One solution to this problem comes from the introduction of multiple indicator measurement models, a possibility we have not considered here (but see McArdle & Prindle, 2009). The current longitudinal latent growth models presented here also have important limitations (see McArdle, 2008). The best scenario when model testing is that the likelihood changes will indicate a model is fundamentally flawed and should be rejected. It is important to remember that good fit does not necessarily mean a good model and there are some assumptions that are relatively difficult to overcome.

Statistical Genetics and Statistical Modeling

It is somewhat ironic that the technological advances that have enabled the precise automated genotyping of hundreds of thousands of markers for thousands of samples may contribute to greater imprecision in the literature from “findings” that represent Type I error. It is generally recognized that testing of multiple markers requires altering the alpha-test levels. However, such correction is often done by applying a published “rule” (e.g., $P < 10^{-7}$), without considering the specific situation, such as how many dependent variables are being analyzed, their inter-correlation, and how many statistical tests are conducted per marker.

Adding interaction terms to the list of tested parameters greatly increases the potential for serendipitous findings. One strategy commonly used in regression analyses for limiting the number of possible tests has been to test for interactions only in the case of significant main effects. However, there are theoretical reasons and empirical counter-examples for not doing this in the case of $G \times E$ (e.g., Table 6 here, Caspi et al., 2003). However, the potential for

concluding there is evidence for *GxE* becomes increasingly likely when researchers scan through hundreds (or thousands) of all possible combinations of allelic and genotypic statistical tests looking for significant *GxE* effects regardless of the genetic main effects. Of course, this is common practice in many areas of science, so these observations are not isolated to the search for measured gene effects.

On the other hand, when gene-environment interaction truly does exist, there may be many reasons our studies are unable to detect it. We may have low statistical power due to low frequency of the risk allele or small effect sizes of the genetic, environmental and/or *GxE* effects. We may be conducting so many statistical tests that we would need a huge effect size to overcome our corrections for multiple testing. There may be genetic-environmental covariance present. Even in the case (as in some RCT and RFT studies) when genotypes are randomly assigned to environmental interventions, unaccounted for environmental factors that are confounded with genotype could be the actual basis of differences attributed to genotype.

We suggest a preferable scenario is to select a set of *a priori* hypotheses, based on biological mechanisms and previous findings in other samples. These sources of information may also contain hypotheses about the *direction* of the effect, increasing the power still further. These confirmatory analyses could be complemented by exploratory analyses that include appropriate corrections for multiple testing. A trade-off between Type I and Type II error can be achieved by using a multi-stage design, which includes a built-in replication stage (e.g., van den Oord, 2008).

A number of power issues in *GxE* studies derive from the measurement – of the outcome, the exposure, and the genotype. If the outcome is categorical (e.g., having a disorder or not), power depends directly on the frequency distribution of the categories. Other things being equal, power to detect an outcome with 5% prevalence in a sample will be much smaller than to detect an outcome that occurs more frequently. But even in the case of an outcome with 50% frequency (as in most case-control designs), achieving power comparable to that for a continuous outcome will require several times as many participants. Careful thought should also be given to measurement of the environmental exposure variable. If the environmental variable is categorical, power will depend on what proportion of participants has been exposed to the risk condition. For ordinal and continuous variables, scaling in a way that is relevant for interacting with genotype may not correspond to typical assumptions (e.g., Kendler, Kuhn, Vittum, Prescott, & Riley, 2005).

Choice of the appropriate genetic variable for analysis may be most complex of all. For the simplest markers there are three possible genotypes (e.g., AA, Aa, aa). However, in practice, there are often more than three possible genotypes because some markers have more than two forms or information from multiple markers is combined (i.e., into haplotypes). For example, a dominant model (as employed in the APOE4 examples here) presumes that the same effect is obtained from having one or two copies of the risk allele. With the additive model, genetic information is coded in terms of the number of risk alleles the individuals has (i.e., 0, 1 or 2 copies). The impact of using the incorrect genetic model (i.e., fitting additive when in fact the alleles act in a dominant fashion) depends on the prevalence of the risk allele. In the case of APOE, which has a relatively low risk allele frequency (about 14% in Plassman et al., 2008), the two models will yield similar results unless the sample is very large. However, in cases where the risk allele is more common, using the incorrect model can lead to incorrectly concluding there is evidence for *GxE*. An implication for designing RCT and RFT studies is that it may be useful to screen and select participants based on genotype to ensure sufficient numbers of individuals with each genotype are assigned to receive each treatment.

