Improving accuracy and precision of heritability estimation in twin studies: Reassessing the measurement error assumption

Gang Chen*[a], Dustin Moraczewski[b], and Paul A. Taylor[a]

[a]Scientific and Statistical Computing Core, National Institute of Mental Health, USA
[b]Data Science and Sharing Team, National Institute of Mental Health, USA

Abstract

In this study, we demonstrate the need for improvement in the conventional ACE model used for estimating heritability when applied to trait data with measurement errors. The critical issue revolves around an assumption concerning measurement errors in twin studies. In cases where traits are measured using samples, data is aggregated during preprocessing, with only a centrality measure (e.g., mean) being used for modeling. Additionally, measurement errors resulting from sampling are assumed to be part of the nonshared environment and are thus overlooked in heritability estimation. Consequently, the presence of intra-individual variability remains concealed. Moreover, recommended sample sizes (e.g., 600 twin pairs) are typically based on the assumption of no measurement errors. We argue that measurement errors are an inherent limitation of finite sampling and should not be regarded as part of the nonshared environment. More importantly, recent studies have provided converging evidence indicating that the intra-individual variability of psychometric effects is much larger than the inter-individual counterpart. To accurately capture the underlying data generating process, we propose a hierarchical model that explicitly incorporates measurement errors.

To demonstrate the appropriateness and advantages of our modeling approach, we utilize simulations as well as a real dataset from the ABCD (Adolescent Brain Cognitive Development) study. Our findings have direct implications for study design and interpretation of results. Specifically, we showcase the following analytical insights for data containing non-negligible measurement errors:

1) The conventional ACE model may underestimate heritability.
2) A hierarchical model provides a more accurate assessment of heritability.
3) Large samples, exceeding 100 observations or thousands of twins, may be necessary to reduce ambiguity.

In summary, our study sheds light on the impact of measurement errors on heritability estimation and proposes a hierarchical model as a more accurate alternative. These findings have important implications for understanding individual differences and for the design and analysis of twin studies.

1 Introduction

As an indication of potential predictability, heritability is an important statistical concept in assessing individual differences. Defined as the proportion of trait variability that can be ascribed to genetics, heritability offers a unique perspective for quantitatively understanding the role of genetics in complex traits (Downes and Turkheimer, 2022; Robette et al., 2022). Twins provide a hypothetically well-controlled scenario where genetics, environment, and their interaction can be statistically separated and apportioned. For example, if

*Corresponding author. E-mail address: gangchen@mail.nih.gov

a trait's variation is largely attributed to genetics, monozygotic (MZ) twins are more likely to be correlated with each other than dizygotic (DZ) twins, resulting in a relatively high heritability. Conversely, if the trait's variation is mostly due to the environment, MZ twins will share a similar extent of resemblance or difference to DZ twins, indicating low heritability. Referred to as the "first law" of behavior genetics (Turkheimer, 2000), heritability measures the degree of similarity, similar to the concept of test-retest reliability. While test-retest reliability addresses the extent to which individual differences are reliable, heritability examines the extent to which individual differences are heritable.

As a population-level metric, heritability is traditionally assessed and quantified using twins or family members. Over the past half-century, thousands of publications have explored around 20,000 traits, with an average observed heritability of approximately 50% (Polderman et al., 2015). The field of behavior genetics and individual differences has matured, with the development and investigation of various modeling approaches and estimation methods. These efforts have contributed to our understanding of individual differences, spanning from physical traits (e.g., height, weight, eye color) to behavioral, neurological, cognitive, cultural, and psychological characteristics (e.g., IQ, school performance, educational attainment) as well as common disorders (e.g., diabetes, ADHD, depression, schizophrenia) (Giangrande et al., 2022).

1.1 Heritability estimation: ACE model

Conventional twin studies are typically conceptualized with three levels of data structure (Fig. 1). Specifically, traits, behavioral measures, and disorders are represented by individual-level scores. These individual measures are nested within families, which are further categorized as either MZ or DZ twins.

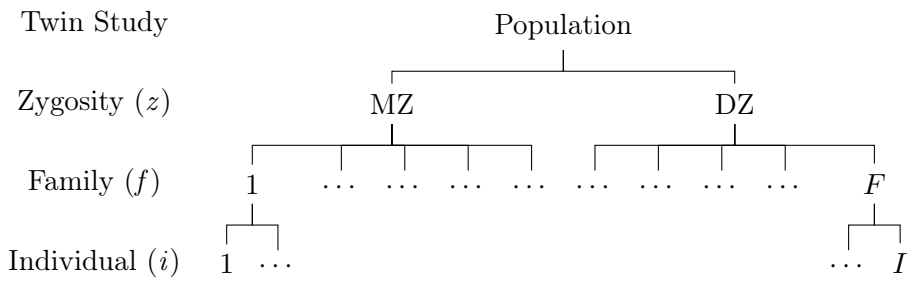

Figure 1: Hierarchical data structure in conventional twin studies. Three levels are involved in a nesting stature: a trait is measured for each individual in a family with either MZ or DZ twins. The three indices i, f, and z are respectively indicators for I individuals, F families, and the two zygosities of MZ and DZ twins.

A model can be formulated for a quantitative trait of interest that is measured at the individual level. In the popular ACE formulation (Maes, 2005; Downes and Matthews, 2020; Hunter, 2021), the trait data $y_{i(f(z))}$ is expressed as the combination of latent components at three levels – individual ($i = 1, 2, ..., I$), family ($f = 1, 2, ..., F$), and zygosity ($z = \text{MZ, DZ}$) – as follows:

$$\text{individual: } y_{i(f(z))} = \alpha + A_{i(f(z))} + C_{i(f(z))} + E_{i(f(z))}. \tag{1}$$

Each pair of parentheses indicates a nesting structure among the subscripts. For example, each individual i belongs to a family f with a zygosity z. The intercept α captures the overall trait effect at the population level. The acronym for the ACE model reflects the three latent sources of variability. $A_{i(f(z))}$ represents the additive genetic effects, and $C_{i(f(z))}$ represents the common or shared environmental effects. In addition, $E_{i(f(z))}$ characterizes the unique or nonshared environmental effects. The model (1) can be easily extended to include twin- and individual-level covariates such as sex and age.

The variances associated with the three latent components are crucial parameters in twin studies. One

Term	Description
$\sigma^2_{A_z}$	variance of the additive genetic effects, for a given zygosity z
$\sigma^2_{C_z}$	variance of the common (shared) environment effects, for a given zygosity z
$\sigma^2_{E_z}$	variance of the nonshared genetic effects, for a given zygosity z
$\sigma^2_A, \sigma^2_C, \sigma^2_E$	variance components (above) under homogeneity assumption (equal variance across zygosity)
ρ_z	relatedness for additive genetic effects between two twins
h^2	heritability, proportion of total variability attributed to additive genetic effects
c^2	proportion of total variability attributed to common environment
e^2	proportion of total variability attributed to nonshared environment
r_z	correlation between two twins of zygosity z
σ^2_z	(HME) individual-level trait variance for zygosity z
τ^2_z	(HME) family-level variance for zygosity z
ω^2_0	(HME) theoretical cross-trial variance of observations; intra-individual variability
ω^2	(HME) cross-trial sampling variance
R^2_v	(HME) ratio of variances: intra-individual to sum of inter-individual and inter-family
U	(HME) scaling factor for correlation r_z, containing measurement error bias

Table 1: A reference table of major variables and parameters used with heritability modeling. Quantities which originate in hierarchical modeling estimation (HME) are noted explicitly; the rest originate in the ACE model.

may make the following assumptions for two twins i_1 and i_2 within a family f (Arbet et al., 2020),

$$\begin{aligned}
\begin{bmatrix} A_{i_1(f(z))} \\ A_{i_2(f(z))} \end{bmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_z \\ \rho_z & 1 \end{bmatrix} \sigma^2_{A_z} \right), \\
\begin{bmatrix} C_{i_1(f(z))} \\ C_{i_2(f(z))} \end{bmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \sigma^2_{C_z} \right), \\
\begin{bmatrix} E_{i_1(f(z))} \\ E_{i_2(f(z))} \end{bmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \sigma^2_{E_z} \right),
\end{aligned} \tag{2}$$

where $\mathcal{N}(\cdot,\cdot)$ denotes a Gaussian distribution. The relatedness ρ_z for the additive genetic effects between two twins i_1 and i_2 in a family is assumed to be directly related to zygosity:

$$\rho_z = \begin{cases} 1, & \text{if } z = \text{MZ}; \\ 0.5, & \text{if } z = \text{DZ}. \end{cases} \tag{3}$$

The six variances, $\sigma^2_{A_z}$, $\sigma^2_{C_z}$, and $\sigma^2_{E_z}$ ($z = \text{MZ, DZ}$), represent the level of variability associated with the three latent components. As a side note, the main notations used in this paper are listed in Table 1.

A crucial feature of the ACE model is the inclusion of the *homogeneity assumption*. This assumption is necessary when estimating variances within the model framework in (2). Six variances, namely $\sigma^2_{A_z}$, $\sigma^2_{C_z}$, and $\sigma^2_{E_z}$ for $z = \text{DZ}$ and MZ, need to be estimated, resulting in an undetermined system. To resolve this identifiability issue, the homogeneity assumption is employed, assuming that the variances are homogeneous across both MZ and DZ twins (Arbet et al., 2020):

$$\sigma^2_{A_{\text{MZ}}} = \sigma^2_{A_{\text{DZ}}} = \sigma^2_A, \qquad \sigma^2_{C_{\text{MZ}}} = \sigma^2_{C_{\text{DZ}}} = \sigma^2_C, \qquad \sigma^2_{E_{\text{MZ}}} = \sigma^2_{E_{\text{DZ}}} = \sigma^2_E. \tag{4}$$

The assumption effectively reduces the number of variance parameters by half. However, this also leads to having homogeneity of total variance between MZ and DZ: $\sigma^2_{A_{\text{MZ}}} + \sigma^2_{C_{\text{MZ}}} + \sigma^2_{E_{\text{MZ}}} = \sigma^2_{A_{\text{DZ}}} + \sigma^2_{C_{\text{DZ}}} + \sigma^2_{E_{\text{DZ}}} = \sigma^2_A + \sigma^2_C + \sigma^2_E$. One subset of the homogeneity assumption (4) is referred to as the *equal environment assumption*:

$$\sigma^2_{C_{\text{MZ}}} = \sigma^2_{C_{\text{DZ}}} = \sigma^2_C. \tag{5}$$

In practical application, the equal environment assumption is considered particularly controversial on interpretational grounds due to the likelihood of MZ twins sharing more similar environments than DZ ones (Richardson and Norgate, 2005; Willoughby et al., 2023).

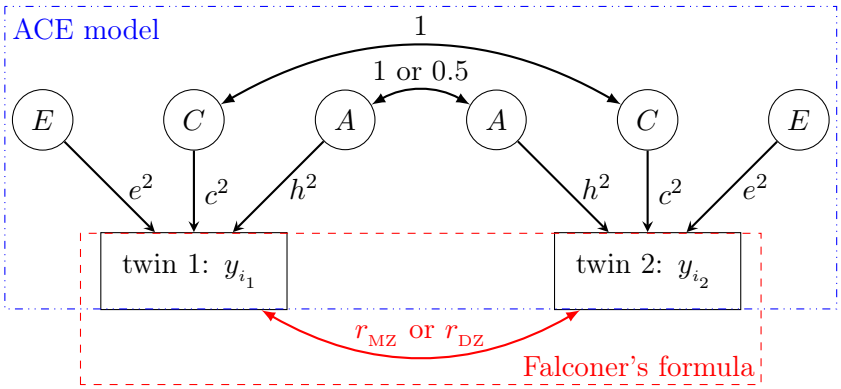

Figure 2: Path diagram (reticular action model) for the ACE formulation (1). Subscript indices for individual (i), family (f), and zygosity (z) are mostly dropped from the nodes for brevity. Trait y is considered as directly observable effects are represented as rectangles. In contrast, latent (unobservable) effects (A, C and E components) are represented by eclipses. A directed path, represented by a single-headed arrow, indicates predictability. An undirected path, represented by a double-headed curved arrow, indicates a covarying relationship. The value on each path shows the correlation coefficient. Note that h^2, c^2, and e^2 can be estimated through effect partitioning into the three latent components (A, C, and E), or through Falconer's formula (10) when r_z are available.

The heritability in a twin study can be defined under the homogeneity assumption (4). The classical methodology is to adopt structural equation modeling (e.g., Rijsdijk and Sham, 2002; Neale et al., 2016; Holst et al., 2016; Bates et al., 2019) to estimate the three variances. Then, the heritability, conventionally denoted as h^2 (or a^2), can be quantified as the proportion of total variability attributed to additive genetic effects,

$$h^2 = \frac{\sigma^2_A}{\sigma^2_A + \sigma^2_C + \sigma^2_E}. \tag{6}$$

Along the same vein, the proportion of trait variability due to common environment is expressed as,

$$c^2 = \frac{\sigma^2_C}{\sigma^2_A + \sigma^2_C + \sigma^2_E}. \tag{7}$$

Naturally, the complementary component of total variability to the above two proportions shows the amount of contribution from nonshared environment,

$$e^2 = 1 - h^2 - c^2 = \frac{\sigma^2_E}{\sigma^2_A + \sigma^2_C + \sigma^2_E}. \tag{8}$$

Effect decomposition under the ACE model (1) and the associated estimation of heritability can be visually represented as a path diagram (Fig. 2), which is analogous to a directed acyclic diagram in causal inference. The values of h^2, c^2, and c^2 in common practice are typically reported with their point estimates, accompanied by their uncertainty expressed through standard errors or 95% quantile intervals.

