

Are Extended Twin Family Designs Worth the Trouble? A Comparison of the Bias, Precision, and Accuracy of Parameters Estimated in Four Twin Family Models

Matthew C. Keller · Sarah E. Medland ·
Laramie E. Duncan

Received: 23 June 2009 / Accepted: 24 November 2009
© Springer Science+Business Media, LLC 2009

Abstract The classical twin design (CTD) uses observed covariances from monozygotic and dizygotic twin pairs to infer the relative magnitudes of genetic and environmental causes of phenotypic variation. Despite its wide use, it is well known that the CTD can produce biased estimates if its stringent assumptions are not met. By modeling observed covariances of twins' relatives in addition to twins themselves, extended twin family designs (ETFDS) require less stringent assumptions, can estimate many more parameters of interest, and should produce less biased estimates than the CTD. However, ETFDS are more complicated to use and interpret, and by attempting to estimate a large number of parameters, the precision of parameter estimates may suffer. This paper is a formal investigation into a simple question: Is it worthwhile to use more complex models such as ETFDS in behavioral genetics? In particular, we compare the bias, precision, and accuracy of estimates from the CTD and three increasingly complex ETFDS. We find the CTD

does a decent job of estimating broad sense heritability, but CTD estimates of shared environmental effects and the relative importance of additive versus non-additive genetic variance can be biased, sometimes wildly so. Increasingly complex ETFDS, on the other hand, are more accurate and less sensitive to assumptions than simpler models. We conclude that researchers interested in characterizing the environment or the makeup of genetic variation should use ETFDS when possible.

Keywords Behavior genetics · Model misspecification · Extended twin family design · Classical twin design · Parameter indeterminacy

Introduction

The observed covariances of twins, adoptees, and their family members are often used to understand the relative importance of genetic and environmental causes of phenotypic variation. The most commonly used genetically informative design is the Classical Twin Design (CTD), which compares the monozygotic (MZ) twin covariance to the dizygotic (DZ) twin covariance to estimate the variation in a trait due to unique environmental effects (V_E) as well as any two of the three variance components—additive genetic (V_A), dominance genetic (V_D), and common environmental (V_C)—that cause familial similarity.

There are several appeals to the CTD. For example, MZ and DZ twins serve as natural controls to one another, their data is relatively simple to collect, and shared environmental effects are not confounded with genetic effects, as they are in non-twin familial studies (Martin et al. 1997). Nevertheless, it has long been understood that the CTD suffers from several important limitations (Eaves et al.

Edited by Gitta Lubke.

M. C. Keller · L. E. Duncan
Department of Psychology and Neuroscience,
University of Colorado, Boulder, CO, USA

M. C. Keller · L. E. Duncan
Institute for Behavioral Genetics, University of Colorado,
Boulder, CO, USA

S. E. Medland
Queensland Institute for Medical Research, Brisbane,
QLD, Australia

M. C. Keller (✉)
Department of Psychology and Neuroscience,
Muenzinger Hall, 345 UCB, Boulder, CO 80309, USA
e-mail: matthew.c.keller@gmail.com
URL: www.matthewckeller.com

1978). For one, \hat{V}_A , \hat{V}_D , and \hat{V}_C are mutually confounded in the CTD, allowing only two of these three parameters to be estimated.¹ This follows from the fact that it is impossible to simultaneously estimate three parameters (\hat{V}_A , \hat{V}_D , and \hat{V}_C) from just two pieces of relevant information (the MZ and DZ covariances). To circumvent this under-identification problem, behavioral geneticists using the CTD routinely assume that either $V_D = 0$ or that $V_C = 0$. However, these are simply assumptions, untestable using twins alone, born from the mathematical necessity of making the CTD identified. To the degree these assumptions are violated, CTD estimates of V_A tend to be biased upward and estimates of V_D and V_C tend to be biased downward (Grayson 1989; Heath et al. 1985; Keller and Coventry 2005). Second, the CTD does not model the effects of assortative mating or gene-environment covariance, the presence of which will create biases in estimates (e.g., \hat{V}_A will be too low). Third, the CTD has nothing to say about the etiology of the shared environmental effects (contributing to V_C): to what degree are they passed culturally from parent to offspring and to what degree are they due to non-parental factors such as peer influences? Finally, the CTD does not use information efficiently: for every twin pair recruited (two new subjects), only a single additional bit of information (one covariance estimate) is gained useful to modeling the causes of familial similarity.

For these and other reasons, in the 1970s researchers began exploring extended twin family designs (ETFDS), which require less stringent assumptions and produce less biased estimates than the CTD (Fulker 1982). These alternative designs use data on parents of twins (Eaves et al. 1978; Neale and Fulker 1984) and offspring of twins (Nance and Corey 1976) to better reveal genetic non-additivity and the role of parental environmental effects, and use parents of twins and spouses of twins (Eaves 1979) to model the effects of assortative mating. Cloninger et al. (1979) first described how to use all of these relative types together in a single model. Their model is the forerunner to the three ETFDS described in this paper: the Nuclear Twin Family Design (NTFD) (Heath et al. 1985), the *Stealth* design (Truett et al. 1994), and the *Cascade* design (Keller et al. 2009). For a more thorough history of twin and family designs, see Eaves (2009).

Extended twin family designs (ETFDS) address the limitations of the CTD described above. Compared to the CTD, ETFDS allow for finer grained descriptions of the causes of phenotypic variation, they produce less biased parameter estimates, and more information (increasing statistical power) is gained per additional subject in ETFDS (Posthuma and Boomsma 2000). Yet, the reduction in bias and more

detailed information associated with ETFDS comes at the cost of greatly increased complexity. This complexity is a major problem for instantiating the model into code. For example, such scripts written in Mx (Neale 1999) can stretch for 50 printed pages or more, making human errors a virtual certainty regardless of how vigilant the error checking is. We note, however, that a new version of Mx, OpenMx (<http://openmx.psyc.virginia.edu/>), will be available as a package for the R statistical language in early 2010, and changes in the OpenMx syntax should significantly simplify ETFD code. Nevertheless, the complexity of ETFDS may also obscure logical errors at the heart of the designs; certain expectations may simply have been wrong at the modeling stage. Furthermore, as with all models, ETFDS also must make assumptions in order for their models to be identified, and it is possible that they may perform as bad or worse than simpler models when these assumptions are violated. Finally, the complexity of ETFD models and the number of parameters they attempt to estimate may lead to an unacceptable level of imprecision in estimates caused by the high covariation between the large numbers of estimated parameters (multicollinearity problems). For these reasons, some researchers in behavioral genetics remain skeptical of the value of ETFDS and favor the use of simpler, time-tested models such as the CTD, which are easy to use and interpret and require less data collection.

The goal of this paper is to explore these trade-offs. In particular, we use simulations to gauge the bias, precision, and accuracy of parameters estimated using the CTD and three ETFDS in order to understand whether they work as intended, under what circumstances their estimates are biased, if the increase in information in ETFDS comes at an unacceptable cost in precision, and how violations of assumptions affect parameter estimates. In addition to identifying the central tendency of the parameter estimates, we also explore their spread, covariation, and distributional shapes. Such results can help researchers interpret CTD and ETFD findings with proper circumspection. In summary, this paper is a formal investigation into a simple question: Is it worthwhile to use more complex models such as ETFDS in behavioral genetics?

Method

General strategy

We seek four properties—the bias, precision, accuracy, and distributions—of parameter estimates derived from the CTD, NTFD, *Stealth*, and *Cascade* designs. Of course, the parameter bias, precision, accuracy, and distributions for a given design change depending on the scenario, so we need to measure these properties under several scenarios that

¹ We follow the convention that \hat{V} is the estimate of the population parameter V .

might occur in nature. A given scenario, for example, might simulate specified levels of additive genetic, dominant genetic, and common environmental effects on some hypothetical trait. These scenarios should also violate assumptions of the four designs to check their sensitivities to assumptions. To accomplish these goals, the first author created a program, *GeneEvolve*, that simulates twin family data. The user supplies input for various parameters (e.g., the amount of variation in a phenotype due to various types of genetic and environmental effects) to simulate different scenarios. We obtained simulated twin family data from *GeneEvolve* under several different scenarios that might occur in real life and ran Mx models from the four designs above on this data. We then compared the estimated variance parameters (denoted by \hat{V} .) derived from Mx, to the true variance parameters (denoted by V .) simulated using *GeneEvolve*. We iterated this process 500 times for each of 10 different scenarios. In total, 20,000 Mx models were fit (500 iterations \times 4 models per iteration \times 10 scenarios), taking a total of \sim 13,000 h of CPU time.

Description of the three extended twin family designs (ETFDs)

Table 1 gives the interpretations of the variance parameters discussed in this paper as well as which designs can estimate which variance parameters. For a description of the CTD, see Plomin et al. (2001), and for a more detailed description and explanation of these three ETFDs, including algebraic expectations, see Keller et al. (2009).

