Behav Genet
DOI 10.1007/s10519-014-9698-y

ORIGINAL RESEARCH

# Replication of a Gene–Environment Interaction Via Multimodel Inference: Additive-Genetic Variance in Adolescents' General Cognitive Ability Increases with Family-of-Origin Socioeconomic Status

**Robert M. Kirkpatrick · Matt McGue · William G. Iacono**

Received: 24 February 2014 / Accepted: 7 December 2014


**Abstract** The present study of general cognitive ability attempts to replicate and extend previous investigations of a biometric moderator, family-of-origin socioeconomic status (SES), in a sample of 2,494 pairs of adolescent twins, non-twin biological siblings, and adoptive siblings assessed with individually administered IQ tests. We hypothesized that SES would covary positively with additive-genetic variance and negatively with shared-environmental variance. Important potential confounds unaddressed in some past studies, such as twin-specific effects, assortative mating, and differential heritability by trait level, were found to be negligible. In our main analysis, we compared models by their sample-size corrected AIC, and base our statistical inference on model-averaged point estimates and standard errors. Additive-genetic variance increased with SES—an effect that was statistically significant and robust to model specification. We found no evidence that SES moderated shared-environmental influence. We attempt to explain the inconsistent replication record of these effects, and provide suggestions for future research.

Edited by Danielle Posthuma.

**Electronic supplementary material** The online version of this article (doi:10.1007/s10519-014-9698-y) contains supplementary material, which is available to authorized users.

R. M. Kirkpatrick · M. McGue · W. G. Iacono
Department of Psychology, University of Minnesota, 75 E. River Rd, Minneapolis, MN 55455, USA

*Present Address:*
R. M. Kirkpatrick (✉)
Virginia Institute for Psychiatric & Behavioral Genetics, Virginia Commonwealth University, Richmond, VA 23298-0126, USA
e-mail: rkirkpatrick2@vcu.edu

**Keywords** Gene–environment interaction · SES · Multimodel inference · General cognitive ability · IQ · Twin study · Adoption study

## Background

### Biometric modeling of general cognitive ability

Gene–environment interaction ($G \times E$) occurs when the phenotypic effect of genetic factors varies as a function of one or more environmental variables. The present work is concerned with an extension of the $G \times E$ concept: estimating how much the magnitudes of *all* biometric variance components depend upon one or more observable variables. We will use "biometric moderation" to refer to the phenomenon that the biometric decomposition of a phenotype varies as a function of some observable variable, the "biometric moderator." We will specifically be concerned with biometric moderation in general cognitive ability (GCA, the phenotype) by family-of-origin socioeconomic status (SES, the moderator). We will attempt to replicate the result of Turkheimer et al. (2003): increasing additive-genetic variance and decreasing shared-environmental variance with increasing SES.

GCA is that ability which is tapped by all cognitively demanding tasks. Often identified with Spearman's (1904) *g*, it can be operationalized as a composite score from a battery of tests that adequately samples the domain of cognitive tasks and specific abilities—for example, full-scale IQ (FSIQ) from an individually administered IQ test. Decades of research (to say the least—see Galton 1869) have made clear that general cognitive ability is a substantially heritable trait. Estimates of its heritability typically range from 0.50 to 0.70 (Bouchard and McGue 1981,

Published online: 25 December 2014

2003; Deary et al. 2006), and are sometimes as high as ~0.80 (Rijsdijk et al. 2002).

As we described above, an important principle in contemporary behavior-genetic research is that the magnitude of a biometric variance component may depend upon other variables (moderators). The heterogeneity of heritability estimates for GCA across studies may reflect the influence of such moderators. The role of one of them, age, has been well replicated: the general trend is that, from early childhood through late adolescence or early adulthood, IQ's heritability increases while its shared-environmentality decreases (Bouchard and McGue 2003; Deary et al. 2006). A more tentative biometric moderator is family-of-origin SES. Two theoretical perspectives—those of Sandra Scarr (1992) and of Bronfenbrenner and Ceci (1994)—predict that cognitive abilities will be more heritable among children from higher-SES families. The two theories make that prediction for somewhat different reasons: Scarr's theory emphasizes active gene–environment correlation ($r_{GE}$; Plomin et al. 1977), whereas Bronfenbrenner and Ceci emphasize parental facilitation of "proximal processes" in development. Scarr (Scarr-Salapatek 1971) was the first to investigate whether the heritability of children's GCA might vary as a function of their family SES. This and other earlier studies (Fischbein 1980; Van den Ooord and Rowe 1998; Rowe et al. 1999) are reviewed in Supplementary Note #1 (Online Resource).

## Turkheimer et al. (2003): $A \times SES$ and $C \times SES$ effects

In an important study that has generated much interest, Turkheimer et al. (2003) applied the continuous-moderator model of Purcell (2002) in a small sample of 319 pairs of 7-year-old twins. They found that the biometric decomposition of FSIQ (from the Wechsler Intelligence Scale for Children) varied as a function of parental SES. They operationalized SES as a composite of parental education, income, and occupational status. At the upper extreme of the SES variable, IQ variance decomposed into ~ 80 % additive-genetic variance and near-zero shared-environmental variance, whereas at the lower extreme of the SES variable, it decomposed into near-zero additive-genetic variance and ~ 60 % shared-environmental variance. Further, unshared-environmental variance decreased with SES. However, judging by what is mentioned in the title and abstract of Turkheimer et al.'s article, it is the moderation of genetic variance (a specific form of $G \times E$, which we will designate as $A \times SES$) that is of primary interest, with the moderation of shared-environmental variance (shorthand, $C \times SES$) of secondary interest.

It is important to recognize that, since family SES is the same for both twins in a pair, irrespective of zygosity, it is effectively part of the shared environment as far as twin models are concerned. However, the association between family SES and children's GCA is surely at least partly genetically mediated, as evident from the larger associations between family characteristics and offspring ability in biological families vis-à-vis adoptive families (e.g., Scarr and Weinberg 1978; Kirkpatrick et al. 2009). This is an example of passive $r_{GE}$ (Plomin et al. 1977): parental cognitive ability and SES are positively correlated, and higher-ability parents pass on their trait-relevant genes to their children as well as provide them with an enriched rearing environment. Because $r_{GE}$ can result in spurious detection of $G \times E$, it is advisable to incorporate SES into the moderation model as a fixed regressor, which will partial out any phenotypic variance due to correlation between SES and $A$ (Purcell 2002; Van der Sluis et al. 2012).

We are aware of five studies of GCA interpretable as attempts at replicating Turkheimer et al.'s (2003) $A \times SES$ and $C \times SES$ effects (Harden et al. 2007; Van der Sluis et al. 2008; Grant et al. 2010; Hanscombe et al. 2012; Bates, Lewis, and Weiss 2013). The effects' replication record among these studies is mixed, possibly due to heterogeneity among the studies with respect to participant age (child, adolescent, adult) and country (USA, UK, Netherlands). The studies also vary with regard to how SES was operationalized. SES is not completely temporally stable[1]; parental income and occupational status, in particular, can change with the vicissitudes of the labor economy. Only Hanscombe et al. had the advantage of repeated measures of SES (although the adult participants in Grant et al.'s (2010) study were asked for the highest education level their parents ever achieved). Details concerning the five replication studies are available in Supplementary Note #2 (Online Resource).

Of course, the $A \times SES$ and $C \times SES$ effects could be spurious. The $A \times SES$ element seems less plausible from sample-size considerations alone, since it is only supported in samples of fewer than 1000 twin pairs (Turkheimer et al. 2003; Harden et al. 2007; Bates et al. 2013). Several phenomena can lead to detection of spurious $G \times E$. One of these is differential heritability (or shared-environmentality) by phenotype level. If the influence of $A$ increases, or the influence of $C$ decreases, with increasing GCA, then this heterogeneity may appear to be a biometric moderation effect of SES, simply because SES and GCA are positively correlated.[2] Another complication is if there is greater assortative mating for GCA at lower SES levels (Loehlin

[1] We are grateful to two anonymous referees for calling to our attention the points made in this paragraph concerning stability of SES.

[2] See (Tucker-Drob et al. 2009) and McCallum and Mar (1995) for discussion of how quadratic trends may be mistaken for multiplicative interactions.

et al. 2009). Because assortative mating deflates twin-based estimates of additive-genetic variance and commensurately inflates estimates of shared-environmental variance, it would then appear that additive-genetic variance increases with SES.