Another set of issues relevant to designing *GxE* studies relates to the amount of measurement error. As indicated by the results of our example analyses, the numeric results of the longitudinal analyses were similar to those based on a single occasion. However, accounting for the occasion-specific variation substantially increases the ability to detect *GxE*. A simulation study by Wong and colleagues (2003) considered the impact of measurement error on power to detect *GxE*. Some relevant conclusions were: (1) The sample size needed to detect *GxE* is dependent upon the magnitude of the interaction, the allele frequency and the strength of the association between exposure and outcome. (2) The required sample size is highly dependent upon the measurement error in the exposure and outcome measures. (3) Studies employing imprecise exposure and outcome assessment may need to be 20 times larger than studies that utilize repeated and more precise measurement. Consequently, investment in better measurement may be a more cost-effective strategy for the detection of this form of *GxE* interaction than simply increasing sample size.

Current Limitations of Formal Statistical Modeling

There have been many recent discussions about the limits of applying a linear model approach in the context of *GxE* interactions (see Reynolds, Gatz, Berg & Pedersen, 2007; Shanahan & Hofer, 2005). These serve to remind us, as Freeman (1973) cautioned:

“...what can one say about the main effects of genotypes and environments when there is *GxE* interaction? When interactions are present, estimates of main effects are conditional: that is, one can only validly assert that genotypic effects are as observed in this particular set of environments, not over all possible environments.” (p. 345–346).

Another concern is that relying on analysis techniques that are widely available can also pose barriers to careful consideration of the hypothesized mechanisms of genetic and environmental effects. For example, the use of ANOVA methods and thinking may create an overly simplified view that does not adequately convey the complexity of the process:

“What genes really determine are the reaction ranges exhibited by individuals with more or less similar genes over the gamut of possible environments ... Heredity is not a status but a process. Genetic traits are not performed in the sex cells, but emerge in the course of development, when potentialities determined by the genes are realized in the process of development in certain environments..” (Dobzhansky, 1973; p.8).

This concept of the nature of genetic impacts, acting together with the environmental impacts in specific sequences is not clear from our initial equation, or even the other models presented here. The use of these models, although quite standard in the field, does not immediately suggest that the genotype limits the “reaction range” of phenotypes or that the environment is responsible for the measured behaviors within this range. Although concepts about more complex theoretical nonlinear sequences have been discussed (e.g., Waddington, 1962; Finch & Kirkwood, 2000) these are rarely utilized in behavioral sciences. However, in one sense, these concepts of restriction of range or “canalization” do seem to coincide with the use of multiple group trajectories – where the latent means change differentially over genotypes and the latent variances within each genotype decrease over time as well. Restriction of range and canalization are powerful ideas that are certainly worthy of more attempts at formalization. These processes imply that researchers designing RCT and RFT studies of *GxE* should consider the developmental stage at which these effects may occur or be detectable. The challenge of contemporary modeling is to create even more complex models which map directly onto these interesting ideas and help us reach our goal of characterizing the unfolding of genetic and environmental effects through development.

Acknowledgments

This work was supported by NIA Grant AG-7137-20 to the first author. We thank our colleagues David Conti (University of Southern California), Ken Langa (University of Michigan), Norman Henderson (Oberlin College) and the Editor, David Reiss (George Washington University), for their important contributions to this work. Kristen Fong assisted with manuscript preparation.