Nuclear twin family design (NTFD)

The NTFD (Fig. 1) uses data on MZ twins, DZ twins, and their parents. These three relative classes provide four pieces of information from which parameters are estimated: the covariance between MZ twins, $C\hat{V}(\text{MZ}, \text{MZ})$, the covariance between DZ twins, $C\hat{V}(\text{DZ}, \text{DZ})$, the covariance between parents, $C\hat{V}(\text{spouse})$, and the covariance between parents and children, $C\hat{V}(\text{Par}, \text{Child})$. This additional information allows the NTFD to estimate \hat{V}_A , \hat{V}_D , and \hat{V}_C simultaneously, allows the effects of assortative mating on parameter estimates to be accounted for, and allows passive gene-environment covariance to be differentiated from the effects of \hat{V}_C . While there are many ways the NTFD can be parameterized, we focus here on a parameterization (Fig. 1) that divides \hat{V}_C into the variance of effects shared between siblings and twins but not parents (\hat{V}_S) and the variance of effects that are transmitted via vertical transmission from parents to offspring (\hat{V}_F). Because only three pieces of data, $C\hat{V}(\text{MZ}, \text{MZ})$, $C\hat{V}(\text{DZ}, \text{DZ})$, and $C\hat{V}(\text{Par}, \text{Child})$, provide information on four parameters (\hat{V}_A , \hat{V}_D , \hat{V}_S , and \hat{V}_F), one of these

parameters (typically \hat{V}_F or \hat{V}_S) must be set to 0 in any NTFD model. Latent variances that are not shown in Fig. 1, 2, and 3 are equal to 1.

Stealth design

By using data from MZ and DZ twins and their siblings, parents, offspring, and spouses, 88 sex-specific relative covariances can be estimated. Many of these 88 relative classes are identical except for sex-specific pathways. For example, nephew-aunt covariances between sons of DZ females and their female DZ co-twins are differentiated from nephew-aunt covariances that are between sons of DZ males and female DZ co-twins. The *Stealth* uses these 88 covariance observations to simultaneously estimate sex-specific \hat{V}_A , \hat{V}_D , \hat{V}_S , \hat{V}_F , \hat{V}_T , and \hat{V}_E (see Table 1 for their interpretations) as well as additive genetic variation unique to males/females, the effects of assortative mating, and A–F covariance. The *Stealth* model used in this paper is simplified by excluding sex effects, reducing the number of relative classes from 88 to 17. The path diagram for this *Stealth* model is shown in Fig. 2, and is identical to Fig. 1 except that spouses of twins and children of twins have been added. To keep the diagram uncluttered, siblings of twins are not shown.

Cascade design

Like the *Stealth*, the *Cascade* uses information on twins and their siblings, parents, spouses, and children to model all of the variance components modeled by the *Stealth*. However, a limitation of the *Stealth* is that it models only one type of mating (primary phenotypic mating) and only one type of vertical transmission (from parental phenotype to offspring F). The purpose of the *Cascade* is to provide a general framework for relaxing the assumptions regarding mate choice and vertical transmission made by the *Stealth*. This is done through the use of latent phenotypes upon which spouses mate or upon which parents influence their children. To keep the number of model comparisons manageable, we focus here on the mating aspect of the *Cascade* rather than the vertical transmission aspects of it. The only difference between Fig. 2 (the *Stealth* model) and Fig. 3 (the *Cascade* model) is the addition of the latent phenotype (\tilde{P}) upon which mates assort. Depending on the type of mating or vertical transmission model being used, the path coefficients to \tilde{P} are set to either be equal to the path coefficients to P or to be equal to zero. For example, to model social homogamy, all genetic path coefficients to \tilde{P} are set to zero ($\tilde{a} = 0$ and $\tilde{d} = 0$) and all environmental path coefficients to \tilde{P} are constrained to be equal to the values of the corresponding path coefficients to P ($\tilde{f} = f$, $\tilde{s} = s$, $\tilde{t} = t$, and $\tilde{e} = e$). To understand whether social

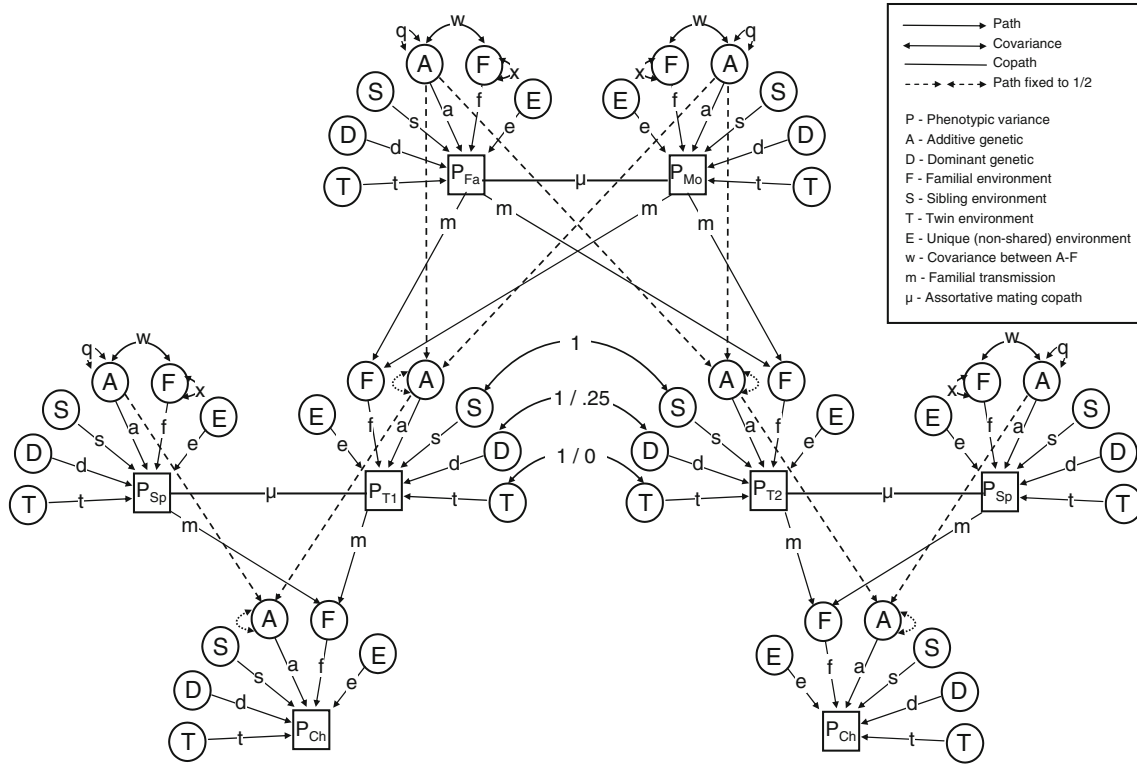


Fig. 2 Stealth path diagram

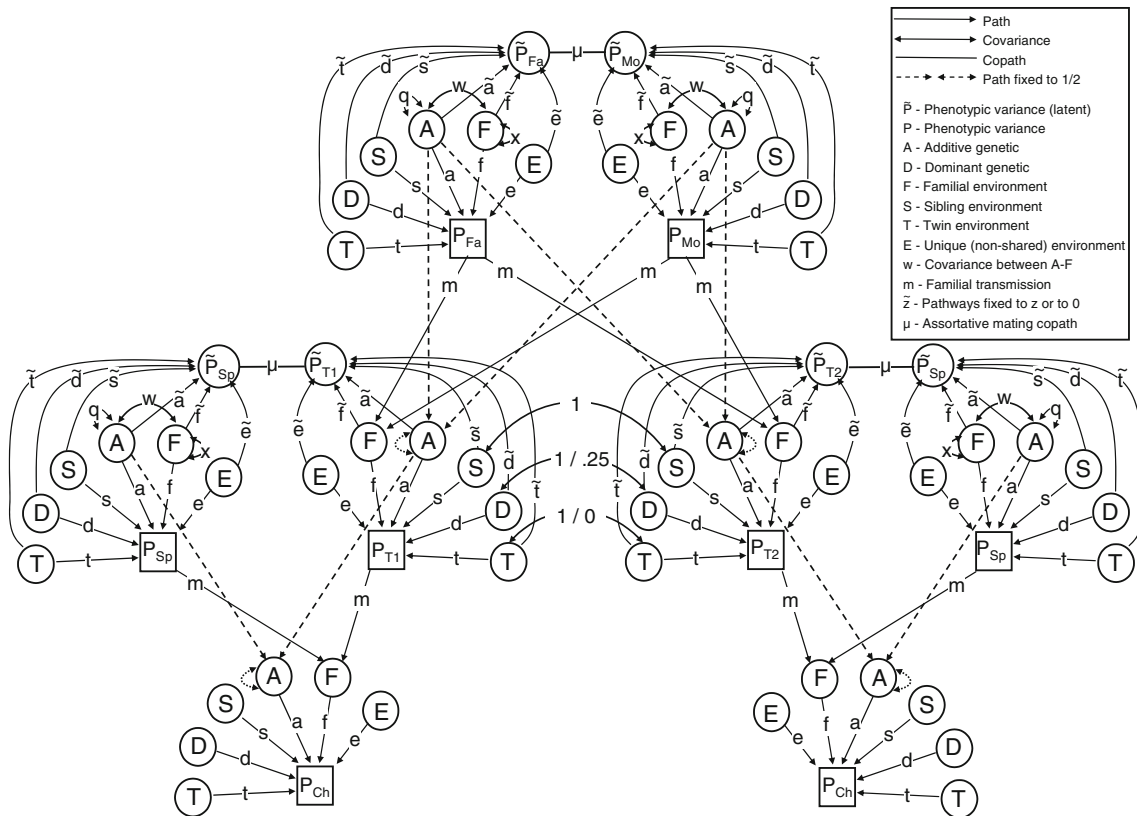


Fig. 3 Cascade path diagram

Core Team 2009) and available at www.matthewkeller.com. *GeneEvolve* accurately simulates genetically informative data as well as complex dynamics in evolutionary genetics. With complicated scenarios, it is difficult or impossible to find expected equilibrium parameter values analytically (e.g., the equilibrium additive-by-additive epistatic genetic variation in a population mating assortatively). Doing so through simulation, however, is straightforward. Given user input, *GeneEvolve* simulates the effects of alleles and environments on individuals' traits in a population, and allows this population to evolve (meet, mate, and have offspring, who meet, mate, and have offspring, etc....) for many generations, until parameters reach equilibrium. Currently, *GeneEvolve* allows user input of 48 different parameters, including 21 variance and covariance parameters, 3 different types of assortative mating, and 3 different types of vertical transmission.