Finally, there is the issue of the specificity of biometric-moderation effects. Under the continuous-moderator model, biometric moderation may be thought of simply as heteroskedasticity in the regression of the phenotype onto the putative moderator. The specificity issue concerns how well an analysis can resolve which and how many biometric-moderation effects are nonzero—that is, which biometric variance components are heteroskedastic. Purcell (2002) remarked on this issue when discussing a substantial estimate of a $C \times SES$ effect in simulated data when the true generating model only had an $A \times SES$ effect. Both Turkheimer et al. (2003) and Hanscombe et al. (2012) refer to the issue as well. It is also evident in Harden et al.'s Table 5: the estimate of the $A \times SES$ effect when the $C \times SES$ effect was fixed to zero was very similar to the estimate of the $C \times SES$ effect when the $A \times SES$ effect was fixed to zero. This calls to mind an essential fact: inference about a parameter from a given model is *ipso facto* model-dependent; specifically, it depends upon which other parameters are free to be estimated in the model at hand. We wish to bring attention to one specific questionable practice that is widespread in behavior genetics: that of selecting one single model that is best (by some criterion), and then basing inference only on that model, as though no others had ever been considered. Breiman (1992, p. 738) has called this practice "a quiet scandal." We instead provide an alternative approach: inference about parameters can be based on *multiple* models; in fact, one can (in a sense) select many or even *all* models under consideration, each only to the extent that it is supported by the data. Much of the present study is conducted using methods of multimodel inference. Awareness of these methods is not as widespread as we believe it should be, which is why we describe them in the Appendix. Our description mostly follows that of Burnham and Anderson (2001, 2002, 2004), whose work we recommend for further details.

### Study overview

Our study, which attempts to replicate the $A \times SES$ and $C \times SES$ effects of Turkheimer et al. (2003), improves upon previous replication attempts in several ways. First, our large sample is composed of twins, non-twin biological siblings, and adoptive siblings, assessed at a range of ages spanning the teenage years. A prior study of IQ in a substantially identical sample has been reported (Kirkpatrick et al. 2009). The presence of adoptees provides us with a "backstop" against artifacts stemming from passive $r_{GE}$ and assortative mating, and allows us to directly estimate shared-environmental variance (and, in principle, variance due to covariance between the $A$ and $C$ factors). Second, we also have parental phenotype—IQ scores for the parents of the twins and siblings—and therefore can estimate assortative mating, both SES-independent and SES-dependent. Third, we have data on the same three SES indices used in the original Turkheimer et al. (2003) report.

Our primary analysis attempts to replicate the $A \times SES$ and $C \times SES$ moderation effects. We will compare performance of SES-moderation models when the age-moderation effects established in the literature, $A \times Age$ and $C \times Age$, are included versus when they are not. In addition, we conduct three preliminary analyses prior to the primary analysis, and one exploratory analysis subsequently to it. The first preliminary analysis serves to test for a source of spurious moderation effects, SES-dependent assortative mating among parents (i.e., IQ correlation between mothers and fathers being dependent upon their SES). In our second preliminary analysis, we identify the sources of variance that should be represented in our model. The ACE model is quite plausible a priori from existing literature (reviewed above), especially for an adolescent (rather than adult) sample, and in light of the dearth of evidence for non-additive genetic variance in the domain of cognitive abilities (Bouchard 2004). However, we can estimate more than two sources of familial variance in our sample. One possibility would be twin-specific environmental effects ("twin effects"), which would contribute to between-family variance among twins but not among non-twin siblings. Another possible source of variance is assortative mating. We need not assume that the additive-genetic correlation between full siblings is 0.5—we can estimate it from the data, because an ACE model would be identified by MZ twins and adoptees alone. Once the sources of variance are identified, our third preliminary analysis will determine whether the biometric decomposition of IQ varies as a function of trait level (that is, whether the influence of heredity and shared environment differs across the IQ distribution). This constitutes another test for a source of spurious moderation effects. Finally, after the main analysis, we will explore the possibility that the SES-moderation effects are age-dependent; we hypothesize that they will weaken through adolescence.

## Methods

### Sample

The primary sample ($N = 4{,}973$ from 2,494 sibling pairs) consisted of twins from the Minnesota Twin Family Study ("MTFS"; Iacono et al. 1999; Iacono and McGue 2002;

Keyes et al. 2009), and non-twin sibling pairs from the Sibling Interaction and Behavior Study ("SIBS"; McGue et al. 2007). In addition to this primary sample, one of our secondary analyses used a sample of 3,916 parents from MTFS and SIBS. The primary sample is substantially identical to that of Kirkpatrick et al. (2009), and MTFS and SIBS, their cognitive ability testing, and their zygosity determination and inclusion criteria have been described there and elsewhere (e.g., Kirkpatrick et al. 2014). We have therefore relegated many details concerning the sample and measurements to a Supplementary Methods section (Online Resource).

For the present study, we used parental data only from parents who were the "original rearing" parents in the family. Usually, the original rearing parents would be the biological parents of the family's offspring, unless it was known that one of them had limited contact with the children while they were growing up (due to divorce, etc.). In the case of families with only adopted offspring, the original rearing parents would be those with whom the offspring were originally placed for adoption, unless again it was known that one of them had limited contact with the children.

## SES

Our analysis used three family-level SES variables: (1) the higher of the parents' occupational statuses, (2) midparental educational attainment, and (3) annual household income. We only used the occupational and educational data of the original rearing parents. If data were available only for one of the parents, we took that parent's occupation and education as the higher occupational status and the average education level of the couple, respectively. After exclusions, at least one family-level SES variable was observed for 2,501 families.

Mothers' and fathers' occupational status was assessed during the recruitment phone interview with families' mothers. Occupational status was coded on the Hollingshead scale (Hollingshead 1957). We reverse-scored the Hollingshead scale so that higher values, on a scale of 1–7, represent higher status. We coded as missing the occupational status of those who did not work full-time in their reported occupation, those who reported their occupation as "homemaker," and those reported to be retired, disabled, or institutionalized.

Mothers' and fathers' educational attainment was also assessed during the phone interview. We harmonized educational attainment from the slightly different phone interviews given to different subsamples into a five-point scale (1 = less than high school, 2 = high school, 3 = some post-secondary education, 4 = four-year college degree, 5 = graduate/professional degree).

Annual household income was collected by parental report at the intake assessment of MTFS, and at the first follow-up visit of SIBS. Income was measured on an ordinal scale representing income brackets: 0 = "less than $10,000," 1 = "$10,001–$15,000," and so forth, up to a maximum of 12 = "Over $80,000."

Of the 2,501 families, the percentages missing data on each family-level SES variable were 7.4 % for occupational status, 0.6 % for educational attainment, and 8.4 % for household income. Around 85 % of families had no missing observations, 14 % had one missing observation, and 1 % had two missing observations. As did Turkheimer et al. (2003) and Myrianthopoulos and French (1968), we converted each family's score on the three SES variables into a cumulative proportion (from that variable's empirical CDF), and then averaged the available proportions, producing an SES score for each family (if only one proportion was available, it was taken as the family's SES score). There were 2,494 families having both an SES score and FSIQ for at least one of the offspring. There were 2,382 families in which SES and at least one parent's FSIQ score were available.

## Analyses

Unless stated otherwise, all analyses were conducted in *OpenMx* (Boker et al. 2011), via full-information maximum-likelihood (FIML) estimation from raw data. In most of our analyses, the endogenous variable was offspring IQ, which is assumed to follow a bivariate normal distribution (conditional on age, sex, and SES).

For model comparison and multimodel inference (see Appendix), we used Hurvich and Tsai's (1989) sample-size-corrected version of Akaike's Information Criterion, AICc:

$$AICc = -2logL\left(\hat{\theta}|M, x\right) + 2k + \frac{2k(k+1)}{N-k-1} \quad (1)$$

In large samples, AICc differs little from AIC. However, some (e.g., Burnham and Anderson 2004) argue that AICc should always be used in practice, and that AIC's reputation for overfitting has resulted partly from failure to use AICc in simulation studies.

We proceeded by fitting models, assessing model performance via AICc, and using the performance of previously fitted models to guide specification of subsequent ones. We refrain from reporting parametric inference until all models informative about a particular parameter have been fitted, and—so that Akaike weights can be used—until all AIC-comparable models have been fitted as well. At that point, if more than one model informative about the

parameter had been fitted, we computed model-averaged point estimates, with confidence intervals and $p$ values[3] from the model-averaged standard error, under the assumption of normal sampling distribution. We also obtained a 95 % confidence set for the best-approximating model. Details concerning Akaike weights, model-averaging, and the confidence set are provided in the Appendix. *Very* briefly, Akaike weights are AICcs transformed to proportions so that smaller AICcs have larger weights. These weights are used to compute averages of parameter estimates and their standard errors across models. The confidence set is expected to contain the "best" model with probability 0.95 over repeated sampling, and helps to quantify model-selection uncertainty due to sampling error.

## Results

We first estimated IQ standard deviations and sibling correlations, separately by family type, while correcting for age and sex (McGue and Bouchard 1984), which is especially important in the present case since members of a sibling pair from SIBS were not necessarily the same age and sex, whereas MTFS twins were. From these estimates (Table 1), we can see that the DZ-twin correlation and SD were greater than those of the non-twin full sibs, suggesting the possibility of twin effects. The presence of adoptees also enables us to estimate $r_{GE}$. However, it is evident that the phenotypic variance among adoptees was greater, not less than, the variance among biological offspring (which includes the twins), in which case the estimated correlation between $A$ and $C$ ($r_{AC}$) would be *negative*—in other words, that a typical person's genes and shared environment affect IQ in opposite directions. On its face, this is a difficult conclusion to accept. We therefore decided not to fit any models including an $r_{AC}$ estimate. In any event, the four standard-deviation parameters in Table 1 were not significantly different from one another (LRT $\chi^2(3) = 4.25$, $p = 0.2353$), which is not suggestive of significant $r_{AC}$.