References

- Aguinis H, Beaty JC, Boik RJ, Pierce CA. Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology* 2005;90:94–107. [PubMed: 15641892]
- Albert M, Blacker D, Moss MB, Tanzi R, McArdle JJ. Longitudinal change in cognitive performance among individuals with mild cognitive impairment. *Neuropsychology* 2007;21:158–169. [PubMed: 17402816]
- Bäckman L, Jones S, Berger AK, Laukka EJ, Small BJ. Cognitive impairment in preclinical Alzheimer's Disease: A meta-analysis. *Neuropsychology* 2005;19:520–531. [PubMed: 16060827]
- Baker LA, Treloar SA, Reynolds CA, Heath AC, Martin NG. Genetics of Educational Attainment in Australian Twins: Sex Differences and Secular Changes. *Behavior Genetics* 1996;26:89–102. [PubMed: 8639155]
- Ball K, Berch DB, Helmers KF, Jobe JB, Leveck MD, Marsiske M, Morris JN, Rebok GW, Smith DM, Tennstedt SL, Unverzagt FW, Willis SL. Effects of cognitive training interventions with older adults: a randomized controlled trial. *Journal of the American Medical Association* 2002;288:2271–2281. [PubMed: 12425704]
- Baltes PB, Dittmann-Kohli F, Kliegl R. Reserved capacity of the elderly in aging-sensitive tests of fluid intelligence: Replication and extension. *Psychology and Aging* 1986;1:172–177. [PubMed: 3267394]
- Baltes, PB.; Nesselroade, JR. History and rationale of longitudinal research. In: Nesselroade, JR.; Baltes, PB., editors. *Longitudinal research in the study of behavior and development*. New York: Academic Press; 1979. p. 1-39.
- Bauer DJ, Curran PJ. Distributional assumptions of growth mixture models: Implications for over-extraction of latent trajectory classes. *Psychological Methods* 2003;8:338–363. [PubMed: 14596495]
- Baxter LC, Caselli RJ, Johnson SC, Reiman E, Osborne D. Apolipoprotein E epsilon 4 affects new learning in cognitively normal individuals at risk for Alzheimer's disease. *Neurobiol Aging* 2003;24:947–952. [PubMed: 12928055]
- Bertram L, Tanzi RE. Thirty years of Alzheimer's disease genetics: The implications of systematic meta-analyses. *Nature Reviews Neuroscience* 2008;9:768–778.
- Blalock, HM. *Causal models in panel and experimental designs*. New York: Aldine Publishing; 1985.
- Caselli RJ, Reiman EM, Osborne D, Hentz JG, Baxter LC, Hernandez JL, Alexander GE. Longitudinal changes in cognition and behavior in asymptomatic carriers of the APOE e4 allele. *Neurology* 2004;62:1990–1995. [PubMed: 15184602]
- Caselli RJ, Dueck AC, Osborne D, Sabbagh MD, Connor DJ, Ahern GL, Baxter LC, Rapcsak SZ, Shi J, Woodruff BK, Locke DEC, Snyder CH, Alexander GE, Rademakers R, Reiman EM. Longitudinal modeling of age-related memory decline and the APOE e4 effect. *New England Journal of Medicine* 2009;361:255–263. [PubMed: 19605830]
- Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, Harrington HL, McClay J, Mill J, Martin J, Braithwaite R, Poulton R. Influence of life-stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science* 2003;301:386–389. [PubMed: 12869766]
- Cnaan A, Laird NM, Slasor P. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine* 1997;16:2349–2380. [PubMed: 9351170]
- Cooper RM, Zubek JP. Effects of enriched and restricted early environments on the learning ability of bright and dull rats. *Can J Psych* 1958;12:159–164.
- Cronbach, LJ.; Snow, RE. *Aptitudes and instructional methods: A handbook for research on interactions*. Oxford, England: Irvington; 1977.