GeneEvolve has an option to create twin and twin relative phenotypes during the final generation of the simulation. We used this option to write out the phenotypic scores of twins and their siblings, spouses, parents, and offspring to flat files (one row per family), which were then used as input into Mx (see below). Each flat file contained a total of ~15,000 families (6,500 MZ families and 8,500 DZ families). Although there were a total of 18 potential relative types in each family (two twins, two parents, four siblings, one spouse of twin 1, one spouse of twin 2, four children of twin 1, and four children of twin 2), families had an average of about five non-missing phenotypic scores and each flat file contained a total ~70,000 individuals. These numbers were chosen to reflect the sample sizes and missingness patterns in the combined Australia and Virginia extended twin databases (see Medland and Keller 2009), which is the largest extended twin family dataset in existence. Missingness in extended twin datasets arises through difficulties in ascertainment as well as variation in

age of death and number of children within families. Sample sizes of this magnitude are necessary for making fine-grained distinctions between parameters, especially with respect to sex-specific pathways (Heath et al. 1985; Medland and Keller 2009), although more modest datasets are adequate for differentiating models that do not require sex differentiated pathways.

Table 2 shows how each of the ten scenarios examined in this project was defined. V_E was set to .3 for each scenario, and all other variance parameters not shown in Table 2 were set to zero. The variance components inherited by offspring— V_A , V_F , $V_{A \times A}$, and $V_{A \times Age}$ —tend to increase across generations as a function of assortative mating and/or vertical transmission, and reach equilibrium values within 5–10 generations. We ran each *GeneEvolve* simulation for 20 generations to ensure that these parameters reached equilibrium. It can be difficult to predict the equilibrium values of these variance components at the beginning of a simulation. Our strategy was therefore to begin each *GeneEvolve* simulation such that all variance components summed to unity ($V_p = 1$) at the first generation, and to allow the variance components and V_p to increase to whatever their equilibrium values were. The equilibrium values for each variance component (from the 20th generation) are shown in Table 2; values in parentheses are the start values if different than the equilibrium values. Thus, the equilibrium variance components did not sum to unity for five of the models.

We simulated three different modes of assortative mating (see rows 5–8, Table 2). Phenotypic homogamy (also called “primary phenotypic assortment”) occurs when ‘like mates with like’ based on the manifest phenotype. For example, if tall people choose other tall people because they are tall, this would classify as phenotypic homogamy. This is the most commonly modeled type of assortative mating in the behavioral genetics and evolutionary genetics literatures.

Table 2 Simulated variance parameters associated with 10 different scenarios

| Scenario | V_A | V_D | V_F | V_S | V_T | $V_{A \times A}$ | $V_{A \times Age}$ | r (spouse) | A.M. type |
|-----------------------------|-----------|-------|-----------|-------|-------|------------------|--------------------|--------------|---------------------|
| ADE | .50 | .20 | 0 | 0 | 0 | 0 | 0 | 0 | N/A |
| ASE | .50 | 0 | 0 | .20 | 0 | 0 | 0 | 0 | N/A |
| ADSE | .40 | .15 | 0 | .15 | 0 | 0 | 0 | 0 | N/A |
| ADFE | .40 | .15 | .20 (.15) | 0 | 0 | 0 | 0 | 0 | N/A |
| ADFE + assortative mating | .49 (.40) | .15 | .36 (.15) | 0 | 0 | 0 | 0 | .30 | Phenotypic homogamy |
| ADFE + assortative mating | .42 (.40) | .15 | .27 (.15) | 0 | 0 | 0 | 0 | .30 | Social homogamy |
| ADFE + assortative mating | .51 (.40) | .15 | .34 (.15) | 0 | 0 | 0 | 0 | .30 | Genetic homogamy |
| ADFSTE + assortative mating | .35 (.30) | .10 | .18 (.10) | .10 | .10 | 0 | 0 | .30 | Phenotypic homogamy |
| ASE + A × A epistasis | .40 | 0 | 0 | .15 | 0 | .15 | 0 | 0 | N/A |
| ASE + A × Age Int. | .40 | 0 | 0 | .15 | 0 | 0 | .15 | 0 | N/A |

Numbers in parentheses are variance parameters at the first generation, which may change by the final (here 20th) generation if vertical transmission or assortative mating occurs (see text)

V_E .30 for all scenarios; A.M. assortative mating

Social homogamy refers to mate similarity arising from similar environmental backgrounds. For example, if people marry within religions and choice of religion is not heritable, than any similarity between spouses due to religion (e.g., similar views on abortion) would be due to social homogamy rather than primary phenotypic assortment.

A third possibility, genetic homogamy, occurs if mates choose each other based on the heritable aspect of their phenotypes rather than on their manifest phenotypes (Fisher 1918; Thiessen and Gregg 1980). Although seemingly implausible, there are two ways this might occur. The first is if people attempt to control for the effects of the environment when making mate choices (e.g., “He/she is really smart given the environment they come from”). The second is if people base mate choice on some third variable (e.g., overall mate value) that is related to the phenotype of interest purely genetically. This would be an extreme form of ‘good genes’ theories of human mate choice (Miller and Todd 1998). Consider, for example, assortative mating for intelligence. If people choose mates solely based on mate value (e.g., the first principal component of traits such as health, athleticism, height, facial attractiveness, bodily attractiveness, intelligence, and so forth), and if the inter-relationship between these mate value components is genetic in nature, then similarity between spouses on intelligence would be due to genetic homogamy. Our point is not to argue that genetic homogamy is or is not a likely mode of mate similarity, but rather to note that it is a viable option that should be tested empirically. Of the four twin-family designs discussed here, only the *Cascade* can model genetic and social homogamy.

We also simulated two scenarios that include parameters that could not be estimated in any model (rows 9–10, Table 2). These two scenarios allowed us to test the sensitivity to assumptions for all designs, including the *Stealth* and *Cascade*.

Model fitting with Mx

The authors wrote Mx scripts for the CTD (137 lines of code), the NTFD (189 lines of code), and the *Cascade* design (2,717 lines of code); the script for the *Stealth* design (2,780 lines of code) was written by H. Maes (Maes et al. 2009). These scripts are available at <http://www.matthewkeller.com/html/cascade.html>. An advantage of the *Stealth* script, not yet instantiated in the *Cascade* script, is that it is set up to fit multivariate data. The advantage of the *Cascade* design, and its original purpose, is the additional flexibility in modeling assortative mating and vertical transmission.

For each simulated dataset run using the NTFD, *Stealth*, and *Cascade* scripts, both a full and reduced model were fit (no reduced models were necessary for the CTD). The full NTFD model estimated \hat{V}_A , \hat{V}_D , \hat{V}_T , \hat{V}_E , and either \hat{V}_F (if familial effects existed in the scenario) or \hat{V}_S (if sibling

effects existed).² The full *Stealth* and *Cascade* models estimated \hat{V}_A , \hat{V}_D , \hat{V}_F , \hat{V}_S , \hat{V}_T , and \hat{V}_E , (note that $C\hat{V}(A, F)$ is technically a non-linear constraint and is not freely estimated; see Keller et al. 2009). The reduced NTFD, *Stealth*, and *Cascade* models estimated only those variance parameters that were truly non-zero in the given scenario. The fitting of both full and reduced models was done to demonstrate the effects of the common practice of dropping non-significant predictors. For example, under the ADE scenario (top row, Table 2), \hat{V}_A , \hat{V}_D , \hat{V}_F , \hat{V}_S , \hat{V}_T , and \hat{V}_E were estimated in the full *Stealth* and *Cascade* models but only \hat{V}_A , \hat{V}_D , and \hat{V}_E were estimated in the reduced models; \hat{V}_F , \hat{V}_S , and \hat{V}_T were dropped (set equal to 0). Our strategy therefore assumed that no type-I errors occurred in choosing the reduced models. While not optimal, creating a program that tested the significance of each estimate individually and dropped non-significant ones would have added enormous complexity and computing time onto a project that already stretched both of these limits. Moreover, estimates would have been incorrectly retained only ~5% of the time (the type-I error rate), and therefore this strategy introduced only minor and probably negligible inaccuracy into our reduced model results.

Finding the bias, precision, and accuracy of parameter estimates

We compared the parameters estimated from Mx for each design to the true parameters from *GeneEvolve* for each simulation run. This allowed us to empirically determine the bias, precision, and accuracy of the parameter estimates, as well as their distributional shapes and covariances (Casela and Berger 1990). The *bias* of a statistic is generally defined as $E(\hat{V} - V)$, the expected (i.e., mean) difference between the estimated parameter, \hat{V} , and the true parameter, V . An alternative is to use the median difference rather than the expected difference, $M(\hat{V} - V)$, which is less influenced by outlier estimates. We chose this latter measure of bias because several outlier \hat{V} 's in our data are probably artifactual due to the automated way the models were run. Although we discarded estimates from models that gave a “Code Red” (IFAIL = 6) in Mx, which occurs when constraints cannot be satisfied and is symptomatic of poorly performing estimation, inspection of Mx output led us to conclude that occasionally (~2–8% of the time, depending on the scenario), Mx poorly recreated the expected covariance matrix and gave bad estimates even when no “Code Red” occurred. Such estimates are

² Strictly speaking, $C\hat{V}(A, F)$ is a nonlinear constraint and is not freely estimated in ETFDs. It is determined by, and helps to determine, estimated parameters by constraining their inter-relationships in a way that keeps the entire model internally consistent.

artificial in the present context because they likely could have been averted in most real life modeling contexts by providing different start values, dropping parameters, or by taking other remedial measures to improve the fit.

