Heritability estimation of cognitive phenotypes in the ABCD Study® using mixed models

Authors: Diana M. Smith[1-3], Rob Loughnan[4], Naomi P. Friedman[5], Pravesh Parekh[6], Oleksander Frei[6], Wesley K. Thompson[7], Ole A. Andreassen[6], Michael Neale[8], Terry L. Jernigan[2,9-11], Anders M. Dale[3,9-12]

Author Affiliations:
[1]Neurosciences Graduate Program, University of California San Diego, La Jolla, CA.
[2]Center for Human Development, University of California, San Diego, La Jolla, CA.
[3]Center for Multimodal Imaging and Genetics, University of California, San Diego School of Medicine, La Jolla, CA.
[4]Population Neuroscience and Genetics Lab, University of California, San Diego, La Jolla, CA.
[5]Institute for Behavioral Genetics and Department of Psychology and Neuroscience, University of Colorado Boulder, Boulder, CO.
[6]NORMENT, Division of Mental Health and Addiction, Oslo University Hospital & Institute of Clinical Medicine, University of Oslo, Oslo, Norway
[7]Center for Population Neuroscience and Genetics, Laureate Institute for Brain Research, Tulsa, OK.
[8]Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA.
[9]Department of Cognitive Science, University of California, San Diego, La Jolla, CA.
[10]Department of Radiology, University of California, San Diego School of Medicine, La Jolla, CA.
[11]Department of Psychiatry, University of California, San Diego School of Medicine, La Jolla, CA.
[12]Department of Neuroscience, University of California, San Diego School of Medicine, La Jolla, CA.

Running head: Heritability estimation in the ABCD Study using mixed models

Corresponding Author:
Diana M. Smith
9452 S Medical Ctr Dr
4th Floor, #4W217 C12
La Jolla, CA 92037
Tel: (617) 360-1280
d9smith@health.ucsd.edu

Abstract

Twin and family studies have historically aimed to partition phenotypic variance into components corresponding to additive genetic effects (*A*), common environment (*C*), and unique environment (*E*). Here we present the ACE Model and several extensions in the Adolescent Brain Cognitive Development Study (ABCD Study®), employed using the new Fast Efficient Mixed Effects Analysis (FEMA) package. In the twin sub-sample (*n* = 924, 462 twin pairs), heritability estimates were similar to those reported by prior studies for height (twin heritability = 0.86) and cognition (twin heritability from 0.00 to 0.61), respectively. Incorporating measured genetic relatedness and using the full ABCD Study® sample (*n* = 9,742) led to narrower confidence intervals for all parameter estimates. By leveraging the sparse clustering method used by FEMA to handle genetic relatedness only for participants within families, we were able to take advantage of the diverse distribution of genetic relatedness within the ABCD Study® sample.

Key words: heritability, twin studies, mixed models, cognition, height, random effects

Introduction

For over a century, researchers have relied on variance partitioning as a statistical method for estimating heritability (Carey 2003). Historically, twin studies provided an avenue by which researchers could model the variance of a given phenotype as comprised of distinct components: e.g., additive genetic effects (*A*), common environmental effects (*C*), and unique environmental effects (also including error or unmodeled unexplained variance; Martin and Eaves 1977; Neale and Maes 2004). Specifically, in a linear mixed-effect (LME) regression model,

$$y_{ij} = \mu + x'_{ij}\beta + A_{ij} + C_{ij} + E_{ij} \qquad (1)$$

where y_{ij} is the trait value of individual j in family i; μ is the overall mean; x_{ij} denotes a vector of covariates; and A_{ij} , C_{ij} , E_{ij} represent latent additive genetic, common environmental and unique environmental random effects, respectively.

Although the ACE model has often been implemented using structural equation model (SEM) software such as OpenMx (Neale et al. 2016), the SEM representation is mathematically equivalent to the LME regression model shown in Equation 1 (Neale and Maes 2004; Visscher et al. 2004; McArdle and Prescott 2005). Indeed, prior applications of the ACE framework have been implemented using LMEs from R and Stata packages (Rabe-Hesketh et al. 2008) as well as SAS (Wang et al. 2011). For studies that incorporate extended family designs with several random effects, Visscher and colleagues (2004) recommended implementation using a LME approach, though SEM methods also exist to model complex family structure (Truett et al. 1994; Keller et al. 2009).

Many of the software packages used to estimate coefficients of mixed-effects models are based on restricted maximum likelihood (REML) estimation to avoid bias introduced by inclusion of many fixed effects (Shaw 1987). With recent advances in genomic sequencing, there has been an influx of methods that use measured genetic data rather than inferred genetic similarity from twin status. For example, genome-wide complex trait analysis (GCTA; Yang et al. 2011) was developed to incorporate a pairwise genetic relatedness matrix (GRM) between individuals using information from single nucleotide polymorphisms (SNPs). Twin studies have subsequently been adapted to incorporate empirical measures of genetic relatedness (Kirkpatrick et al. 2021). However, incorporating a matrix of pairwise relatedness values for each set of participants leads to an increase in the computational time when estimating these model parameters. Various subsequent adaptations have been developed to increase the processing speed of GCTA software (Ge et al. 2015) and to incorporate effects of maternal and/or paternal genotype on the traits within GCTA (Eaves et al. 2014; Qiao et al. 2020; Eilertsen et al. 2021).

Comparison of heritability estimates derived from non-twin versus twin analyses have found that non-twin studies consistently yield lower heritability estimates, an example of the so-called “missing heritability” in genetics research (Kim et al. 2015). Some researchers have suggested that this phenomenon may be due in part to inflated twin heritability estimates, for example due to dominant genetic variation which might be masked by shared environment in twin and family studies (Chen et al. 2015). Indeed, twin and family studies have developed several ways of parsing “common environment”, including using geospatial location information (Heckerman et al. 2016; Fan et al. 2018) and adding a random effect of twin status (T) when including twins and full siblings in the same study (Zyphur et al. 2013).

The Adolescent Brain Cognitive Development Study (ABCD Study®) provides a particularly appealing dataset for estimation of heritability, not only due to its population sampling frame,

large sample size, and longitudinal design, but also because it contains an embedded sub-sample of 840 pairs of same-sex twins recruited through birth registries at four sites (Iacono et al. 2018). The overall sample is thus enriched for genetic relatedness, with families that include siblings, half siblings, dizygotic (DZ) twins, and monozygotic (MZ) twins. The ABCD Study® data therefore requires the application of modeling approaches that take family structure and relatedness into account.

In this study we implemented modeling strategies that account for family structure and pairwise genetic relatedness using the recently developed Fast Efficient Mixed Effects Analysis (FEMA; Fan et al. 2021). We used FEMA to model participants nested within families, where random effects such as genetic relatedness were taken into account for each pair of subjects within a family, and set to zero for individuals who are not in the same family (Fan et al. 2021). FEMA provides a flexible platform for users to specify a wide array of fixed and random effects, which makes it a useful tool for modeling variance components in the ABCD Study®.

We first compared the basic ACE model implemented in FEMA versus OpenMx (Neale et al. 2016). Next, we tested the effect of including measured genetic relatedness (using genotype array data) compared to assigning approximate relatedness based on zygosity (i.e., 1.0 for MZ twins, 0.5 for DZ twins and full siblings). We then progressively expanded our sample size, first going from “twins only” to the full ABCD Study® baseline sample (including non-twin siblings and singletons), and finally to the full sample across multiple timepoints. We compared model estimates for the commonly used *A*, *C*, and *E* components, as well as a subject-level component (*S*) in the longitudinal data, and the twin component (*T*), which captured the variance attributable to variance in the common environment of twin pairs. In addition, we explored the change in model estimates and model fit when adjusting for specific fixed effect covariates, and when excluding the twin sub-sample.

Methods

Sample

The ABCD Study® is a longitudinal cohort of 11,880 adolescents beginning when participants were aged 9-11 years, with annual visits to assess mental and physical health (Volkow et al. 2018). The study sample spans 21 data acquisition sites and includes participants from demographically diverse backgrounds such that the sample demographics approximate the demographics of the United States (Garavan et al. 2018). The sample includes many siblings as well as a twin sub-sample consisting of 840 pairs of same-sex twins recruited from state birth registries at four sites (Garavan et al. 2018). Exclusion criteria for participation in the ABCD Study® were: 1) lack of English proficiency in the child; 2) the presence of severe sensory, neurological, medical or intellectual limitations that would inhibit the child's ability to comply with the study protocol; 3) an inability to complete an MRI scan at baseline. The study protocols were approved by the University of California, San Diego Institutional Review Board. Parent/caregiver permission and child assent were obtained from each participant. The data used in this study were obtained from ABCD Study® data release 4.0.

Statistical analyses were conducted on a sample that included a total of 13,984 observations from 9,742 unique participants across two timepoints (the baseline and year 2 visits). The twin sub-sample used in this study consisted of 462 pairs of twins with complete data (total N = 924). Observations were included in the final sample if the participant had complete data across sociodemographic factors (household income, highest parental education), available genetic data (to provide ancestry information using the top 10 principal components), and the phenotypes of interest. Table 1 shows the baseline demographics of the full sample as well as

the twin sub-sample. Compared to the full sample, the twin sub-sample had a higher percentage of parents with bachelor's degrees (33.3% compared to 26.7% in the full sample), and household income was shifted higher (52.7% with income over $100,000 compared to 42.0% in the full sample).

Measures

Phenotypes of interest

For the present study, we included height as a phenotype of interest due to its common use in twin and family studies (Silventoinen et al. 2003), as well as the availability of larger genetic studies from samples of unrelated participants (Yengo et al. 2022). Several cognitive phenotypes were included from the NIH toolbox cognition battery (Gershon et al. 2013): specifically, we analyzed the raw composite scores measuring fluid and crystallized intelligence, which have been validated against gold-standard measures of cognition (Akshoomoff et al. 2013; Heaton et al. 2014). We also included the uncorrected scores from the flanker task, picture sequence memory task, list sorting memory task, pattern comparison processing speed, dimensional change card sort task (components of fluid cognition); and the oral reading recognition task and picture vocabulary task (components of crystallized cognition). In addition to the NIH Toolbox, we included the matrix reasoning test from the Wechsler Intelligence Scales for Children (WISC-V; Wechsler 2014), the total percent correct from the Little Man visuospatial processing task (Acker 1982), and the total number of items correctly recalled across the five learning trials of the Rey Auditory Verbal Learning Task (RAVLT; Daniel et al. 2014). See Extended Methods for a complete description of each phenotype of interest including data collection procedures.

Covariates

Unless otherwise specified, models were run on data that was pre-residualized for age and sex only, in keeping with common practice for twin studies (Neale and Maes 2004). In models that included pre-residualization for additional covariates, these were chosen based on common practices in cognitive and behavioral research, and included recruitment site, parental education, household income, and the first ten genetic principal components.

Genetic Principal Components and Genetic Relatedness

Methods for collecting genetic data have been described in detail elsewhere (Uban et al. 2018). Briefly, a saliva sample was collected at the baseline visit, as well as a blood sample from twin pairs. The Smokescreen™ Genotyping array (Baurley et al. 2016) was used to assay over 300,000 SNPs. Resulting genotyped and imputed SNPs were used for principal components derivation as well as genetic relatedness calculation.

The genetic principal components were calculated using PC-AiR (Conomos et al. 2015). PC-AiR was designed for robust population structure inference in the presence of known or cryptic relatedness. Briefly, PC-AiR captures ancestry information that is not confounded by relatedness by finding a set of unrelated individuals in the sample that have the highest divergent ancestry and computes the PCs in this set; the remaining related individuals are then projected into this space. This method has been recommended by the Population Architecture through Genomics and Environment Consortium (Wojcik et al. 2019), which is principally concerned with conducting genetic studies in diverse ancestry populations.

PC-AiR was run on using the default suggested parameters from the GENESIS package (Gogarten et al. 2019). We used non-imputed SNPs passing quality control (516,598 variants

and 11,389 individuals). Using the computed kinship matrix, PC-Air was then run on a pruned set of 158,103 SNPs, which resulted in 8,005 unrelated individuals from which PCs were derived – leaving 3,384 related individuals being projected onto this space.

We then computed a GRM using PC-Relate (Conomos et al. 2016). PC-Relate aims to compute a GRM that is independent from ancestry effects as derived from PC-AiR. PC-Relate was run on the same pruned set of SNPs described above using the first two PCs computed from PC-Air.

Data analysis

Pre-residualization

We used R version 3.6.3 for data processing. After obtaining the sample of complete cases for all variables, phenotypes were pre-residualized for age and sex using the *lm* function. For certain models (see Table 2), we additionally included the following covariates during this residualization step: site, parental education, income, and the first ten genetic principal components. The purpose of pre-residualization was to ensure that both FEMA and OpenMx implementations were fitting random effects to the same data. Because our models only fit random effects, and because FEMA implements an unbiased estimation of total variance, the FEMA implementation was therefore mathematically equivalent to the REML estimation in OpenMx. However, it should be noted that there are negligible differences between REML and maximum likelihood (ML) estimates when applied to large sample sizes such as those in the ABCD Study® sample (Browne and Draper 2006).

Previous work has found evidence for a practice effect in some of the cognitive measures from the ABCD Study® (Anokhin et al. 2022). Therefore, in models that included data from baseline and year 2, we included a "practice effect" as a dummy variable in the pre-residualization step. This variable was equal to 0 if the observation was the first instance of data for that participant (i.e., all participants had 0 at baseline), and 1 if the participants were providing data for the second time at the year 2 visit. Most participants (N = 4242, 76.19%) had a value of 1 at the year 2 visit.