1.2 Heritability estimation: Falconer's formula

The three variability proportions of h^2, c^2, and e^2 can alternatively be expressed as the relatedness between the twins of each zygosity. We denote r_{MZ} and r_{DZ} as the correlations between two twins i_1 and i_2 within a

family f for MZ and DZ, respectively. The following can be derived per the ACE model (1),

$$r_z = \text{corr}(y_{i_1(f(z))},\ y_{i_2(f(z))}) = \frac{\rho_z \sigma_A^2 + \sigma_C^2}{\sigma_A^2 + \sigma_C^2 + \sigma_E^2} = \rho_z h^2 + c^2,\ z = \text{ MZ, DZ.} \tag{9}$$

Here r_{MZ} and r_{DZ} are directly related to the notion of intraclass correlation, with the class being MZ or DZ families (e.g., Guo and Wang, 2002; Rijsdijk and Sham, 2002; Chen et al., 2018). In other words, r_{MZ} and r_{DZ} can be conceptualized as the test-retest reliability in the sense that the test and retest are executed between each pair of twins. The expression (9) leads to Falconer's formula (Falconer and MacKay, 1996):

$$h^2 = 2(r_{\text{MZ}} - r_{\text{DZ}}), \qquad c^2 = 2r_{\text{DZ}} - r_{\text{MZ}}, \qquad e^2 = 1 - h^2 - c^2 = 1 - r_{\text{MZ}}. \tag{10}$$

Falconer's formula holds intrinsic importance. Although we have derived it here from the ACE model, Falconer's relation (10) can be derived independently and stands valid on its own. In fact, it provides an alternative approach to estimate heritability based on the difference between the two twin correlations, r_{MZ} and r_{DZ}, without directly relying on the concept of latent effects under the ACE model (1), as depicted in the lower portion of the path diagram in Fig. 2. Moreover, it offers the advantage of estimating heritability without assuming Gaussian distributions for each of the ACE parameters in (2), making it applicable in scenarios where Gaussianity is questionable (e.g., skewed data, presence of outliers). Interestingly, Falconer's formula (10) does not require the assumption of *variance homogeneity* across zygosity (4). In other words, it allows for the variance of each of the three latent effects to differ between MZ and DZ twins. Instead, it relies on a different assumption, *proportion homogeneity*, which assumes that the variance proportions of h^2, c^2, and e^2 are homogeneous across the two zygosities. It is worth noting that variance homogeneity is a stronger assumption compared to proportion homogeneity (Arbet et al., 2020).

The concept of heritability aims to quantify and partition the similarity of traits among relatives. It can be approached from two perspectives. The first is the *proportion perspective*, in which heritability is defined using the ACE model as the amount of genetic variance σ_A^2 relative to the total phenotypic variance $\sigma_A^2 + \sigma_C^2 + \sigma_E^2$, represented as h^2 in formula (6). In other words, heritability measures the extent to which individual genetic differences contribute to variations in phenotypic traits. In the context of unraveling the contributions of "nature" and "nurture," the term $1 - h^2$ represents the complementary component to heritability and is sometimes referred to as *environmentality* (Downes and Matthews, 2020). The second is the *similarity perspective*, which focuses on the degree of association between each pair of twins. As indicated in Falconer's formula (10), heritability is expressed as the doubled difference between the correlation of two MZ twins and the correlation of two DZ twins. In summary, estimating heritability under the ACE model relies on the accuracy of the three latent effect variances and their homogeneity across the two zygosities. In contrast, heritability estimation based on Falconer's formula directly relies on the similarities observed between twins.

1.3 Motivation: addressing the limitations of the conventional ACE framework

We address the challenges of heritability estimation from the perspectives of *accuracy* and *precision*, which are typically the objectives of studies to maximize. These perspectives are complemented by their "negative" counterparts, bias and ambiguity, respectively, which should be minimized as much as possible. Here, we define the accuracy of an estimate as the absence of systematic biases. Inaccuracy, for example, implies an upward or downward shift in a point estimate of effect centrality (e.g., mean, mode, median), a quantile interval, or a posterior distribution. On the other hand, we define ambiguity (the inverse of precision) as the extent of uncertainty in an estimate, which can be quantified as the standard error, a quantile interval, or the width of a posterior distribution. Accuracy and precision are desirable properties in statistics, and they are the focus of our assessment and comparison of model performance.

An implicit assumption is present within the conventional ACE model regarding measurement errors. In the current context, measurement error refers to the random or stochastic fluctuations in a measurement from one instance to another. The standard error is a common metric used to quantify measurement error. Specifically, the ACE model assumes one of two possibilities: 1) the absence of measurement errors in the phenotypic data $y_{i_1(f(z))}$, or 2) the inclusion of measurement errors within the nonshared environment component $E_{i_1(f(z))}$ (along with its associated variance σ_E^2) when they are present. Both assumptions lead to the same practical outcome: data with repeated measures are typically preprocessed by aggregating information through a centrality metric (such as the average) before estimating heritability.

Measurement errors are not a significant concern for many measures, such as physical traits that can typically be assessed with high precision. For example, if adult body height is measured for 100 consecutive days, the standard deviation is generally small (e.g., less than 3 mm). Although the ACE model (1) does not account for measurement errors or intra-individual variability, one could simply average those 100 weight values and use the aggregated data in the ACE model (1), disregarding the measurement errors. The potential impact of ignoring the uncertainty in this case is likely to be minimal.

We focus here on heritability estimation in situations where measurement errors are not negligible. For example, in a Stroop task where a large number of trials (e.g., 100) are presented in the experiment for each congruent and incongruent condition. To utilize the ACE model directly, the data are typically aggregated, and a centrality measure is used as input for heritability estimation. How should one generally assess intra-individual variability? And what occurs when intra-individual variability is not negligible compared to inter-individual variability?

Data aggregation can be necessary and even crucial in analysis. Firstly, the aggregation process consolidates potentially large amounts of information and organizes them into a more representative or computationally manageable form (e.g., sum, average, median, or mode). For example, research typically focuses on the population summary or centrality of a trait measure, rather than individual samples. Without some level of conceptual simplification through extraction, data may retain irrelevant details and impede interpretation and insight. One such example is the Child Behavior Checklist, where an integrative score is computed based on answers from a questionnaire using a three-point Likert scale. Secondly, as a practical consideration, some analytical tools have been historically developed without accounting for sampling errors, which are traditionally treated as a nuisance. Therefore, aggregation to a simple, single measure becomes necessary.

However, data aggregation can also be a double-edged sword, as valuable information can be lost in the winnowing process. One must balance information integration with information loss. The exact consequences will depend on both the types and amounts of information removed, and often on the stage at which it is reduced. The impact of ignoring intra-individual variability has been explored in different contexts. Data aggregation, although a common practice in many fields (e.g., psychometrics, neuroimaging), is intrinsically full of caveats because of information loss. Once condensed, some extent of loss or even distortion (e.g., variability, outliers) is unavoidable, and it is difficult for the adopted model to accurately capture the underlying data generating mechanism. For instance, the issue can be traced back to Spearman (1904) who pointed out the underestimation problem in estimating the correlation between two variables when measurement errors occur. It has also been recently shown that, without proper accountability, test-retest reliability can be substantially underestimated (Rouder and Haaf, 2019; Haines et al., 2020; Chen et al., 2021). Even in heritability estimation, underestimation has been revealed in item response theory due to the adoption of the aggregation process in the form of sum–scores (van den Berg et al., 2007; Schwabe et al., 2019). Recently, the bias problem due to measurement errors has been investigated regarding the reliability of polygenic score and that of a phenotypic trait framed under structural equation modeling (Pingault et al., 2022).

Data aggregation is a common practice in heritability estimation. Examples can be found in the fields of psychometrics (e.g., Rea-Sandin et al., 2023; ; Hung et al., 2023; Gustavson et al., 2023; Vellani et al.,

2022; Viktorsson et al., 2022; Yeom et al., 2022; Wang et al., 2020; Routledge et al., 2018; Stins et al., 2005; Schachar et al., 2010; Fan et al., 2001) and neuroimaging (Kastrati et al., 2022; van Drunen et al., 2021; Chen et al., 2019; Harper et al., 2019; van der Meulen et al., 2018; Anokhin et al., 2017; Blokland et al., 2011; Matthews et al., 2007; Polk et al., 2007). This practice is also popular in test-retest reliability estimation in psychometrics (e.g., Hedge et al., 2018) and neuroimaging (e.g., Elliott et al., 2020). Therefore, it would be desirable to fully investigate both the appropriateness and the consequences of ignoring intra-individual variability and to propose an improved approach.

Here, we employ a hierarchical modeling estimation[1] (HME) approach to account for intra-individual variability directly within heritability modeling. We believe that measurement errors reflect the limitations of the sampling process. Our recommendation contrasts with the current common practice, as we propose that *measurement errors should not be considered part of the nonshared environment component*, but rather partitioned appropriately within the model hierarchy. Therefore, by accurately mapping varying levels of effect granularity to the data structure, we aim to reduce information loss caused by data aggregation in heritability estimation, with a focus on estimation accuracy and precision. In addition, we employ the HME approach to reexamine the conventional ACE framework, with the latter being a special case of the former. Specifically, we demonstrate that the common practice of data aggregation fails to adequately address the impact of intra-individual variability. Instead, it is concealed within the modeling process, leading to a false sense of accuracy and certainty. With simulations and real datasets, we address the following questions:

1) Does disregarding intra-individual variability result in biased estimation? If so, to what extent?
2) To what degree does intra-individual variability contribute to precision in heritability estimation?
3) Are typical twin study sample sizes sufficient to achieve a proper precision of heritability estimation?

2 Estimating heritability under hierarchical framework

A hierarchical model partitions data variability by mapping the stratified structure to effects at various levels. The crucial question in heritability estimation is, *Can the latent effects in the ACE model be sorted and mapped to a nested structure under a hierarchical model?* In this section, we set up and explore the HME approach, illustrating its advantages over the ACE model.

2.1 HME: reformulating the ACE model

We begin by focusing on the data generating structure of an individual-level trait measure rather than on latent components. Specifically, instead of expressing effect decomposition through a seemingly flat structure, as done in the ACE model (1), we decompose the individual-level trait data $y_{i(f(z))}$ into effects directly associated with the three hierarchical levels present: individual ($i = 1, 2, ..., I$), family ($f = 1, 2, ..., F$), and zygosity ($z =$ MZ, DZ). This results in the following two distributions (Guo and Wang, 2002):

$$\begin{aligned} \text{individual: } y_{i(f(z))} &\sim \mathcal{N}(\nu_{f(z)},\ \sigma_z^2), \\ \text{family: } \nu_{f(z)} &\sim \mathcal{N}(\alpha,\ \ \tau_z^2). \end{aligned} \tag{11}$$

There are five parameters to estimate: the trait's overall mean α, variances σ_z^2 and τ_z^2. As in the ACE model (1), covariates can be readily added to the model. The two distributions can be extended to be any probability density function that appropriately characterizes the individual-level effects $y_{i(f(z))}$ and family-level effects $\nu_{f(z)}$ through their respective variances of σ_z^2 and τ_z^2. We note that the individual-level variability is allowed to be heterogeneous between the MZ and DZ zygosities: that is, σ_{MZ}^2 is not necessarily equal to σ_{DZ}^2.

The variances in the hierarchical model (11) correspond to the three sources of variability in the ACE

[1] A popular term for this type of framework is linear mixed-effects modeling. We adopt the term "hierarchical modeling estimation" to emphasize the hierarchical aspect of the data structure.

model. Specifically, the inter-individual variance σ_z^2 is attributed to nonshared environment E, while the inter-family variance τ_z^2 is divided into two latent components associated with additive genetic effects A and shared environment C. Similar to formula (9), the correlation between any two twins, i_1 and i_2, within a family f can be estimated as

$$r_z = \mathrm{corr}(y_{i_1(f(z))},\ y_{i_2(f(z))}) = \frac{\tau_z^2}{\tau_z^2 + \sigma_z^2},\ z = \mathrm{MZ},\ \mathrm{DZ}. \tag{12}$$

The estimates for h^2, c^2, and e^2 can be obtained by plugging in r_{MZ} and r_{DZ} into Falconer's formula (10). This approach allows for heritability estimation without explicit dependence on the three latent effects of A, C, and E, as depicted in the lower portion of the diagram in Fig. 2. In addition to the conventional point estimate and its uncertainty (standard error or quantile interval), the HME approach, within the Bayesian framework, can provide a full posterior distribution that represents the epistemic uncertainty.

2.2 Advantages of the hierarchical framework

There are three advantages to the hierarchical framework. First, it is more general than the conventional ACE model (1) and encompasses it as a special case. If we assume total variance homogeneity between the two zygosities, as in the ACE model (1),

$$\tau_{\mathrm{MZ}}^2 + \sigma_{\mathrm{MZ}}^2 = \tau_{\mathrm{DZ}}^2 + \sigma_{\mathrm{DZ}}^2, \tag{13}$$

the equivalence between the two modeling approaches can be established. Specifically, the equivalence is observed when we equate (a) the total variance of $\sigma_A^2 + \sigma_C^2 + \sigma_E^2$ in the ACE model (1) and that in (13) under the hierarchical framework (11), as well as (b) r_z in (9) and (12):

$$\sigma_A^2 = 2(\tau_{\mathrm{MZ}}^2 - \tau_{\mathrm{DZ}}^2) = 2(\sigma_{\mathrm{DZ}}^2 - \sigma_{\mathrm{MZ}}^2), \qquad \sigma_C^2 = 2\tau_{\mathrm{DZ}}^2 - \tau_{\mathrm{MZ}}^2, \qquad \sigma_E^2 = \sigma_{\mathrm{MZ}}^2 = \sigma_{\mathrm{DZ}}^2 - (\tau_{\mathrm{MZ}}^2 - \tau_{\mathrm{DZ}}^2).$$

It is important to emphasize that the hierarchical model (11) does not assume total variance homogeneity (13). Instead, proportionality homogeneity across zygosities is assumed when estimating heritability using Falconer's formula (10).