- Dobzhansky, T. Genetic diversity and human equality: The facts and fallacies in the explosive genetics and education controversy. New York: Basic Books; 1973.
- Eaves LJ. Genotype $\times$ environment interaction in psychopathology: Fact or artifact. *Twin Research and Human Genetics* 2006;9:1–8. [PubMed: 16611461]
- Finch, CE.; Kirkwood, T. Chance, development, and aging. New York: Oxford University Press; 2000.
- Fisher RA. The general sampling distribution of the multiple correlation coefficient. *Proceedings of the Royal Society, A* 1928;121:654–673.
- Freeman GH. Statistical methods for the analysis of genotype-environment interactions. *Heredity* 1973;31:339–354. [PubMed: 4589174]
- Gatz M, Mortimer JA, Fratiglioni L, Johansson B, Berg S, Andel R, Crowe M, Fiske A, Reynolds CA, Pedersen NL. Accounting for the relationship between low education and dementia: A twin study. *Physiology & Behavior* 2007;92:232–237. [PubMed: 17597169]
- Harden KP, Turkheimer E, Loehlin JC. Genotype by environment interaction in adolescents' cognitive aptitude. *Behavior Genetics* 2007;37:273–283. [PubMed: 16977503]
- Horn JL. State, trait, and change dimensions of intelligence. *The British Journal of Mathematical and Statistical Psychology* 1972;42:159–185.
- Horn JL, McArdle JJ. A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research* 1992;18:117–144. [PubMed: 1459160]
- Jobe JB, Smith DM, Ball K, Tennstedt SL, Marsiske M, Willis SL, Rebok GW, Morris JN, Helmers KF, Leveck MD, Kelnman K. ACTIVE: A cognitive intervention trial to promote independence in older adults. *Control Clinical Trials* 2001;22:453–479.
- Joober R, Sengupta S, Schmitz N. Promoting measured genes and measured environments: On the importance of careful statistical analysis and biological relevance. *Archives of General Psychiatry* 2007;64:377–378. [PubMed: 17339527]
- Kendler KS, Kuhn JW, Vittum J, Prescott CA, Riley B. The interaction of stressful life events and a serotonin transporter polymorphism in the prediction of episodes of major depression: A replication. *Archives of General Psychiatry* 2005;62:529–535. [PubMed: 15867106]
- Kendler, KS.; Prescott, CA. *Genes, Environment and Psychopathology: Understanding the Causes of Psychiatric and Substance Use Disorders*. New York: Guilford Publications; 2006.
- Langa KM, Plassman BL, Wallace RB, Herzog AR, Heeringa SG, Ofstedal MB, Burke JR, Fisher GG, Fultz NH, Hurd MD, Potter GG, Rodgers WL, Steffens DC, Weir DR, Willis RJ. The aging, demographics, and memory study: Study design and methods. *Neuroepidemiology* 2005;25:181–191. [PubMed: 16103729]
- Little, RTA.; Rubin, DB. *Statistical analysis with missing data*. New York: Wiley; 1987.
- McArdle JJ. Latent variable growth within behavior genetic models. *Behavior Genetics* 1986;16:163–200. [PubMed: 3707483]
- McArdle JJ. Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research* 1994;29:409–454.
- McArdle JJ. Current directions in structural factor analysis. *Current Directions in Psychological Science* 1996;5:11–18.
- McArdle JJ. Latent variable modeling of longitudinal data. *Annual Review of Psychology* 2008;60:577–605.
- McArdle JJ, Fisher GG, Kadlec KM. Latent variable analysis of age trends in tests of cognitive ability in the Health and Retirement Survey, 1992–2004. *Psychology and Aging* 2007;22:525–545. [PubMed: 17874952]
- McArdle JJ, Grimm K, Hamagami F, Bowles R, Meredith W. Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods* 2009;14:126–149. [PubMed: 19485625]
- McArdle, JJ.; Koontz, J.; Langa, K.; Plassman, BL. Unpublished manuscript. University of Southern California; 2009. Multivariate mixed-effects modeling of diagnostic differences in age trajectories of episodic memory in the ADAMS/HRS longitudinal panel. under review
- McArdle JJ, Plassman BL. A biometric latent curve analysis of memory decline in older men of the NAS-NRC Twin Registry. *Behavior Genetics* 2009;39:472–495. [PubMed: 19404731]