Model specification

We ran a series of models, described in Table 2. For each model, we specified whether genetic relatedness was "measured" (calculated using PC-AiR and PC-Relate; Conomos et al. 2015, 2016) or "assigned". For assigned relatedness, we used the zygosity data from the twin sub-sample to assign a value of 1 for MZ twins, 0.5 for DZ twins, and 0.5 for all other siblings (under the assumption that there are only full siblings within a family).

Since each phenotype was pre-residualized, we only needed to estimate the random effects components in each LME run within FEMA and OpenMx. These included an effect of family ID (common environment, C), additive effect of genetic relatedness (A), subject (S), twin status (T, calculated by creating a variable "pregnancy ID" that was shared by any two individuals with the same family ID and same birth date), and unique environment/unexplained variance (E).

OpenMx

We first ran an ACE model in the baseline twin sample, using the *OpenMx* package in R (package version 2.20.6; Neale et al. 2016). We elected to use OpenMx as the comparison software due to its widespread use in twin and family studies to estimate heritability. We chose

to use the REML estimator within OpenMx, which differs from ML estimators by a) using an unbiased estimation to calculate total variance, and b) first estimating the random effects iteratively and then estimating the fixed effects coefficients, as opposed to alternating estimation of variances and fixed effects. However, due to the preresidualization step described above, in our models we solely estimated random effects, such that the REML estimator in OpenMx provided a good comparison for FEMA (a ML estimator that uses an unbiased estimation of total variance). We ran *OpenMx* using R version 3.6.3., using the default SLSQP optimizer. Because data were preresidualized for age and sex, we did not fit any additional covariates. OpenMx provides likelihood-based confidence intervals by default (Neale and Miller 1997), which we used to compare with the likelihood-based confidence intervals calculated in FEMA.

Fast Efficient Mixed Effects Analysis (FEMA)

FEMA was developed for the efficient implementation of mass univariate LMEs in high dimensional data (e.g., brain imaging phenotypes; Fan et al. 2021). Whereas the original version of FEMA used a method of moments estimator for increased computational efficiency, we modified the package to allow the user to select a ML estimator. An updated version of FEMA that includes this option is available at the time of this publication (https://github.com/cmig-research-group/cmig_tools). Because FEMA uses an unbiased estimation of total variance, and we were only fitting random effects and not fixed effects, the estimates from the FEMA implementation of ML regression were predicted to be mathematically equivalent to REML. To run FEMA, we passed a design matrix (the design matrix was “empty” because we were not fitting any fixed effects) as well as a file containing a matrix of (measured or assigned) genetic relatedness values. FEMA then used a nested random effects design to create a sparse relatedness matrix, in which the relatedness values for all participants not assigned the same familyID was set to zero. For all models we reported the random effects variances as a percent of the total variance in the residualized phenotype that was explained by

variance in the random effect of interest. As a result, for a given model, the random effects estimates sum to 1 (representing 100% of the variance in the residualized phenotype). For ease of interpretation and comparison to previous literature, in this paper the term “heritability estimate” refers to the percent of residualized phenotypic variance that is explained by variance in genetic relatedness, i.e., variance explained by variance in *A.*

Model Comparison

For comparing two models that used identical samples, we calculated the Akaike Information Criterion (AIC) as (Akaike 1974):

$$AIC = (-2)ln(likelihood) + 2k \quad (2)$$

where *k* represents the number of model parameters. Therefore, the difference in AIC between two models (*ΔAIC*) can be calculated as:

$$\Delta AIC = -2(\Delta LL) + 2(\Delta k) \quad (3)$$

where *ΔLL* represents the difference in log likelihood between the two models and *Δk* represents the difference in the number of parameters between the two models. In models that have the same level of complexity (*Δk* = 0), the *ΔAIC* is equal to -2(*ΔLL*). We chose to use the AIC as opposed to the likelihood ratio test statistic for model comparison because several comparisons were not between nested models.

Results

ACE model (Model 1) in FEMA versus OpenMx

A summary of heritability estimates (i.e., the *A* random effects) from all models is provided in Supplementary Table 1. To compare model estimates between OpenMx and FEMA, we fit the same ACE model in each, using the same sample of 462 complete twin pairs from the twin sub-sample (i.e., pairs in which each twin had complete data for all phenotypes). Figure 1 shows a comparison of the two results as well as the parameter estimates using FEMA. The difference in heritability estimates between the two software packages was less than 0.001 for all phenotypes. On comparing these models (Figure 1B), we found that the difference in AIC was less than 0.05 for all phenotypes, indicating that there was no difference in the model fit. Because the model estimates and model fits were practically the same, we elected to use the LME implementation in FEMA for all further analyses.

Effect of including measured genetic relatedness (Model 2)

To assess the difference in parameter estimates when including measured versus assigned genetic relatedness, we fit two versions of the ACE model in the baseline twin sample. Model 1 (the ACE model described above, implemented in FEMA) used a matrix of assigned relatedness values (1.0 for MZ twins and 0.5 for DZ twins) whereas Model 2 used a matrix of measured relatedness values.

The models provided equivalent heritability estimates, with differences in A estimates ranging from -0.01 (Little Man Task) to 0.03 (pattern comparison; Figure 2A). On inspecting the differences in the AIC between the two models, we found that using measured GRM led to small

improvements in the overall model fit. This improvement was most pronounced for height (*ΔAIC* = -1.04) but less so for the cognitive phenotypes (Figure 2B). Random effects variance component estimates are presented in Figure 2C.

Effect of increased sample size (Model 3)

We next tested the change in model estimates when moving from the twin sub-sample (*n* = 924) to the full baseline sample (*n* = 8,242). Model 2 (from previous analysis) and Model 3 both used the measured GRM values and included *A*, *C*, and *E* random effects. The two models are therefore equivalent except for the much larger sample fit in Model 3. As described in Methods, the sparse clustering method within FEMA ignored the genetic relatedness among individuals with different family IDs. In practice, this meant that the sample of 8,242 unique subjects at baseline was clustered into 7,136 families, and genetic relatedness values were only used for individuals within families.

Figure 3 shows the estimates from Model 3 and their comparison to Model 2. The increased sample size led to much smaller confidence intervals for all random effects estimates calculated in Model 3 (Figure 3A). In general, using the full sample led to smaller estimates of A and larger estimates of C compared to Model 2. The changes in heritability estimates ranged from -0.31 (NIH Toolbox Fluid Cognition) to +0.04 (height). The estimated total variance was larger in Model 3 for most phenotypes (Figure 3B), with the largest increase in variance in total composite cognition (24.36% increase in total variance), crystallized cognition (23.76% increase), and oral reading recognition (24.08% increase). Because the two models were fit to different samples, it was not possible to directly compare their AIC model fit from the likelihood statistics. Figure 3C shows the random effects variances from Model 3.

Adding a Twin random effect (Model 4)

Given that the full sample analysis included singletons, half siblings, and adopted siblings, as well as twins and triplets, we next tested the addition of a random effect of twin status (*T*). We calculated a "pregnancy ID" that was shared by individuals who had the same family ID and the same birth date. We then used this "pregnancy ID" to code for the *T* random effect in an ACTE Model (Model 4). Figure 4 shows the model estimates from Model 4 as well as a comparison to Model 3; the two models are equivalent with the exception of the *T* random effect.

For most phenotypes, the addition of the *T* random effect did not lead to a change in parameter estimates (i.e., *T* was estimated to be 0). The largest change in parameter estimates was in matrix reasoning (heritability estimate decreased by 0.07, *T* estimated at 0.07; Figure 4A). Model comparison found that the difference in the AIC was at or near 2.0 for all phenotypes except for the RAVLT (ΔAIC = 1.58) and matrix reasoning (ΔAIC = 0.85). Because the AIC was calculated as $-2\Delta LL$ plus double the difference in model parameters (Equation 3), the consistent values of 2.0 reflect that the $-2\Delta LL$ statistic was approximately 0 before the penalization for the additional parameter in Model 4 (Figure 4B). The random effects variances for Model 4 are shown in Figure 4C.

Incorporation of two timepoints (Model 5, 6)

We next moved from examining the full sample at baseline (Model 4) to the full sample at baseline and Year 2 (Model 5). Models 4 and 5 were equivalent except for the difference in sample size (i.e., Model 5 did not account for nesting of data within subjects, in order to directly assess this effect in Model 6). Because not all phenotypes were available at the Year 2 visit, models that included baseline and Year 2 data only included pattern comparison processing speed, flanker task performance, picture sequence memory, picture vocabulary, oral reading

recognition, crystallized cognition, RAVLT, Little Man Task, and height. To account for nesting of multiple visits within subjects, we added a random effect of subject (*S*) in Model 6. Figure 5 shows the change in random effects variances moving from Model 4 to Model 5 and from Model 5 to Model 6.

Adding the second visit led to overall increases in the heritability estimates for the cognitive phenotypes, with changes ranging from -0.06 (Little Man Task) to +0.38 (pattern comparison; Figure 5A). Conversely, the heritability estimate for height decreased by 0.21. Adding the random effect of subject led to minimal change (<0.001) in the heritability estimate for height, Little Man Task, and RAVLT, but a decrease in heritability estimates across the other cognitive phenotypes (changes ranging from -0.19 to -0.05) compared to estimates from Model 5. The total difference in heritability estimates going from Model 4 to Model 6 ranged from -0.21 (height) to +0.22 (pattern comparison; Figure 5C).

Model comparison between Model 5 and Model 6 found that the model was substantially improved for crystallized cognition (ΔAIC = -5.26), oral reading recognition (ΔAIC = -7.45), picture vocabulary (ΔAIC = -5.61), flanker (ΔAIC = -11.33), and pattern comparison (ΔAIC = -25.40). However, the difference in the fit was smaller for height (ΔAIC = +2.00), picture sequence memory (ΔAIC = +0.91), the RAVLT (ΔAIC =+2.00), and the Little Man Task (ΔAIC = +2.00; Figure 5D). Figure 5B and 5E show the random effects variances from Model 5 and Model 6.

Effect of assigning genetic relatedness in large samples (Model 7-9)

For the next set of models, we tested whether the parameter estimates for models using the full sample (Models 3-5) changed in the absence of measured genetic relatedness. We used a matrix of assigned genetic relatedness (assigning 1.0 for MZ twins from the twin sub-sample, and 0.5 for DZ twins from the twin sub-sample and all other individuals in the same family). The assigned relatedness value therefore assumed that all non-twins in the same family, as well as twins who were not part of the twin sub-sample, were full siblings.

Figure 6 compares the ACTSE longitudinal model with an equivalent model that used assigned genetic relatedness. Supplementary Figure 1 shows the same question of assigned versus measured relatedness applied to Models 3 and 4. Overall, the random effects estimates were largely unchanged with the use of assigned GRM, with the largest changes in the ACTSE model occurring in flanker (ΔA = 0.09) and pattern comparison (ΔA = -0.09; Figure 6A, Supplementary Figure 1A,D). Model comparison using ΔAIC found that the ACTSE model using measured GRM had better model fit for height (ΔAIC = -34.25), crystallized cognition (ΔAIC = -15.69), oral reading recognition (ΔAIC = -19.43), picture vocabulary (ΔAIC = -6.87), and pattern comparison (ΔAIC = -8.67) compared to the model using assigned GRM; the difference in model fit was less pronounced for picture sequence memory (ΔAIC = +0.84) and flanker (ΔAIC = +0.45; Figure 6B, Supplementary Figure 1B,E). Figure 6C and Supplementary Figure 1C and 1F show the random effects variances for models using assigned genetic relatedness.

Residualizing for additional covariates (Models 10-14)

While it is common in twin and family analyses to include only age and sex as fixed effects, behavioral scientists often include additional fixed effects such as sociodemographic variables or recruitment site as covariates. To test whether the inclusion of such variables led to changes in our random effects estimates, we ran several of our original models with additional variables included in the pre-residualization step (i.e., site, parental education, income, and the first ten genetic principal components). Figure 7 shows the results of this model comparison applied to Model 1 (the “classic” ACE model). Supplementary Figure 2 shows the same pre-residualization and model comparison applied to Models 2-4 and 6. In the classic ACE model, the *A* estimate tended to decrease and the *C* estimate tended to decrease in the models that included additional covariates (Figure 7A). Residualizing for additional covariates led to a decrease in the total residual variance across all phenotypes, with decreases ranging from -2.67% (RAVLT) to -26.02% in the ACE model(crystallized cognition; Figure 7B). Because the two models were run on different datasets (pre-residualized for different covariates), we did not calculate the difference in AIC between the two models. Figure 7C and Supplementary Figure 2C, 2F, 2I, and 2L show the random effects variances for the models that were residualized for additional covariates.

Effect of removing the twin-enriched sample (Models 15-16)

The size and structure of the ABCD Study® cohort, with its embedded twin sub-sample as well as the large number of related participants, led us to test the degree to which the model fit depended on having a large subset of MZ and DZ twins. As a proxy for the general population, we removed the twin sub-sample. This left a small number of twins and triplets recruited through the general recruitment pipeline (168 twin pairs and 6 sets of triplets, with 57 pairs of participants with genetic relatedness > 0.9 across the full sample). The number of twin and

triplet sets in this sample (174 out of 8131 pregnancies, 2.14%) was less than the 3.11% twin birth rate reported in the general population of the United States (Osterman et al. 2021). We therefore assumed that the ABCD Study® sample excluding the embedded twin sub-sample was a proxy for a population sample with a naturally occurring number of twins. We then fit an ACSE model, applied to the full sample excluding the twin sub-sample, at baseline and year 2, to represent the "best" model possible of those explored thus far, excluding the *T* random effect (Model 16). We compared this model to the same ACSE model applied to the full sample, inclusive of twins (Model 15).