The second advantage of HME is its adaptability and flexibility in accommodating different distribution types, unlike the ACE model, which requires Gaussianity. As shown in the model formulation (11), the individual- and family-level variances in the two Gaussian distributions $\mathcal{N}(\cdot,\ \sigma_z^2)$ and $\mathcal{N}(\cdot,\ \tau_z^2)$ can be expanded to include a wider range of distribution options, such as Student's t distribution, exponentially-modified Gaussian distribution, and (shifted) log-normal distribution. This flexibility in choosing distributions can greatly enhance the quality of the model, especially when working with datasets that exhibit heavy tails, positive-only quantities, truncated or bounded values.

Finally, and perhaps most importantly, HME has the ability to explicitly capture intra-individual variability, rather than grouping measurement errors with the nonshared environment component. This approach offers the advantage of appropriately partitioning variability within the data hierarchy. We delve into this feature extensively through simulations and real data analysis in the subsequent section.

2.3 Consistency between ACE formulation and HME

Compared to the conventional ACE model, how does a hierarchical model perform on a simple dataset with negligible measurement errors from a twin study? To address this question, we utilized a publicly available dataset of body mass index (BMI) from the `R` package `mets` (Holst et al., 2016) to validate the hierarchical modeling approach. Despite having only one BMI measurement per individual, the presence of intra-individual variability can be considered negligible. In summary, the BMI data comprised $I = 11188$ twins from $F = 6917$

families in Finland, including 3665 MZ and 7523 DZ twins aged between 32 and 61 years. Information on each individual's age and sex was also included in the dataset. Both the data and the code for this example are available at https://github.com/afni/apaper_heritability.

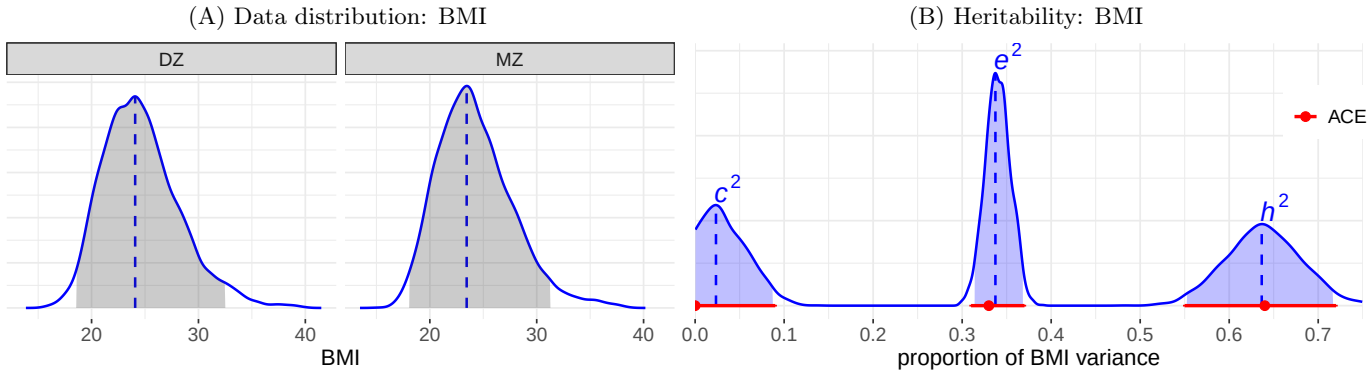

Figure 3: (A) Data distribution. BMI exhibits a right-skewed distribution, with slightly greater dispersion among DZ twins compared to MZ twins. (B) Heritability estimation of BMI data. The proportions of BMI variability attributed to each of the three components are depicted. The distributions for h^2, c^2, and e^2 are estimated using the hierarchical model (11) and presented in blue. The shaded area under each distribution represents the 95% highest density interval, while the vertical dashed line indicates the mode (peak). For comparison, the point estimates (dot) and their corresponding 95% quantile intervals (horizontal line) for the ACE model (1) are displayed in red.

Heritability estimation for the BMI dataset was performed using the ACE model with the following specifications. Alongside the three latent components A, C, and E in formulation (1), we incorporated the following covariates: zygosity and a nonlinear age effect for sex using third-order B-spline bases. The ACE model was implemented using the `twinlm` function from the `R` package `mets` (Holst et al., 2016). The ACE model yielded the following point estimates: a heritability estimate of $h^2 = 64\%$, a shared environment estimate of $c^2 = 0\%$, and a non-shared environment estimate of $e^2 = 34\%$. The 95% uncertainty intervals for these estimates are indicated in red in Fig. 3B.

The HME approach provided comparable estimation for heritability (Fig. 3B). Specifically, we adopted the model (11) with log-normal and normal density for the individual- and family-level distributions, respectively. The choice of log-normal distribution was based on the tendency of the BMI data skewed to the right (3A). The following covariates were included in the hierarchical model: zygosity and nonlinear age effect for each sex using smooth splines with thin plate bases. The hierarchical model (11) was implemented under the Bayesian framework using the `R` package `brms` (Bürkner, 2017). The resulting estimation for h^2, c^2, and e^2 was largely consistent with the ACE model (Fig. 3B), both in terms of point estimate and uncertainty range values.

While the point estimates and ranges of uncertainty obtained from both methods are largely similar, the Bayesian HME approach offers four distinct advantages over ACE. First, nonlinearity can be effectively captured through smooth splines without the need to explicitly specify the extent of nonlinearity, such as the number of degrees for polynomials. Second, the Bayesian framework provides a wider range of distribution options, as demonstrated here with a log-normal density for the BMI dataset. Third, more robust estimation can be achieved for parameter values that are close to numerical boundaries. In some cases, when a variance is very small, conventional modeling algorithms may yield unrealistic estimates due to numerical singularities (e.g., zero variance). In contrast, random draws from Monte Carlo simulations within the Bayesian framework typically provide more realistic estimations. Lastly, posterior distributions obtained from Bayesian modeling offer a more intuitive interpretation of parameter uncertainty compared to conventional confidence intervals. Bayesian approaches provide a comprehensive view of the uncertainty by presenting the full posterior distribution, allowing for a more nuanced understanding of parameter estimation (Morey et al., 2016).

3 HME: accounting for intra-individual variability

The precision of an estimate is a crucial aspect for evaluating its reliability, often quantified by its associated standard error or quantile interval. Within the hierarchical framework, we will employ numerical simulations to systematically investigate the influence of trial and participant sample sizes on precision, and determine their required orders of magnitude to achieve a satisfactory level of precision. The insights obtained from these simulations will be further validated by applying the hierarchical modeling framework to a behavioral dataset involving twins from the ABCD study.

3.1 Measurement errors: part of nonshared environment component?

Measurement errors have traditionally been regarded as part of the nonshared environment component in the heritability model, either implicitly or explicitly (e.g., Maes, 2005; Germine et al., 2015). In other words, it has been considered appropriate to include any measurement errors in the trait measurement $y_{i(f(z))}$ within the variance component σ_E^2 under the ACE model (1).

While measurement errors are expected to be inherent in any study, we contend that an ideal modeling approach should appropriately allocate these errors rather than grouping them together with the nonshared environment. Suppose that the observation $y_{i(f(z))t}$ $(t = 1, 2, ..., T)$ in the tth trial is sampled from a Gaussian distribution with a mean effect $\theta_{i(f(z))}$ and a standard error ω_0,

$$\text{trial: } y_{i(f(z))t} \sim \mathcal{N}(\theta_{i(f(z))},\ \omega_0^2). \tag{14}$$

Thus, the sample mean $\widehat{\theta}_{i(f(z))} = \sum_{t=1}^{T} y_{i(f(z))t}/T$ carries a sampling variance $\omega^2 = \omega 0^2/T$, which represents the precision of the estimate. Therefore, on average, a dataset with $T = 50$ trials would have twice the sample variance of another dataset with $T = 100$ trials. In common practice, when data is aggregated, only the sample mean $\widehat{\theta}_{i(f(z))}$ is utilized in the ACE model, while measurement errors (e.g., sampling variance) are not explicitly accounted for. Consequently, measurement errors are implicitly treated as part of the nonshared environment in the ACE model (1). Therefore, the sampling variance $\omega^2 = \omega_0^2/T$ remains embedded as part of σ_E^2, and the estimation of heritability in (6) becomes dependent on the sampling precision. As heritability aims to measure differences among individuals rather than within individuals, it is more conceptually sensible to construct a model where the sample size impacts the precision of the estimated variance, rather than its accuracy, in heritability estimation.

Measurement errors can be appropriately accounted for as a separate component from the nonshared environment. To maintain logical consistency of latent effects under the ACE model (1), we consider the individual-level trait effect $\theta_{i(f(z))}$ as a latent variable that is not directly observable. For the corresponding estimate $\widehat{\theta}_{i(f(z))}$, we characterize the measurement error through its standard error $\widehat{\omega}$. In other words, we do not solely partition the trait into the three latent components of A, C, and E. Instead, we also treat the true trait effect $\theta_{i(f(z))}$ as another latent effect, as depicted in the path diagram shown in Figure 4. Additionally, we emphasize that the standard error $\widehat{\omega}$ is not conceptualized as part of the (latent) nonshared environment component E, but rather as a separate entity that is directly observable. Therefore, the original ACE model (1) is augmented to include two levels,

$$\begin{aligned} &\text{aggregation: } \widehat{\theta}_{i(f(z))} \sim \mathcal{N}(\theta_{i(f(z))},\ \widehat{\omega}^2); \\ &\text{individual: } \theta_{i(f(z))} = \alpha + A_{i(f(z))} + C_{i(f(z))} + E_{i(f(z))}. \end{aligned} \tag{15}$$

The same distribution assumptions in (2) apply to the three latent effects of A, C, and E here. Solving this augmented ACE model (15) directly is challenging. However, in Subsection 3.4, we will present an approximate approach to heritability estimation 3.4.

Measurement errors can also be incorporated into the hierarchical modeling framework, similar to (15).

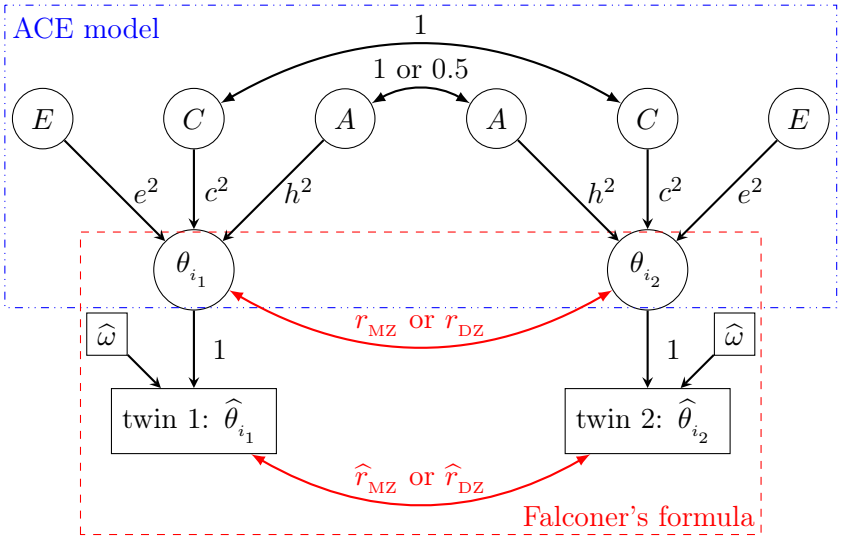

Figure 4: Path diagram (reticular action model) for the augmented ACE formulation (15). Subscript indices for individual i, family f, and zygosity z are mostly omitted from the nodes for brevity. Unlike the path diagram in Fig. 2, the trait measures y are treated as latent effects and are represented as eclipses, similar to other latent effects (e.g., components A, C, and E). Only the sample means, $\widehat{\theta}_{i_1}$ and $\widehat{\theta}_{i_2}$, are observable. A directed path, indicated by a single-headed arrow, represents predictability, while an undirected path, shown by a double-headed curved arrow, represents a covarying relationship. The value on each path indicates the correlation coefficient.

In particular, we augment the hierarchical model (11) to

$$\begin{aligned} \text{aggregation: } \widehat{\theta}_{i(f(z))} &\sim \mathcal{N}(\theta_{i(f(z))}, \ \widehat{\omega}^2); \\ \text{individual: } \theta_{i(f(z))} &\sim \mathcal{N}(\nu_{f(z)}, \quad \sigma_z^2); \\ \text{family: } \nu_{f(z)} &\sim \mathcal{N}(\alpha, \qquad \tau_z^2). \end{aligned} \tag{16}$$

When the standard error $\widehat{\omega}$ is available, this augmented hierarchical model (16) can be analyzed under the Bayesian framework. Alternatively, it can be approximately solved for heritability estimation, as discussed in Subsection 3.4.

Framing the presence of measurement errors as a distinction between observed and latent (theoretical) effects helps in understanding the associated impact. As illustrated in the path diagram (Fig. 4), one can estimate heritability directly based on the correlations $\widehat{r}_z$ using the sample means $\widehat{\theta}_{i(f(z))}$ through Falconer's formula (10). Long ago, Spearman Spearman (1904) highlighted a bias issue: the correlation between two quantities becomes attenuated when measurement errors are not accounted for. Similarly, when measurement errors are disregarded, $\widehat{r}_z$ would be underestimated compared to their counterparts r_z based on the latent effects $\theta_{i(f(z))}$ (Fig. 4). In the next subsection, we construct a hierarchical model at the observation level to fully illustrate the issues associated with data aggregation.