- McArdle, JJ.; Plassman, BL. Influences of Educational Attainment in a biometric latent curve analysis of memory decline in older men of the NAS-NRC Twin Registry. 2010. Manuscript submitted for publication
- McArdle JJ, Prindle JJ. A latent change score analysis of a randomized clinical trial in reasoning training. *Psychology and Aging* 2008;23:702–719. [PubMed: 19140642]
- McLachlan, GJ.; Peel, D. Finite mixture models. New York: Wiley; 2000.
- Monroe SM, Reid MW. Gene-environment interactions in depression research: Genetic polymorphisms and life-stress polyprocedures. *Psychological Science* 2008;19:947–956. [PubMed: 19000200]
- Mosteller, F.; Boruch, R., editors. Evidence matters: Randomized trials in education research. Washington, D.C: Brookings Institution Press; 2002.
- Muthén, L.K.; Muthén, B.O. M plus, the comprehensive modeling program for applied researchers user's guide. Los Angeles, CA: Muthen & Muthen; 2002.
- Plassman BL, Langa KM, Fisher GG, Heeringa SG, Weir DR, Ofstedal MB, Burke JR, Hurd MD, Potter GG, Rodgers WL, Steffens DC, McArdle JJ, Willis RJ, Wallace RB. Prevalence of cognitive impairment without dementia in the United States. *Annals of Internal Medicine* 2008;148:427–434. [PubMed: 18347351]
- Reynolds CA, Gatz M, Berg S, Pedersen NL. Genotype-environment interactions: Cognitive aging and social factors. *Twin Research and Human Genetics* 2007;10:241–254. [PubMed: 17564514]
- Reynolds CA, Finkel D, McArdle JJ, Berg S, Peersen NL. *Developmental Psychology* 2005;41:3–16. [PubMed: 15656733]
- Risch N, Herrell R, Lehner T, Liang KY, Eaves L, Hoh J, Griem A, Kovacs M, Ott J, Merikangas KR. Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression: A meta-analysis. *Journal of the American Medical Association* 2009;301:2462–2471. [PubMed: 19531786]
- Rubin, DB. Matched sampling for causal effects. New York, NY: Cambridge University Press; 2006.
- Scarr S, McCartney K. How people make their own environments: A theory of genotype → environment effects. *Child Development* 1983;54:424–435. [PubMed: 6683622]
- Scarr-Salapatek S. Race, social class, and IQ. *Science* 1971;174:1285–1295. [PubMed: 5167501]
- Shadish, W.; Cook, TD.; Campbell, DT. Experimental and quasi-experimental design for generalized causal inference. Boston, MA: Houghton-Mifflin; 2002.
- Shanahan MJ, Hofer SM. Social context in gene-environment interactions: Retrospect and prospect. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 2005;60:65–76.
- Silventoinen K, Kaprio J, Lahelma E. Genetic and environmental contributions to the association between body height and educational attainment: A study of adult Finnish twins. *Behavior Genetics* 2000;30:477–445. [PubMed: 11523706]
- Sörbom D. An alternative to the methodology for analysis of covariance. *Psychometrika* 1979;43:381–396.
- Thomas, D. Statistical methods in genetic epidemiology. Oxford: Oxford University Press; 2004.
- Thurstone, LL. Multiple-Factor Analysis. Chicago: University of Chicago Press; 1947.
- Tryon RC. The inheritance of maze-learning ability in rats. *Journal of Comparative Psychology* 1940;4:1–8.
- Tukey J. One degree of freedom for additivity. *Biometrics* 1949;5:232–243.
- Vandenberg SG, Falkner F. Heredity factors in human growth. *Human Biology* 1965;37:357–365.
- van den Oord EJ. Controlling false discoveries in genetic studies. *American Journal of Medical Genetics, Neuropsychiatric Genetics* 2008;147B:637–644.
- Waddington, CH. New Patterns in Genetics and Development. New York: Columbia University Press; 1962.
- Willis SL, Tennstedt SL, Marsiske M, Ball K, Elias J, Koepke KM, Morris JN, Rebok GW, Unverzagt FW, Stoddard AM, Wright E. Long-term effects of cognitive training on everyday functional outcomes in older adults. *Journal of the American Medical Association* 2006;296:2805–2814. [PubMed: 17179457]

- Wilson, S.; Jones, L.; Coussens, C.; Hanna, K., editors. Cancer and the environment: Gene-environment interaction. Washington, DC: National Academy of Sciences Press; 2002.
- Wong MY, Day NE, Luan JA, Chan KP, Wareham NJ. The detection of gene-environment interaction for continuous traits: Should we deal with measurement error by bigger studies of better measurement? *International Journal of Epidemiology* 2003;23:51–57. [PubMed: 12690008]

Figure 1.
Model-Expected Episodic Memory over Age for Three Diagnostic Groups (from McArdle, et al., 2009)

Figure 2.
Path models of alternative linear regression models with and without GxE interaction (after McArdle & Prescott, 1992)

Figure 3.
Path model of measured gene longitudinal model allowing linear latent curve for Episodic Memory over all ages and data waves (after McArdle & Prescott, 1992)

Figure 4.
Expected Means of Episodic Memory over Age by APOE4 Genotype and Educational Level
in ADAMS/HRS Sample

Table 1

Estimated Summary Statistics for the ADAMS/HRS Episodic Memory Time 1 Data assuming Missing At Random (MAR)

<i>(a) Univariate Statistics</i>	Estimated Mean	Estimated Variance	Minimum	Maximum
EM[1] (0–10)	4.76	4.27	0	10
Age[1] (years)	72.16	3.56	73	93
APOE4 (0 or 1)	0.26	0.19	0	1
Education (years)	11.74	12.44	8	14
APOE4 × Educ-12	−0.07	3.56	0	2
<i>(b) Estimated Correlations</i>	EM[1]	Age[1]	APOE4	Education
EM[1]	1.000			
Age[1]	−0.368	1.000		
APOE4	−0.055	−0.030	1.000	
Education	0.410	−0.135	−0.002	1.000
APOE4 × Education	0.147	−0.028	−0.064	0.533

Notes:

N=842. $L^2=-6675$ for test of no associations.