The *precision* of estimates measures the spread of the estimates around their center, and is typically measured by the standard deviation or variance of the parameter estimates, e.g., $\sqrt{\frac{1}{n-1} \sum_{i=1}^n (\hat{V}_i - E(\hat{V}))^2}$. An alternative which we use for the same reasons mentioned above—namely, that we wish to downweight outliers that are likely to be artificial—is the median absolute deviation, or MAD, which is equal to $M(|\hat{V}_i - M(\hat{V})|)$.

The *accuracy* of a statistic combines information on both bias and precision to gauge how far away from the true value an estimate typically is. Thus, an estimate can be precise but nevertheless inaccurate if it is biased, or can be unbiased but inaccurate if it is imprecise. As with precision, accuracy is often measured using the variance or standard deviation, except that estimates are judged by how far away they are from the *value of the true parameter* rather than the values of the mean estimates, e.g.,

$\sqrt{\frac{1}{n-1} \sum_{i=1}^n (\hat{V}_i - E(V))^2}$. In this situation, $accuracy^2 = bias^2 + precision^2$ using the first of each of the definitions above. In the present study, we use the median absolute error, $M(|\hat{V}_i - V|)$, to measure accuracy so as to lessen the impact of outliers.

Results

Bias, precision, and accuracy of parameter estimates

The distributions of four of the parameter estimates for each of the ten scenarios described in Table 2 are shown in Figs. 4, 5, 6, 7, 8, 9, 10, 11, 12, and 13. These figures do not show \hat{V}_T , \hat{V}_E , or $CV(A, F)$ because these estimates tend to be of less interest. These figures also place CTD estimates of \hat{V}_C into the column reserved for \hat{V}_F or \hat{V}_S , whichever is appropriate given the scenario. As noted above, no reduced CTD models needed to be fit, and so reduced CTD estimates are not shown.

Results for the ADE and ASE scenarios, which did not violate assumptions in any of the four designs, are shown in Figs. 4 and 5. A few things should be noted. First, when assumptions of the CTD are not violated (i.e., V_C in the ADE scenario and V_D in the ASE scenario), estimates from the CTD are unbiased and have decent precision. Second, the reduced models from the three ETFDs are also unbiased, and they have greater precision than the CTD estimates. Reduced ETFD estimates are more precise because they are based on much more information (covariance observations) than the CTD estimates. Third, the full models for the three ETFDs show varying degrees of bias and poorer precision than the other models. The bias in the ETFD full models occurs for the same fundamental reason that bias exists in Cholesky models (Carey 2005): variance estimates are forced to be non-negative. By chance, the ETFD full models pick up

Fig. 4 ADE scenario

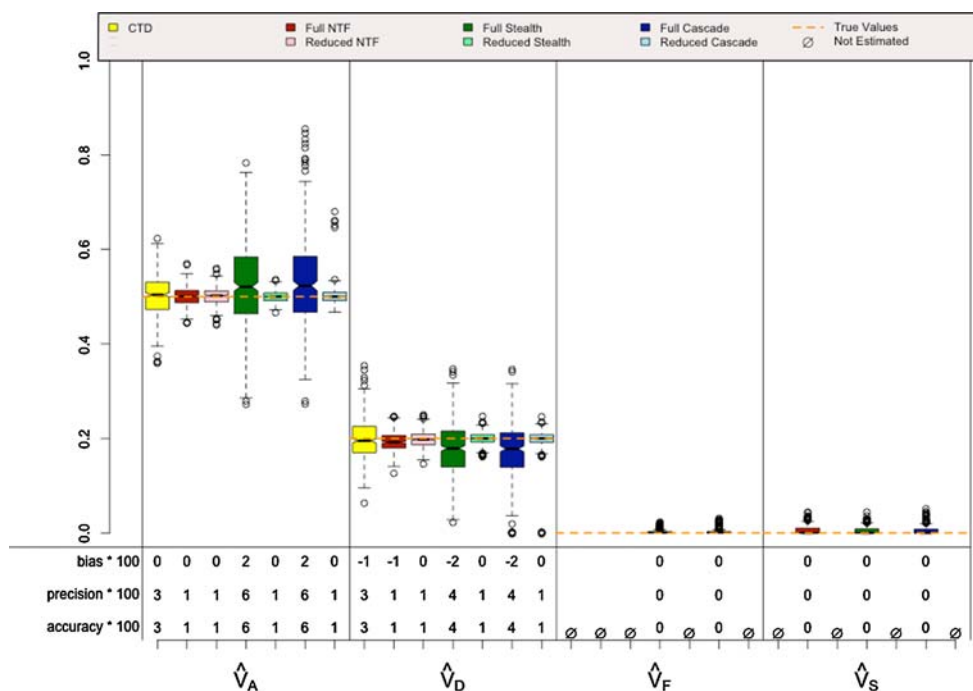


Fig. 5 ASE scenario

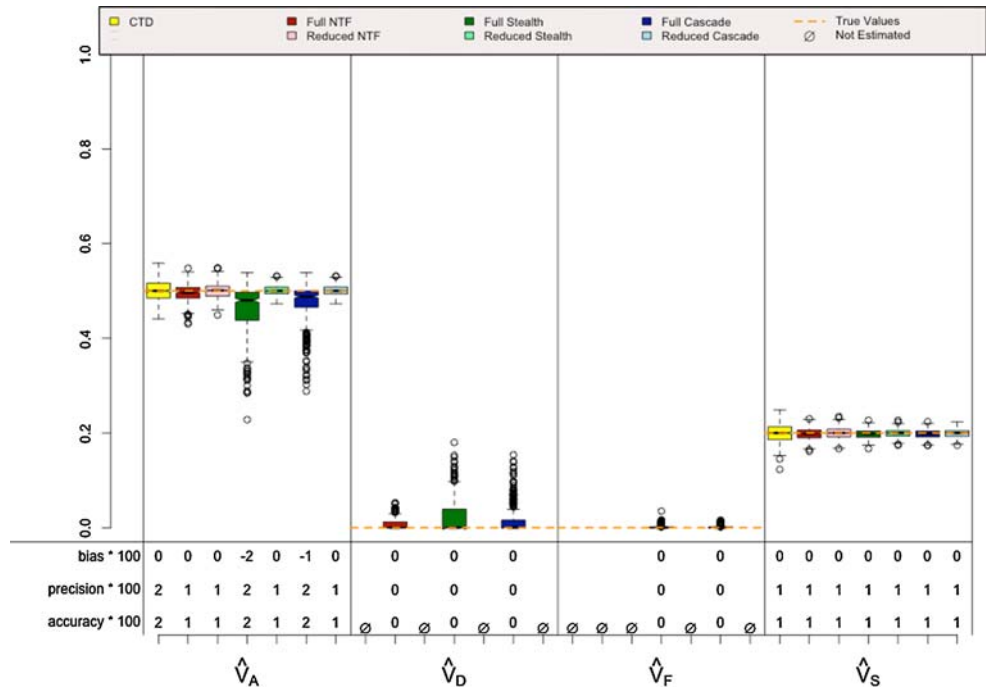
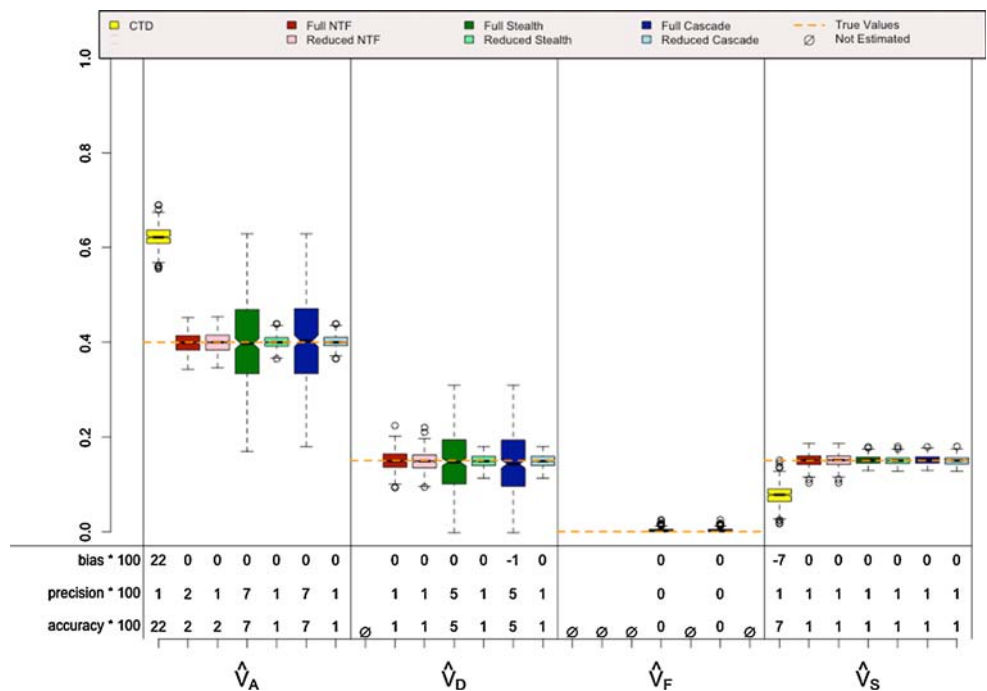


Fig. 6 ADSE scenario



slight evidence for non-zero variance parameters that, in truth, are actually zero (e.g., V_F and V_S in the ADE scenario). If the evidence suggests that these estimates are positive, ETFD models estimate them freely, but if negative, these estimates hit the zero boundary. This imbalance pulls the other estimated parameters (e.g., \hat{V}_A and \hat{V}_D in the ADE scenario) in only one direction, causing bias. This source of bias, though minor, could be removed if the ETFD models allowed variance estimates

to be negative. The lack of precision in full ETFDs, on the other hand, cannot be so easily rectified, but rather is a natural consequence of attempting to estimate so many more parameters in ETFDs, especially in the *Stealth* and *Cascade* designs.