A comparison of the parameter estimates is shown in Figure 8A. The model excluding the twin sub-sample led to a difference in *A* estimates of -0.19 (picture sequence memory task) to +0.12 (pattern comparison). Excluding the twin sub-sample led to an increase in the total residual variance across all phenotypes, with changes ranging from +0.24% (pattern comparison) to +17.55% (Little Man Task; Figure 8B). Because the two models were fit to different samples, it was not possible to directly compare model fit from the likelihood statistics. Figure 8C shows the random effects variances from the model that omitted the twin sub-sample participants.

Discussion

In this paper we present results from different modeling strategies for implementing the ACE model using LMEs, as implemented in FEMA. FEMA is capable of applying the ACE model as well as incorporating additional features such as using a sparse matrix of within-family genetic relatedness and a random effect of subject to model longitudinal data. Notably, the use of FEMA to incorporate relatedness across all subjects within a family allows for the flexibility to include the full ABCD Study® sample, rather than restricting analysis to the twin sub-sample. After expanding our analyses to include the full sample, even when genetic relatedness was

assumed rather than measured, and in the absence of a twin-enriched sample, changes in the model estimates of heritability and other random effects were generally small.

We first applied the ACE model in the baseline twin sample. FEMA and OpenMx found nearly equivalent estimates for all random effects variances, demonstrating the equivalence of the two models being fitted. We estimated the heritability of height at 0.86, which is near the top of the range of twin heritability estimates reported by a comparative study of twin cohorts in eight countries (ranging from 0.68 to 0.87; Silventoinen et al. 2003). Our twin heritability estimate for height was higher than the SNP heritability, which was recently estimated to be 40% of phenotypic variance in European ancestry populations and 10%-20% in other ancestries (Yengo et al. 2022). Of the cognitive phenotypes, we found the highest twin heritability estimate for total composite cognition (0.61) and oral reading recognition (0.58), consistent with prior findings that heritability estimates tend to be higher for more “crystallized” and culturally sensitive measures of cognition (Kan et al. 2013). Interestingly, the picture vocabulary test had a relatively lower heritability estimate in this model (0.24) compared to the reading recognition test (0.58), which may reflect a difference in the cultural sensitivity of the two “crystallized” cognition tasks. The NIH Toolbox tasks comprising fluid cognition (flanker task, picture sequence memory task, list sorting, pattern comparison, and dimensional card sort) ranged in heritability estimates from 0.22 (flanker) to 0.41 (picture), which is within the wide range of heritability estimates for similar tasks in children (approximately 0-0.6; see Kan et al. 2013). Interestingly, the RAVLT had near-zero estimates for all random effects variances in all models, indicating that this task may be exceptionally unreliable in this sample, or perhaps particularly prone to variance in measurement.

We next tested the change in model fit and parameter estimation when using measured genetic relatedness rather than assigned relatedness based on twin zygosity. Parameter estimates

were largely unchanged, reflecting that in a twin sample, the assigned relatedness values of 0.5 and 1 are sufficient to arrive at similar random effects estimates compared to models using measured relatedness (though the model fit was improved with the measured relatedness values).

Perhaps one of the most exciting applications comes when extending the model to the full ABCD Study® sample. By leveraging the sparse clustering method used by FEMA to handle genetic relatedness only for participants within families, we were able to take advantage of the diverse distribution of genetic relatedness, ranging from 0 (e.g. adopted siblings) to 1 (i.e., MZ twins) for any pair of participants within a family. Unlike the large computational load generated by other similar genome-based REML regressions, the use of sparse clusters allowed FEMA to dramatically cut the computational time (Fan et al. 2021), allowing all the analyses in this paper to be fit on a single machine without the use of parallel computing. Using the full sample, first at baseline then with the addition of the Year 2 data, led to narrower confidence intervals, as shown in Figure 3. Inclusion of the full sample led to lower heritability estimates for several cognitive phenotypes, which may be related to the relative homogeneity of the twin sub-sample leading to potential for overestimation of heritability. Of note, though singletons (participants who are the sole members of their family cluster) did not contribute to estimation of the random effects variances themselves, they did contribute to the estimation of the total variance, which allows the model to leverage the full ABCD Study® sample.

After expanding the model to include the full sample, we tested the effect of an added random effect of twin status (i.e., “pregnancy ID”). We found evidence for a *T* effect in matrix reasoning, with a compensatory decrease in the heritability estimate when *T* was included in the model. This *T* effect could include any components of the environment that are shared between twins but not among siblings. Examples could include shared uterine environment and prenatal

factors, such as gestational age; or the fact that twins experience the same environmental events at exactly the same time. To illustrate this point, a pair of twins might experience a global pandemic at exactly the same age, causing them to experience any effects of the event in similar ways. In contrast, if two siblings are different ages at the time of the event, it might have a different age-dependent effect on each of them (despite the fact that it is occurring as part of their "common environment"). Future work could further investigate additional effects, such as gestational age or specific age × environment interactions, to tease apart the multifactorial influences that relate to shared twin environments.

We next used the complete sample across multiple timepoints, for a total of over 13,000 observations (Figure 5). Adding the second timepoint led to a substantial decrease in the heritability estimate for height, with a similar increase in the *E* component for height. This may be due to several factors, including possible nonadditive genetic effects (e.g., Silventoinen et al. 2008). Conversely, many of the cognitive phenotypes (with the exception of the Little Man Task and the RAVLT) saw an increase in heritability estimates when modeled across multiple timepoints (Figure 5E). It is possible that this phenomenon is related to the well documented increase in apparent heritability of cognitive traits with age (Davis et al. 2009; Haworth et al. 2010), which may be due in part to the gene × environment correlation (Loughnan et al. 2019). The estimation of the *S* variance component varied by phenotype; height had a negligible *S* component, which may be due to the large amount of variance that was already explained by genetic and environmental effects. On the other hand, the NIH Toolbox tasks each had a variance component explained by subject-level variance, indicating that variance in these phenotypes may be relatively more stable for a given participant over time. For these tasks, including *S* in our model allows for better explanation of variance that would otherwise be part of the *E* component. The Little Man Task and the RAVLT did not exhibit subject-specific variance,

which may be related to higher noise in these measures as evidenced by the large *E* components for both tasks (.79 and .97 in Model 6, respectively).

Of all the models described in this paper, the model including A, C, T, S, and E, fit across the full ABCD Study® sample and using all timepoints (Model 6), represented the "most complete" model. However, we employed a series of model comparisons to assess the effect of various study design considerations on the random effects variances. First, we examined the change in our model results when using only "assigned" genetic relatedness, to approximate a study design in which genetic data are not readily available. We found that, as expected, the model fit was worse in this model, but the parameter estimates were generally similar. Of note, we deliberately used the twin sub-sample data to "assign" relatedness values, meaning that for these analyses the twins recruited through the general population were assumed to have a relatedness value of 0.5. Despite this deliberate attempt to increase the error in our model, estimates remained relatively similar, with inflated estimates for the *T* variance component that seemed to compensate for the induced error in relatedness values. These results indicate that when using assigned relatedness, variance that would have been attributed to increased genetic relatedness is "shifted" into the *T* component.

We next tested whether including additional covariates in our pre-residualization step would lead to a change in random effects estimates. In general, residualizing for sociodemographic and genetic ancestry covariates led to a decrease in the total residual variance as well as the common environment (*C*) parameter estimate. This was expected, as adjusting for additional covariates led to a better model fit; the improved model fit is accompanied by a smaller amount of residual variance that is not accounted for by the fixed effects, and any variance that would have been partitioned into *C* was already attributed to the covariates such as household income or parental education. Notably, adjusting for genetic principal components is an attempt to

include potential influences of population stratification in the model, and does not have an effect on the estimation of genetic relatedness. Due to the nesting structure of the random effects, the FEMA package only uses the pairwise genetic similarity between individuals within the same family. This is in contrast to the genetic principal components, which are used to estimate a fixed effect across the whole sample that may represent population stratification and other effects of genetic ancestry.

Finally, we tested whether omitting the twin sub-sample led to a difference in model results. Overall, the model estimates only slightly changed for most phenotypes, with the exception of the picture sequence memory task which saw a decrease of 0.19 in its heritability estimate. The confidence intervals generated by the two models were similar, suggesting that a large study sample with many siblings is capable of generating model estimates that are similar to those in a twin-enriched sample. Due to the difference in recruitment strategy between the twin sub-sample and the general population sample (recruited primarily through schools; Garavan et al. 2018), it is possible that these groups differed in ways that could lead to different heritability estimates.

The results from this study should be considered in light of certain limitations. Generally, LMEs are used to partition the variance in a phenotype of interest into components modeled by random effects; however, models are often built with the assumption that the random effects are mutually independent and follow the normal distributions with mean 0 (Neale and Maes 2004; Wang et al. 2011). Additionally, LMEs represent a “top-down” heritability estimation method that can be biased by several factors including gene–environment correlations, selection, non-random mating, and inbreeding (Zaitlen and Kraft 2012; Zhang and Sun 2022). Furthermore, we did not explore non-additive genetic effects, which can attenuate bias of heritability estimates

(Wang et al. 2011); nor did we model any gene × environment interactions, which are likely to exist for some of the phenotypes of interest (Loughnan et al. 2019).

This work describes many of the modeling techniques available for researchers interested in applying the ACE model and its extensions to a large sample with high relatedness such as the ABCD Study® sample. Notably, the FEMA package provides a tool for mass univariate estimation of LMEs, and its current implementation does not allow for bivariate mixed models. SEM and other implementations of bivariate linear mixed models may provide an avenue to address questions involving genetic and environmental correlations between variables. Bivariate models may provide some insight into questions of innovation, i.e., whether the set of genes that influence a given phenotype changes over time.

The last several years have seen the development of several new techniques that can be used to model additional relationships, such as random effect × time interaction (He et al. 2016), random effect × covariate interaction (Arbet et al. 2020), covariance among random effects (Zhou et al. 2020; Dolan et al. 2021), and allowing random effects estimates to vary as a function of the phenotype (Azzolini et al. 2022). The sparse clustering design employed in the FEMA package leads to improved computational efficiency compared to other LME implementation software (Fan et al. 2021); future work will investigate the use of FEMA to estimate random effects estimates in more high-dimensional datasets, such as the brain imaging data present in the ABCD Study®, and compare with other computationally efficient implementations of the ACE Model such as Accelerated Permutation Inference for the ACE Model (APACE; Chen et al. 2019) and positive semidefinite ACE (PSD-ACE; Risk and Zhu 2021). More broadly, as stated by (Zyphur et al. 2013), “top down” heritability estimates should serve as just one piece of the puzzle connecting genes and the environment, where current

techniques at the molecular and single-gene level may be useful in filling in the gaps from the bottom up.

Declarations

Funding

Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development Study (ABCD Study®; https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children aged 9-10 and follow them over 10 years into early adulthood. The ABCD Study® is supported by the National Institutes of Health, USA, and additional federal partners under award numbers U01DA041022, U01DA041028, U01DA041048, U01DA041089, U01DA041106, U01DA041117, U01DA041120, U01DA041134, U01DA041148, U01DA041156, U01DA041174, U24DA041123, U24DA041147, U01DA041093, and U01DA041025. A full list of supporters is available at https://abcdstudy.org/federal-partners. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/consortium_members. ABCD Study® consortium investigators designed and implemented the study and/or provided data but did not all necessarily participate in analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD Study® consortium investigators. The ABCD Study® data repository grows and changes over time. The data were downloaded from the NIMH Data Archive ABCD Study® Collection Release 4.0 (DOI: 10.15154/1523041). This work was supported by Kavli Institute for Brain and Mind Innovative Research Grant 2022-2195.

Acknowledgements and Conflicts of Interest

The authors wish to thank the youth and families participating in the ABCD Study® and all staff involved in data collection and curation. Dr. Dale reports that he was a Founder of and holds equity in CorTechs Labs, Inc., and serves on its Scientific Advisory Board. He is a member of the Scientific Advisory Board of Human Longevity, Inc. He receives funding through research grants from GE Healthcare to UCSD. The terms of these arrangements have been reviewed by and approved by UCSD in accordance with its conflict of interest policies. The remaining authors have no conflicts of interest.

Ethics approval

The ABCD Study® protocols were approved by the University of California, San Diego Institutional Review Board.

Consent to Participate

Parent/caregiver permission and child assent were obtained from each participant.

Consent for Publication

Not applicable.

Availability of Data and Material

ABCD Study® data release 4.0 is available for approved researchers in NIMH Data Archive (NDA; http://dx.doi.org/10.15154/1523041).

Code Availability

Code for this project is available at https://github.com/dmysmith/behav-genet-2022. FEMA is available at https://github.com/cmig-research-group/cmig_tools.

Authors' contributions

Conceptualization: Diana M. Smith, Anders M. Dale; **Methodology:** Diana M. Smith, Rob Loughnan, Naomi P. Friedman, Oleksander Frei, Wesley K. Thompson, Michael Neale, Anders M. Dale; **Formal analysis and investigation:** Diana M. Smith; **Writing - original draft preparation:** Diana M. Smith; **Writing - review and editing:** Rob Loughnan, Naomi P. Friedman, Pravesh Parekh, Wesley K. Thompson, Ole A. Andreassen, Michael Neale, Terry L. Jernigan, Anders M. Dale; **Funding acquisition:** Diana M. Smith, Terry L. Jernigan; **Resources:** Oleksander Frei, Michael Neale, Anders M. Dale; **Supervision:** Wesley K. Thompson, Ole A. Andreassen, Michael Neale, Terry L. Jernigan, Anders M. Dale.