EM[1] – Episodic Memory score at assessment 1;

Age[1] – age in years at assessment 1;

APOE4 – coded 0 if no ε4 alleles, coded 1 if one or two ε4 alleles;

Education – years of educational attainment;

APOE4 × Educ-12 – product of APOE4 variable and Education centered around 12 (Educ-12).

Correlations with APOE4 are point-biserial values.

Table 2

Results for SEM Regression Models applied to the ADAMS/HRS Episodic Memory Obtained at Time 1

	(a) Linear Age	(b) + APOE4	(c) + Education	(d) + APOE4 × Education
	Parameter Estimate (T-value)			
$\beta_0 = 1 \rightarrow \text{EM}[1]$	6.29 (27)	6.32 (26)	6.24 (26)	6.24 (26)
$\beta_1 = \text{Age}[1] \rightarrow \text{EM}[1]$	-1.23 (8.2)	-1.24 (8.3)	-1.11 (7.5)	-1.10 (7.4)
$\beta_2 = \text{APOE4} \rightarrow \text{EM}[1]$	=0	-0.32 (1.5)	-0.31 (1.5)	-0.32 (1.6)
$\beta_3 = \text{Educ} \rightarrow \text{EM}[1]$	=0	=0	+0.21 (10)	+0.24 (9.8)
$\beta_4 = \text{Educ} \times \text{APOE4} \rightarrow \text{EM}[1]$	=0	=0	=0	-0.09 (1.93)
Unique ψ^2	3.65	3.55	3.10	3.08
	Model Fit Information			
Explained Variance R^2	13%	14%	25%	29%
Likelihood (L^2)	-6740	-6738	-6675	-6674
ΔL^2	+57	+2	+63	+1
Scale	2.21	2.19	2.19	2.15
# Parameters	14	15	16	17

Notes:

 $N=842$, MAR

EM[1] is Episodic Memory score at time 1.

Age= (age in years - 60)/10; Educ = (# years education-12)/4; APOE4 is coded 0 if no $\epsilon 4$ alleles, coded 1 if one or two $\epsilon 4$ alleles;

=0 indicates parameters fixed to 0;

Baseline model of intercept (β_0) only: $\Delta L^2 = -6797$, scale=2.24. ΔL^2 = change in likelihood relative to comparison model (a vs baseline; b vs a; c vs b; d vs c)

Results for Multiple Group SEM Regression Models applied to the ADAMS/HRS Episodic Memory Data Obtained at Time 1

Table 3

Groups:	(a) Invariant Parameters		(b) + Free Regressions		(c) + Free Residuals	
	(a0) $\epsilon_4=0$	(a1) $\epsilon_4=1$	(b0) $\epsilon_4=0$	(b1) $\epsilon_4=1$	(c0) $\epsilon_4=0$	(c1) $\epsilon_4=1$
Group %	73.7%	26.3%	73.7%	26.3%	73.7%	26.3%
Missclassification	0%	0%	0%	0%	0%	0%
Parameter Estimate (T-value)						
$\beta_0 = 1 \rightarrow \text{EM}[1]$	6.16 (27)	6.16 (27)	6.36 (24)	5.51 (12)	6.36 (24)	5.50 (12)
$\beta_1 = \text{Age}[1] \rightarrow \text{EM}[1]$	-1.10 (7.4)	-1.10 (7.4)	-1.20 (7.2)	-0.74 (2.4)	-1.20 (7.2)	-0.74 (2.4)
$\beta_3 = \text{Educ} \rightarrow \text{EM}[1]$	+0.21 (9.8)	+0.21 (9.8)	+0.24 (9.7)	+0.15 (3.7)	+0.24 (9.7)	+0.15 (3.7)
Unique ψ^2	3.12	3.12	3.07	3.07	3.06	3.08
Model Fit Information						
Explained Variance R^2	27%	27%	31%	15%	31%	15%
Likelihood (L^2)	-1737		-1730		-1729	
ΔL^2	--		+7		+1	
Scale	1.88		1.86		1.90	
# Parameters	10		13		14	