Figures 6, 7, and 8 show results for three scenarios in which CTD assumptions are violated because both shared environmental and non-additive genetic effects influence a trait simultaneously and, in the final scenario, because

Fig. 7 ADFE scenario

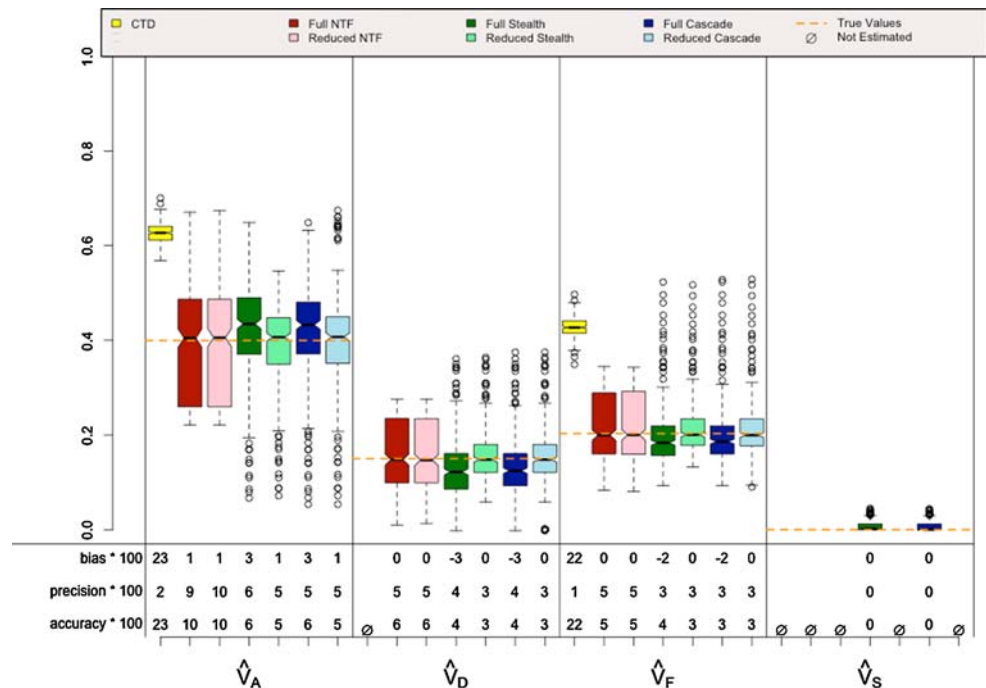
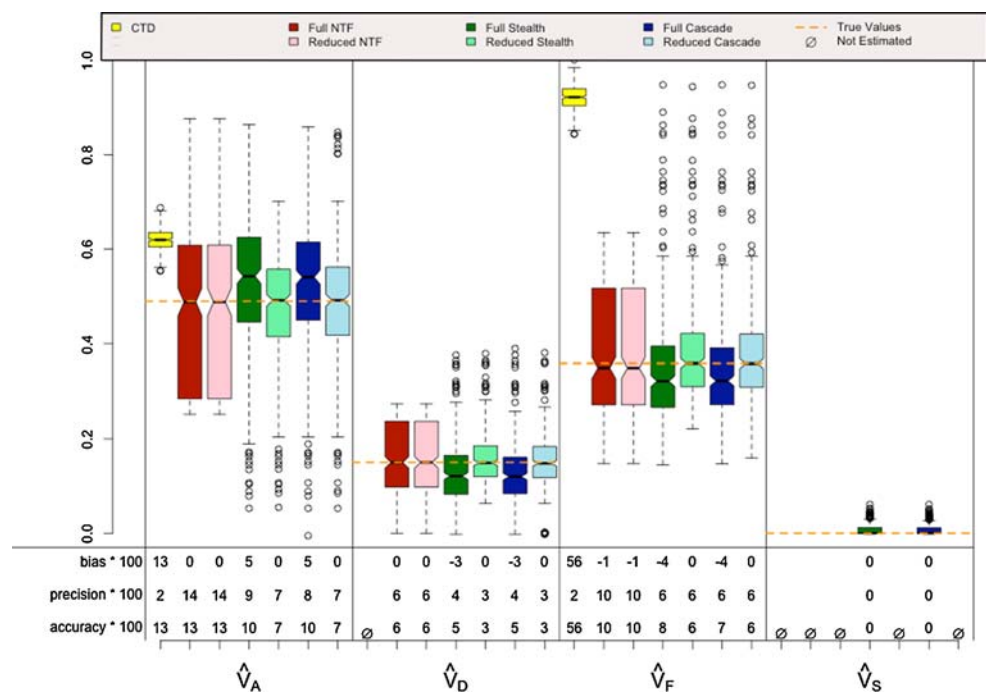


Fig. 8 ADFE & primary phenotypic mating ($r = .3$) scenario



assortative mating exists. However, these scenarios do not violate assumptions for any ETFD. The CTD estimates are highly biased in the expected directions (Grayson 1989; Keller and Coventry 2005), with additive genetic effects being overestimated by about 50% in these examples and non-additive genetic effects ignored because, for reasons of identifiability, they could not be estimated. Shared environmental effects are underestimated by the CTD in the ADSE scenario, but are overestimated in the ADFE and

ADFE & Primary Assortative Mating scenarios. This overestimation is also predictable, and occurs because of the substantial $CV(A, F)$ that is induced by vertical transmission, which mimics shared environment in the CTD (Eaves et al. 1989). As expected, the reduced ETFD models do not show bias whereas the full ETFD models show slight biases for the same reason discussed above. The *Stealth* and *Cascade* estimates are quite accurate in these scenarios, typically being within .05 points of the true parameters. NTFD

Fig. 9 ADFSTE & primary phenotypic mating ($r = .3$) scenario

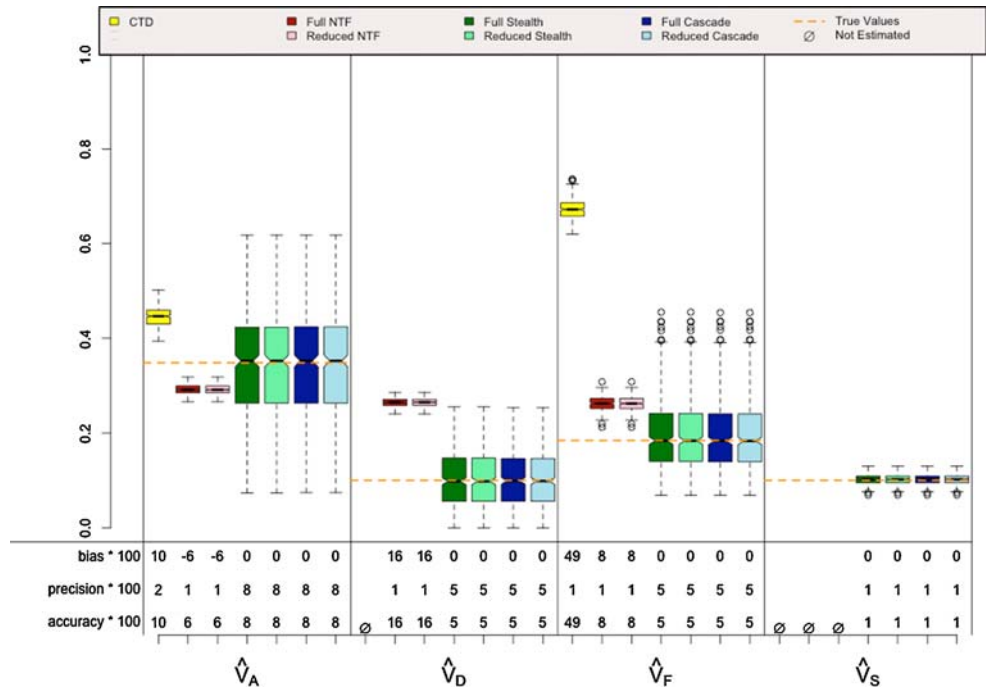
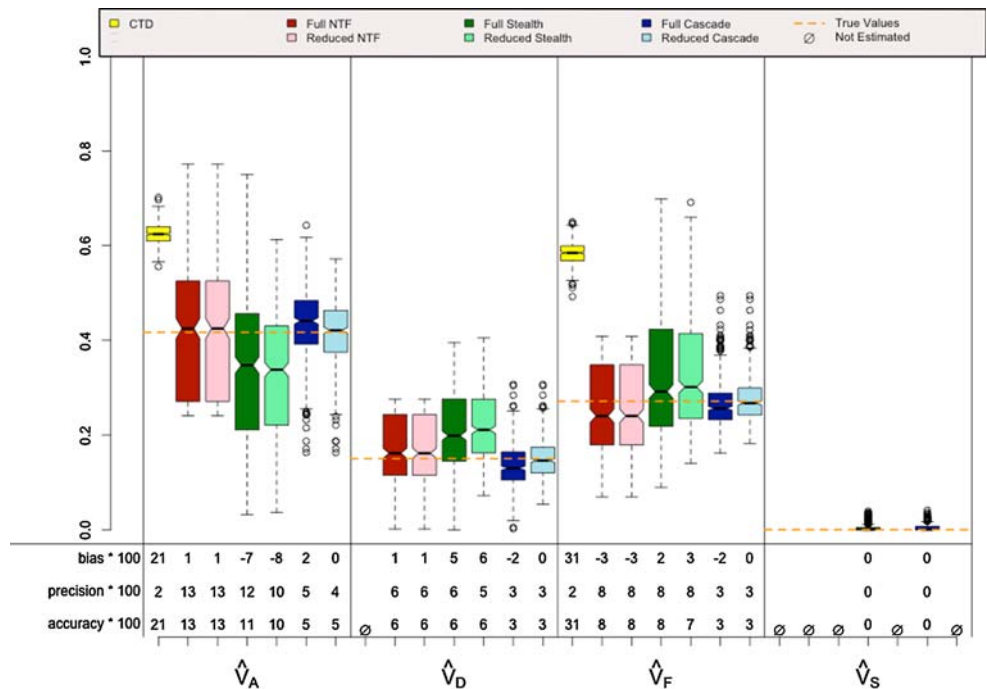