References

Acker W (1982) A computerized approach to psychological screening—The Bexley-Maudsley Automated Psychological Screening and The Bexley-Maudsley Category Sorting Test. Int J Man-Mach Stud 17:361–369. https://doi.org/10.1016/S0020-7373(82)80037-0

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19:716–723. https://doi.org/10.1109/TAC.1974.1100705

Akshoomoff N, Beaumont JL, Bauer PJ, et al (2013) VIII. NIH Toolbox Cognition Battery (CB): composite scores of crystallized, fluid, and overall cognition. Monogr Soc Res Child Dev 78:119–132

Anokhin AP, Luciana M, Banich M, et al (2022) Age-related changes and longitudinal stability of individual differences in ABCD Neurocognition measures. Dev Cogn Neurosci 54:101078. https://doi.org/10.1016/j.dcn.2022.101078

Arbet J, McGue M, Basu S (2020) A robust and unified framework for estimating heritability in twin studies using generalized estimating equations. Stat Med 39:3897–3913. https://doi.org/10.1002/sim.8564

Azzolini F, Berentsen GD, Skaug HJ, et al (2022) The heritability of BMI varies across the range of BMI—a heritability curve analysis in a twin cohort. Int J Obes 46:1786–1791. https://doi.org/10.1038/s41366-022-01172-6

Baurley JW, Edlund CK, Pardamean CI, et al (2016) Smokescreen: a targeted genotyping array for addiction research. BMC Genomics 17:145. https://doi.org/10.1186/s12864-016-2495-7

Browne WJ, Draper D (2006) A comparison of Bayesian and likelihood-based methods for fitting multilevel models. Bayesian Anal 1:473–514. https://doi.org/10.1214/06-BA117

Carey G (2003) Human Genetics for the Social Sciences. SAGE Publications, Inc., Thousand Oaks, CA

Chen X, Formisano E, Blokland GAM, et al (2019) Accelerated estimation and permutation inference for ACE modeling. Hum Brain Mapp 40:3488–3507. https://doi.org/10.1002/hbm.24611

Chen X, Kuja-Halkola R, Rahman I, et al (2015) Dominant genetic variation and missing heritability for human complex traits: Insights from twin versus genome-wide common SNP models. Am J Hum Genet 97:708–714. https://doi.org/10.1016/j.ajhg.2015.10.004

Conomos MP, Miller MB, Thornton TA (2015) Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genet Epidemiol 39:276–293

Conomos MP, Reiner AP, Weir BS, Thornton TA (2016) Model-free estimation of recent genetic relatedness. Am J Hum Genet 98:127–148

Daniel MH, Wahlstrom D, Zhang O (2014) Equivalence of Q-interactive and paper administrations of cognitive tasks: WISC-V. Q-Interact Tech Rep 8:

Davis OSP, Haworth CMA, Plomin R (2009) Dramatic increase in heritability of cognitive development from early to middle childhood: An 8-year longitudinal study of 8,700 pairs of twins. Psychol Sci 20:1301–1308. https://doi.org/10.1111/j.1467-9280.2009.02433.x

Dolan CV, Huijskens RCA, Minică CC, et al (2021) Incorporating polygenic risk scores in the ACE twin model to estimate A–C covariance. Behav Genet 51:237–249. https://doi.org/10.1007/s10519-020-10035-7

Eaves LJ, Pourcain BSt, Smith GD, et al (2014) Resolving the effects of maternal and offspring genotype on dyadic outcomes in Genome Wide Complex Trait Analysis (“M-GCTA”).

Behav Genet 44:445–455. https://doi.org/10.1007/s10519-014-9666-6
Eilertsen EM, Jami ES, McAdams TA, et al (2021) Direct and indirect effects of maternal, paternal, and offspring genotypes: Trio-GCTA. Behav Genet 51:154–161. https://doi.org/10.1007/s10519-020-10036-6
Fan CC, McGrath JJ, Appadurai V, et al (2018) Spatial fine-mapping for gene-by-environment effects identifies risk hot spots for schizophrenia. Nat Commun 9:5296. https://doi.org/10.1038/s41467-018-07708-7
Fan CC, Palmer CE, Iversen JR, et al (2021) FEMA: Fast and efficient mixed-effects algorithm for population-scale whole-brain imaging data. bioRxiv. https://doi.org/10.1101/2021.10.27.466202
Garavan H, Bartsch H, Conway K, et al (2018) Recruiting the ABCD sample: Design considerations and procedures. Dev Cogn Neurosci 32:16–22. https://doi.org/10.1016/j.dcn.2018.04.004
Ge T, Nichols TE, Lee PH, et al (2015) Massively expedited genome-wide heritability analysis (MEGHA). Proc Natl Acad Sci 112:2479–2484. https://doi.org/10.1073/pnas.1415603112
Gershon RC, Wagster MV, Hendrie HC, et al (2013) NIH toolbox for assessment of neurological and behavioral function. Neurology 80:S2–S6. https://doi.org/10.1212/WNL.0b013e3182872e5f
Gogarten SM, Sofer T, Chen H, et al (2019) Genetic association testing using the GENESIS R/Bioconductor package. Bioinformatics 35:5346–5348
Haworth CMA, Wright MJ, Luciano M, et al (2010) The heritability of general cognitive ability increases linearly from childhood to young adulthood. Mol Psychiatry 15:1112–1120. https://doi.org/10.1038/mp.2009.55
He L, Sillanpää MJ, Silventoinen K, et al (2016) Estimating modifying effect of age on genetic and environmental variance components in twin models. Genetics 202:1313–1328. https://doi.org/10.1534/genetics.115.183905
Heaton RK, Akshoomoff N, Tulsky D, et al (2014) Reliability and Validity of Composite Scores from the NIH Toolbox Cognition Battery in Adults. J Int Neuropsychol Soc 20:588–598. https://doi.org/10.1017/S1355617714000241
Heckerman D, Gurdasani D, Kadie C, et al (2016) Linear mixed model for heritability estimation that explicitly addresses environmental variation. Proc Natl Acad Sci 113:7377–7382. https://doi.org/10.1073/pnas.1510497113
Iacono WG, Heath AC, Hewitt JK, et al (2018) The utility of twins in developmental cognitive neuroscience research: How twins strengthen the ABCD research design. Dev Cogn Neurosci 32:30–42. https://doi.org/10.1016/j.dcn.2017.09.001
Kan K-J, Wicherts JM, Dolan CV, van der Maas HLJ (2013) On the nature and nurture of intelligence and specific cognitive abilities: The more heritable, the more culture dependent. Psychol Sci 24:2420–2428. https://doi.org/10.1177/0956797613493292
Keller MC, Medland SE, Duncan LE, et al (2009) Modeling extended twin family data I: Description of the cascade model. Twin Res Hum Genet 12:8–18. https://doi.org/10.1375/twin.12.1.8
Kim Y, Lee Y, Lee S, et al (2015) On the estimation of heritability with family-based and population-based samples. BioMed Res Int 2015:1–9. https://doi.org/10.1155/2015/671349
Kirkpatrick RM, Pritikin JN, Hunter MD, Neale MC (2021) Combining structural-equation modeling with genomic-relatedness-matrix restricted maximum likelihood in OpenMx. Behav Genet 51:331–342. https://doi.org/10.1007/s10519-020-10037-5
Loughnan RJ, Palmer CE, Thompson WK, et al (2019) Gene-experience correlation during cognitive development: Evidence from the Adolescent Brain Cognitive Development (ABCD) Study. bioRxiv. https://doi.org/10.1101/637512
Martin NG, Eaves LJ (1977) The genetical analysis of covariance structure. Heredity 38:79–95.

https://doi.org/10.1038/hdy.1977.9
McArdle JJ, Prescott CA (2005) Mixed-effects variance components models for biometric family analyses. Behav Genet 35:631–652. https://doi.org/10.1007/s10519-005-2868-1
Neale MC, Hunter MD, Pritikin JN, et al (2016) OpenMx 2.0: Extended structural equation and statistical modeling. Psychometrika 81:535–549. https://doi.org/10.1007/s11336-014-9435-8
Neale MC, Maes HHM (2004) Methodology for genetic studies of twins and families. Kluwer Academic Publishers B.V., Dordrecht, NL
Neale MC, Miller MB (1997) The use of likelihood-based confidence intervals in genetic models. Behav Genet 27:113–120. https://doi.org/10.1023/A:1025681223921
Osterman M, Hamilton B, Martin J, et al (2021) Births: Final Data for 2020. National Center for Health Statistics (U.S.)
Qiao Z, Zheng J, Helgeland Ø, et al (2020) Introducing M-GCTA a Software Package to Estimate Maternal (or Paternal) Genetic Effects on Offspring Phenotypes. Behav Genet 50:51–66. https://doi.org/10.1007/s10519-019-09969-4
Rabe-Hesketh S, Skrondal A, Gjessing HK (2008) Biometrical modeling of twin and family data using standard mixed model software. Biometrics 64:280–288. https://doi.org/10.1111/j.1541-0420.2007.00803.x
Risk BB, Zhu H (2021) ACE of space: Estimating genetic components of high-dimensional imaging data. Biostatistics 22:131–147. https://doi.org/10.1093/biostatistics/kxz022
Shaw RG (1987) Maximum-likelihood approaches applied to quantitative genetics of natural populations. Evolution 41:812–826. https://doi.org/10.1111/j.1558-5646.1987.tb05855.x
Silventoinen K, Pietiläinen KH, Tynelius P, et al (2008) Genetic regulation of growth from birth to 18 years of age: The Swedish young male twins study. Am J Hum Biol 20:292–298. https://doi.org/10.1002/ajhb.20717
Silventoinen K, Sammalisto S, Perola M, et al (2003) Heritability of Adult Body Height: A Comparative Study of Twin Cohorts in Eight Countries. Twin Res 6:399–408. https://doi.org/10.1375/136905203770326402
Truett KR, Eaves LJ, Walters EE, et al (1994) A model system for analysis of family resemblance in extended kinships of twins. Behav Genet 24:35–49. https://doi.org/10.1007/BF01067927
Uban KA, Horton MK, Jacobus J, et al (2018) Biospecimens and the ABCD study: Rationale, methods of collection, measurement and early data. Dev Cogn Neurosci 32:97–106. https://doi.org/10.1016/j.dcn.2018.03.005
Visscher PM, Benyamin B, White I (2004) The use of linear mixed models to estimate variance components from data on twin pairs by maximum likelihood. Twin Res Hum Genet 7:670–4. https://doi.org/10.1375/twin.7.6.670
Volkow ND, Koob GF, Croyle RT, et al (2018) The conception of the ABCD study: From substance use to a broad NIH collaboration. Dev Cogn Neurosci 32:4–7. https://doi.org/10.1016/j.dcn.2017.10.002
Wang X, Guo X, He M, Zhang H (2011) Statistical inference in mixed models and analysis of twin and family data. Biometrics 67:987–995. https://doi.org/10.1111/j.1541-0420.2010.01548.x
Wechsler D (2014) WISC-V: Technical and interpretive manual. NCS Pearson, Incorporated
Wojcik GL, Graff M, Nishimura KK, et al (2019) Genetic analyses of diverse populations improves discovery for complex traits. Nature 570:514–518
Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: A tool for Genome-wide Complex Trait Analysis. Am J Hum Genet 88:76–82. https://doi.org/10.1016/j.ajhg.2010.11.011
Yengo L, Vedantam S, Marouli E, et al (2022) A saturated map of common genetic variants associated with human height. Nature. https://doi.org/10.1038/s41586-022-05275-y
Zaitlen N, Kraft P (2012) Heritability in the genome-wide association era. Hum Genet 131:1655–

1664. https://doi.org/10.1007/s00439-012-1199-6

Zhang L, Sun L (2022) Linear mixed-effect models through the lens of Hardy–Weinberg disequilibrium. Front Genet 13:856872. https://doi.org/10.3389/fgene.2022.856872

Zhou X, Im HK, Lee SH (2020) CORE GREML for estimating covariance between random effects in linear mixed models for complex trait analyses. Nat Commun 11:4208. https://doi.org/10.1038/s41467-020-18085-5

Zyphur MJ, Zhang Z, Barsky AP, Li W-D (2013) An ACE in the hole: Twin family models for applied behavioral genetics research. Leadersh Q 24:572–594. https://doi.org/10.1016/j.leaqua.2013.04.001

Table 1

	Full sample	Twin sub-sample
N	8239	924
Age (months; mean (SD))	118.94 (7.55)	121.79 (6.61)
Parental Education (%)		
< HS Diploma	320 (3.9)	18 (1.9)
Bachelor	2196 (26.7)	308 (33.3)
HS Diploma/GED	680 (8.3)	38 (4.1)
Post Graduate Degree	2892 (35.1)	314 (34.0)
Some College	2151 (26.1)	246 (26.6)
Household Income (%)		
< $50,000	2404 (29.2)	170 (18.4)
>= $100,000	3461 (42.0)	487 (52.7)
>= $50,000 & < $100,000	2374 (28.8)	267 (28.9)

Table 1. Sample information at baseline. All samples include complete cases only.