[3] We consider effect sizes and their interval estimates to be more scientifically interesting and informative than hypothesis tests. However, our confidence intervals only have a *marginal* 95 % coverage probability; their joint coverage probability is presumably smaller. Also, not every free parameter we estimated is an easily interpretable effect size, and further, the null hypothesis is indeed of interest and somewhat plausible for certain parameters. We therefore report $p$-values as well, and when making decisions about null hypotheses, compare them to the conventional significance level of $\alpha = 0.05$. $P$ values are also easier than confidence intervals for the reader to adjust for "multiple testing." We report 17 of them altogether. A Bonferroni correction would certainly be conservative, but skeptical readers are free to hold our results to its standard of $\alpha = 0.0029$.

### Preliminary analyses

To test for SES-dependent assortative mating, we modeled parental IQ with a bivariate normal distribution, having a different mean (which was conditioned on SES via regression) and standard deviation for mothers and fathers. We fit two models, one in which the spousal correlation was allowed to vary linearly with SES, and one in which it was constant with respect to SES. The former model estimated that the spousal correlation would be 0.41 at the bottom of the SES distribution, and 0.30 at the top—a change of −0.11 (95 % CI: −0.29, 0.06), which was not statistically distinguishable from zero (LRT $\chi^2(1) = 1.56$, $p = 0.2114$). The estimate of the spousal correlation from the latter model (constant across SES) was moderate, and very close to the meta-analytic average reported over 30 years ago (Bouchard and McGue 1981): $r = 0.35$ (95 % CI: 0.30, 0.39). Obviously, it differed significantly from zero (LRT $\chi^2(1) = 186.43$, $p = 1.908 \times 10^{-42}$).

This analysis indicated that parental assortative mating is moderate in magnitude, and is not SES-dependent, which rules out one possible source of spurious $G \times E$. As explained by Kirkpatrick et al. (2009, footnote 1), if we assume a high heritability for adult IQ, that spouses select mates for psychometric IQ *per se*, and that the phenotypic spousal correlation perfectly reflects a genetic spousal correlation, then the classical twin model would underestimate heritability by about 28 %, and commensurately overestimate shared-environmentality. However, these are "worst-case scenario" assumptions, and are generally not true. Further, in our dataset, the ACE variance components are identified by the adoptees and MZ twins alone, whose covariances are not affected by the true genetic correlation between full siblings. As described in the next section, we actually estimated this genetic correlation.

To decide which sources of variance to include in our biometric models, we fit Models #1 through #4 (collectively, "Block #1"). These four models represented the four combinations of the twin effects path $\gamma_{T0}$ fixed (to zero) versus free, and full-sib genetic correlation $r_A$ fixed (to 0.5) versus free. All four included the main effects of sex and age. Additionally, we estimated a separate intercept ($\beta_0$) for twins, biological SIBS offspring, and adoptees, and a separate SES main effect ($\beta_{SES}$) for biological offspring (including twins) and adoptees. If we were to apply the ACE model to our dataset without SES main effects, variance due to SES *per se* would otherwise be variance due to $C$, but this is not the case for variance due to $A$-SES correlation, as it does not contribute to variance among adoptees nor to covariance between unrelated siblings reared together. Hence, estimating separate SES main effects for biological children and adoptees is a prudent

**Table 1** Age- and Sex-corrected FSIQ correlations and standard deviations, by type of sibling relationship

| | MZ twins | DZ twins | Ado sibs | Bio sibs | Mixed sibs |
|---|---|---|---|---|---|
| SD (SE) | 13.69 (0.25) | 13.72 (0.30) | 14.09 (0.38) | 12.92 (0.42) | [a] |
| *r* (SE) | 0.77 (0.01) | 0.50 (0.03) | 0.11 (0.06) | 0.36 (0.06) | 0.24 (0.07) |

*Ado* both siblings adopted, *Bio* both siblings are biological offspring of parents, *Mixed* one sibling is adopted and one is biological offspring

[a] In mixed families, the standard deviation of biological offspring was constrained equal to that of bio sibs, and the standard deviation of adoptees was constrained equal to that of ado sibs

**Table 2** Model-fitting results from Block #1: AICcs (underline) and Akaike weights

| | Free $\gamma_{T0}$ | Fix $\gamma_{T0} = 0$ |
|---|---|---|
| Free $r_A$ | 38631.89 | 38629.88 |
| | $6.75 \times 10^{-6}$ | $1.85 \times 10^{-5}$ |
| | (Model #1) | (Model #2) |
| Fix $r_A = 0.5$ | 38629.91 | 38627.9 |
| | $1.82 \times 10^{-5}$ | $4.99 \times 10^{-5}$ |
| | (Model #3) | (Model #4) |

The models in this block differ by whether they have $r_A$ and $\gamma_{T0}$ as fixed or free parameters. $r_A$ is the correlation between latent factors $A_1$ and $A_2$ for full siblings (including DZ twins). $\gamma_{T0}$ is the path loading for twin-specific environmental effects. AICcs are underlined; Akaike weights are proportions. Smaller AICcs and greater Akaike weights both correspond to a more-preferable model. A model's Akaike weight is interpretable as the posterior probability that the model is the best at approximating full reality in the population, given the size of the sample and the set of models under consideration (see Appendix)

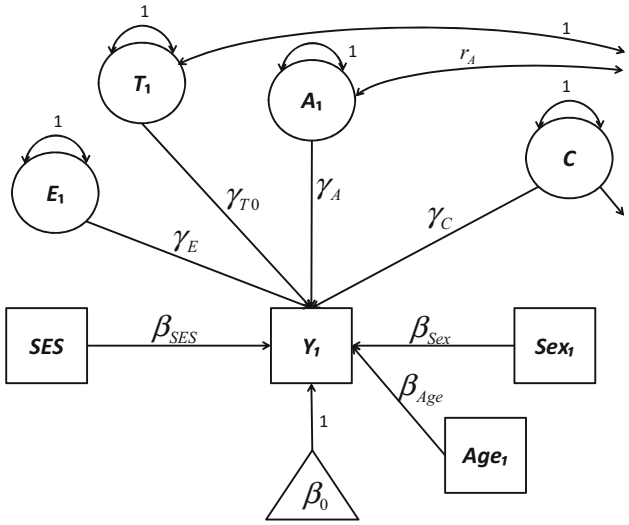


**Fig. 1** Biometric moderation model—general form for Blocks #1, #2, and #3. For ease of presentation, only twin #1's side of the diagram is shown. The path loadings onto the latent *A*, *C*, and *E* variables are allowed to depend upon moderators, and might be written thus: $\gamma_A = \gamma_{A0} + \gamma_{A1}(Age_1) + \gamma_{A2}(SES) + \gamma_{A3}(SES \times Age_1)$, $\gamma_C = \gamma_{C0} + \gamma_{C1}(Age_1) + \gamma_{C2}(SES) + \gamma_{C3}(SES \times Age_1)$, and $\gamma_E = \gamma_{E0} + \gamma_{E1}(Age_1) + \gamma_{E2}(SES)$. For example, $\gamma_{A0}$ is the main effect of *A*, $\gamma_{A1}$ is the $A \times Age$ effect, $\gamma_{A2}$ is the $A \times SES$ effect, and $\gamma_{A3}$ is the $A \times Age \times SES$ effect. In Block #1, only main effects ($\gamma_{A0}$, $\gamma_{C0}$, $\gamma_{E0}$, $\gamma_{T0}$) were estimated. In Block #2, moderation effects of age ($\gamma_{A1}$, $\gamma_{C1}$, $\gamma_{E1}$) and SES ($\gamma_{A2}$, $\gamma_{C2}$, $\gamma_{E2}$) were introduced, and in Block #3, the interactions ($\gamma_{A3}$, $\gamma_{C3}$) were introduced. The twin-effects parameter $\gamma_{T0}$ was only ever estimated in Block #1, and the loading onto *T* was never conditioned on moderators. Separate values of $\beta_0$ were estimated for twins, biological SIBS offspring, and adoptees. Separate values of $\beta_{SES}$ were estimated for biological offspring of parents (including twins) and for adoptees

way to control for *A*-SES correlation (a form of passive $r_{GE}$). Since the association between parental SES and offspring IQ partly reflects this correlation in the case of biological children, but not for adoptees, we naturally anticipate a larger main effect of SES for biological offspring.