3.2 HME with trial-level data under one task condition

We begin by extending the hierarchical model (11) to a case where data are collected at the observation level with repeated measures through trials under a single task condition. The data $y_{i(f(z))t}$ are represented using four indices. as illustrated in Fig.(5): family ($f = 1, 2, ..., F$), zygosity ($z = \mathrm{MZ}, \mathrm{DZ}$), individual ($i = 1, 2, ..., I$), and trial ($t = 1, 2, ..., T$). Instead of utilizing aggregated information as in (15) and (16), we formulate a hierarchical model at three hierarchical levels:

$$\begin{aligned} \text{trial: } y_{i(f(z))t} &\sim \mathcal{N}(\theta_{i(f(z))}, \ \omega_0^2); \\ \text{individual: } \theta_{i(f(z))} &\sim \mathcal{N}(\nu_{f(z)}, \quad \sigma_z^2); \\ \text{family: } \nu_{f(z)} &\sim \mathcal{N}(\alpha, \qquad \tau_z^2). \end{aligned} \tag{17}$$

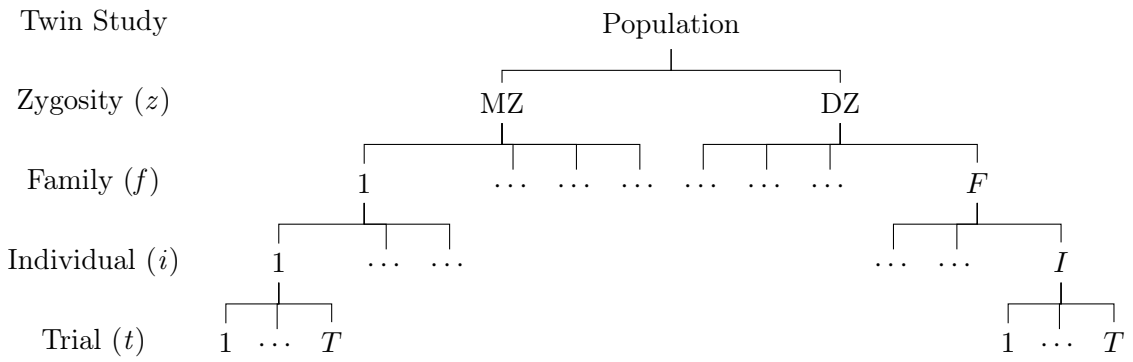

Figure 5: Hierarchical structure with trial-level data in twin studies. The nesting structure involves four levels: the trait is measured for each individual in a family, who can be either a MZ or DZ twin. Additionally, multiple instantiations, trials, or repeated measures are utilized to assess the trait effect. The four indices t, i, f, and z are indicators for trial (or observation), individual, family, and zygotic type.

Here, the formulation (17) can be extended to include various distributions. For instance, the trial-level effects can be characterized through Bernoulli, binomial, exponential, gamma, and Poisson distributions, which allow for modeling different types of data (count, binary, skewed) using appropriate link functions.

Heritability can be estimated using the hierarchical model (17) for trial-level data as follows. Similar to the situation with HME (Hierarchical Multiple Exposures) for the conventional ACE model in (11), we compute the correlations between two twins within a family using the inter-individual and inter-family variances, σ_z^2 and ${\tau_z}^2$, through the formula (12). Then, the three variance proportions h^2, c^2, and e^2 are obtained using Falconer's formula (10).

3.3 Consequences of data aggregation under the ACE model

Now we examine the common practice of data aggregation in light of the hierarchical model (17). Lurking within the traditional approach of data aggregation is an accountability issue regarding intra-individual variability. One important aspect of the hierarchical model (17) is the separation between heritability estimation and the intra-individual variability ω_0^2. Conceptually, heritability characterizes the proportion of additive genetic effects that distinguish MZ twins from DZ twins. In other words, the assessment of additive genetic effects occurs at the individual and family levels through the respective variances (σ_z^2 and τ_z^2); therefore, they should be independent of the variability at the trial level. This separation is evident in the fact that both inter-individual and inter-family variances, σ_z^2 and τ_z^2, are formulated independently from the intra-individual variability ω_0^2, which is captured through the trial-level distribution $\mathcal{N}(\cdot,\, \omega_0^2)$. In comparison, common practice often involves aggregating trial-level data by computing averages across trials, and the resulting summarized data are then used in the conventional ACE model.

What is the impact of intra-individual variability on heritability estimation? For traits such as body mass index that exhibit relatively small intra-individual variability compared to inter-individual variability, both the ACE model (1) and the reduced hierarchical model (11), as approximate special cases, can be reasonably employed. However, it is important to emphasize that data aggregation does not eliminate the intra-individual variability; rather, it conceals it within the aggregated data. As illustrated below, this leads to a bias in the estimation of heritability, which can be seen as either erroneously attributing measurement error to the nonshared environment component (ACE model) or improperly combining intra-individual variability with inter-individual variability (HME model).

To trace the bias, we first assess the magnitude of intra-individual variability ω_0^2. Recent studies (Rouder and Haaf, 2019; Chen et al., 2021) have demonstrated that r_{MZ} and r_{DZ}, as intraclass correlations, can be underestimated when the relative magnitude of intra-individual variability ω_0^2 is large within the framework

(11). To capture the crucial role of intra-individual variability ω_0^2 across different phenotypic traits, we define a dimensionless measure of the variability ratio for each zygosity:

$$R_{v,z} = \frac{\omega_0}{\sqrt{\tau_z^2 + \sigma_z^2}}, \ z = \text{MZ, DZ}. \tag{18}$$

This ratio characterize the impact of intra-individual variability ${\omega_0}^2$. Under the assumption of homogeneity (13), $R_{v,\text{MZ}} = R_{v,\text{DZ}}$. However, in real applications, the ratio $R_{v,z}$ may slightly differ between the two zygosities within the hierarchical framework (17).

The variability ratio $R_{v,z}$ captures the fundamental aspect of heritability: the proportion of inter-family and inter-individual variance relative to intra-individual variance. Under the assumption of total variance homogeneity, $R_{v,z}$ is equal between MZ and DZ twins. For simplicity, we drop the superscript Z and denote their average as R_v. In subsequent sections, we will demonstrate the crucial role of this ratio R_v in comparing model performance, understanding the level of uncertainty in heritability estimation, and determining sample size requirements in twin studies.

How to quantify the bias of heritability estimation when intra-individual variability is not properly accounted for? When the trial-level data $y_{i(f(z))t}$ are aggregated across trials with their average $\bar{y}_{i(f(z))} = \frac{1}{T}\sum_{t=1}^{T} y_{i(f(z))t}$, the model (11) becomes

$$\begin{aligned} \text{individual: } \bar{y}_{i(f(z))} &\sim \mathcal{N}(\nu_{f(z)}, \ \tilde{\sigma}_z^2), \ \tilde{\sigma}_z^2 = \sigma_z^2 + \frac{\omega_0^2}{T}; \\ \text{family: } \nu_{f(z)} &\sim \mathcal{N}(\alpha, \quad \tau_z^2). \end{aligned} \tag{19}$$

In comparing the model (19) with aggregated data to its counterpart (11) for data without measurement errors, we note that ignoring intra-individual variability leads to its combination with the inter-individual variance σ_z^2 into $\tilde{\sigma}_z^2$. As a result, the correlations r_{MZ} and r_{DZ} in (12) are updated to

$$\tilde{r}_z = \frac{\tau_z^2}{\tau_z^2 + \tilde{\sigma}_z^2} = \frac{\tau_z^2}{\tau_z^2 + \sigma_z^2 + \frac{1}{T}\omega_0^2} = r_z U, \tag{20}$$

where the introduced bias into r_{MZ} and r_{DZ} is quantified by the dimensionless quantity

$$U = \frac{1}{1 + \frac{1}{T}R_v^2}. \tag{21}$$

It is worth noting that $U < 1$ for nonzero ω_0, indicating that $\tilde{r}z$ are consistently downward biased due to the presence of intra-individual variability. In the limit where $\omega_0^2/[T(\tau_z^2 + \sigma_z^2)] \to 0$, which can occur with decreasing standard error ω_0 or increasing trial size T, then $\tilde{r}_z \to r_z$. According to Falconer's formula (10), both h^2 and c^2 would be underestimated by a factor of U, while e^2 would be overestimated by the same factor. For example, with $T = 100$ trials, a small intra-individual variability such as $R_v = 1$ has negligible impact on heritability estimation ($U \approx 0.99$), whereas a large intra-individual variability with $R_v = 10$ substantially underestimates h^2 and c^2 by 50%. Conversely, if $R_v = 3$, biases cannot be disregarded even with $T = 20$ trials unless T approaches or exceeds 100.

3.4 Ameliorating the biases in the ACE model

The biases induced in the ACE model can be theoretically corrected by introducing an adjustment term in the denominator of (20) to counteract the contaminating effect of $\frac{1}{T}\omega_0^2$ under the model (19). (19),

$$r_z = \frac{\tau_z^2}{\tau_z^2 + \tilde{\sigma}_z^2 - \frac{1}{T}\omega_0^2}. \tag{22}$$

Similarly, decontamination can be achieved for the formulas (6-8) and (12). However, the success of these adjustments relies on the availability of the intra-individual variance ω_0^2, which is not directly accessible once the data are aggregated.

Nevertheless, the biases can be practically mitigated. Once the trial-level data has been aggregated, it becomes challenging to directly estimate the intra-individual variance ω_0^2 under the model (19). However, we can use the available cross-trial variances $s^2_{i(f(z))}$ for $y_{i(f(z))t}$ to approximate the estimate $\widehat{\omega}_0^2 = \frac{1}{I}\sum_{i=1}^{I} s^2_{i(f(z))}$ in (22). This allows us to obtain the following adjusted estimates as reasonable approximations,

$$\widehat{r}_z = \frac{\tau_z^2}{\tau_z^2 + \tilde{\sigma}_z^2 - \frac{1}{T}\widehat{\omega}_0^2}. \tag{23}$$

Similarly, as σ_E inherently contains the additive contribution of measurement error. we can directly adjust the biases in the ACE estimates to,

$$\widehat{h}^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_C^2 + \sigma_E^2 - \frac{1}{T}\widehat{\omega}_0^2}, \quad \widehat{c}^2 = \frac{\sigma_C^2}{\sigma_A^2 + \sigma_C^2 + \sigma_E^2 - \frac{1}{T}\widehat{\omega}_0^2}, \quad \widehat{e}^2 = \frac{\sigma_E^2 - \frac{1}{T}\widehat{\omega}_0^2}{\sigma_A^2 + \sigma_C^2 + \sigma_E^2 - \frac{1}{T}\widehat{\omega}_0^2} = 1 - \widehat{h}^2 - \widehat{c}^2. \tag{24}$$

These approximate adjustments (23) and (24) offer a solution to the augmented ACE model (15) and its hierarchical counterpart (16). We will further explore and validate their effectiveness as approximate adjustments.

An intriguing aspect of biased estimation for heritability is its analogy to the phenomenon of correlation attenuation in the presence of measurement errors. Spearman (1904) recognized the problem of bias caused by measurement errors and proposed an adjustment method to disattenuate the correlation between two variables. In (20), the term U serves a similar purpose to the reliability coefficient or separation index in classical test theory. Consequently, it is interesting to note that the decontamination formula (23) and its approximation (24) employ a similar adjustment strategy as suggested by Spearman (1904).

3.5 HME with trial-level data under two task conditions

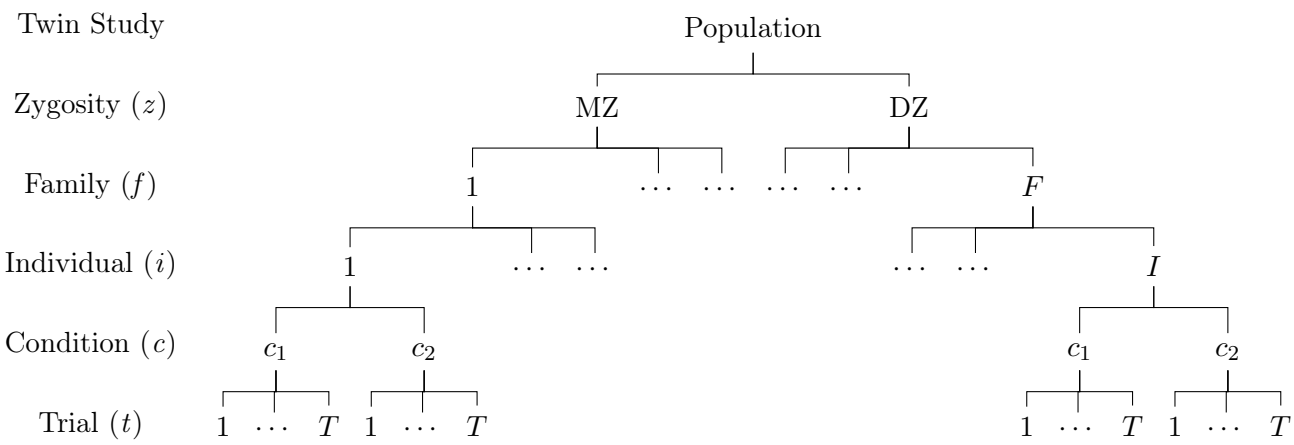

Figure 6: Hierarchical structure with trial-level data in twin studies under two conditions. The figure illustrates the nesting structure involving five levels: a trait is measured for each individual under two conditions in a family with either MZ or DZ twins. Furthermore, multiple trials are conducted to assess the effects at the condition level. The figure indicates the four indices t, c, i, f, and z, which represent the trial, condition, individual, family, and zygosity, respectively.