Table 2

Model	Sample	Timepoints	N_{obs}	GRM values	Covariates	Random Effects
Model 1	Twin	Baseline	924	Assigned	Age and sex	A, C, E
Model 2	Twin	Baseline	924	Measured	Age and sex	A, C, E
Model 3	Full	Baseline	8242	Measured	Age and sex	A, C, E
Model 4	Full	Baseline	8242	Measured	Age and sex	A, C, T, E
Model 5	Full	Baseline and Y2	13984	Measured	Age and sex	A, C, T, E
Model 6	Full	Baseline and Y2	13984	Measured	Age and sex	A, C, T, S, E
Model 7	Full	Baseline	8242	Assigned	Age and sex	A, C, E
Model 8	Full	Baseline	8242	Assigned	Age and sex	A, C, T, E
Model 9	Full	Baseline and Y2	13984	Assigned	Age and sex	A, C, T, S, E
Model 10	Twin	Baseline	924	Assigned	All covariates	A, C, E
Model 11	Twin	Baseline	924	Measured	All covariates	A, C, E
Model 12	Full	Baseline	8242	Measured	All covariates	A, C, E
Model 13	Full	Baseline	8242	Measured	All covariates	A, C, T, E
Model 14	Full	Baseline and Y2	13984	Measured	All covariates	A, C, T, S, E
Model 15	Full	Baseline and Y2	13984	Measured	Age and sex	A, C, S, E
Model 16	Full minus twins	Baseline and Y2	11835	Measured	Age and sex	A, C, S, E

Table 2. List of model specifications. GRM = genetic relatedness matrix; N_{obs} = number of observations; Y2 = year 2 follow-up visit. All models pre-residualized for age and sex; when specified, “all covariates” includes these as well as site, parental education, household income, and first ten genetic principal components. Random effects: A = additive genetic relatedness, C = common environment, S = subject, T = twin status (shared pregnancy ID), E = unexplained variance / error.

Figure Captions

Figure 1. ACE Model in FEMA versus OpenMx. A) Comparison of model estimates. Horizontal error bars represent confidence interval calculated in FEMA; vertical error bars represent confidence intervals calculated in OpenMx. B) Difference in Akaike Information Criterion in FEMA versus in OpenMx. C) Random effects estimates from FEMA.

Figure 2. ACE Model using assigned versus measured GRM. A) Comparison of model estimates. Horizontal error bars represent confidence interval calculated in Model 2; vertical error bars represent confidence intervals calculated in Model 1. B) Difference in Akaike Information Criterion in Model 2 versus in Model 1. C) Random effects estimates from Model 2.

Figure 3. ACE Model using full baseline sample compared to twin sub-sample. A) Comparison of model estimates. Horizontal error bars represent confidence interval calculated in Model 3; vertical error bars represent confidence intervals calculated in Model 2. B) Random effects estimates from Model 3.

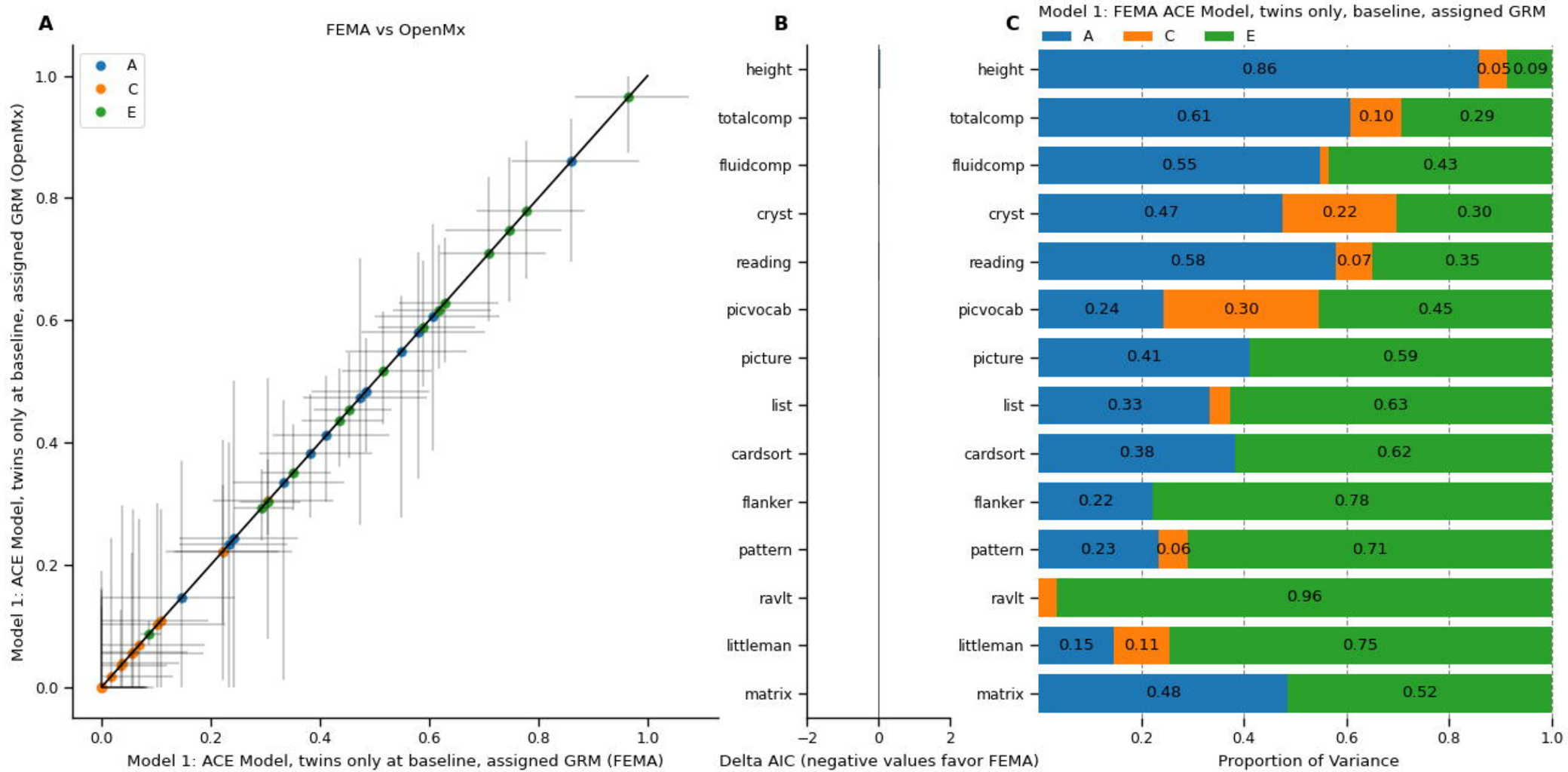
Figure 4. ACE Model versus ACTE model using full baseline sample. A) Comparison of model estimates. Horizontal error bars represent confidence interval calculated in Model 4; vertical error bars represent confidence intervals calculated in Model 3. B) Difference in Akaike Information Criterion in Model 4 versus in Model 3. C) Random effects estimates from Model 4.

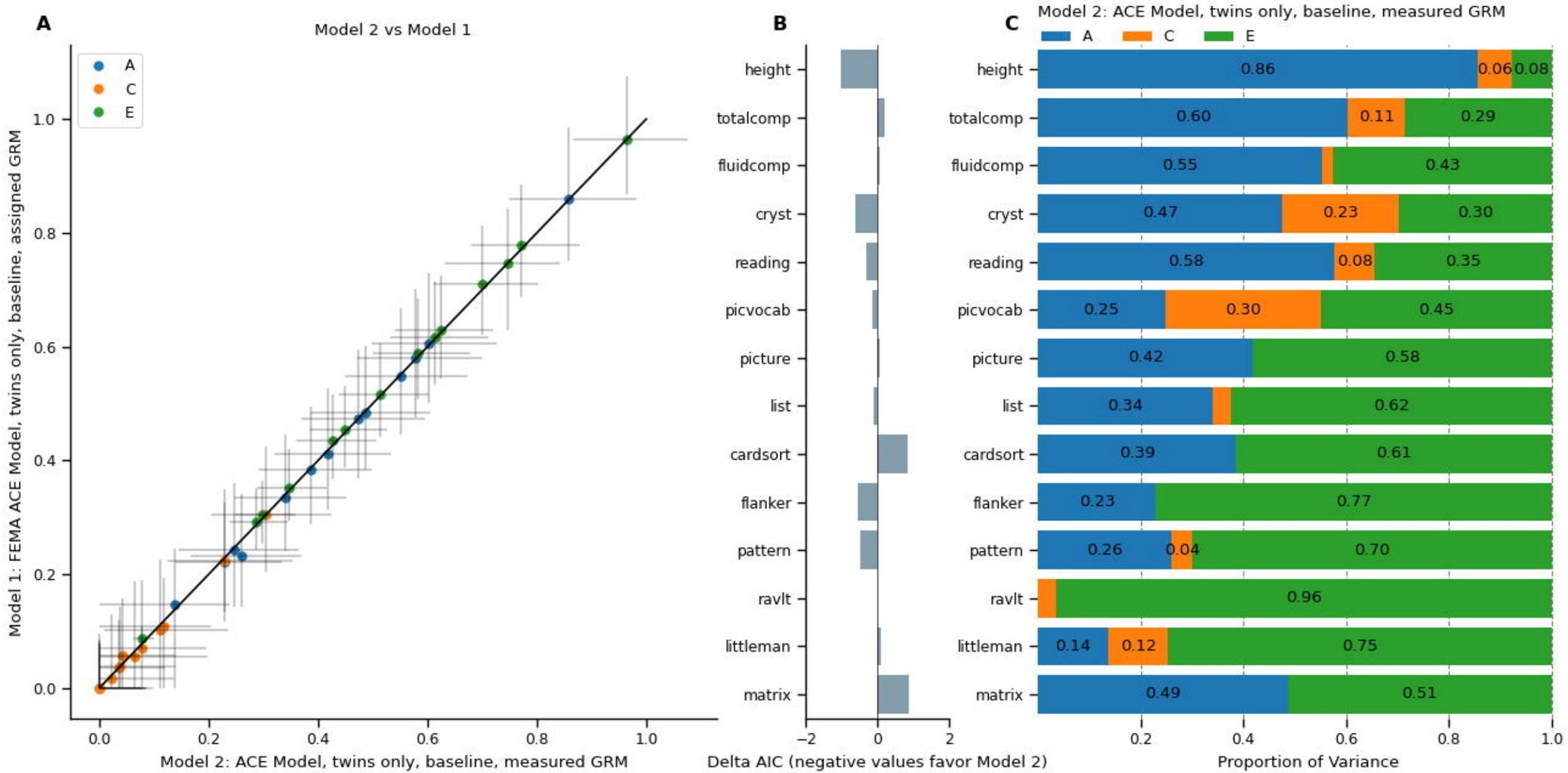
Figure 5. ACTE and ACTSE Model in baseline versus longitudinal sample. A) Comparison of estimates from Model 5 versus Model 4. Horizontal error bars represent confidence interval calculated in Model 5; vertical error bars represent confidence intervals calculated in Model 4. B) Random effects estimates from Model 5. C) Comparison of estimates from Model 6 versus Model 5. Horizontal error bars represent confidence interval calculated in Model 6; vertical error bars represent confidence intervals calculated in Model 5. D) Difference in Akaike Information Criterion in Model 6 versus in Model 5. C) Random effects estimates from Model 6.

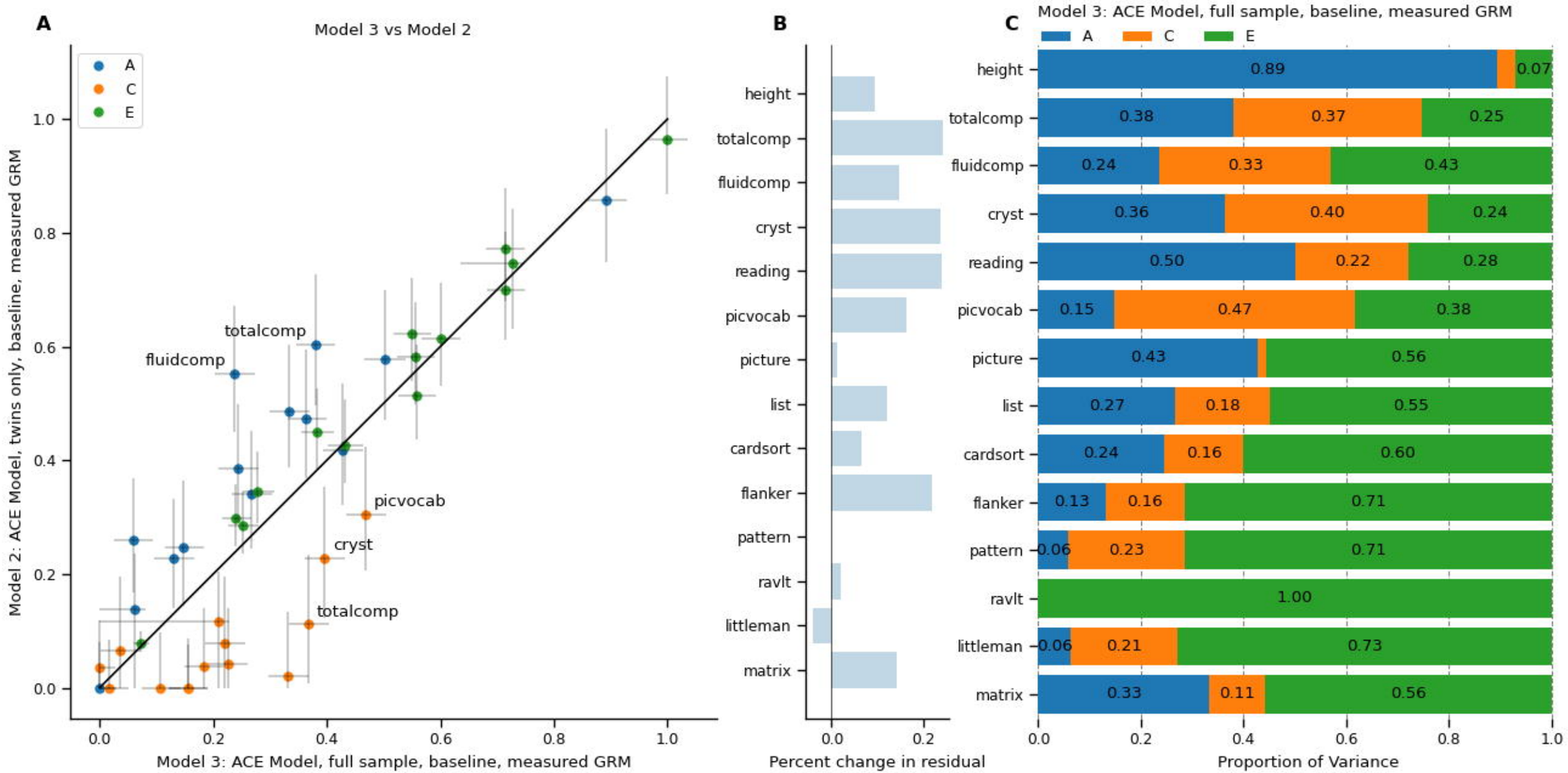
Figure 6. ACTSE model using assigned versus measured genetic relatedness. A) Comparison of estimates from Model 9 versus Model 6. Horizontal error bars represent confidence interval calculated in Model 9; vertical error bars represent confidence intervals calculated in Model 6. B) Difference in Akaike Information Criterion in Model 9 versus in Model 6. C) Random effects estimates from Model 9.