The four models' AICcs are presented in Table 2. Because Block #1 was the first part of a series of comparable models (the general form of which is depicted in Fig. 1), Table 2 also includes their Akaike weights, which are calculated relative to the AICcs of all models in this comparable set. From Table 2, it can be seen that the best-approximating model within this block is #4, which has both $r_A$ and $\gamma_{T0}$ fixed to their null values. As anticipated, we conclude from Block #1 that the biometric ACE components are sufficient to describe our data, and that fixing both $r_A$ and $\gamma_{T0}$ to their null values improves model efficiency. In the previous section, we conclude that the spousal correlation for IQ is not SES-dependent, and we report here that the genetic correlation for full sibs differs unimportantly from 0.5. On the basis of the foregoing, we resolved here to assume in further analyses that the effects of assortative mating are negligible.

The models of Block #1 are the only ones that provide single estimates for the ACE variance components, since models with a biometric-moderation effect estimate, in a sense, different component values at different levels of the moderator. The model-averaged point estimates from Block #1 of additive-genetic variance, shared-environmental variance, and unshared-environmental variance are respectively 109.06, 23.25, and 43.19, which sum to total residual variance 175.50, and respectively yield standardized estimates of 0.62, 0.13, and 0.25.

To assess whether the influence of heredity and shared environment depend upon trait level, we used DeFries-Fulker regression (DeFries and Fulker 1985, 1988) with double-entered data (Rodgers & McGue 1994; Rodgers and Kohler 2005) which has been used for similar purposes in other studies (e.g., Cherny et al. 1992). With double-entered data, phenotype scores are mean-centered within kinship groups, and then each sibling pair (twins being a

special case of siblings) is entered into the dataset twice, with the labels "sibling #1" and "sibling #2" reversed for each entry. Since our data support the use of a model with the ACE biometric components, the DeFries-Fulker regression equation we used is

$$K_1 = b_1 K_2 + b_2(K_2 R) + b_3\left(K_2^2\right) + b_4\left(K_2^2 R\right) + b_5(Age_1) + b_6(Sex_1) \quad (2)$$

where $K_1$ is the phenotype score of sibling #1, $K_2$ is the phenotype score of sibling #2, $R$ is the coefficient of relationship (1 for MZ twins, 0.5 for full siblings, and 0 for adoptive siblings), $Age_1$ is the age of sibling #1, and $Sex_1$ is a dummy variable for whether or not sibling #1 is female. In this model, the interaction coefficients $b_3$ and $b_4$ represent how much the shared-environmentality and heritability, respectively, depend upon trait level.

This DeFries-Fulker regression requires complete data within sibling pairs. There were 2,479 pairs in which FSIQ was available for both members. We conducted the regression represented by Eq. (2) via an implementation of Kohler and Rodgers' (2001) "efficient DF estimation" in the R statistical computing language. The interaction estimates were both small and statistically indistinguishable from zero: $\hat{b}_3 = -2.87 \times 10^{-5}$ (95 % CI: $-2.36 \times 10^{-3}$, $2.30 \times 10^{-3}$; $p = 0.9807$) and $\hat{b}_4 = 7.40 \times 10^{-4}$ (95 % CI: $-1.72 \times 10^{-3}$, $3.21 \times 10^{-3}$; $p = 0.5560$). Further, the joint test of the two interactions was not significant (Wald $\chi^2(2) = 1.05$, $p = 0.5913$). This DeFries-Fulker regression required exclusion of incomplete sibling pairs, and was only informative about the standardized, not raw, additive-genetic and shared-environmental variance components. Nonetheless, we regard it as reasonably good evidence that the additive-genetic and shared-environmental components do not linearly vary across the FSIQ continuum, ruling out another possible source of spurious $G \times E$.

## Primary analysis: can we replicate SES-moderation effects?

To address our research question, we fit Block #2, consisting of Models #5 through #19. These models comprise the eight combinations of $A \times SES$, $C \times SES$, and $E \times SES$ effects being included or excluded. Each such combination was fitted twice: once including $A \times Age$ and $C \times Age$ effects, and again with them dropped.

The AICcs and Akaike weights of this block are reported in Table 3, from which we draw several conclusions. For one, the inclusion of *any* kind of SES-moderation effect improved model efficiency, indicating that the regression of IQ onto SES is heteroskedastic. More importantly, models that included an $A \times SES$ effect clearly fared better than those that did not, and those that included an $E \times SES$ effect fared slightly better than those that did not. But, the $C \times SES$ effect appeared quite extraneous. Further, the AICc rank-orders within each column of Table 3 are nearly identical, indicating that the SES-moderation effects' contributions to relative model efficiency depended little on whether or not age-

**Table 3** Model-fitting results of Block #2: AICcs (underlined) and Akaike weights

| SES-moderation effects | Age-moderation effects | |
|---|---|---|
| | AC | None |
| ACE | 38612.34 | 38613.93 |
| | 0.1188 | 0.0538 |
| | (Model #5)[b] | (Model #13)[b] |
| AC | 38612.75 | 38614.08 |
| | 0.0972 | 0.0498 |
| | (Model #6)[b] | (Model #14)[b] |
| AE | 38612.21 | **38611.91** |
| | 0.1268 | **0.1478** |
| | (Model #7)[b] | **(Model #15)**[b] |
| CE | 38615.26 | 38616.38 |
| | 0.0277 | 0.0158 |
| | (Model #8)[b] | (Model #16) |
| A | 38612.46 | 38612.1 |
| | 0.1120 | 0.1339 |
| | (Model #9)[b] | (Model #17)[b] |
| C | 38619.31 | 38619.78 |
| | 0.0036 | 0.0029 |
| | (Model #10) | (Model #18) |
| E | 38621.47 | 38620.83 |
| | 0.0012 | 0.0017 |
| | (Model #11) | (Model #19) |
| None | 38628.91 | 38627.9 |
| | $3.00 \times 10^{-5}$ | $4.99 \times 10^{-5}$ |
| | (Model #12) | (Model #4)[a] |

AICcs are underlined; Akaike weights are proportions. Smaller AICcs and greater Akaike weights both correspond to a more-preferable model. The overall preferred model, #15, is bolded. A model's Akaike weight is interpretable as the posterior probability that the model is the best at approximating full reality in the population, given the size of the sample and the set of models under consideration (see Appendix). "Age moderation effects" are those latent biometric factors the loadings of which were allowed to be moderated by age; "none" indicates that no age-moderation effects were included, whereas "AC" indicates that both $A \times Age$ and $C \times Age$ effects were included. "SES Moderation Effects" are those latent biometric factors the loadings of which were allowed to be moderated by SES. For example, the models in the row marked "CE" included $C \times SES$ and $E \times SES$ effects

[a] Model #4 is part of Block #1 (see Table 2)

[b] Model is in the 95 % confidence set for best-approximating model (see Appendix)

moderation effects were included. From these results, we concluded that our data support only an $A \times SES$ effect, but not a $C \times SES$ effect.

### Exploratory analysis

Perhaps the $A \times SES$ effect apparent in our data weakens with age. Perhaps there is a small $C \times SES$ effect lurking in our data that is only operative among younger participants. Certainly, if these SES-moderation effects decline with age, it would help to explain why attempts to replicate them in adults (van der Sluis et al. 2008; Grant et al. 2010) failed. To investigate these possibilities, we fit Block #3, composed of Models #20 through #22. Both age- and SES-moderation effects for $A$ and $C$ should be included, since we were considering the moderation effects of an age $\times$ SES interaction. We also included the SES-moderation effect on $E$, since it received limited support in Block #2. Model #20 included the $A \times Age \times SES$ and $C \times Age \times SES$ effects, Model #21 only the former, and Model #22 only the latter. Except where these three-way interactions are concerned, we do not utilize point estimates or standard errors from Models #20, #21, and #22 in model-averaging, partly because of these models' exploratory nature, but primarily because the parameters of greatest interest in our study are age- and SES-moderation effects, which lose interpretability once the three-way interactions are included.

The three models' AICcs and Akaike weights are reported in Table 4. None of the interaction effects contributed to model performance. On this basis alone, we conclude that there is no age-dependent SES-moderation. But, we are now ready to draw inferences about those interaction parameters, and a number of other parameters of interest as well.

**Table 4** Model-fitting results of Block #3

| Model number (Free interaction parameters) | AICc | Akaike weight |
|---|---|---|
| Model #20 ($A \times Age \times SES$, $C \times Age \times SES$) | 38616.27 | 0.0167 |
| Model #21 ($A \times Age \times SES$ only) | 38614.25 | 0.0457 |
| Model #22 ($C \times Age \times SES$ only) | 38614.31 | 0.0444 |
| Model #5[a] (none) | 38612.34 | 0.1188 |

A model's Akaike weight is interpretable as the posterior probability that the model is the best at approximating full reality in the population, given the size of the sample and the set of models under consideration (see Appendix). Smaller AICcs and greater Akaike weights both correspond to a more-preferable model

[a] Model #5 is part of Block #2 (see Table 3)

[b] Model is in the 95 % confidence set for best-approximating model (see Appendix)

### Overall results

Table 5 lists model-averaged parameter estimates, plus corresponding confidence intervals and $p$ values based on the assumption of normal sampling distribution. The estimates of neither three-way interaction from Block #3 differed significantly from zero. Consistent with existing literature, we did observe a significant increase in additive-genetic variance, and a significant decline in shared-environmental variance, with increasing age. Most interestingly, we replicated only the $A \times SES$ effect of Turkheimer et al. (2003): additive-genetic variance varied positively with family SES. The $C \times SES$ effect was not in the hypothesized direction and was estimated with little statistical precision. Finally, although the AICcs provided some support for an $E \times SES$ effect, the model-averaged results show that we do not have sufficient evidence to conclude that it differs from zero.