Fig. 10 ADFE & social homogamy ($r = .3$) scenario



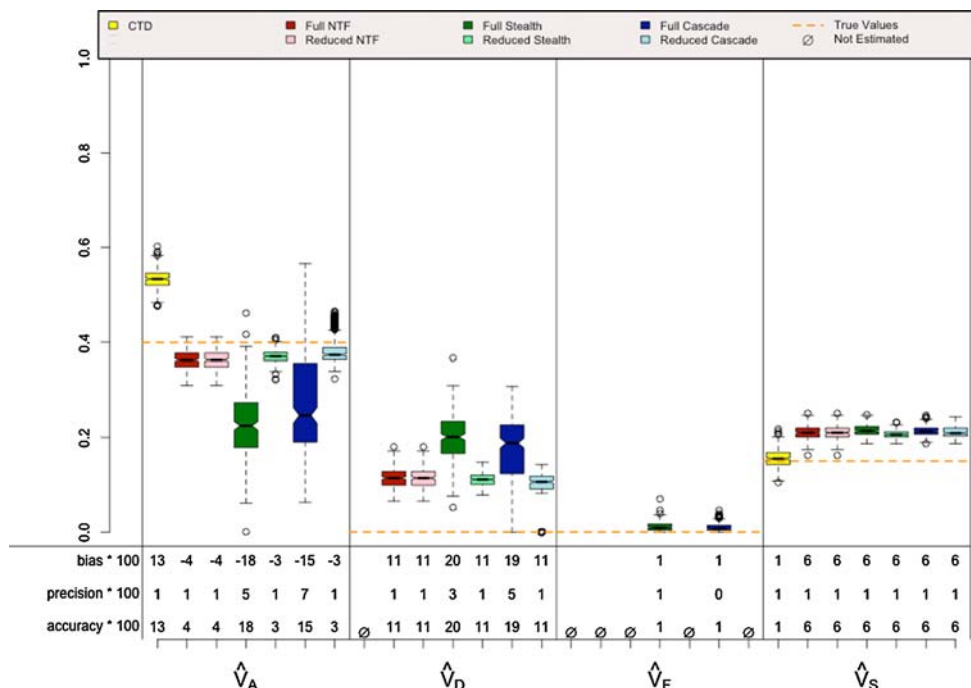
estimates are less accurate when both \hat{V}_A and \hat{V}_F are estimated simultaneously; this is due to the very high correlation between these two estimates (see below).

Figure 9 shows results for a complicated scenario in which V_A , V_D , V_F , V_S , V_T , V_E , and $CV(A, F)$ all contribute to phenotypic variance in the context of primary phenotypic assortative mating. Here, the NTFD assumption that either V_F , or V_S is zero is violated, causing estimates that, although precise, are quite biased. Because all parameters were

retained in the reduced model, the results for the full and reduced ETFD models are identical. All *Stealth* and *Cascade* estimates are unbiased; however, \hat{V}_A shows a fairly high degree of imprecision due to the correlation between \hat{V}_A and \hat{V}_D , and between \hat{V}_A and \hat{V}_F (see next section).

Figures 10 and 11 show results for scenarios identical to that depicted in Fig. 8 except that spousal similarity is due to social homogamy (Fig. 10) or genetic homogamy (Fig. 11). Thus, these two scenarios violate assumptions

Fig. 13 ASE & A × Age Interaction (var = .15) scenario



overestimate V_A and slightly underestimate V_S . However, the overall level of genetic variation ($V_A + V_D + V_{A \times A}$) tends to be only slightly underestimated. Moreover, if \hat{V}_D is considered a broad estimate of non-additive genetic variance rather than an estimate of dominance variance only, estimates of non-additive genetic variation are only slightly underestimated.

Non-scalar gene-by-age interactions (Fig. 13) can be conceptualized as different genes ‘turning on’ at different ages, and as opposed to scalar gene-by-age interactions, tend to decrease genetic covariation between relatives as a function of the age difference between them. Because siblings and twins tend to be close in age to one another, it is sensible that non-scalar gene-by-age interactions lead to overestimation of V_T (not shown) and V_S and underestimation of V_A in ETFDs. Another interesting ramification of such non-scalar gene-by-age interactions is that they can lead to negative vertical transmission pathways in ETFDs (creating positive \hat{V}_F but decreasing similarity between parents and offspring), a not uncommon observation in empirical ETFD studies. In the CTD, on the other hand, non-scalar gene-by-age interactions cause overestimations of V_A . Although we are aware of no models that have been written to do so, ETFDs should be able to model non-scalar gene-by-age interactions due to the wide variation in ages within families used in ETFDs. For example, in Mx, age differences between each pair of family members could be calculated from definitional variables, and these age differences could be used to moderate the expected additive genetic covariances between relative types.

Relationships between parameter estimates

The information required to estimate parameters is often partially redundant. For example, both V_A and V_F cause within-family similarity that drops off as a function of how distant a relative pair is, and so \hat{V}_A and \hat{V}_F tend to be negatively related: as one estimate increases and explains a given pattern of observed covariances, there is less information ‘left over’ for the other estimate to explain. Figure 14 shows that \hat{V}_A and \hat{V}_F , and \hat{V}_A and \hat{V}_D use partially redundant information in the *Cascade* design and so are highly negatively related. \hat{V}_D and \hat{V}_F are positively related, but only in models that also estimate \hat{V}_A : as \hat{V}_A increases, both \hat{V}_D and \hat{V}_F decrease. \hat{V}_S , on the other hand, is nearly independent of \hat{V}_A , \hat{V}_D and \hat{V}_F in the *Cascade*. Information to estimate \hat{V}_S comes primarily from the comparison between twin and sibling covariances versus parent–offspring covariances, and thus does not use information that overlaps with any of the other estimates.

A linear regression model predicting \hat{V}_A in the *Cascade* from \hat{V}_D and \hat{V}_F under the scenario depicted in Fig. 9 has an $r^2 = .969$, which translates to a *variance inflation factor* of $\frac{1}{1-r^2} = 32.6$. Thus, the variance of \hat{V}_A in the *Cascade* model is 32.6 times higher, and the standard error of \hat{V}_A is 5.7 times higher, than in models in which both \hat{V}_D and \hat{V}_F are dropped. Similarly, the standard errors of \hat{V}_D and \hat{V}_F are 4.4 and 4.8 times higher, respectively, than they are in models in which they are estimated alone. Similar findings occur for the other two ETFD models. This effect can be seen in Figs. 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, and 13, which show that the distributions of parameter