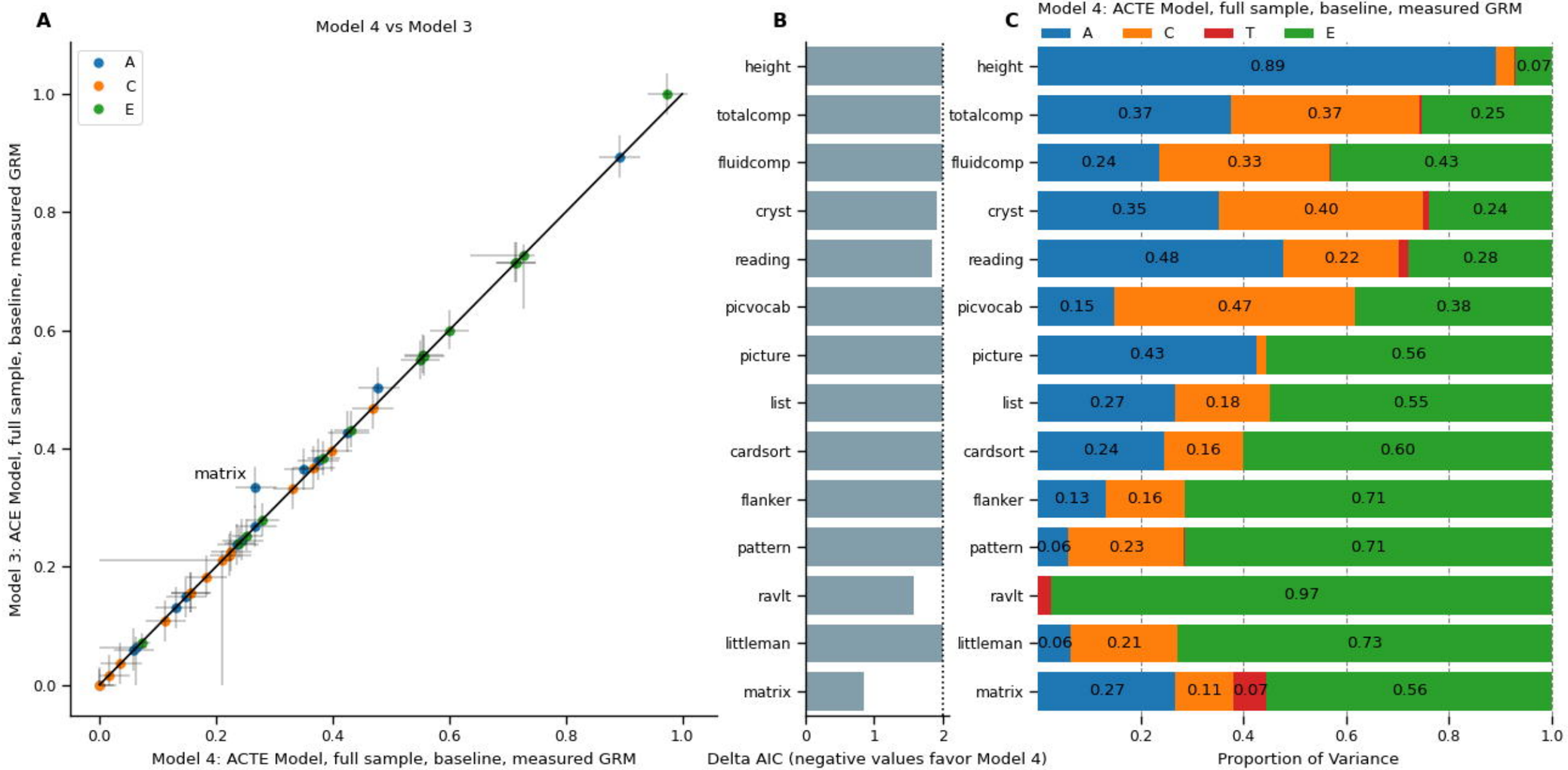
Figure 7. ACE model in baseline twin sample, residualizing for all covariates versus age and sex only. A) Comparison of estimates from Model 10 versus Model 1. Horizontal error bars represent confidence interval calculated in Model 10; vertical error bars represent confidence intervals calculated in Model 1. B) Difference in total residual variance in Model 10 versus in Model 1. C) Random effects estimates from Model 10.

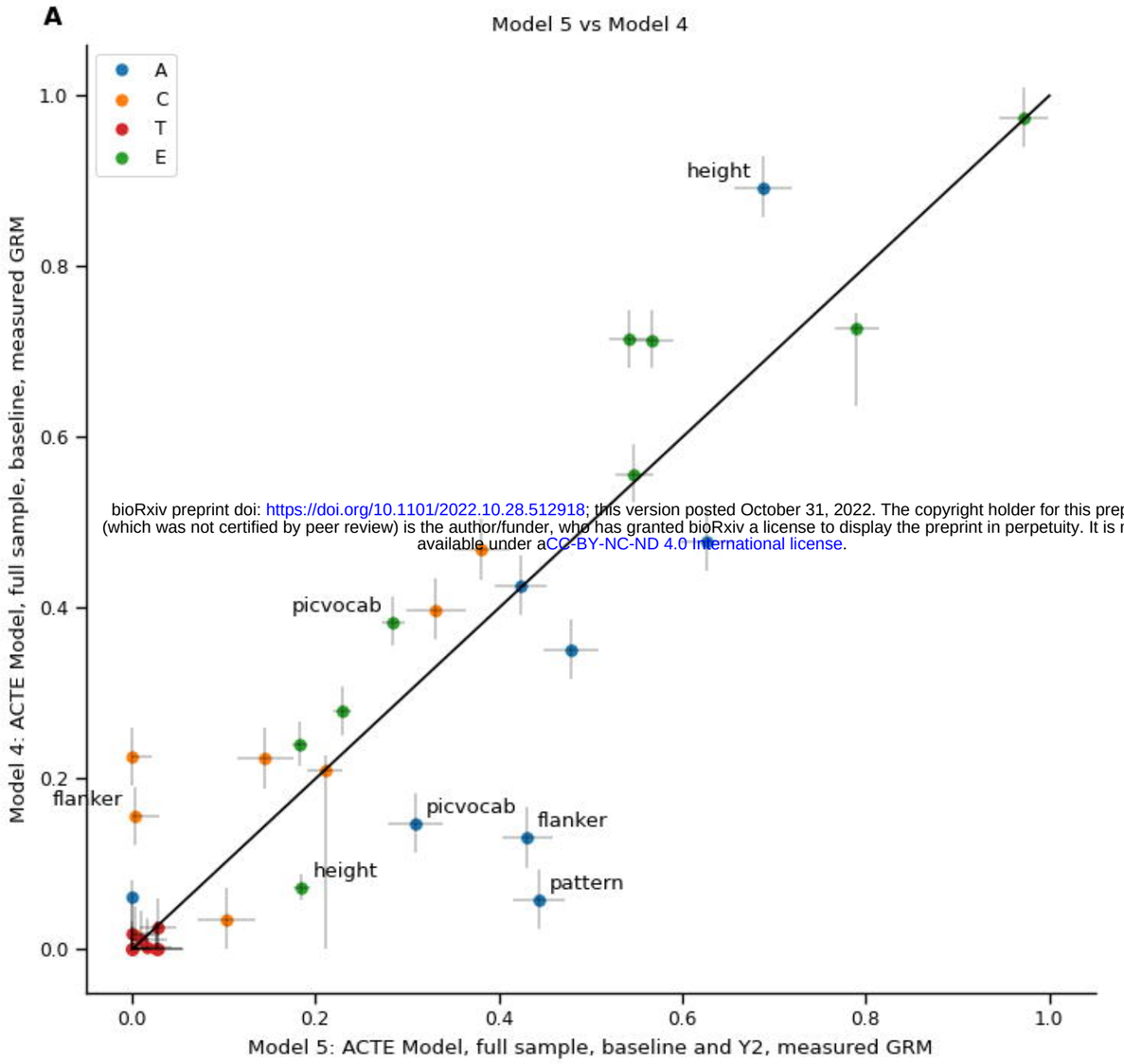
Figure 8. ACSE model in longitudinal sample, in a sample that excludes twin registry participants. Comparison model is equivalent but includes the full sample inclusive of twin registry participants. A) Comparison of estimates from Model 16 versus Model 15. Horizontal error bars represent confidence interval calculated in Model 16; vertical error bars represent confidence intervals calculated in Model 15. B) Difference in total residual variance in Model 16 versus in Model 15. C) Random effects estimates from Model 16.

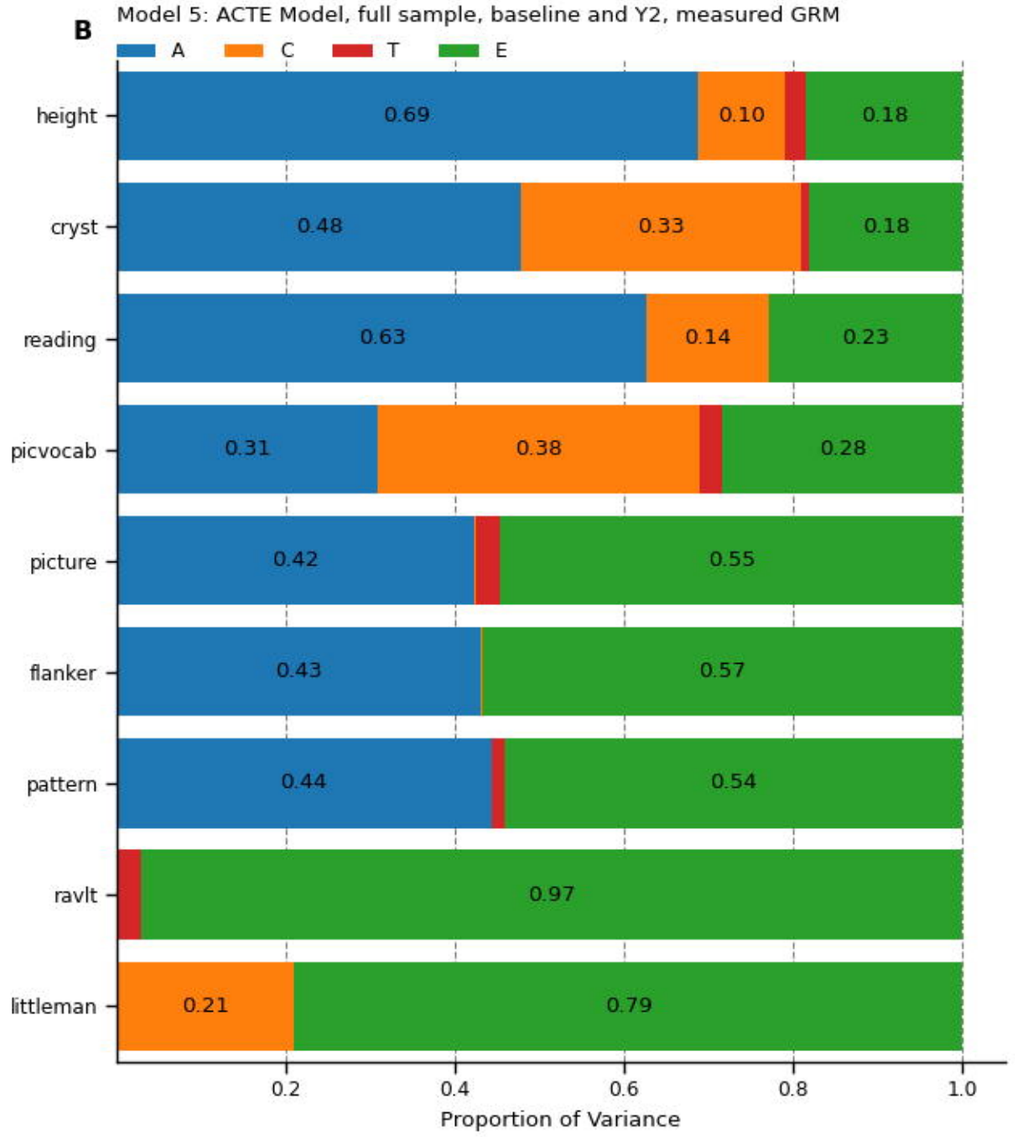

A
FEMA vs OpenMx
A
C
E
Model 1: ACE Model, twins only at baseline, assigned GRM (OpenMx)
Model 1: ACE Model, twins only at baseline, assigned GRM (FEMA)
B
height
totalcomp
fluidcomp
cryst
reading
picvocab
picture
list
cardsort
flanker
pattern
ravlt
littleman
matrix
Delta AIC (negative values favor FEMA)
C
Model 1: FEMA ACE Model, twins only, baseline, assigned GRM
A
C
E
height 0.86 0.05 0.09
totalcomp 0.61 0.10 0.29
fluidcomp 0.55 0.43
cryst 0.47 0.22 0.30
reading 0.58 0.07 0.35
picvocab 0.24 0.30 0.45
picture 0.41 0.59
list 0.33 0.63
cardsort 0.38 0.62
flanker 0.22 0.78
pattern 0.23 0.06 0.71
ravlt 0.96
littleman 0.15 0.11 0.75
matrix 0.48 0.52
Proportion of Variance