Although model-averaging is well-suited for inference about one parameter at time, it does not necessarily make for easy interpretation. Consider the model-averaged estimate of the $A \times SES$ effect, 2.969. Because the SES variables were scaled to the interval [0, 1], this value means that for the highest-SES families, the loading onto $A$ is greater than that for the lowest-SES families by 2.969. But to really interpret this value, one would need a value for the main-effect of $A$, which is not a parameter of interest. Sometimes, a meritorious model can tell a complete story in a way that model-averaging cannot easily do. For this reason, we also report point estimates and standard errors from the two most AICc-favored models, Model #15 ($A \times SES$, $E \times SES$, no age-moderation) and Model #17 ($A \times SES$ only), respectively (Table S2 in Online Resource). It can be seen that the model-conditional estimates of free parameters do not differ drastically from the corresponding model-averaged estimates. In Fig. 2, we graph how the biometric decomposition would vary by SES according to the estimates from Model #15, expressed in raw variance components and in normalized variance proportions.

## Discussion

Guided by existing data and theory, we fit a number of biometric models to a relatively large dataset collected from twins, non-twin biological siblings, and adoptive siblings. We compared models by a sample-size corrected version of AIC, the AICc (Hurvich and Tsai 1989). We compared

**Table 5** Multimodel inference from Blocks #1 through #3

| Parameter | Point estimate (CI) | *p* value |
|---|---|---|
| Full-sib genetic correlation ($r_A$) | 0.489 (0.380, 0.599) | 0.8483[a] |
| Twin effects ($\gamma_{T0}$) | $1.27 \times 10^{-5}$ (−3.593, 3.593)[b] | 1.000 |
| $A \times SES$ Effect | 2.969 (1.095, 4.843) | 0.0019 |
| $C \times SES$ Effect | 1.299 (−2.762, 5.360) | 0.5307 |
| $E \times SES$ Effect | 0.891 (−0.244, 2.027) | 0.1264 |
| $A \times Age$ Effect | 0.318 (0.024, 0.611) | 0.0339 |
| $C \times Age$ Effect | −1.437 (−2.254, −0.621) | 0.0006 |
| $A \times SES \times Age$ Effect | 0.160 (−0.383, 0.703) | 0.5635 |
| $C \times SES \times Age$ Effect | −0.272 (−1.794, 1.250) | 0.7260 |
| SES Main Effect, adoptees ($\beta_{SES,A}$) | 6.961 (1.617, 12.305) | 0.0107 |
| SES Main Effect, bio offspring ($\beta_{SES,B}$) | 16.047 (13.892, 18.202) | $3.073 \times 10^{-48}$ |

Models #20, #21, and #22 (Block #3) are only included in calculating model-averaged inference for the three-way interactions ($A \times SES \times Age$ and $C \times SES \times Age$; explanation in text). Otherwise, point estimates and standard errors for each parameter were calculated from all models among Models #1 through #19 in which the parameter was freely estimated. Confidence intervals and *p* values were calculated from point estimates and standard errors, assuming a normal sampling distribution. Signs on moderation effects are reported so that a negative value indicates that the loading on the latent biometric factor becomes more negative as the moderator becomes more positive

[a] Null parameter value for $r_A$ is 0.5

[b] The sign of the twin-effects parameter ($\gamma_{T0}$) is arbitrary, since the actual corresponding variance component is $\gamma_{T0}^2$. The 95 % profile-likelihood confidence interval for $\gamma_{T0}^2$, from Model #3 ($\gamma_{T0}$ free, $r_A$ fixed), is (0, 12.27)

models' AICcs to first resolve basic questions of specification, then to attempt to replicate the SES effects of primary interest, and finally to explore the possibility of age-dependent SES effects. We first resolved that an a priori plausible ACE model would suffice for our purposes, and that the effects of assortative mating and of differential heritability/shared-environmentality by trait level were negligible. We fit models with various SES-moderation effects, both including and excluding two age-moderation effects identified in the literature. We observed support for the hypothesized $A \times SES$ effect, weakly suggestive evidence of an $E \times SES$ effect, but none for the hypothesized $C \times SES$ effect. Our exploratory analysis did not provide any evidence for age-dependent SES-moderation effects. Thus, our study shows that additive-genetic variance in GCA increases with family-of-origin SES. This replication of the $A \times SES$ effect is robust to model specification: what all models belonging to the 95 % confidence set (marked with superscript "b" in Tables 3 and 4), save one, have in common is a free $A \times SES$ parameter. The effect is also statistically significant (Table 5): it would survive conservative Bonferroni correction for the 17 *p*-values we report.

The $A \times SES$ and $C \times SES$ interactions from Turkheimer et al. (2003) are the biometric-moderation effects of primary interest in this study, and although they have generated much interest, they have not been replicated together in any study of general cognitive ability applying Purcell's continuous-moderation model. They have failed replication twice (Grant et al. 2010; van der Sluis et al. 2008), and the $C \times SES$ component has been replicated once (Hanscombe et al. 2012). Our study constitutes the third replication of the $A \times SES$ element, after Harden et al. (2007) and Bates et al. (2013). Interestingly, the $A \times SES$ effect has only been observed in U.S. samples in which parental income was available as an SES variable. It has not replicated in European samples nor in an American sample in which only parental education was available. In public health, it has been shown that income and education each provide different information about health-relevant aspects of an individual's SES, and are usually not so highly correlated that entering both into a regression analysis produces multicollinearity problems (Braveman et al. 2005). Further, a given SES variable's relations with other variables can differ by country, and by demographic strata and regions within countries (Uher et al. 2006; Braveman et al. 2005). Possibly, the $A \times SES$ effect is a distinct moderation effect of family income in the United States. More research is needed to evaluate this tentative proposition. In the present study, we could have conducted analyses to gauge how much each of the three SES variables contributed to the $A \times SES$ effect. However, this would be a greater undertaking than it might seem, since rigorously gauging variables' relative importance can be rather involved in multiple regression (Azen and Budescu 2003), let alone in a structural equation model involving interactions with latent variables.

Our study, Bates et al. (2013), Harden et al. (2007), and Turkheimer et al. (2003) were all conducted in samples of

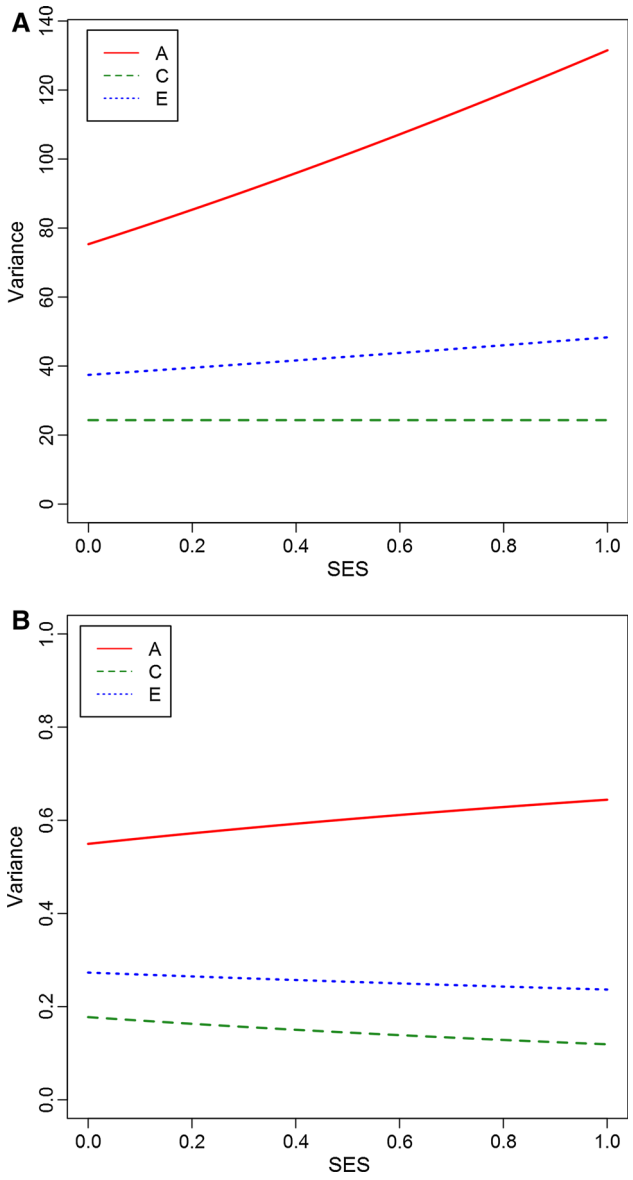


**Fig. 2** Biometric variance components (**a**) and variance proportions (**b**) as function of SES, based on estimates from best-approximating Model #15. At a given point on the abscissa in panel **b**, the ordinate positions of each curve sum to unity. SES is a composite of parental educational attainment, parental occupational status, and household income, transformed to cumulative proportions (mean = 0.58, SD = 0.24). Model #15 included $A \times SES$ and $E \times SES$ effects

Americans in which parental income was available, but the $C \times SES$ effect only occurs alongside the $A \times SES$ effect in the original 2003 study. We offer a speculative explanation for why this is so. Our sample, Harden et al.'s, Bates et al.'s, and Grant et al.'s (2010) are predominantly Caucasian, but Turkheimer et al.'s is mostly (54 %) African-American. Perhaps low SES is not enough to produce the extreme deprivation that, according to Scarr (1992), is necessary to amplify the differential effect of the rearing environment; perhaps low SES must be combined with membership in a disadvantaged minority group whose place in and experience of American society is unique due to the historical legacy of slavery.