We now extend the hierarchical model (17) to accommodate two task conditions. In fields such as psychometrics and neuroimaging, the focus often lies in comparing and analyzing the contrast between two conditions. One illustrative example is the contrast between incongruent and congruent conditions in a cognitive inhibition experiment, such as the Stroop task. We expand the previous hierarchical model (17) in Subsection 2.2, resulting in the following formulation that accommodates five levels (Fig. 6) – family ($f = 1, 2, ..., F$),

zygosity ($z = \text{MZ, DZ}$), individual ($i = 1, 2, ..., I$), condition ($c = c_1, c_2$), and trial ($t = 1, 2, ..., T$):

$$\begin{aligned}
\text{trial:} \quad & y_{ci(f(z))t} && \sim \mathcal{N}(\theta_{ci(f(z))},\ \omega_0^2); \\
\text{individual:} \quad & \begin{bmatrix} \theta_{c_1 i(f(z))} \\ \theta_{c_2 i(f(z))} \end{bmatrix} && \sim \mathcal{N}\left(\begin{bmatrix} \nu_{c_1 f(z)} \\ \nu_{c_2 f(z)} \end{bmatrix}, \begin{bmatrix} \sigma^2_{c_1,g} & * \\ * & \sigma^2_{c_2,g} \end{bmatrix} \right); \\
\text{family:} \quad & \begin{bmatrix} \nu_{c_1 f(z)} \\ \nu_{c_2 f(z)} \end{bmatrix} && \sim \mathcal{N}\left(\begin{bmatrix} \alpha_{c_1} \\ \alpha_{c_2} \end{bmatrix}, \begin{bmatrix} \tau^2_{c_1,g} & * \\ * & \tau^2_{c_2,g} \end{bmatrix} \right).
\end{aligned} \tag{25}$$

The differences from subsection 2.2 are twofold: (a) the presence of two intercepts, α_c, one for each condition, and (b) the individual- and family-level distributions being bivariate instead of univariate. While the covariances for the individual- and family-level distributions are not of interest in the current context, we acknowledge their presence by denoting them with an asterisk $*$ in the respective variance-covariance matrix.

Heritability estimation is straightforward for each condition under the hierarchical model (25). First, we calculate the correlation between two twins within a family for condition c_k ($k = 1, 2$) using

$$\begin{aligned}
r_{c_k,\text{MZ}} &= \frac{\tau^2_{c_k,\text{MZ}}}{\tau^2_{c_k,\text{MZ}} + \sigma^2_{c_k,\text{MZ}}}, \\
r_{c_k,\text{DZ}} &= \frac{\tau^2_{c_k,\text{DZ}}}{\tau^2_{c_k,\text{DZ}} + \sigma^2_{c_k,\text{DZ}}}.
\end{aligned} \tag{26}$$

Then, we estimate heritability for each condition by plugging $r_{c_k,\text{MZ}}$ and $r_{c_k,\text{DZ}}$ into Falconer's formula (10).

Two approaches are available for estimating heritability and the variability ratio for the contrast between two conditions. One approach involves reparameterizing the model (25) by using an indicator variable to dummy-code the two conditions,

$$x_{c_k} = \begin{cases} \frac{1}{2}, & \text{if } k = 1; \\ -\frac{1}{2}, & \text{if } k = 2. \end{cases} \tag{27}$$

This reparameterization replaces the two conditions, c_1 and c_2, in the model (25) with two new parameters representing the average and the contrast between the two conditions, respectively. Alternatively, within the Bayesian framework, one can directly obtain the distribution of the contrast by formulating it based on each posterior draw from the hierarchical model (25).

The discussions on biased estimation due to data aggregation and its correction in the previous subsection also apply to the scenario with two conditions. For each condition and their contrast, the variability ratio R_v can be similarly defined. The only modification for the contrast is to replace $\frac{1}{T}\sigma_0^2$ with $\frac{2}{T}\sigma_0^2$ in (18). Similarly, the discussions regarding the biases in heritability estimation for each condition in the preceding subsection can be directly applied here. However, when considering the contrast, the bias requires replacing $\frac{1}{T}\sigma_0^2$ in (19-24) with $\frac{2}{T}\sigma_0^2$.

To summarize, we argue that the HME approach offers several advantages over the conventional ACE model in heritability estimation. By closely reflecting the actual data structure, a hierarchical model appropriately accounts for different sources of data variability. In cases where data are collected through repeated measures, they are often aggregated during preprocessing, allowing for convenient fitting to models like the ACE formulation (1) or the hierarchical platform (11). However, based on the theoretical reasoning presented thus far, we expect biases to occur when data are aggregated, with the severity dependent on the ratio R_v and the trial sample size T. Furthermore, these biases could theoretically be corrected if the intra-individual variance σ_0^2 were known a priori. However, the extent of ambiguity in heritability estimation remains unclear, as there is no analytical quantification available.

3.6 Assessing model performance through simulations

Simulations are an effective approach for assessing model performance as they allow for controlled experiments with known ground truth. They serve to validate, inform, and expand upon theoretical insights. Simulations are particularly valuable when theoretical insights are absent, incomplete, or in doubt. In our case, we will employ simulations to investigate the following questions and gain further understanding:

1) *Does the ACE model truly provide biased estimation in the presence of measurement errors?* If so, what factors contribute to the bias and ambiguity?
2) *Does the HME approach provide the expected unbiased estimation in the presence of measurement errors?* Conversely, can biases resulting from data aggregation under ACE be rectified as theoretically derived?
3) *How can we assess the precision of heritability estimation?* What role does the variability ratio play in the precision? What insights can we gain regarding the required sample sizes of twins and trials?

We conducted simulations in two scenarios: 1) a study with one task condition, and 2) a study with the contrast between two conditions. For each scenario, we simulated two approaches: the ACE model (1), which aggregated data across trials by averaging within each condition, and the hierarchical formulation (17) or (25). The simulations involved manipulating three parameters: the number of families (F), the number of trials per condition (T), and the variability ratio R_v. Additionally, two bias adjustments for the ACE model, referred to as ACE1 and ACE2 in Fig. 7, were performed. Detailed information about the specific numerical schemes employed can be found in Appendix A.

The summarized findings for the contrast of two conditions (Fig. 7) are as follows. The results for a single condition, although not presented here, exhibit similar patterns.

1) **HME provides virtually unbiased estimation, while the ACE model exhibits biases.** The distributions of estimated h^2, c^2, and e^2 show peaks around their simulated values. In contrast, the ACE model underestimates h^2 and c^2 to varying degrees while overestimating e^2. The biases become more pronounced as the variability ratio R_v increases and/or the trial sample size T decreases.
2) **Intra-individual variability significantly affects the ACE model's performance.** The bias in ACE estimates worsens with higher values of R_v. When $R_v \lesssim 1$, heritability estimation can be reasonably approximated using the ACE model by aggregating data across trials. However, as R_v increases, noticeable bias emerges, and when $R_v \gg 1$, the bias becomes substantial.
3) **Intra-individual variability impacts estimation precision.** Smaller values of R_v lead to more concentrated estimates with smaller standard errors, indicating greater precision. Conversely, larger values of R_v result in more diffuse and uncertain estimation across both models and sample sizes.
4) **Family sample size has a greater impact than trial sample size.** Increasing the number of families or trials reduces estimation uncertainty. However, the impact of family sample size is more substantial. For instance, when $R_v \lesssim 1$, an appropriate level of uncertainty can be achieved with 50 trials and 1,000 families. When $R_v \gg 1$, several thousand families may be necessary.
5) **Biases under the ACE model can be effectively adjusted.** After data aggregation, adjustments can be made using empirical or theoretical standard errors through (23) or (22). Both adjustments show satisfactory agreement with the HME results.

To recapitulate, the role of intra-individual variability is crucial. The variability ratio R_v imposes significant constraints on heritability estimation. The simulations indicate that when $R_v \gg 1$, several issues may arise: 1) the ACE model with data aggregation tends to underestimate heritability; 2) achieving precise estimations becomes challenging; 3) larger sample sizes than those commonly employed in the field may be required. However, for certain traits like BMI, intra-individual variability is likely negligible ($R_v \ll 1$). This raises an important question: What is the relative magnitude of intra-individual variability, or R_v, for psychometric traits such as reaction time and BOLD responses in functional magnetic resonance imaging? Addressing this question will contribute to improved study design and facilitate informed modeling choices.

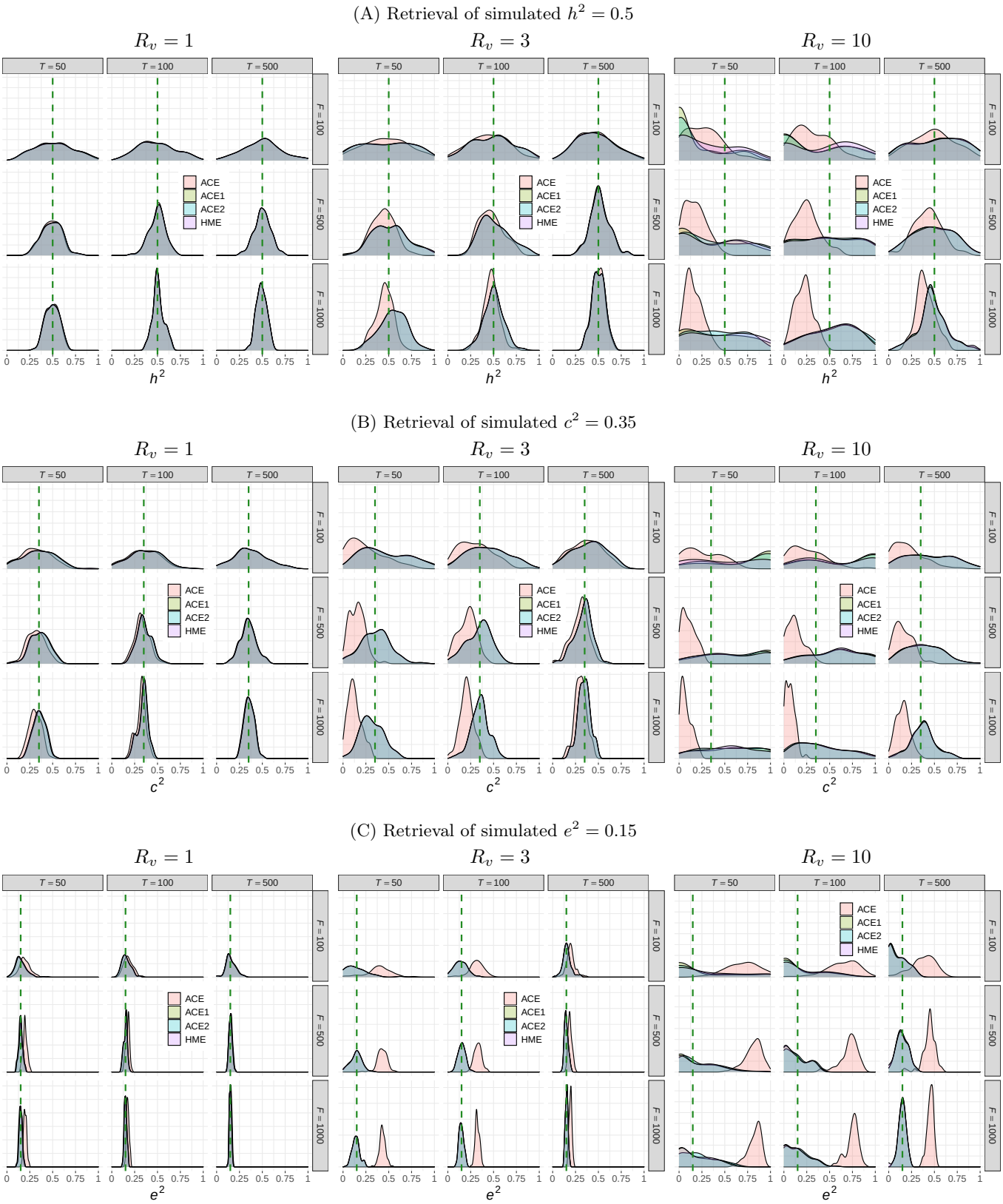

Figure 7: Simulation results of a trial-based experiment for the contrast between two conditions. Four factors were considered in the simulations: two models (ACE and HME), three variability ratios ($R_v = 1, 4, 10$), three family sizes ($F = 100, 500, 1000$), and three trial sample sizes ($T = 50, 100, 500$). ACE1 and ACE2 represent the adjusted estimates obtained through the decontamination formulas (23) and (22), respectively. Each curve represents the density of parameter estimation from 200 repetitions, and each vertical dashed line indicates the simulated parameter value. The curves for HME, ACE1, and ACE2 are nearly indistinguishable from each other.

4 Applying hierarchical modeling to an experimental dataset

We apply the HME approach to an experimental dataset to address two primary questions. First, do the insights gained from numerical simulations in the previous section align with the findings when real data is analyzed? Second, what is the range of the relative magnitude of intra-individual variability, as indicated by the ratio R_v, in commonly encountered empirical datasets?

In this section, we utilize a behavioral dataset obtained from an experiment conducted as part of the ABCD study. The experiment focused on investigating selective attention during adolescence using an emotional Stroop task (Smolker et al., 2022), with reaction time serving as a phenotypic trait. The data was collected from 21 sites during the 1-year follow-up visit and is publicly accessible through the 2020 ABCD Annual Curated Data Release 4.0 (https://nda.nih.gov/study.html?id=1299). The analysis scripts can be found online at: https://github.com/afni/apaper_heritability.

4.1 Data description

A subset of the original dataset, specifically containing twins, was utilized for the analysis. Refer to Table 2 for detailed information on participant counts and demographic data. The subset consisted of 1,102 twins (including some triplets) from 555 families and was selected from a larger dataset of 11,876 participants (Iacono et al., 2018). Among the included twins, there were 461 monozygotic (MZ) and 641 dizygotic (DZ) participants from four different sites.