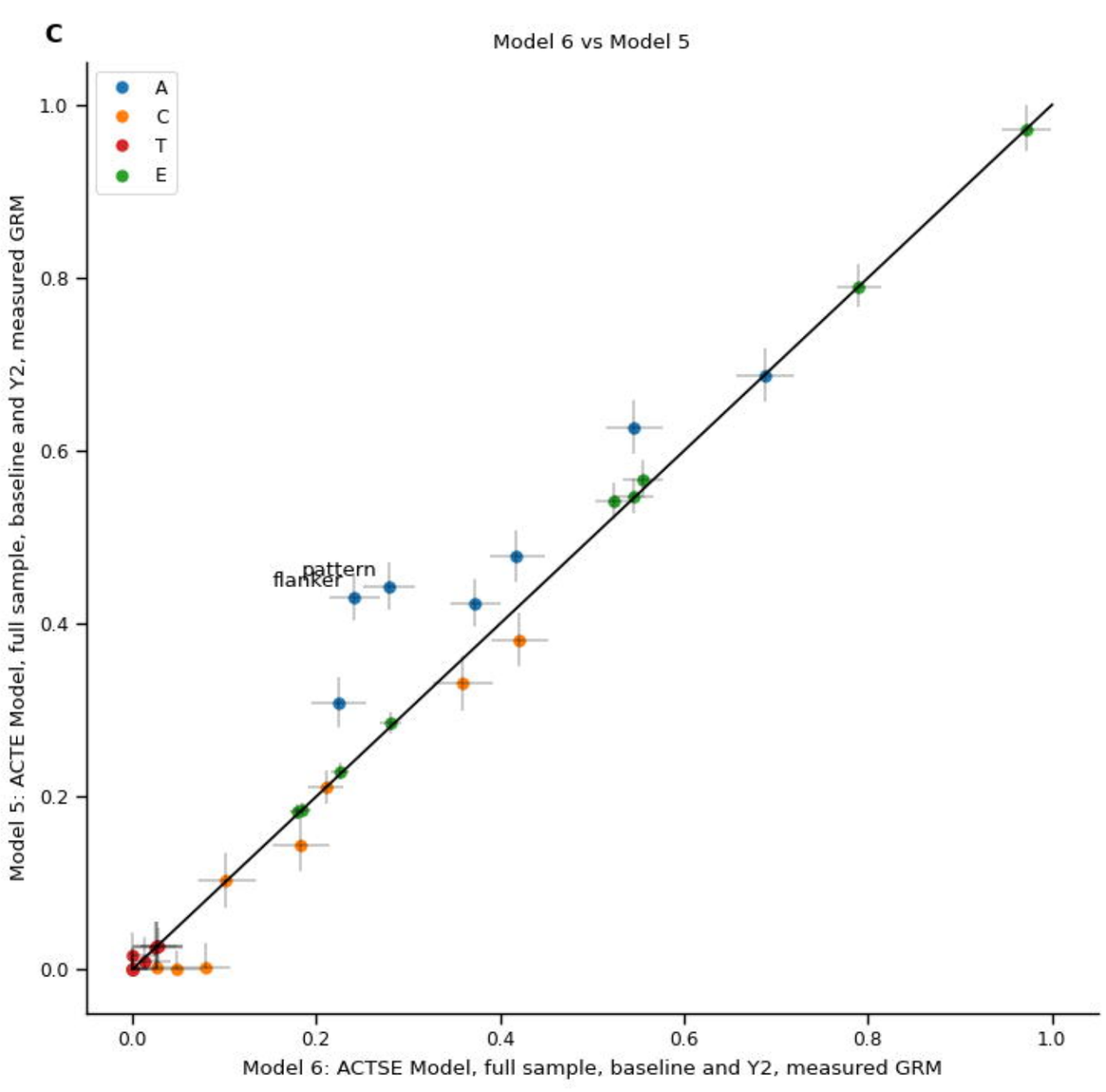

A
Model 2 vs Model 1
A
C
E
Model 1: FEMA ACE Model, twins only, baseline, assigned GRM
Model 2: ACE Model, twins only, baseline, measured GRM
B
height
totalcomp
fluidcomp
cryst
reading
picvocab
picture
list
cardsort
flanker
pattern
ravlt
littleman
matrix
Delta AIC (negative values favor Model 2)
C
Model 2: ACE Model, twins only, baseline, measured GRM
A
C
E
height 0.86 0.06 0.08
totalcomp 0.60 0.11 0.29
fluidcomp 0.55 0.43
cryst 0.47 0.23 0.30
reading 0.58 0.08 0.35
picvocab 0.25 0.30 0.45
picture 0.42 0.58
list 0.34 0.62
cardsort 0.39 0.61
flanker 0.23 0.77
pattern 0.26 0.04 0.70
ravlt 0.96
littleman 0.14 0.12 0.75
matrix 0.49 0.51
Proportion of Variance

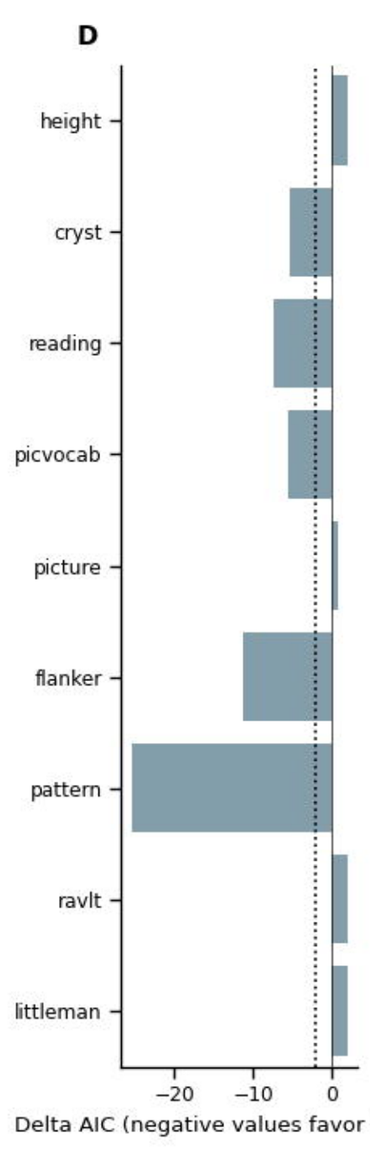

A
Model 3 vs Model 2
A
C
E
totalcomp
fluidcomp
picvocab
cryst
totalcomp
Model 2: ACE Model, twins only, baseline, measured GRM
Model 3: ACE Model, full sample, baseline, measured GRM
B
height
totalcomp
fluidcomp
cryst
reading
picvocab
picture
list
cardsort
flanker
pattern
ravlt
littleman
matrix
Percent change in residual
C
Model 3: ACE Model, full sample, baseline, measured GRM
A
C
E
height 0.89 0.07
totalcomp 0.38 0.37 0.25
fluidcomp 0.24 0.33 0.43
cryst 0.36 0.40 0.24
reading 0.50 0.22 0.28
picvocab 0.15 0.47 0.38
picture 0.43 0.56
list 0.27 0.18 0.55
cardsort 0.24 0.16 0.60
flanker 0.13 0.16 0.71
pattern 0.06 0.23 0.71
ravlt 1.00
littleman 0.06 0.21 0.73
matrix 0.33 0.11 0.56
Proportion of Variance

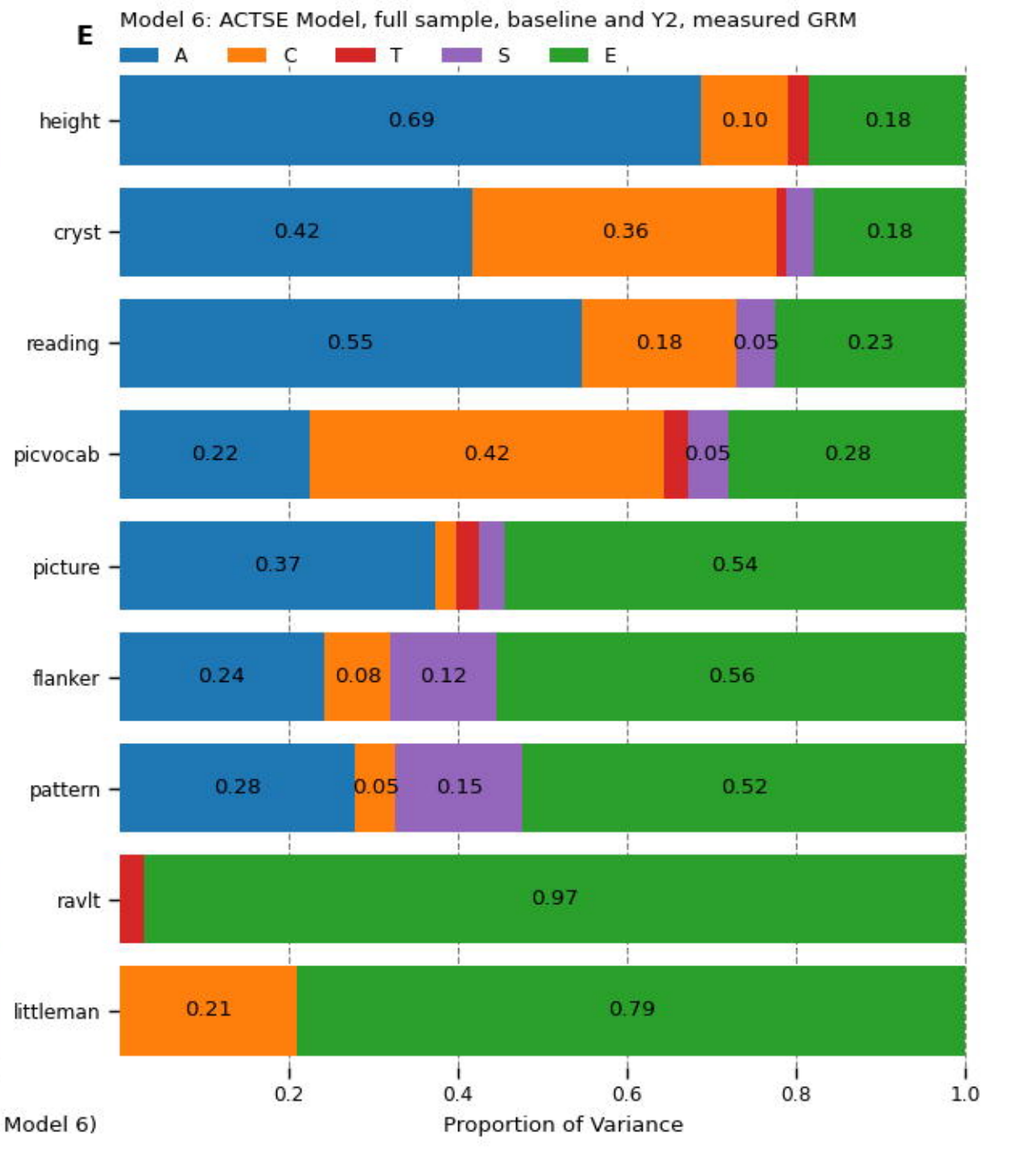

A
Model 4 vs Model 3
A
C
E
matrix
Model 3: ACE Model, full sample, baseline, measured GRM
Model 4: ACTE Model, full sample, baseline, measured GRM
B
height
totalcomp
fluidcomp
cryst
reading
picvocab
picture
list
cardsort
flanker
pattern
ravlt
littleman
matrix
Delta AIC (negative values favor Model 4)
C
Model 4: ACTE Model, full sample, baseline, measured GRM
A
C
T
E
0.89 0.07
0.37 0.37 0.25
0.24 0.33 0.43
0.35 0.40 0.24
0.48 0.22 0.28
0.15 0.47 0.38
0.43 0.56
0.27 0.18 0.55
0.24 0.16 0.60
0.13 0.16 0.71
0.06 0.23 0.71
0.97
0.06 0.21 0.73
0.27 0.11 0.07 0.56
Proportion of Variance