The fact that the $A \times SES$ effect has failed replication in adults suggests that it could be age-dependent. But, Hanscombe et al.'s (2012) graphs and point estimates show no clear age-related trend; further, we tested this hypothesis directly, and it was not supported. The availability of IQ data at different ages, which allowed us to directly estimate the age-dependence of SES-moderation effects, is one of several advantages our study has over some existing ones. Another advantage is that we were able to empirically check for possible sources of spurious results, including assortative mating, and differential heritability/shared-environmentality by trait level. Still another advantage was the availability of adoptees, whose data are informative about shared-environmental variance, without bias due to assortative mating, passive $r_{GE}$, or violations of the "equal environments assumption" for twins. We were also able to calculate different SES main effects for adoptees and biological children. The one for adoptees shows that family SES has a moderate, *environmental* effect on children's cognitive functioning, equal to a 7-point IQ advantage for children from the highest-SES families versus the lowest-SES families. Finally, we consider our use of multimodel inference to be a major advantage of our study, because it enables us to produce point estimates and confidence intervals based on all fitted models informative about a parameter, each to the extent that AICc favors it over others. This avoids the bias resulting from conditioning one's parametric inference only upon a single model (Lukacs et al. 2009).

We wish to temper our endorsement of multimodel inference with a few caveats. First, we must emphasize that Model #15 ($A \times SES$, $E \times SES$, no age-moderation) is not necessarily most likely to be the true model because it has the smallest AICc. Likewise, a model's Akaike weight is not the posterior probability that the model is the true model. AIC is not intended to discover the "true" model in the first place. Instead, as stated by Browne (2000, p. 129), AIC is "not appropriate for selecting the best-fitting model in some general sense independent of sampling error, but…for indicating models whose calibrations can be trusted given a specified sample size."

Second, our conclusions depend upon the candidate set of models under consideration.[4] We wanted to obtain estimates of each SES-moderation effect from models in which other moderation effects were variously present or

[4] Readers certainly can think of models we could have fitted, but did not. Some readers may be interested in Table S3 (Online Resource), which, for the sake of completeness, reports point estimates and standard errors from a post hoc, "full" model in which all parameters under consideration were freely estimated.

absent. We had to balance that objective with the needs to preserve interpretability and a manageable scope, to avoid blindly empirical "data fishing," and keep our analyses relevant to our research objectives. It slightly complicates matters that our candidate model set evolved as our analyses proceeded, in that we used the results from previously fitted models to guide specification of subsequent ones. Also, for the sake of interpretability and maintaining a manageable scope, we proceeded from simpler to more-complicated models. In these respects, our approach bears some resemblance to stepwise forward-selection. However, we deliberately avoided some of the most objectionable aspects of stepwise analyses. We did not conduct a purely data-driven, blindly empirical analysis. Our analysis was guided by subject-matter knowledge, each block of models was intended to address a specific question, and we saved the most exploratory analyses for last. Further, we did not use significance testing for model selection, nor did we base our conclusions solely upon the final model.

One restriction we imposed upon the candidate set is that all the biometric-moderation models we considered are of the form of Purcell's (2002) continuous-moderator model. There are other model formulations arguably more appropriate for estimating $G \times E$ in the presence of $r_{GE}$, such as others described by Purcell (2002), and those of Rathouz et al. (2008) or of Price and Jaffee (2008)—all of which involve biometrically decomposing the putative moderator in some way. We decided to retain the Purcell formulation because existing studies of SES-moderation have used it, and our study is intended as a replication study of Turkheimer et al. (2003). Nonetheless, inclusion of SES main effects in our models is a rather vexing problem. If one thinks of the path diagram in, say, Fig. 1 as a simultaneous regression of IQ onto both observable and latent variables, then clearly the main effect of SES must be included if any interactions of SES with latent variables are to be included as well. With data from twins only, SES will necessarily account for variance otherwise attributable to $C$ (or to $r_{AC}$, which would appear as variance due to $C$). Our data enabled us to separately estimate the $\beta_{SES}$ path coefficient for adoptees and biological offspring; both effect sizes are nontrivial, and possibly, enough shared-environmental variance was partialled out that the $C \times SES$ effect was rendered impossible. On the other hand, including the two SES main effects allows us to be reasonably certain that our $A \times SES$ result is not an artifact of correlation between SES and latent variable $A$. Because we conditioned our models upon SES (as a fixed regressor in the definition of the model-expected phenotypic mean), any phenotypic variance due to SES or to covariance between SES and latent variable $A$ would be partialled out (Purcell 2002; Van der Sluis et al. 2012).

Our study raises several other questions that can guide future research. We have already suggested three: to what extent are SES-moderation effects dependent upon country, SES measure, or ethnic minority status? Future studies could attempt to test specific hypotheses made by the Scarr (1992) and Bronfenbrenner and Ceci (1994) theories about SES-moderation. For instance, Scarr's theory predicts that $C \times SES$ effects are only likely to be observed when the lowest echelons of SES are represented in the sample. Similarly, Bronfenbrenner and Ceci emphasize the importance of environmental stability for effective development. Since family SES is correlated with stability of the rearing environment (Evans 2004), perhaps stability is what really drives SES-moderation effects. It would also be interesting to investigate another correlate of SES—parental *phenotype*, that is, parental cognitive ability—as a biometric moderator. Finally, behavior geneticists could attempt to replicate the $A \times SES$ effect when genetic factors are not latent, but measured as molecular-genetic data. Exciting avenues of $G \times E$ research remain to be explored.

## Appendix: the information-theoretic approach and multimodel inference

Kullback & Leibler's important 1951 paper concerns, *inter alia*, derivation of a metric representing how well one probability distribution is approximated by another. Specifically, it is the expected amount of information (in Kullback & Leibler's generalized Shannon-Wiener sense) lost when one probability distribution is approximated by another. This metric has become known as Kullback–Leibler (KL) divergence. A sensible objective of model selection, then, is to choose the model that has the smallest KL divergence from full reality. Full reality, of course, is not known, and may not even be knowable in principle; possibly, any complete description of full reality would be infinitely long. If we accept the possibility that no statistical model can completely describe full reality, then the premise of a "true model" that generated the data becomes rather dubious. These issues pose no problem, however, if one is only interested in the *relative* divergence of different models, since the unknown constants depending upon full reality cancel out from subtraction.

In a series of important contributions in the 1970 s, Hirotugu Akaike[5] showed that the maximized joint log-likelihood of a model's parameters estimates how relatively "close" (in a KL-divergence sense) the model is to full reality, except that this estimator is biased upward, because it represents the fit of the model in the same data

[5] Unfortunately, several important primary sources by Akaike are inaccessible to us, due to being conference presentations or being written in Japanese. We do not cite sources we cannot read. Here, we rely on secondary sources by Burnham and Anderson (2001, 2002, 2004) and Pawitan (2013).

from which its parameters were estimated. Akaike further showed that, in large samples, the magnitude of this bias is in fact approximately equal to $k$, the number of free parameters. Subtracting $k$ from the loglikelihood thus serves to estimate the expected loglikelihood of the model when "plugging in" parameter estimates previously obtained from a separate, independent sample of the same size. Akaike multiplied this bias-adjusted loglikelihood by $-2$ (to turn it into a bias-adjusted deviance), obtaining what has become known as Akaike's Information Criterion,

$$AIC = -2logL\left(\widehat{\theta}|M,\ x\right) + 2k \tag{3}$$

where $\widehat{\theta}$ is the vector of maximum-likelihood estimates of model $M$'s $k$ parameters, as estimated from dataset $x$. In theory, the candidate model with the smallest AIC is expected to be the model that best approximates full reality, conditional on sample size $N$ and the set of candidate models considered. The expected relative KL divergence of two candidate models may be estimated simply by subtracting their AICs.