Twin	$I = 1102$ twins; $F = 555$ families; 3 families with DZ triplets; 11 families with available data from only one twin
Zygosity	461 MZ twins, 641 DZ twins
Sex	550 males, 552 females
Race	720 white, 152 black, 112 Hispanic, 3 Asian, 114 others
Age	108 to 132 months; mean: 121 months, standard deviation: 6 months

Table 2: Demographic information of twins in a Stroop experiment from the ABCD Study

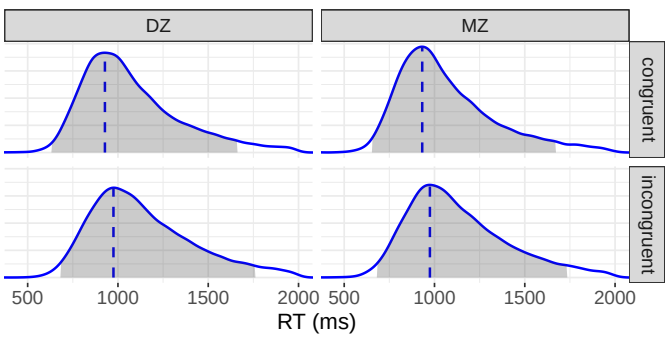

Figure 8: Reaction time (RT) distributions. The shaded area under each density represents the 95% highest density interval, while the vertical dashed line indicates the mode. The RT distribution is 1) right-skewed, 2) slightly right-shifted under the incongruent condition compared to the congruent condition, and 3) slightly more dispersed for DZ twins than MZ twins.

We focus on the reaction time (RT) data from the behavioral dataset as the phenotypic trait for estimating heritability. The RT was measured for two levels of congruency in a Stroop task: congruent and incongruent. Each participant was instructed to respond to a total of 48 trials, consisting of 24 congruent and 24 incongruent trials. The response time window for each trial was set to 2000 ms. Out of the 1,102 twins, a total of 49,524 trials were included in the analysis after excluding incorrect responses. This resulted in an overall correct response rate of approximately 93.6%. The distribution of RTs is heavily right-skewed (Fig. 8), with a mode of 962 ms and a 95% highest density interval ranging from 634 ms to 1,827 ms.

4.2 Model comparisons with real data

The RT data was analyzed using two different approaches: HME and ACE. For the HME approach, the trial-level data was fitted using the formulation (25). Site, sex, race, age, and zygosity were included as

covariates. To account for the right-skewness of the RT data (as shown in Fig. 8), a log-normal distribution was used for the trial-level effects in the HME model. The Bayesian framework was employed to implement the HME approach, utilizing the `brms` package in `R` (Bürkner, 2017). Heritabilty was estimated for each of the three RT effects: the congruent condition, incongruent condition, and the Stroop effect (the contrast between incongruent and congruent conditions).

The ACE model was utilized with aggregated data. Specifically, the RT data was aggregated across trials for congruent and incongruent condition at the individual level. Similar to the HME approach, covariates including site, sex, race, age, and zygosity were included. The ACE model was implemented using the `twinlm` function from the `mets` package in `R` (Holst et al., 2016).

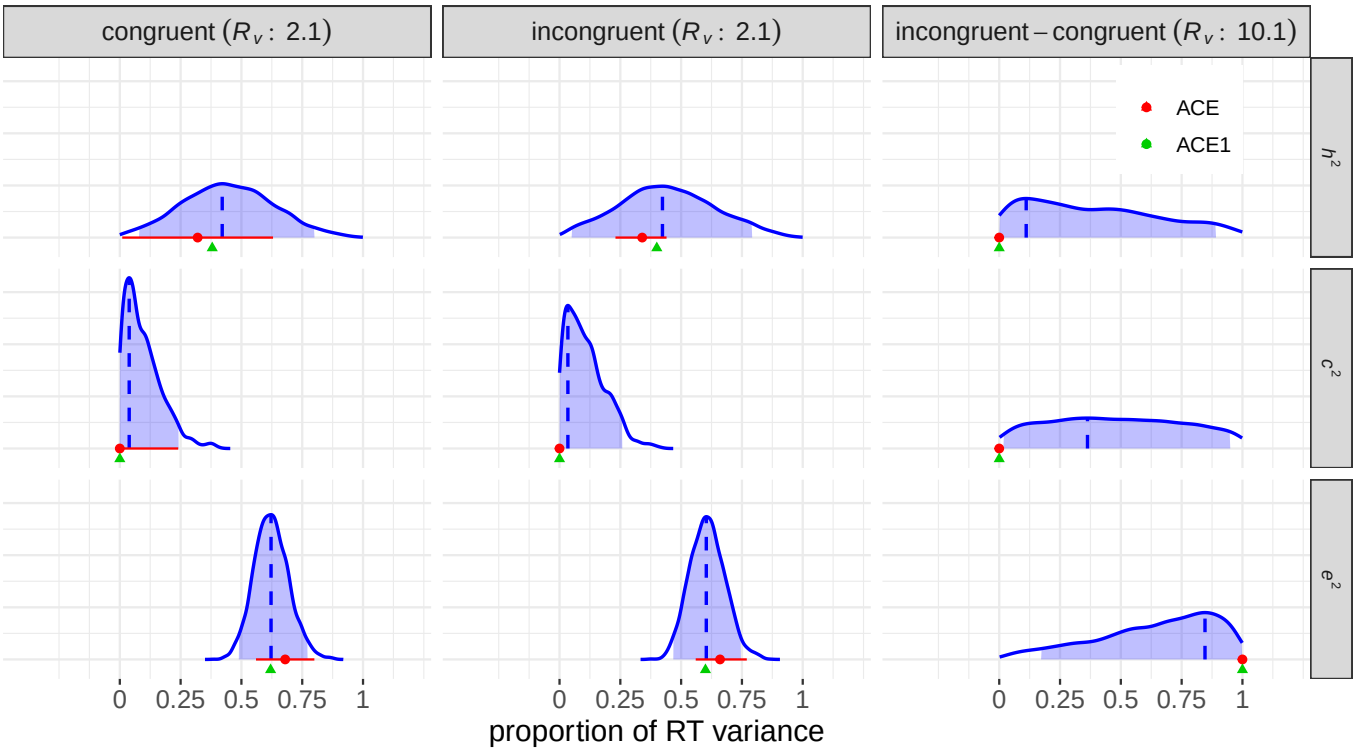

Figure 9: Estimated heritability for RT data. The three columns represent the two conditions (congruent and incongruent), as well as their contrast, with the corresponding variability ratio R_v indicated in the column labels. The three rows correspond to h^2, c^2, and e^2. In each panel, the HME result is represented by a solid blue density curve, derived from random draws from posterior chains. The mode is marked by a vertical blue dashed line, and the shaded blue region represents the 95% uncertainty interval. The ACE counterparts are also displayed in each plot, with the point estimate depicted as a red dot and its 95% uncertainty interval represented by a horizontal red line. Notably, the ACE point estimates tend to be smaller than their HME counterparts for h^2 and c^2 (while larger for e^2).Adjustments for the ACE estimates, using the formula (24), are denoted as ACE1 and shown as green triangles.

The estimation results are presented in Figure 9. The performance of HME relative to ACE, as summarized below, aligns with our simulations in subsection 3.6. Overall, both ACE and HME exhibited a significant amount of uncertainty in estimating heritability for both conditions. ACE displayed noticeable underestimation of h^2 and c^2 (first two columns, Fig. 9). However, neither modeling approach provided satisfactory estimation for the contrast (third column, Fig. 9). This estimation challenge arises from the combination of three factors encapsulated by a large intra-individual variability ratio $R_v \approx 10.1$: 1) a much smaller effect size, 2) an extremely limited trial sample size, and 3) a relatively small twin sample size.

Below are a few detailed elaborations:

1) **Impact of relative intra-individual variability on estimation precision**. First, under each of the individual congruent and incongruent conditions, the observed intra-individual variability was not very large ($R_v \approx 2.1$), resulting in moderate uncertainties for HME estimates of h^2, c^2, and e^2. However, the ACE model showed difficulty in accurately assessing uncertainty near the parameter boundaries. For

instance, the ACE model's small uncertainty interval (0, 0.01) for c^2 likely stemmed from numerical singularity issues in the traditional statistical framework. In contrast, regularization in hierarchical modeling (Chung et al., 2013) yielded a more reasonable uncertainty interval (0, 0.28) for c^2 under the Bayesian framework. Second, the intra-individual variability for the Stroop effect (contrast between the two conditions) was large ($R_v \approx 10.1$). Consequently, the uncertainties for h^2, c^2 under HME, and e^2 were very large, with the estimated density of c^2 resembling a uniform distribution, making it challenging to locate a peak. The ACE model did not provide any meaningful estimates either.

2) **Impact of relative intra-individual variability on estimation bias**. Under both congruent and incongruent conditions, the intra-individual variability is moderate ($R_v \approx 2.1$), resulting in small underestimations of h^2 and c^2 by the ACE model. Correspondingly, the overestimation of e^2 was also small. However, the large intra-individual variability for the Stroop effect ($R_v \approx 10.1$) led to more noticeable underestimations of h^2 and c^2 by the ACE model.
3) **Impact of relative intra-individual variability on sample sizes**. The larger R_v for the contrast ($R_v \approx 10.1$) is consequential. Simulation results in subsection 3.6 indicate that larger sample sizes, especially in terms of family count, would be required to reduce the large uncertainty. We note that the observed range of R_v values aligns with psychometric data from individual studies in test-retest reliability estimation (Rouder and Haaf, 2019; Chen et al., 2021; Baker et al., 2021).
4) **Bias adjustment for ACE estimates**. The biases under the ACE model, due to data aggregation, adjusted using formula (23), were reduced. The adjusted estimates for $\hat{h}^2$ under the congruent, incongruent, and contrast conditions were 0.38, 0.40, and 0.0, respectively (green triangles, Fig. 9). These adjustments effectively reduced bias, although they remained slightly biased compared to HME, which could be attributed to deviations from the Gaussian assumption under ACE.

We also explored the HME approach for the ABCD-Stroop data using a conventional linear mixed-effects modeling framework instead of a Bayesian approach. The `lmer` function from the `lme4` package in R (Bates et al., 2015) was utilized to fit the models (17) and (25) for the congruent and incongruent conditions, as well as their contrast, respectively. The covariates included site, sex, race, age, and zygosity, while the two conditions were dummy-coded using the indicator variable (27). However, the `lme4` package has limitations as it only accommodates the Gaussian distribution, making it challenging to handle situations involving data skewness and outliers. Although the point estimates (not shown here) for h^2, c^2, e^2, and R_v under the congruent and incongruent task conditions were largely consistent with the values obtained under the Bayesian framework, the numerical solver in `lme4` failed to converge for the contrast between the incongruent and congruent conditions due to the relatively small inter-individual variances ($R_v \approx 10.1$). Moreover, the conventional modeling framework does not readily provide uncertainty information for these point estimates of h^2, c^2, and e^2.

5 Discussion

Heritability estimation based on data with a non-negligible intra-individual variability necessitates a model that accurately represents the underlying hierarchical structure. When assessing a phenotypic trait with repeated measures, we propose a hierarchical model that encompasses all relevant levels, allowing for the incorporation, propagation, and separation of intra-individual measurement error from parameter estimation at higher levels. Through numerical simulations and a real behavioral dataset, we have demonstrated the advantages of hierarchical modeling over the conventional ACE framework. These advantages include: 1) avoidance of estimation bias, 2) the ability to account for the significant influence of intra-individual variability on heritability estimation, and 3) enhanced interpretability and explanatory power of results, such as identifying the challenges associated with reducing estimation uncertainty due to sample size limitations.

Standard ACE modeling is prone to estimation bias and underestimation of uncertainty. Many phenotypic

traits of interest exhibit considerable inter-trial variability that has not been adequately addressed. Heritability, as a quantitative measure of individual differences, shares this vulnerability and faces similar challenges in the presence of intra-individual variability, as demonstrated by recent studies investigating test-retest reliability (Rouder and Haaf, 2019; Haines et al., 2020; Chen et al., 2021).

5.1 The importance of modeling data generating mechanism

Data reduction through aggregation is a commonly used in statistical applications, particularly in studying individual differences. One reason for this is that many classical concepts in the field were developed in a context where measurement errors were minimal or nonexistent. For example, the traditional concept of intraclass correlation for test-retest reliability in individual studies was first developed before considering the notion of intra-individual variability. A similar trajectory can be observed in heritability estimation. Consequently, the existing modeling and conceptual frameworks do not explicitly account for the level of measurement units (e.g., individual trials) and instead simplify the data through preliminary aggregation steps, such as averaging. Moreover, data aggregation helps alleviate the computational burden of analysis.

However, it is important to recognize that data aggregation, although all models are Procrustean to some degree, carries the risk of information loss and, as we have demonstrated, estimation bias. Preliminary data reduction may distort the partitioning of effects among different sources of data variability. For instance, recent findings have shown that without properly accounting for intra-individual variability, the traditional modeling framework for intraclass correlation can lead to substantial underestimation of test-retest reliability (Rouder and Haaf, 2019; Haines et al., 2020; Chen et al., 2021). This oversight of ignoring the importance of intra-individual variability likely contributes to paradoxically low reliability estimates in the literature (e.g., Hedge et al., 2018; Elliott et al., 2020). In this study, we demonstrate that the same concern applies to heritability estimation in twin studies.

In contrast, hierarchical modeling provides a unifying framework and serves as a preferable alternative to data aggregation. Through numerical simulations and an analysis of an experimental dataset, we have illustrated the significance of accurately capturing the data structure using a hierarchical model. This approach prevents intra-individual variability from contaminating parameter estimation beyond the measurement level (e.g., trials). Conceptually, the hierarchical framework encompasses the conventional ACE model as a special case, as demonstrated by the formulations in (1) and (11), and supported by numerical evidence in Fig. 3.

5.2 Biases, ambiguity and challenge of heritability estimation

There are two aspects of accuracy compromise that need to be considered in heritability estimation. The first aspect pertains to estimation biases. As demonstrated in this study, failure to fully incorporate the data structure can lead to biased estimates of heritability. The second aspect concerns the ambiguity in heritability estimation. In addition to providing a point estimate for the effect of interest, it is equally important to quantify its uncertainty. This uncertainty can be characterized through measures such as standard error, a quantile interval (e.g., 95%), or even a full distribution (as depicted in Fig. 9). However, the concept of uncertainty is often not well emphasized in common practice. In some cases, only the central tendency (e.g., mean) of heritability estimation is reported, without conveying any information about uncertainty. However, to truly comprehend the generalizability of results, understanding uncertainty is crucial. One of the benefits of the HME framework is its ability to directly generate posterior distributions that illustrate estimate precision.