Notes:
 $N=842$, MAR; Group $\epsilon_4=0$ $N=620$; Group $\epsilon_4=1$ $N=222$; ΔL^2 = change in likelihood relative to prior model

Results for Two-Class Latent Mixture SEM Regression Models applied to the ADAMS/HRS Episodic Memory Data Obtained at Time 1

Table 4

	(a) Free Intercept		(b) + Free Regressions		(c) +Free Residuals	
	(a1) Class 1	(a2) Class 2	(b1) Class 1	(b2) Class 2	(c1) Class 1	(c2) Class 2
Class %	85.3%	14.7%	60.7%	39.3%	51.4%	48.5%
Misclassification	31.3%	10.4%	28.2%	26.5%	29.5%	30.3%
	Parameter Estimate (T-value)					
$\beta_0 = 1 \rightarrow \text{EM}[1]$	5.71 (19)	6.16 (27)	5.51 (12)	8.27 (13)	5.51 (12)	7.91 (13)
$\beta_1 = \text{Age}[1] \rightarrow \text{EM}[1]$	-1.02 (5.8)	-1.02 (5.8)	-0.53 (2.0)	-1.99 (7.2)	-0.39 (1.3)	-1.87 (6.2)
$\beta_3 = \text{Educ} \rightarrow \text{EM}[1]$	+0.21 (9.6)	+0.21 (9.6)	+0.21 (7.3)	+0.24 (4.5)	+0.20 (6.8)	+0.23 (5.1)
Unique ψ^2 (R^2)	2.43 (30%)	2.43 (30%)	2.23 (24%)	2.23 (52%)	1.95 (24%)	2.63 (45%)
Relation to APOE4	+1.59 (1.1)		-0.24 (0.4)		-0.13 (0.2)	
	Model Fit Information					
Likelihood (L^2)	-2221		-2209		-2207	
ΔL^2	--		+12		+2	
Scale	2.08		1.84		1.77	
# Parameters	12		14		15	

Notes: ΔL^2 = change in likelihood relative to prior model

Table 5

Results for Longitudinal Latent Curve/Mixed Effects Models Applied to Longitudinal HRS/ADAMS Episodic Memory Data

	(a) Age	(b) + APOE4	(c) + Educ	(d) + APOE4 × Educ
Parameter Estimate (T-value)				
<i>Fixed Effects</i>				
$\alpha_0=1 \rightarrow f_0$	5.62 (35)	5.70 (30)	5.72 (32)	5.72 (32)
$\beta_0=1 \rightarrow f_1$	-0.59 (6.9)	-0.60 (6.1)	-0.59 (6.2)	-0.59 (6.2)
$\alpha_1= \text{APOE4} \rightarrow f_0$	=0	-0.28 (0.8)	-0.33 (1.0)	-0.37 (1.1)
$\beta_1= \text{APOE4} \rightarrow f_1$	=0	+0.01 (0.8)	+0.19 (0.2)	+0.09 (0.5)
$\alpha_2= \text{Educ} \rightarrow f_0$	=0	=0	0.27 (7.1)	0.31 (7.1)
$\beta_2= \text{Educ} \rightarrow f_1$	=0	=0	-0.05 (2.8)	-0.08 (3.8)
$\alpha_3= \text{GxE} \rightarrow f_0$	=0	=0	=0	-0.19 (2.3)
$\beta_3= \text{GxE} \rightarrow f_1$	=0	=0	=0	+0.12 (2.9)
<i>Random Effects</i>				
Intercept σ_0^2	3.66 (5.3)	3.56 (5.3)	2.75 (5.1)	2.70 (4.9)
Slope σ_1^2	0.68 (3.9)	0.63 (3.9)	0.59 (4.2)	0.60 (4.2)
Covariance σ_{01}	-1.74 (3.6)	-1.74 (3.6)	-1.00 (3.8)	-1.00 (3.7)
Unique Variance Ψ^2	1.94 (23)	1.94 (23)	1.94 (23)	1.93 (23)
Model Fit Information				
Likelihood (L^2)	-2242	-2239	-2236	-2144
ΔL^2	--	+3	+3	+92
Scale	2.15	2.08	2.05	1.97
# Parameters	12	14	16	18

Age= (age in years - 60)/10; Educ = (# years education-12)/4; APOE4 is coded 0 if no $\epsilon 4$ alleles, and 1 if one or two $\epsilon 4$ alleles; ΔL^2 = change in likelihood relative to prior model