As is evident from the previous paragraph, AIC is a penalized fit index. The unpenalized model deviance, $-2logL\left(\widehat{\theta}|M,\ x\right)$, by itself is a poor measure of a model's merit, as it may be made arbitrarily small by adding parameters and increasing model complexity. AIC's penalty is the approximate amount by which model deviance is underestimated when assessing the model in the same sample in which its parameters are being estimated. In other words, AIC has deep theoretical connections to cross-validation (discussed further by Stone 1977; Shao 1997; and Browne 2000). Specifically, in large samples, it is expected to select that model in the candidate set which minimizes error of prediction in new samples of the same size from the population, where error is based on a log-likelihood function (Hastie et al. 2009). Since maximizing normal likelihood is equivalent to minimizing quadratic loss, and since many analyses assume (at least implicitly) a normal distribution, in many contexts AIC is expected to select that model in the candidate set which minimizes *mean squared* error of prediction. We therefore phrase our interpretations of AIC in terms of "efficiency" or "performance"—shorthand for *expected relative* efficiency or performance—rather than "fit," because, again, one can just add more parameters to improve model fit to the data at hand.

However, one of AIC's appealing qualities is that it allows the expected relative efficiency of *all* the models in the candidate set to be compared to one another. Unlike the likelihood ratio test (LRT), AIC can be used to compare multiple models to one another and rank them in terms of their merit; they need not be a sequence of nested models. In fact, different models' AICs will be comparable to one another provided that the models all: (1) are fitted to the same dataset (and in particular, have the same $N$); (2) have the same endogenous variable(s) (which are no longer considered "the same" if they have been transformed); and (3) either have likelihood functions from the same family of distributions *or* use fully normalized densities as likelihoods (Burnham and Anderson 2002).

We now describe how AIC can be used to weight the results of multiple models under consideration, and obtain model-averaged point estimates and sampling variances. Let $AIC_{min}$ denote the smallest AIC in a set of $m$ comparable models. Then, those models' AICs can be re-expressed relative to $AIC_{min}$. For some model $l$, let $\Delta_l = AIC_l - AIC_{min}$. Then, model $l$'s Akaike weight can be calculated as

$$w_l = \frac{\exp(-\Delta_l/2)}{\sum_{i=1}^{m}\exp(-\Delta_i/2)} \tag{4}$$

Do this for all models $l = 1, \ldots, m$. The resulting Akaike weights are normalized (sum to 1); each is interpretable as the posterior probability that its model is the one that minimizes KL divergence from full reality in the population (again conditional on $N$ and the candidate set of comparable models; Burnham and Anderson 2002). The implicit prior probability on each model in the set calculated is not equal for all models. Instead, it is a "savvy prior" that takes into consideration the number of free parameters relative to sample size (see Burnham and Anderson 2004).

Once Akaike weights are computed for all comparable models in the candidate set, a pragmatic way to proceed is to average each parameter's estimates, and their corresponding sampling variances, across those models in which the parameter is free to be estimated[6] (Burnham and Anderson 2002). For purposes of model-averaged estimates, the Akaike weights need to be re-normalized so that they sum to 1 within the subset of models in which the parameter of interest is free. If some parameter $\theta$ is a free parameter in some subset $\mathcal{S}$ of the comparable set of models, then for some model $l$ within that subset, the re-normalized Akaike weight $w_l^*$ equals

[6] It may be objected that basing inference about a parameter only upon those models in which it is freely estimated ignores evidence about the parameter conveyed by those models in which it is fixed. If one's objective is regression prediction rather than inference, Burnham and Anderson (2002) do recommend calculating the model-averaged regression coefficient from models in which it is fixed, as well as those in which it is free. However, as Bartels (1997, footnote 11) points out, a model-averaged estimate computed in this way will not have a normal sampling distribution, which complicates its use for statistical inference.

$$w_l^* = \frac{w_l}{\sum_{i \in \mathcal{S}} w_i} \quad (5)$$

Do this for all models $l$, $l \in \mathcal{S}$. With the re-normalized weights, the model-averaged point estimate of $\theta$ can be calculated:

$$\widehat{\theta}_. = \sum_{i \in \mathcal{S}} w_i^* \widehat{\theta}_i \quad (6)$$

where $\widehat{\theta}_i$ is the maximum-likelihood estimate of $\theta$, conditional on model $i$. In a sense, when computing $\widehat{\theta}_.$, one is "integrating out" the model-dependence of the point estimates by averaging across models informative about the parameter, each contributing to the average in proportion to its relative weight-of-evidence. The model-averaged point estimate $\widehat{\theta}_.$ has estimated sampling variance equal to (Burnham and Anderson 2004):

$$\widehat{\text{var}}\left(\widehat{\theta}_.\right) = \sum_{i \in \mathcal{S}} w_i^* \left[\widehat{\text{var}}_i\left(\widehat{\theta}_i\right) + \left(\widehat{\theta}_i - \widehat{\theta}_.\right)^2\right] \quad (7)$$

where $\widehat{\text{var}}_i\left(\widehat{\theta}_i\right)$ is the estimated sampling variance of the MLE of $\theta$, conditional on model $i$. Thus, the model-averaged sampling variance represents a weighted average of within-model variance estimates and between-model variance estimates. In the simplest application, one uses the square root of $\widehat{\text{var}}\left(\widehat{\theta}_.\right)$ as the standard error to form confidence intervals and test null hypotheses, assuming asymptotic normality of $\widehat{\theta}_.$, which is what we do herein.

The chief advantage of multimodel inference is that it enables the researcher to base inference about parameters on all models under consideration, allowing each model to contribute in proportion to how well it is supported by the data (Burnham and Anderson 2002). Even if, say, the best-approximating model has the shared-environmental effect fixed to zero, it does not necessarily follow that the best estimate of the effect is zero, especially if other models under consideration had AICs close to that of the best model. The multimodel approach attempts to avoid the biased estimation and inference that result from conditioning one's conclusions on a single best model (Lukacs et al. 2009). In applied contexts, information-theoretic model-averaging can also improve predictive accuracy (e.g., Kapetanios et al. 2008).

We acknowledge, though, that model-averaged estimates are not always easily interpretable, whereas a set of parameter estimates, taken together from the single "best" model, can tell a coherent "story," and help the investigator form a gestalt whose whole may be greater than the sum of its parts. But, whatever criteria were used to select the "best" model are prone to sampling error. With this in mind, some way of quantifying model-selection uncertainty is desirable. Akaike weights can be applied to form a confidence set for the best-approximating model, expected to contain, with a given probability over repeated sampling, the model in the candidate set that minimizes KL divergence in the population. For this purpose, we adopt a simple but easily understood method: sum Akaike weights from greatest to least until the cumulative sum first equals or exceeds the desired coverage probability; the confidence set is composed of those models whose Akaike weights contributed to the cumulative sum at that stopping point (Burnham and Anderson 2002).

**Acknowledgments** This research was supported in part by USPHS Grants from the National Institute on Alcohol Abuse and Alcoholism (AA09367 and AA11886), the National Institute on Drug Abuse (DA05147, DA13240, and DA024417), and the National Institute on Mental Health (MH066140). The first author (RMK) was supported by a Doctoral Dissertation Fellowship from the University of Minnesota Graduate School and by grant DA026119 from the National Institute on Drug Abuse. The authors acknowledge the assistance of Niels G. Waller and Saonli Basu, who provided helpful comments on an early draft of this paper. The first author gives his special thanks to Scott I. Vrieze and Joshua D. Isen for thought-provoking discussion of model-selection and of the main effects of SES, respectively.

**Conflict of interest** Robert M. Kirkpatrick, Matt McGue, and William G. Iacono declare that they have no conflict of interest.

**Human and Animal Rights and Informed Consent** The MTFS and SIBS studies were reviewed and approved by the Institutional Review Board at the University of Minnesota. Written informed assent or consent was obtained from all participants, with parents providing written consent for their minor children.