Our findings demonstrate that the biases resulting from data aggregation can be mitigated to some extent if variability can be determined separately (e.g., through repeated measures), as indicated by formulas (23) or (24). This insight stems from hierarchical modeling, which closely represents the actual data structure. Our simulations indicate that the bias adjustments are generally satisfactory when assuming a Gaussian distribution (Fig. 7). However, in practical applications, these adjustments are suboptimal due to distributional

deviations, as demonstrated in our example using the Stroop dataset. Hence, a comprehensive hierarchical modeling framework remains the preferred choice.

Uncertainty provides a measure of precision for a heritability estimate in addition to its point estimate. The dataset we used for demonstration, exemplifying a cognitive inhibition study, suggests that reasonable levels of uncertainty can be achieved with sample sizes of less than 1000 families and less than 100 trials per individual condition (congruent and incongruent). The estimated heritability of approximately 40% (first two columns, Fig. 9) aligns with the general range observed in typical phenotypic traits in the literature (Polderman et al., 2015). However, the contrast between conditions (e.g., Stroop effect) is often the focal point of interest. Even with HME estimation, the uncertainty of heritability for this contrast remains unresolved (third column, Fig. 9), creating ambiguity regarding its magnitude. In other words, despite attempts by the Consortium (Iacono et al., 2018; Smolker et al., 2022) to address the sample size issue, the dataset from the ABCD Study (consisting of 461 MZ and 641 DZ twin pairs, with less than 48 trials per condition) does not provide sufficient certainty for estimating the heritability of the Stroop effect. Achieving a reasonable level of precision may require impractical sample sizes (e.g., hundreds of trials and thousands of individuals). Consequently, conducting a detailed assessment of uncertainty allows for a more realistic evaluation of result generalizability.

The relative magnitude of intra-individual variability, as quantified by the ratio R_v, serves as an informative indicator in heritability modeling. As a dimensionless parameter, it influences not only the accuracy and uncertainty of heritability estimation but also those of test-retest reliability (Rouder et al., 2019; Chen et al., 2021). Historical power analyses in twin studies have suggested a minimum sample size of 600 twin pairs (Martin et al., 1978; Sham et al., 2020). However, our simulations demonstrate that, in the presence of measurement errors, a large Rv ratio poses a significant challenge for future studies in the field of individual differences, particularly when examining effect contrasts and higher-order interactions (e.g., Stroop effect between two emotional valences, happy and sad). Additionally, this ratio highlights the relative importance of trial sample size compared to participant sample size across various experimental modalities, such as functional magnetic resonance imaging, magnetoencephalography, electroencephalography, and psychometrics. In all these cases, the R_v ratio exceeds 1, and sometimes even surpasses 10 (Baker et al., 2021; Chen et al., 2021; Chen et al., 2022). Due to this substantial ratio, the sample size of trials can be nearly as crucial as the number of participants in terms of experimental efficiency in neuroimaging and psychometrics.

5.3 Heritability estimation in neuroimaging

To date, there has been an increasing number of twin studies utilizing task-based fMRI. In these studies, it has been a common practice to aggregate data across trials during fMRI data analysis, resulting in the neglect of intra-individual variability, which is neither accounted for nor reported. For instance, Polk et al. (2007) reported a small heritability estimate of BOLD response ($h^2 \sim 0.2$) for face and house processing (the specific contrasts were not analyzed), but negligible heritability for pseudowords and chairs in the ventral visual cortex, based on an fMRI experiment involving 13 MZ and 11 DZ twins, with 90 trials per condition. Similarly, Matthews et al. (2007) revealed a moderate heritability ($h^2 = 0.37$; 90% interval: (0, 0.74)) for the interference effect in the dorsal ACC during a multi-source interference task with congruent and incongruent conditions, involving 20 MZ and 20 DZ twins, with 144 trials per condition. The heritability estimates for other regions, such as ventral ACC, posterior cingulate, and left and right insula, were negligible. In contrast, the heritability of reaction time was moderate for the congruent condition ($h^2 = 0.45$; 90% interval: (0, 0.76)), but negligible for the incongruent condition and the interference effect. Additionally, Blokland et al. (2011) found moderate to high heritability estimates ($h^2 = 0.40$ to 0.65) in more than ten regions during an n-back working memory experiment involving 150 MZ and 132 DZ twins, with 128 trials per condition.

Is intra-individual variability a concern for heritability estimation in neuroimaging? The aforementioned task-based fMRI experiments have primarily relied on a large number of trials to obtain reliable estimates

of condition-level effects. This is similar to the Stroop experimental dataset we investigated here, with the distinction that the focus is on BOLD response rather than reaction time. However, the family sample size has often been relatively small, leading to larger uncertainty ranges in the estimates. This issue is particularly pronounced because the relative intra-individual variability across the brain, as reported in the literature, tends to be substantial, with $R_v \gg 1$ (Chen et al., 2021; Baker et al., 2021; Chen et al., 2022). Thus, heritability estimation in neuroimaging is at least as challenging as, if not more so than, typical traits such as psychometric data.

5.4 Limitations of heritability estimation through hierarchical modeling

The hierarchical modeling approach comes with additional costs. Firstly, introducing an extra level in the data hierarchy significantly increases the complexity of the model structure. Secondly, and perhaps the greater challenge, this increased complexity brings along numerical burdens. Traditional tools like linear mixed-effects estimation are not well-suited for solving hierarchical models of heritability. Instead, resorting to a Bayesian approach may be necessary to handle the numerical challenges (e.g., singularity).

There is always room for improvement in modeling. For example, the full details of the underlying mechanism and framework of cognitive inhibition involved in the reaction time of the Stroop effect are not fully known to researchers. Therefore, no model can fully replicate their structure. However, HME attempts to model as much as is known and observed in a study. Model fitting can be improved by incorporating auxiliary information, such as accommodating abnormalities like skewness, outliers, and truncation through more inclusive and adaptive distributions (e.g., log-normal, ex-Gaussian). Additionally, one could reconsider the chosen partitioning into three components of h^2, c^2, and e^2 in twin studies and other assumptions (Robette et al., 2022): the additivity of genetic effects, the absence of assortative mating, the nonexistence of genetic dominance or epistasis, the generalizability from twins to the rest of the population, equal environment impact between MZ and DZ twins, and the absence of gene-environment correlation or interaction.

The interpretation of heritability is subtle and sometimes controversial. Our focus here is solely on the technical aspects of heritability estimation. Nevertheless, we emphasize caution in its interpretation. As a statistical metric, heritability captures variation and/or correlation rather than causation. Therefore, one must not confuse the extent of phenotype variability with the contribution of genetic factors. The concept of heritability effectively pertains to the population level and cannot be realistically applied to a particular individual. On the other hand, the information provided by heritability lies in its potential predictability. It can probabilistically predict, but not causally determine, the extent of phenotypic variability. Similar to the predictability of an explanatory variable in a regression model, a high heritability for a phenotypic trait may warrant further investigation into the underlying complex genetic mechanisms or the etiology of genetic risk factors, such as biomarkers. This perspective highlights the need to complement heritability research of variance partitioning with mechanism elucidation (Downes and Turkheimer, 2022).

6 Conclusions

We propose a hierarchical modeling approach to improve heritability estimation in twin studies when the phenotypic trait is measured with multiple samples. The methodology aims to separate measurement errors from the variations of interest within the heritability formulation and addresses issues such as information loss due to data reduction, distribution violations, and uncertainty characterization in current modeling approaches. Heritability reflects the similarity between twins and is influenced by inter-individual variability, while measurement error defines trial-level variability, which includes intra-individual variability. These sources of variability exist at different levels of the data hierarchy, and combining them is conceptually inconsistent and introduces bias to the heritability estimation.

We demonstrated that the conventional ACE model is likely to underestimate heritability when intra-individual variability is moderate to high (which is common in many real-world scenarios). We supported this finding with simulations and an experimental dataset from the ABCD study, validating the performance of the hierarchical modeling approach. Our simulation results suggest that traits with small effect sizes, which are often the focus of studies, may require much larger sample sizes than currently practiced. We continue to recognize the intriguing role of intra-individual variability in studying individual differences and experimental design, as characterized by the variability ratio R_v. We also hope that understanding the magnitude of intra-individual variability, in contrast to inter-individual variability, will provide further insights from cognitive and neuronal perspectives in the future.

Acknowledgments

GC and PT were supported by the NIMH Intramural Research Program (ZICMH002888) of the NIH/HHS, USA. DM was also supported by the NIMH IRP (ZICMH002960) of the NIH/HHS, USA. Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA).

Appendices

A Simulation specifications

A.1 One condition with trials

Simulations were conducted with the following specifications under the hierarchical formulation (17) compared to the conventional ACE model (1). All distributions were assumed to be Gaussian, and the data were first aggregated across trials when applied to the ACE model. The overall effect at the population level was fixed at $\alpha = 0$. The sum of inter-individual and inter-family variance $(\sigma_z^2 + \tau_z^2)$ was fixed at 1 for MZ and DZ twins without loss of generality. Four factors were varied to create the set of simulations, with the following value ranges:

(a) Three family sample sizes: $F = 100, 500, 1000$, with half being MZ or DZ zygosities.
(b) Three trial sample sizes: $T = 50, 100, 500$.
(c) Three ratios of $R_v = 1, 5, 10$ (and we note that in each case $\omega_0^2 = R_v$, due to the inter-individual and -family variance scaling).
(d) Two modeling approaches: ACE and HME.

Each of these $3 \times 3 \times 3 \times 2 = 54$ combinations was simulated with 200 repetitions. For ACE, the function `lme` in the `R` package `nlme` (Pinheiro et al., 2022) was used to estimate variances. For HME, the function `lmer` in the `R` package `lme4` (Bates et al., 2015) was adopted with the following iterative steps:

1) For each twin pair (i_1, i_2) within family f of zygosity z, individual-level effects were obtained by randomly sampling from the following bivariate distribution:

$$\begin{bmatrix} \theta_{i_1(f(z))} \\ \theta_{i_2(f(z))} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & r_z \\ r_z & 1 \end{bmatrix} \right).$$

Here, we chose $r_{MZ} = 0.6$ and $r_{DZ} = 0.35$ so that the targeted values of h^2, c^2, and e^2 would be 0.5, 0.1, and 0.4, respectively, as per Falconer's formula (10).
2) Randomly draw trial-level data $y_{i(f(z))}$ from the distribution $\mathcal{N}(\alpha + \theta_{i(f(z))},\ \sigma_0^2)$, where $\omega_0^2 = R_v^2$.
3) Estimate σ_z^2 and τ_z^2. For ACE, apply the model (15) using the function `lme` from the `R` package `nlme`. For HME, apply the hierarchical model (17) using the function `lmer` from the `R` package `lme4`.
4) Obtain r_z using the formula (12). Additionally, make two adjustments using formulas (22) and (23).
5) Estimate h^2, c^2, and e^2 according to Falconer's formula (10).
6) Estimate bias adjustment for ACE: apply (22) to estimate ACE1 and (23) to estimate ACE2.

A.2 Two conditions with trials

Simulations for the two conditions were largely similar to the case with one condition in the previous subsection. All distributions involved were assumed to be Gaussian, and the data were aggregated across trials for each condition when applied to the ACE model. The two overall effects under the hierarchical (25) were fixed at $\alpha_c = 0$ ($c = c_1, c_2$). Since the standard deviation is essentially a scaling factor, the sum of inter-individual and inter-family variances was fixed as $\sigma^2_{cg} + \tau^2_{cg} = 1$ ($c = c_1, c_2$; z = MZ, DZ) without loss of generality. Four factors were changed to create the set of simulations across the following ranges of values:

(a) Three family sample sizes $F = 100, 500, 1000$, half of which were MZ (and DZ) zygosities.
(b) Three trial sample sizes $T = 50, 100, 500$.
(c) Three ratios of $R_v = 1, 4, 10$, leading to $\sigma^2_{cg} = R_v$.
(d) Two modeling approaches: ACE and HME.

Each of these $3 \times 3 \times 3 \times 2 = 54$ combinations was simulated with 200 repetitions. For ACE, the function `lme` in the `R` package `nlme` (Pinheiro et al., 2022) was used to estimate variances. For HME, the function `lmer` in the `R` package `lme4` (Bates et al., 2015) was adopted with the following iterative steps. The condition contrast was parameterized through the indicator variable defined in (27).

1) For each twin pair (i_1, i_2) within family f of zygosity z, obtain their effects at the two conditions $c1$ and c_2 through randomly sampling from the following quadrivariate distribution:

$$\begin{bmatrix} \theta_{c_1 i_1(f(z))} \\ \theta_{c_1 i_2(f(z))} \\ \theta_{c_2 i_1(f(z))} \\ \theta_{c_2 i_2(f(z))} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \gamma_1 & \gamma_2 & \rho_z \\ \gamma_1 & 1 & \rho_z & \gamma_2 \\ \gamma_2 & \rho_z & 1 & \gamma_1 \\ \rho_z & \gamma_2 & \gamma_1 & 1 \end{bmatrix} \right).$$

Here, γ_1 and γ_2 were fixed at 0.8 and 0.5, respectively, and ρz was chosen as 0.375 when z = MZ and 0.5 when z = DZ, so that the simulated correlation r_z between a twin pair within a family regarding the condition contrast would be 0.85 for z = MZ and 0.45 for z = DZ, respectively. Thus, the targeted heritability h^2 would be 0.5 per Falconer's formula (10), while c^2 and e^2 would be 0.35 and 0.15.
2) Randomly draw trial-level data $y_{ci(f(z))}$ from $\mathcal{N}(\alpha_c + \theta_{ci(f(z))},\ \sigma^2_0)$ ($c = c_1, c_2$), where $\sigma^2_0 = R^2_v$.
3) Estimate σ^2_z and τ^2_z. For ACE, obtain the average across trials for each condition and for each individual, and then apply the individual-level contrast between the conditions to the formulation (11) using the function `lme` from the `R` package `nlme`. For HME, apply the model (25) using the function `lmer` from the `R` package `lme4`.
4) Obtain r_z using the formula (26). In addition, make two adjustments using formulas (22) and (23) with $1/T$ replaced by $2/T$.
5) Estimate h^2, c^2, and e^2 using Falconer's relationship (10).
6) Estimate bias adjustments for the ACE model: apply (23) with an extra factor of 2 to estimate ACE1, and (22) with an extra factor of 2 to estimate ACE2.