Table 6

Results from Longitudinal Multiple Groups Latent Curve Models Applied to Longitudinal HRS/ADAMS Episodic Memory Data

	(a) Regressions		(b) + Covariances	
	(a0) $\epsilon 4=0$	(a1) $\epsilon 4>0$	(b0) $\epsilon 4=0$	(b1) $\epsilon 4>0$
Class %	73.7%	26.3%	73.7%	26.3%
Misclassification	0%	0%	0%	0%
Parameter Estimate (T-value)				
<i>Fixed Effects</i>				
$\alpha_0=1 \rightarrow f_0$	5.73 (32)	5.36 (19)	5.74 (32)	5.29 (18)
$\beta_0=1 \rightarrow f_1$	-0.60 (6.2)	-0.51 (3.1)	-0.61 (6.3)	-0.47 (2.6)
$\alpha_2= \text{Educ} \rightarrow f_0$	0.31 (7.1)	0.13 (1.9)	0.31 (7.1)	0.12 (1.8)
$\beta_2= \text{Educ} \rightarrow f_1$	-0.08 (3.8)	0.03 (1.0)	-0.08 (3.8)	0.04 (1.1)
<i>Random Effects</i>				
Intercept σ_0^2	2.67 (4.9)	2.67 (4.9)	2.86 (4.5)	2.13 (2.0)
Slope σ_1^2	0.58 (4.1)	0.58 (4.1)	0.61 (3.9)	0.59 (1.8)
Covariance σ_{01}	-0.98 (3.7)	-0.98 (3.7)	-1.08 (3.5)	-0.76 (1.4)
Unique Variance Ψ^2	1.94 (23)	1.94 (23)	1.93 (23)	1.94 (23)
Model Fit Information				
ΔL^2	+6		+3	
Scale	1.89		1.76	
Parameters	16		19	

Notes:

α_0 is expected EM score at age 60 ; Educ = (# years education-12)/4; APOE4 is coded 0 if no $\epsilon 4$ alleles, and 1 if one or two $\epsilon 4$ alleles; ΔL^2 is change in likelihood: for model a relative to baseline model of a single class; for model b, relative to model a.

Table 7

Results of Longitudinal Latent Curve Mixture Models Applied to HRS/ADAMS Episodic Memory Data

	(a1) Class 1 Regressions	(a2) Class 2 Regressions	(b1) Class 1 +Covars	(b2) Class 2 +Covars
Class %	84.3%	15.7%	92.8%	7.1%
Misclassification	18.9%	35.0%	9.9%	1.4%
Parameter Estimate (T-value)				
<i>Fixed Effects</i>				
$\alpha_0 = 1 \rightarrow f_0$	5.87 (18)	5.04 (8.8)	5.60 (18)	1.37 (1.5)
$\beta_0 = 1 \rightarrow f_1$	-0.53 (3.4)	-0.53 (3.4)	-0.53 (5.9)	0.08 (0.2)
$\alpha_2 = \text{Educ} \rightarrow f_0$	0.27 (7.0)	0.27 (7.0)	0.31 (5.2)	-0.26 (2.3)
$\beta_2 = \text{Educ} \rightarrow f_1$	-0.05 (2.8)	-0.05 (2.8)	0.04 (1.1)	0.08 (1.5)
<i>Random Effects</i>				
Intercept σ_0^2	2.69 (4.4)	2.69 (4.4)	2.71 (4.8)	2.58 (3.7)
Slope σ_1^2	0.60 (4.1)	0.60 (4.1)	0.59 (4.1)	0.43 (3.9)
Covariance σ_{01}	-1.07 (3.9)	-1.07 (3.9)	-1.00 (3.7)	-1.00 (3.8)
Unique Variance Ψ^2	1.93 (23)	1.93 (23)	1.94 (23)	1.94 (23)
Relation to APOE4 Groups	+1.36 (1.3)		+2.60(14.5)	
ΔL^2				
Scale	+12		+77	
# Parameters	1.50		1.32	
	16		19	

Notes:

α_0 is expected EM score at age 60; Educ = (# years education -12)/4; APOE4 is coded 0 if no $\epsilon 4$ alleles, and 1 if one or two $\epsilon 4$ alleles; ΔL^2 is change in likelihood: for model a relative to baseline model of a single class; for model b, relative to model a.