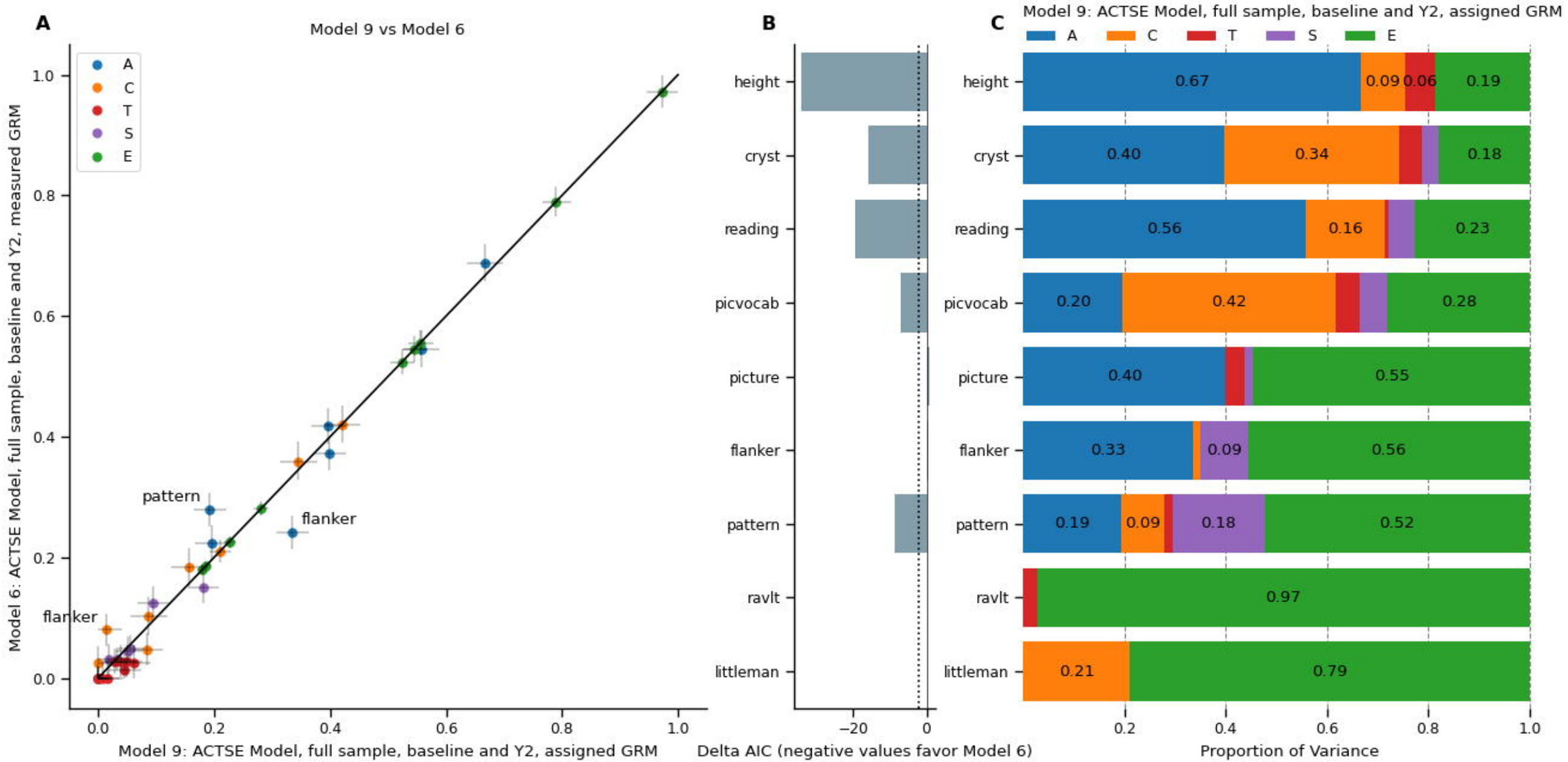


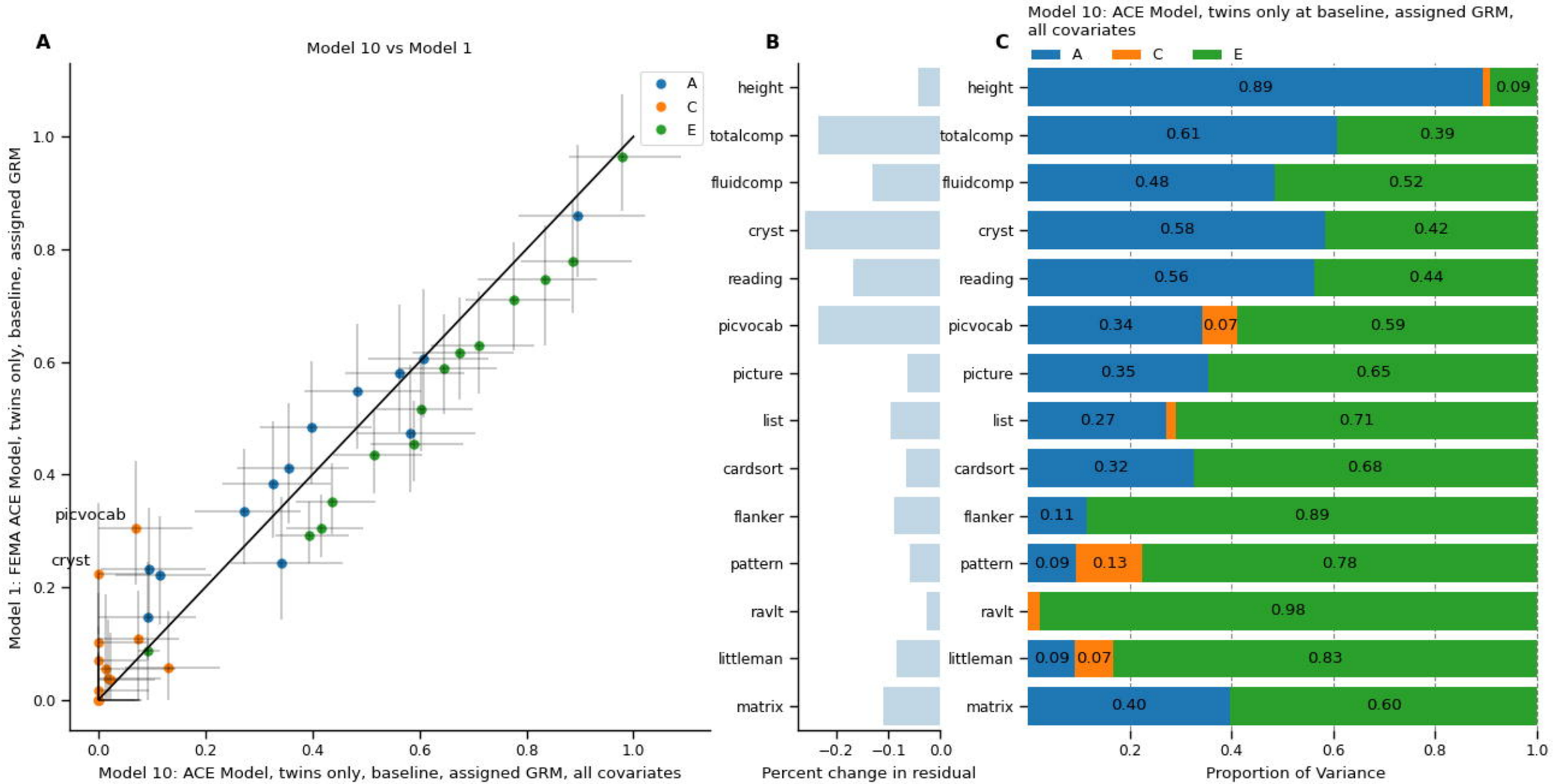


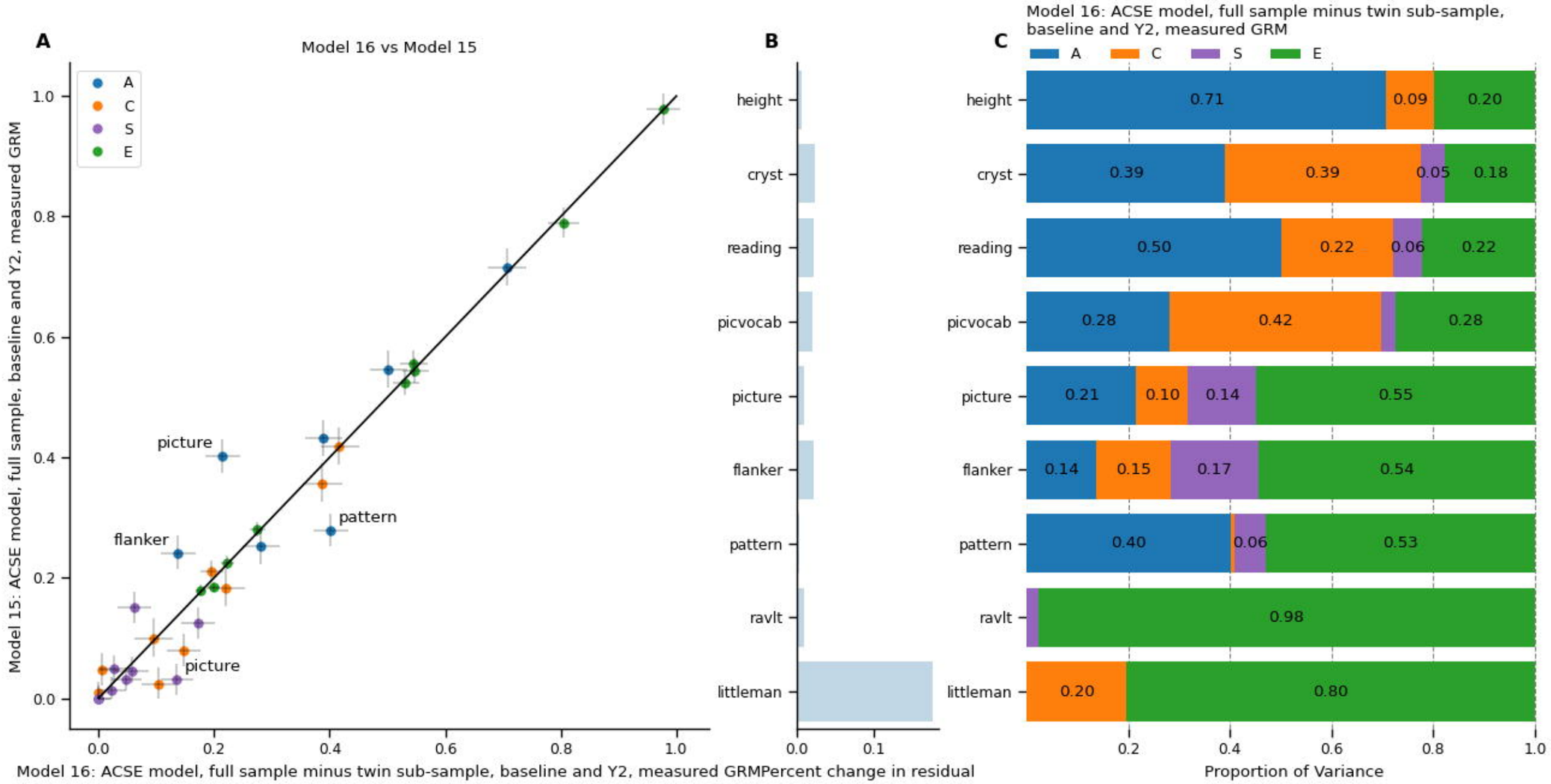







A
Model 9 vs Model 6
A
C
T
S
E
pattern
flanker
flanker
Model 6: ACTSE Model, full sample, baseline and Y2, measured GRM
Model 9: ACTSE Model, full sample, baseline and Y2, assigned GRM
B
height
cryst
reading
picvocab
picture
flanker
pattern
ravlt
littleman
−20
0
Delta AIC (negative values favor Model 6)
C
Model 9: ACTSE Model, full sample, baseline and Y2, assigned GRM
A
C
T
S
E
height 0.67 0.09 0.06 0.19
cryst 0.40 0.34 0.18
reading 0.56 0.16 0.23
picvocab 0.20 0.42 0.28
picture 0.40 0.55
flanker 0.33 0.09 0.56
pattern 0.19 0.09 0.18 0.52
ravlt 0.97
littleman 0.21 0.79
Proportion of Variance


A
Model 10 vs Model 1
A
C
E
picvocab
cryst
Model 1: FEMA ACE Model, twins only, baseline, assigned GRM
Model 10: ACE Model, twins only, baseline, assigned GRM, all covariates
B
height
totalcomp
fluidcomp
cryst
reading
picvocab
picture
list
cardsort
flanker
pattern
ravlt
littleman
matrix
Percent change in residual
C
Model 10: ACE Model, twins only at baseline, assigned GRM, all covariates
A
C
E
0.89
0.09
0.61
0.39
0.48
0.52
0.58
0.42
0.56
0.44
0.34
0.07
0.59
0.35
0.65
0.27
0.71
0.32
0.68
0.11
0.89
0.09
0.13
0.78
0.98
0.09
0.07
0.83
0.40
0.60
Proportion of Variance


A
Model 16 vs Model 15
A
C
S
E
picture
pattern
flanker
picture
Model 15: ACSE model, full sample, baseline and Y2, measured GRM
0.0
0.2
0.4
0.6
0.8
1.0
Model 16: ACSE model, full sample minus twin sub-sample, baseline and Y2, measured GRM
B
height
cryst
reading
picvocab
picture
flanker
pattern
ravlt
littleman
0.0
0.1
Percent change in residual
C
Model 16: ACSE model, full sample minus twin sub-sample, baseline and Y2, measured GRM
A
C
S
E
height 0.71 0.09 0.20
cryst 0.39 0.39 0.05 0.18
reading 0.50 0.22 0.06 0.22
picvocab 0.28 0.42 0.28
picture 0.21 0.10 0.14 0.55
flanker 0.14 0.15 0.17 0.54
pattern 0.40 0.06 0.53
ravlt 0.98
littleman 0.20 0.80
0.2
0.4
0.6
0.8
1.0
Proportion of Variance