References

Anokhin, A.P., Golosheykin, S., Grant, J.D., Heath, A.C., 2017. Heritability of brain activity related to response inhibition: A longitudinal genetic study in adolescent twins. International Journal of Psychophysiology 115, 112–124.

Arbet, J., McGue, M., Basu, S., 2020. A robust and unified framework for estimating heritability in twin studies using generalized estimating equations. Statistics in Medicine 39, 3897–3913.

Baker, D.H., Vilidaite, G., Lygo, F.A., Smith, A.K., Flack, T.R., Gouws, A.D., Andrews, T.J., 2021. Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. Psychological Methods 26, 295–314.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67, 1–48.

Bates, T.C., Maes, H., Neale, M.C., 2019. Umx: Twin and Path-Based Structural Equation Modeling in R. Twin Research and Human Genetics 22, 27–41.

Blokland, G.A.M., McMahon, K.L., Thompson, P.M., Martin, N.G., de Zubicaray, G.I., Wright, M.J., 2011. Heritability of Working Memory Brain Activation. The Journal of Neuroscience 31, 10882–10890.

Bürkner, P.C., 2017. Brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software 80, 1–28.

Chen, G., Pine, D.S., Brotman, M.A., Smith, A.R., Cox, R.W., Haller, S.P., 2021. Trial and error: A hierarchical modeling approach to test-retest reliability. NeuroImage 245, 118647.

Chen, G., Pine, D.S., Brotman, M.A., Smith, A.R., Cox, R.W., Taylor, P.A., Haller, S.P., 2022. Hyperbolic trade-off: The importance of balancing trial and subject sample sizes in neuroimaging. NeuroImage 247, 118786.

Chen, G., Taylor, P.A., Haller, S.P., Kircanski, K., Stoddard, J., Pine, D.S., Leibenluft, E., Brotman, M.A., Cox, R.W., 2018. Intraclass correlation: Improved modeling approaches and applications for neuroimaging. Human Brain Mapping 39, 1187–1206.

Chen, X., Formisano, E., Blokland, G.A.M., Strike, L.T., McMahon, K.L., de Zubicaray, G.I., Thompson, P.M., Wright, M.J., Winkler, A.M., Ge, T., Nichols, T.E., 2019. Accelerated estimation and permutation inference for ACE modeling. Human Brain Mapping 40, 3488–3507.

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., Liu, J., 2013. A Nondegenerate Penalized Likelihood Estimator for Variance Parameters in Multilevel Models. Psychometrika 78, 685–709.

Downes, S.M., Matthews, L., 2020. Heritability, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. spring 2020 ed.. Metaphysics Research Lab, Stanford University.

Downes, S.M., Turkheimer, E., 2022. An Early History of the Heritability Coefficient Applied to Humans (1918–1960). Biological Theory 17, 126–137.

Elliott, M.L., Knodt, A.R., Ireland, D., Morris, M.L., Poulton, R., Ramrakha, S., Sison, M.L., Moffitt, T.E., Caspi, A., Hariri, A.R., 2020. What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. Psychological Science 31, 792–806.

Falconer, D.S.D.S., MacKay, T.F.C., 1996. Introduction to Quantitative Genetics. Harlow : Prentice Hall.

Fan, J., Wu, Y., Fossella, J.A., Posner, M.I., 2001. Assessing the heritability of attentional networks. BMC Neuroscience 2, 14.

Germine, L., Russell, R., Bronstad, P.M., Blokland, G.A., Smoller, J.W., Kwok, H., Anthony, S.E., Nakayama, K., Rhodes, G., Wilmer, J.B., 2015. Individual aesthetic preferences for faces are shaped mostly by environments, not genes. Current biology : CB 25, 2684–2689.

Giangrande, E.J., Weber, R.S., Turkheimer, E., 2022. What Do We Know About the Genetic Architecture of Psychopathology? Annual Review of Clinical Psychology 18, 19–42.

Guo, G., Wang, J., 2002. The Mixed or Multilevel Model for Behavior Genetic Analysis. Behavior Genetics 32, 37–49.

Gustavson, D.E., Nayak, S., Coleman, P.L., Iversen, J.R., Lense, M.D., Gordon, R.L., Maes, H.H., 2023. Heritability of Childhood Music Engagement and Associations with Language and Executive Function: Insights from the Adolescent Brain Cognitive Development (ABCD) Study. Behavior Genetics .

Haines, N., Kvam, P.D., Irving, L.H., Smith, C., Beauchaine, T.P., Pitt, M.A., Ahn, W.Y., Turner, B., 2020. Learning from the Reliability Paradox: How Theoretically Informed Generative Models Can Advance the Social, Behavioral, and Brain Sciences. Preprint. PsyArXiv. doi:10.31234/osf.io/xr7y3.

Harper, J., Malone, S.M., Iacono, W.G., 2019. Target-related parietal P3 and medial frontal theta index the genetic risk for problematic substance use. Psychophysiology 56, e13383.

Hedge, C., Powell, G., Sumner, P., 2018. The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. Behavior Research Methods 50, 1166–1186.

Holst, K.K., Scheike, T.H., Hjelmborg, J.B., 2016. The liability threshold model for censored twin data. Computational Statistics & Data Analysis 93, 324–335.

Hung, I.T., Ganiban, J.M., Saudino, K.J., 2023. Using the Flanker Task to Examine Genetic and Environmental Contributions in Inhibitory Control Across the Preschool Period. Behavior Genetics 53, 132–142.

Hunter, M.D., 2021. Multilevel Modeling in Classical Twin and Modern Molecular Behavior Genetics. Behavior Genetics 51, 301–318.

Iacono, W.G., Heath, A.C., Hewitt, J.K., Neale, M.C., Banich, M.T., Luciana, M.M., Madden, P.A., Barch, D.M., Bjork, J.M., 2018. The utility of twins in developmental cognitive neuroscience research: How twins strengthen the ABCD research design. Developmental Cognitive Neuroscience 32, 30–42.

Kastrati, G., Rosén, J., Thompson, W.H., Chen, X., Larsson, H., Nichols, T.E., Tracey, I., Fransson, P., Åhs, F., Jensen, K.B., 2022. Genetic Influence on Nociceptive Processing in the Human Brain—A Twin Study. Cerebral Cortex 32, 266–274.

Maes, H.H., 2005. ACE Model, in: Encyclopedia of Statistics in Behavioral Science. John Wiley & Sons, Ltd.

Martin, N.G., Eaves, L.J., Kearsey, M.J., Davies, P., 1978. The power of the classical twin study. Heredity 40, 97–116.

Matthews, S.C., Simmons, A.N., Strigo, I., Jang, K., Stein, M.B., Paulus, M.P., 2007. Heritability of anterior cingulate response to conflict: An fMRI study in female twins. NeuroImage 38, 223–227.

Morey, R.D., Hoekstra, R., Rouder, J.N., Lee, M.D., Wagenmakers, E.J., 2016. The fallacy of placing confidence in confidence intervals. Psychonomic Bulletin & Review 23, 103–123.

Neale, M.C., Hunter, M.D., Pritikin, J.N., Zahery, M., Brick, T.R., Kirkpatrick, R.M., Estabrook, R., Bates, T.C., Maes, H.H., Boker, S.M., 2016. OpenMx 2.0: Extended Structural Equation and Statistical Modeling. Psychometrika 81, 535–549.

Pingault, J.B., Allegrini, A.G., Odigie, T., Frach, L., Baldwin, J.R., Rijsdijk, F., Dudbridge, F., 2022. Research Review: How to interpret associations between polygenic scores, environmental risks, and phenotypes. Journal of Child Psychology and Psychiatry 63, 1125–1139.

Pinheiro, J., Bates, D., R Core Team, 2022. nlme: Linear and Nonlinear Mixed Effects Models. URL: `https://CRAN.R-project.org/package=nlme`. r package version 3.1-159.

Polderman, T.J.C., Benyamin, B., de Leeuw, C.A., Sullivan, P.F., van Bochoven, A., Visscher, P.M., Posthuma, D., 2015. Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nature Genetics 47, 702–709.

Polk, T.A., Park, J., Smith, M.R., Park, D.C., 2007. Nature versus Nurture in Ventral Visual Cortex: A Functional Magnetic Resonance Imaging Study of Twins. Journal of Neuroscience 27, 13921–13925.

Rea-Sandin, G., Clifford, S., Doane, L.D., Davis, M.C., Grimm, K.J., Russell, M.T., Lemery-Chalfant, K., 2023. Genetic and environmental links between executive functioning and effortful control in middle childhood. Journal of Experimental Psychology: General , No Pagination Specified–No Pagination Specified.

Richardson, K., Norgate, S., 2005. The equal environments assumption of classical twin studies may not hold. British Journal of Educational Psychology 75, 339–350.

Rijsdijk, F.V., Sham, P.C., 2002. Analytic approaches to twin data using structural equation models. Briefings in Bioinformatics 3, 119–133.

Robette, N., Génin, E., Clerget-Darpoux, F., 2022. Heritability: What's the point? What is it not for? A human genetics perspective. Genetica .

Rouder, J., Kumar, A., Haaf, J.M., 2019. Why Most Studies of Individual Differences With Inhibition Tasks Are Bound To Fail. Technical Report. PsyArXiv. doi:`10.31234/osf.io/3cjr5`.

Rouder, J.N., Haaf, J.M., 2019. A psychometrics of individual differences in experimental tasks. Psychonomic Bulletin & Review 26, 452–467.

Routledge, K.M., Williams, L.M., Harris, A.W.F., Schofield, P.R., Clark, C.R., Gatt, J.M., 2018. Genetic correlations between wellbeing, depression and anxiety symptoms and behavioral responses to the emotional faces task in healthy twins. Psychiatry Research 264, 385–393.

Schachar, R.J., Forget-Dubois, N., Dionne, G., Boivin, M., Robaey, P., 2010. Heritability of Response Inhibition in Children. Journal of the International Neuropsychological Society 17, 238–247.

Schwabe, I., Gu, Z., Tijmstra, J., Hatemi, P., Pohl, S., 2019. Psychometric Modelling of Longitudinal Genetically Informative Twin Data. Frontiers in Genetics 10, 837.

Sham, P.C., Purcell, S.M., Cherny, S.S., Neale, M.C., Neale, B.M., 2020. Statistical Power and the Classical Twin Design. Twin Research and Human Genetics 23, 87–89.

Smolker, H.R., Wang, K., Luciana, M., Bjork, J.M., Gonzalez, R., Barch, D.M., McGlade, E.C., Kaiser, R.H., Friedman, N.P., Hewitt, J.K., Banich, M.T., 2022. The Emotional Word-Emotional Face Stroop task in the ABCD study: Psychometric validation and associations with measures of cognition and psychopathology. Developmental Cognitive Neuroscience 53, 101054.

Spearman, C., 1904. The Proof and Measurement of Association between Two Things. The American Journal of Psychology 15, 72–101.

Stins, J.F., de Sonneville, L.M.J., Groot, A.S., Polderman, T.C., van Baal, C.G.C.M., Boomsma, D.I., 2005. Heritability of Selective Attention and Working Memory in Preschoolers. Behavior Genetics 35, 407–416.

Turkheimer, E., 2000. Three Laws of Behavior Genetics and What They Mean. Current Directions in Psychological Science 9, 160–164.

van den Berg, S.M., Glas, C.A.W., Boomsma, D.I., 2007. Variance Decomposition Using an IRT Measurement Model. Behavior Genetics 37, 604–616.

van der Meulen, M., Steinbeis, N., Achterberg, M., van IJzendoorn, M.H., Crone, E.A., 2018. Heritability of neural reactions to social exclusion and prosocial compensation in middle childhood. Developmental Cognitive Neuroscience 34, 42–52.

van Drunen, L., Dobbelaar, S., van der Cruijsen, R., van der Meulen, M., Achterberg, M., Wierenga, L.M., Crone, E.A., 2021. The nature of the self: Neural analyses and heritability estimates of self-evaluations in middle childhood. Human Brain Mapping 42, 5609–5625.

Vellani, V., Garrett, N., Gaule, A., Patil, K.R., Sharot, T., 2022. Quantifying the heritability of belief formation. Scientific Reports 12, 11833.

Viktorsson, C., Lindskog, M., Li, D., Tammimies, K., Taylor, M.J., Ronald, A., Falck-Ytter, T., 2022. Infants' sense of approximate numerosity: Heritability and link to other concurrent traits. Developmental Science n/a, e13347.

Wang, L., Wang, Y., Xu, Q., Liu, D., Ji, H., Yu, Y., Hu, Z., Yuan, P., Jiang, Y., 2020. Heritability of reflexive social attention triggered by eye gaze and walking direction: Common and unique genetic underpinnings. Psychological Medicine 50, 475–483.

Willoughby, E.A., Polderman, T.J.C., Boutwell, B.B., 2023. Behavioural genetics methods. Nature Reviews Methods Primers 3, 1–16.

Yeom, D., Tan, Y.T., Haslam, N., Mosing, M.A., Yap, V.M.Z., Fraser, T., Hildebrand, M.S., Berkovic, S.F., McPherson, G.E., Peretz, I., Wilson, S.J., 2022. Genetic factors and shared environment contribute equally to objective singing ability. iScience 25, 104360.