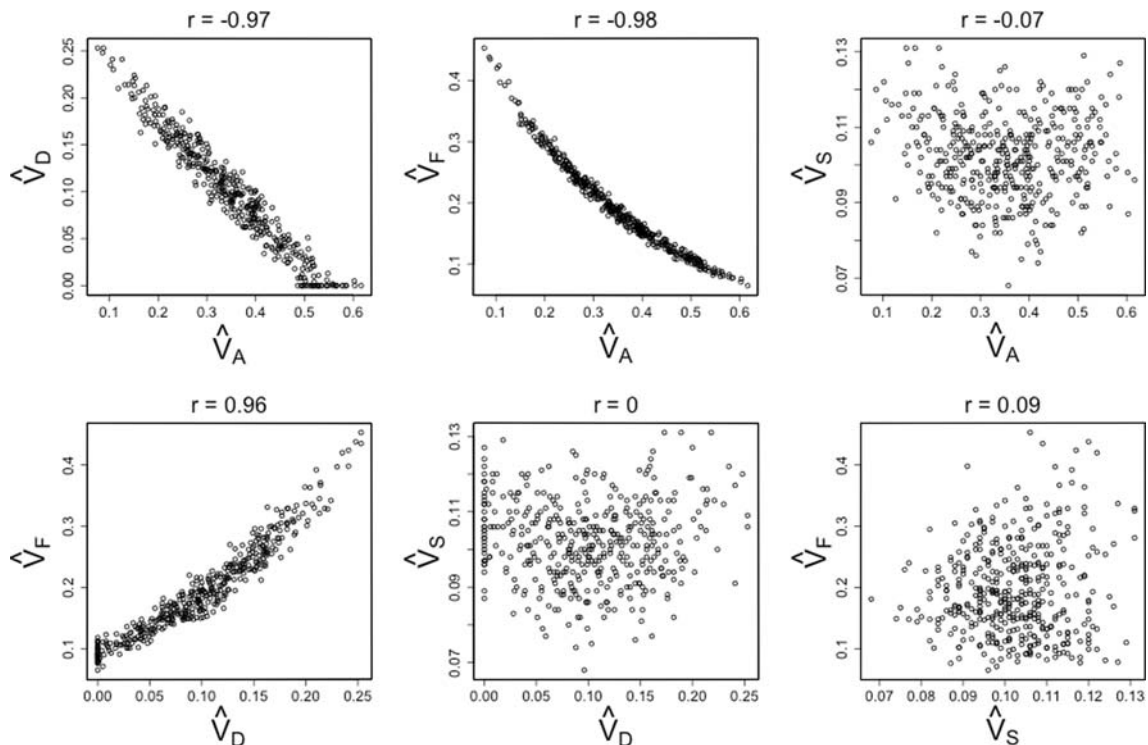


Fig. 14 Parameter correlations from a *Cascade* model estimating parameters under an ADFSTE & primary phenotypic mating ($r = .3$) scenario

Table 3 A small subset of the sources of information for estimating parameters in the *Cascade* and *Stealth* models

| Parameter | Increases as a function of... |
|-------------|---|
| \hat{V}_A | $C\hat{V}(MZ) - C\hat{V}(DZ); C\hat{V}(MZ.avuncular) - C\hat{V}(DZ.avuncular); C\hat{V}(MZ.cous) - C\hat{V}(DZ.cous)$ |
| \hat{V}_D | $C\hat{V}(MZ) - C\hat{V}(DZ); C\hat{V}(DZ) - C\hat{V}(\text{Parent, Offspring})$ |
| \hat{V}_S | $C\hat{V}(DZ) - C\hat{V}(MZ); C\hat{V}(DZ) - C\hat{V}(\text{Parent, Offspring}); C\hat{V}(DZ) - C\hat{V}(MZ.avuncular)$ |
| \hat{V}_F | $C\hat{V}(\text{Parent, Offspring}) - C\hat{V}(MZ.avuncular); C\hat{V}(DZ.cous) - C\hat{V}(MZ.cous)$ |
| \hat{V}_T | $C\hat{V}(DZ) - C\hat{V}(\text{Sib})$ |

estimates for ETFD full models are consistently more spread out than the distributions of parameter estimates for reduced models.

Information for estimating parameters in ETFDs

It is useful to have a sense of how observed covariance estimates translate into estimated parameters. In the CTD, it is obvious that the difference between $C\hat{V}(MZ, MZ)$ and $C\hat{V}(DZ, DZ)$ provides all the information needed to estimate \hat{V}_A and \hat{V}_C (in ACE models) or \hat{V}_A and \hat{V}_D (in ADE models). However, it becomes increasingly difficult to discern how observed covariance estimates influence estimated parameters in increasingly complex ETFDs. For example, which covariance estimates help differentiate \hat{V}_A from \hat{V}_F in the *Stealth* or *Cascade*? What information allows differentiation of social homogamy from primary phenotypic assortment in the *Cascade* model?

Unfortunately, there are no simple answers to these types of questions in ETFDs. A huge number of partially redundant bits of information help estimate the unknown parameters in ETFDs, and the effect of this information depends on the model being fit (e.g., how assortative mating is modeled) as well as on the values of the other simultaneously estimated parameters (e.g., the degree of vertical transmission alters how observed covariances affect \hat{V}_A).

Despite these difficulties, Table 3 provides some insight into how observed covariances are used to estimate parameters in the *Cascade* and *Stealth* models. The table is not exhaustive; for certain parameters (especially \hat{V}_A and \hat{V}_F), nearly every covariance estimate plays some role in their estimation. Rather, Table 3 lists some of the most consistent sources of information across models used in estimating the five variance parameters that cause familial resemblance. $C\hat{V}(MZ.avuncular)$ refers to the covariance

between the children of one MZ twin and the other (avuncular) MZ co-twin, whereas $C\hat{V}(MZ.cous)$ refers to the covariance between cousins whose parents are MZ co-twins. With respect to assortative mating in the *Cascade* model, in-laws are particularly helpful for differentiating social from phenotypic homogamy. For example, under social homogamy, there is no expected difference between MZ in-law correlations and DZ in-law correlations, whereas under phenotypic homogamy, in-law relationships should differ by zygosity status.

Discussion

Our results show that ETFDs work as designed. They are generally unbiased when assumptions are met, and unlike the CTD, they are not overly sensitive to violations of assumptions so long as \hat{V}_D is interpreted broadly, as an estimate of genetic non-additivity in general (including gene-by-age interaction effects) rather than as dominance in particular. Our results also highlight that the key trade-off in using ETFDs is one of complexity versus accuracy. By attempting to estimate a large number of parameters, many of which use overlapping information, the precision of ETFD estimates suffers (see the full ETFD model estimates in Figs. 4, 5, 6, 7, 8, 9, 10, 11, 12, and 13 and parameter covariances in Fig. 14). The ETFD estimates in Fig. 8, for example, are much less precise than those from the CTD. Nevertheless, ETFD estimates tend to be unbiased under a much wider range of scenarios than CTD estimates, and because of this, are almost universally more accurate than are CTD estimates. This improved accuracy can be quantified by empirical researchers using ETFDs by comparing a goodness of fit index of an ETFD only estimating a few parameters (e.g., \hat{V}_A , \hat{V}_D , and \hat{V}_E) versus an ETFD estimating all parameters. The difference between these two fit indices provides an idea of how important using an ETFD is over a simpler model (e.g., the CTD) given the phenotype in question.

The trend of increasing accuracy with increasing complexity repeats itself within the ETFD models: *Stealth* estimates are accurate across a wider range of scenarios than are NTFD estimates (Fig. 6), and *Cascade* estimates are accurate across a wider range of scenarios than are *Stealth* estimates (Figs. 10, 11). For example, the mean accuracy values (lower being more accurate) across the ten scenarios for \hat{V}_A were .140 for the CTD, .069 for the NTFD, .049 for the *Stealth*, and .045 for the *Cascade*. As expected, the *Cascade* and *Stealth* results were virtually identical except in cases where assumptions regarding mating in the *Stealth* were violated.

Nevertheless, the question remains: given the increased difficulty in fitting the models and collecting the requisite

data, is it worth it to use ETFDs? Our results cannot provide an answer to this question, but they do provide guidance. For all the problems associated with the CTD, the combined CTD parameters of $\hat{V}_A + \hat{V}_D$ do provide decent estimates of broad sense heritability. If a researcher's goal is primarily to understand broad sense heritability, or to understand broad sense genetic covariances in a multivariate setting, the CTD is adequate, and using ETFDs is probably not worth the hassle unless extended family data already exists. To the degree that any genetic non-additivity or spousal similarity exists, however, CTD models can wildly under- or over-estimate shared environmental effects (see Figs. 7, 8, 9, 10, and 11). Thus, if one's interest is in characterizing the effects of the environment in any way—including arguing that shared environmental effects are small—the CTD is a singularly bad method. Similarly, if one's interest is in understanding the relative importance of additive versus non-additive genetic variation, the CTD provides little help. In these latter situations, researchers should seriously consider the use of ETFDs. These conclusions are not merely based on the simulation results of this paper. Coventry and Keller (2005) compared the parameter estimates of every available *Stealth* model run up to that time to the estimates that would have been obtained using the CTD on the same data and phenotype. Consistent with prediction, they found that CTD results gave predictably distorted pictures of the makeup of genetic variation and the makeup and importance of the common environment.

For researchers who already have the data needed to fit the *Stealth* or *Cascade* models, our results suggest the *Cascade* model should be used over existing ETFD models. However, an argument could be made from our results that the NTFD represents a good compromise between the accuracy of the *Cascade* and the simplicity of the CTD. NTFD estimates tended to be less precise and slightly more biased than *Cascade* estimates, but these differences were minor compared to the difference between the ETFD estimates as a group and the CTD estimates. Of course, the major limitation of the NTFD is that the source of shared environmental effects (due to sibling effects or vertical transmission from parents) cannot be discerned, and when both shared environmental sources are present, estimated parameters will be biased. In a separate piece (Medland and Keller 2009), we discuss which relative types provide the most power for detecting different parameters in the *Cascade*, which should be of service to investigators interested in collecting new data for any ETFD (see also Heath et al. 1985).

Hill et al. (2008) recently argued that most genetic variance in most traits is additive in nature. If $V_D \sim 0$ for most traits, then CTD estimates of V_A and V_C should be accurate in the absence of assortative mating and vertical transmission, and thus ETFDs would often be overkill.

While we agree with Hill et al.'s (2008) conclusion that genetic variation is likely to be mostly additive in nature for most traits, we disagree with potential conclusions drawn from this paper (e.g., Wahlberg 2009) that non-additive genetic variance is typically small and insignificant. A meta-analysis of results from the *Stealth* design (Coventry and Keller 2005) found that typically $\hat{V}_D \gg 0$ and, on average across 38 phenotypes, \hat{V}_D was nearly as large as \hat{V}_A , being a full 75% of \hat{V}_A . These *Stealth* results showing evidence for substantial non-additive genetic variance are much more convincing than Hill et al.'s (2008) twin-only analysis, in which correlations of monozygotic and dizygotic twins were compared across 86 phenotypes: as we have shown (Figs. 4, 5, 6, 7, 8, 9, 10, 11, 12, and 13), the relative magnitude of V_A versus V_D cannot be accurately ascertained using twins alone. Moreover, because natural selection erodes additive genetic variation much faster than non-additive genetic variance, theory suggests that traits related to Darwinian fitness should have relatively high degrees of non-additive genetic variation (Haldane 1932; Wright 1929), and indeed empirical reviews show that non-additive genetic variance in non-human animals is similar in magnitude to additive genetic variance among fitness-related traits (Crnokrak and Roff 1995). Thus, without empirical investigation, we think it would be premature to take solace in the hope that non-additive genetic variance is low enough for most traits for CTD estimates to be generally unbiased.

There are several limitations with the current approach to understanding the bias, precision, and accuracy of parameter estimates from twin-family designs. First, as mentioned above, our procedure for automating model fitting meant that the results from reduced ETFD models were optimistic. However, as we argued in the “[Methods](#)” section, this probably produced a negligible degree of bias in our results. A more important source of bias in our results, which worked in the opposite direction, is that a human could not guide each fitting process interactively due to the automated way models were fit. A non-negligible number (2% to 8% depending on the scenario) of model runs produced outlier estimates, poorly reproduced the observed covariance matrices, and probably failed to find the true maximum likelihood estimates. An experienced modeler could have detected these situations and taken remedial measures, such as changing start values, to improve the fit of the model. This suggests that the ETFD results presented in this paper appear less precise than they will be when fit interactively on real data.

Another limitation to the current approach was that we investigated only a very small portion of the space of possible parameters that might exist in the real world. For example, we did not investigate alternative modes of vertical transmission or spousal similarity due to convergence,

both of which can be modeled in the *Cascade*. We also did not investigate any number of alternative scenarios that might occur and cause bias in all the models investigated here, such as mixed models of assortative mating (Reynolds et al. 2000), additional types of gene–environment interactions and correlations, higher-order epistasis, in utero effects, and special MZ-twin environments. This latter issue is particularly important. At the heart of all twin models, including ETFDs, is the comparison between MZ and DZ twins. If some non-genetic factor such as in utero effects increases MZ twin resemblance, all models described in this paper will overestimate \hat{V}_A and especially \hat{V}_D . Furthermore, for simplicity, we did not investigate sex-specific estimates in this paper, which would have had similar biases but lower precision than those presented here. Given this, none but the largest sex-specific effects are likely to be detectable with even the largest available extended twin family datasets. A final limitation to our study is that only univariate models were investigated. Although univariate parameter estimates are interesting in their own rights, ETFD models become more interesting in a multivariate context. For example, parental warmth may be negatively associated with adolescent depression in children (Operario et al. 2006), but the reasons for this association are unclear. ETFD models can discern whether this association is due to the same genes affecting both warmth and depression risk or to parental warmth being culturally transmitted to offspring in the form of lower depression risk. Our paper did not assess the parameter characteristics in such multivariate models, although there is no reason to believe that the quality of multivariate parameter estimates would be substantially different than univariate ones. Despite these limitations, the current paper represents the fullest exploration to date of how different real world scenarios affect estimates from twin-family designs.

We have argued that the most commonly used design in behavioral genetics, the CTD, is inadequate for understanding the relative magnitude of shared environmental effects or the ratio of additive to non-additive genetic variation. Our results demonstrate that, irrespective of power or sample size, estimates of these two quantities from CTDs cannot be interpreted with any degree of confidence unless strong assumptions—no assortative mating, no gene–environment covariance, and that either non-additive genetic variance or shared environmental variance is zero—have been verified. ETFDs, on the other hand, provide unbiased and fairly accurate estimates of this information. More complex ETFDs, such as the *Cascade*, are unbiased under an even wider range of scenarios and provide additional details on the makeup of shared environmental effects that may itself be of interest. The principal reasons why ETFDs are rarely used in behavioral

genetics is that they are more difficult to use and that little extended twin family data exists suitable for their use. We hope that the current paper clarifies the rationale for using ETFDs and encourages researchers to collect extended twin family data when circumstances warrant their use.

References

- Carey G (2005) Cholesky problems. *Behav Genet* 35:653–665
- Casela G, Berger RL (1990) *Statistical inference*. Wadsworth, Belmont
- Cloninger CR, Rice J, Reich T (1979) Multifactorial inheritance with cultural transmission and assortative mating II: a general model of combined polygenic and cultural inheritance. *Am J Hum Genet* 31:176–198
- Coventry WL, Keller MC (2005) Estimating the extent of parameter bias in the classical twin design: a comparison of parameter estimates from extended twin-family and classical twin designs. *Twin Res Hum Genet* 8:214–223
- Crnokrak P, Roff DA (1995) Dominance variation: associations with selection and fitness. *Heredity* 75:530–540
- Eaves LJ (1979) The use of twins in the analysis of assortative mating. *Heredity* 43:399–409
- Eaves LJ (2009) Putting the ‘human’ back in genetics: modeling the extended kinship of twins. *Twin Res Hum Genet* 12:1–7
- Eaves LJ, Last KA, Young PA, Martin NG (1978) Model-fitting approaches to the analysis of human behavior. *Heredity* 41:249–320
- Eaves LJ, Eysenck HJ, Martin JM (eds) (1989) *Genes, culture, and personality: an empirical approach*. Academic Press, Londong
- Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Trans Roy Soc Edinb* 52:399–433
- Fulker DW (1982) Extension of the classical twin method. In Bonné-Tamir B, Cohen T, Goodman RM (eds) *Human genetics, part A: the unfolding genome (Progress in clinical and biological research 103A)*. Alan R Liss, New York, pp. 395–406
- Grayson DA (1989) Twins reared together: minimizing shared environmental effects. *Behav Genet* 19:593–604
- Haldane JBS (1932) *The causes of evolution*. Princeton University Press, Princeton, N.J.
- Heath AC, Kendler KS, Eaves LJ, Markell D (1985) The resolution of cultural and biological inheritance: informativeness of different relationships. *Behav Genet* 15:439–465
- Hill WG, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet* 4:1–10
- Keller, M. C. (2007). *PedEvolve: a simulator of genetically informative data implemented in R*. Annual meeting of the behavior genetics association, Amsterdam
- Keller MC, Coventry WL (2005) Quantifying and addressing parameter indeterminacy in the classical twin design. *Twin Res Hum Genet* 8:201–213
- Keller MC, Medland SE, Duncan LE, Hatemi PK, Neale MC, Maes HMM et al (2009) Modeling extended twin family data I: description of the cascade model. *Twin Res Hum Genet* 12:8–18
- Maes HMM, Neale MC, Medland SE, Keller MC, Martin NG, Heath AC et al (2009) Flexible Mx specifications of various extended twin kinship designs. *Twin Res Hum Genet* 12:26–34
- Martin NG, Boomsma DI, Machin G (1997) A twin-pronged attack on complex traits. *Nat Genet* 17:387–392
- Medland SE, Keller MC (2009) Modeling extended twin family data II: power associated with different family structures. *Twin Res Hum Genet* 12:19–25
- Miller G, Todd PM (1998) Mate choice turns cognitive. *Trends Cogn Sci* 2:190–198
- Nance WE, Corey LA (1976) Genetic models for the analysis of data from the families of identical twins. *Genetics* 83:811–826
- Neale MC (1999) *MX: statistical modelling*, 5th edn. Department of Psychiatry, Richmond, VA
- Neale MC, Fulker DW (1984) A bivariate path analysis of fear data on twins and their parents. *Acta Genetica Medica Gemellol (Roma)* 33:273–286
- Operario D, Tschann J, Flores E, Bridges M (2006) Brief report: associations of parental warmth, peer support, and gender with adolescent emotional distress. *J Adolesc* 29(2):299–305
- Plomin R, DeFries JC, McClearn GE, McGuffin P (2001) *Behavioral genetics*, 4th edn. Worth Publishers, New York
- Posthuma D, Boomsma DI (2000) A note on the statistical power in extended twin designs. *Behav Genet* 30:147–158
- R Core Development Team (2009) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria
- Reynolds CA, Baker LA, Pedersen NL (2000) Multivariate models of mixed assortment: phenotypic assortment and social homogamy for education and fluid ability. *Behav Genet* 30(6):455–476
- Thiessen DD, Gregg B (1980) Human assortative mating and genetic equilibrium: an evolutionary perspective. *Ethol Sociobiol* 1:111–140
- Truett KR, Eaves LJ, Walters EE, Heath AC, Hewitt JK, Meyer JM et al (1994) A model system for analysis of family resemblance in extended kinships of twins. *Behav Genet* 24:35–49
- Wahlberg P (2009) *Chicken genomics-linkage and QTL mapping*. Digital comprehensive summaries of Uppsala. Dissertations from the Faculty of Medicine
- Wright S (1929) Fisher’s theory of dominance. *Am Nat* 63:274–279