## References

Azen R, Budescu DV (2003) The dominance analysis approach for comparing predictors in multiple regression. Psychol Methods 8(2):129–148. doi:10.1037/1082-989X.8.2.129

Bartels LM (1997) Specification uncertainty and model averaging. Am J Polit Sci 41(2):641–674

Bates TC, Lewis GJ, Weiss A (2013) Childhood socioeconomic status amplifies genetic effects on adult intelligence. Psychol Sci 24(10):2111–2116. doi:10.1177/0956797613488394

Boker S, Neale M, Maes H, Wilde M et al. (2011) OpenMx: an open source extended structural equation modeling framework. *Psychometrika, 76*(2), 306–317. doi: 10.1007/S11336-010-9200-6. Software and documentation available at http://openmx.psyc.virginia.edu/

Bouchard TJ (2004) Genetic influence on human psychological traits: a survey. Curr Dir Psychol Sci 13(4):148–151

Bouchard TJ, McGue M (1981) Familial studies of intelligence: a review. Science 212(4498):1055–1059

Bouchard TJ, McGue M (2003) Genetic and environmental influences on human psychological differences. J Neurobiol 54:4–45

Braveman PA, Cubbin C, Egerter S, Chideya S, Marchi KS, Metzler M, Posner S (2005) Socioeconomic status in health research: One size does not fit all. J Am Med Assoc 294(22):2879–2888

Breiman L (1992) The little bootstrap and other methods for dimensionality selection in regression: x-fixed prediction error. Journal of the American Statistical Association 87(419):738–754

Bronfenbrenner U, Ceci SJ (1994) Nature-nurture reconceptualized in developmental perspective: a bioecological model. Psychol Rev 101(4):568–586

Browne MW (2000) Cross-validation methods. J Math Psychol 44:108–132. doi:10.1006/jmps.1999.1279

Burnham KP, Anderson DR (2001) Kullback-Leibler information as a basis for strong inference in ecological studies. Wildl Res 28:111–119

Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach, 2nd edn. Springer, New York

Burnham KP, Anderson DR (2004) Multimodel inference: Understanding AIC and BIC in model selection. Sociol Methods Res 33(2):261–304. doi:10.1177/0049124104268644

Cherny SS, Cardon LR, Fulker DW, DeFries JC (1992) Differential heritability across levels of cognitive ability. Behav Genet 22(2):153–162

Deary IJ, Spinath FM, Bates TC (2006) Genetics of intelligence. Eur J Hum Genet 14:690–700. doi:10.1038/sj.ejhg.5201588

DeFries JC, Fulker DW (1985) Multiple regression analysis of twin data. Behav Genet 15(5):467–473

DeFries JC, Fulker DW (1988) Multiple regression analysis of twin data: Etiology of deviant scores versus individual differences. Acta Geneticae Medicae et Gemellologiae 37:205–216

Evans GW (2004) The environment of childhood poverty. Am Psychol 59(2):77–92. doi:10.1037/0003-066X.59.2.77

Fischbein S (1980) IQ and social class. Intelligence 4:51–63

Galton F (1869). Hereditary genius: an inquiry into its laws and consequences. London: MacMillan & Co. Retrieved from http://galton.org/

Grant MD, Kremen WS, Jacobson KC et al (2010) Does parental education have a moderating effect on the genetic and environmental influences of general cognitive ability in early adulthood? Behav Genet 40:438–446. doi:10.1007/s10519-010-9351-3

Hanscombe KB, Trzaskowski M, Haworth CMA, Davis OSP, Dale PS, Plomin R (2012) Socioeconomic status (SES) and children's intelligence (IQ): in a UK-representative sample SES moderates the environmental, not genetic, effect on IQ. PLoS ONE 7(2):e30320. doi:10.1371/journal.pone.0030320

Harden KP, Turkheimer E, Loehlin JC (2007) Genotype by environment interaction in adolescents' cognitive aptitude. Behav Genet 37:273–283. doi:10.1007/s10519-006-9113-4

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer Science + Business Media, New York. doi:10.1007/b94608

Hollingshead AB (1957) Two factor index of social position. Hollingshead, New Haven

Hurvich CM, Tsai C-L (1989) Regression and time series model selection in small samples. Biometrika 76(2):297–307

Iacono WG, McGue M (2002) Minnesota twin family study. Twin Res 5(5):482–487

Iacono WG, Carlson SR, Taylor J, Elkins IJ, McGue M (1999) Behavioral disinhibition and the development of substance-use disorders: findings from the Minnesota twin family study. Dev Psychopathol 11:869–900

Kapetanios G, Labhard V, Price S (2008) Forecasting using Bayesian and information-theoretic model-averaging: An application to U.K. inflation. J Bus Econ Stat 26(1):33–41. doi:10.1198/073500107000000232

Keyes MA, Malone SM, Elkins IJ, Legrand LN, McGue M, Iacono WG (2009) The enrichment study of the Minnesota twin family study: increasing the yield of twin families at high risk for externalizing psychopathology. Twin Res Human Genet 12(5):489–501

Kirkpatrick RM, McGue M, Iacono WG (2009) Shared-environmental contributions to high cognitive ability. Behav Genet 39:406–416. doi:10.1007/s10519-009-9265-0

Kirkpatrick RM, McGue M, Iacono WG, Miller MB, Basu S, Pankratz N (2014) Low-frequency copy-number variants and general cognitive ability: no evidence of association. Intelligence 42:98–106. doi:10.1016/j.intell.2013.11.005

Kohler HP, Rodgers JL (2001) DF-analyses of heritability with double-entry twin data: asymptotic standard errors and efficient estimation. Behav Genet 31(2):179–191

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22(1):79–86

Loehlin JC, Harden KP, Turkheimer E (2009) The effect of assumptions about parental assortative mating and genotype-income correlation on estimates of genotype-environment interaction in the National Merit Twin Study. Behav Genet 39:165–169. doi:10.1007/s10519-008-9253-9

Lukacs PM, Burnham KP, Anderson DR (2009) Model selection bias and Freedman's paradox. Ann Inst Stat Math 62:117–125. doi:10.1007/s10463-009-0234-4

McCallum RC, Mar CM (1995) Distinguishing between moderator and quadratic effects in multiple regression. Psychol Bull 118(3):405–421

McGue M, Bouchard TJ (1984) Adjustment of twin data for the effects of age and sex. Behav Genet 14(4):325–343

McGue M, Keyes M, Sharma A, Elkins I, Legrand L, Johnson W, Iacono WG (2007) The environments of adopted and non-adopted youth: Evidence on range restriction from the Sibling Interaction and Behavior Study (SIBS). Behav Genet 37:449–462. doi:10.1007/s10519-007-9142-7

Myrianthopolous NC, French KS (1968) An application of the U.S. Bureau of the Census socioeconomic index to a large, diversified patient population. Soc Sci Med 2:283–299

Pawitan Y (2013) In all likelihood: statistical modelling and inference using likelihood. Oxford University Press, Oxford

Plomin R, DeFries JC, Loehlin JC (1977) Genotype-environment interaction and correlation in the analysis of human behavior. Psychol Bull 84(2):309–322

Price TS, Jaffee SR (2008) Effects of the family environment: Gene-environment interaction and passive gene-environment correlation. Dev Psychol 44(2):305–315. doi:10.1037/0012-1649.44.2.305

Purcell S (2002) Variance components models for gene-environment interaction in twin analysis. Twin Research 5(6):554–571

Rathouz PJ, Van Hulle CA, Rodgers JL, Waldman ID, Lahey BB (2008) Specification, testing, and interpretation of gene-by-measured environment interaction models in the presence of gene-environment correlation. Behav Genet 38:301–315. doi:10.1007/s10519-008-9193-4

Rijsdijk FV, Vernon PA, Boomsma DI (2002) Application of hierarchical genetic models to Raven and WAIS subtests: a Dutch twin study. Behav Genet 32(3):199–210

Rodgers JL, Kohler HP (2005) Reformulating and simplifying the DF analysis model. Behav Genet 35(2):211–217

Rodgers JL, McGue M (1994) A simple algebraic demonstration of the validity of Defries-Fulker analysis in unselected samples with multiple kinship levels. Behav Genet 24(3):259–262

Rowe DC, Jacobson KC, van den Oord EJCG (1999) Genetic and environmental influences on vocabulary IQ: parental educational level as moderator. Child Dev 70(5):1151–1162

Scarr S (1992) Developmental theories for the 1990s: development and individual differences. Child Dev 63:1–19

Scarr S, Weinberg RA (1978) The influence of "family background" on intellectual attainment. Am Sociol Rev 43(5):674–692

Scarr-Salapatek S (1971) Race, social class, and IQ. Science 174(4016):1285–1295

Shao J (1997) An asymptotic theory for linear model selection. Stat Sin 7:221–264

Spearman C (1904) "General intelligence", objectively determined and measured. Am J Psychol 15(2):201–292

Stone M (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. J R Stat Soc Ser B (Methodol), 39(1):44–47

Tucker-Drob EM, Harden KP, Turkheimer E (2009) Combining nonlinear biometric and psychometric models of cognitive abilities. Behav Genet 39:461–471. doi:10.1007/s10519-009-9288-6

Turkheimer E, Haley A, Waldron M, D'Onofrio B, Gottesman II (2003) Socioeconomic status modifies heritability of IQ in young children. Psychol Sci 14(6):623–628

Uher R, Dragomirecka E, Papezova H (2006) Use of socioeconomic status in health research. J Am Med Assoc 295(15):1770

Van den Ooord EJCG, Rowe DC (1998) An examination of genotype-environment interactions for academic achievement in an U.S. National Longitudinal Survey. Intelligence 25(3):205–228

Van der Sluis S, Willemsen G, de Geus EJC, Boomsma DI, Posthuma D (2008) Gene-environment interaction in adults' IQ scores: measures of past and present environment. Behav Genet 38:348–360. doi:10.1007/s10519-008-9212-5

Van der Sluis S, Posthuma D, Dolan CV (2012) A note on false positives and power in G × E modelling of twin data. Behav Genet 42:170–186. doi:10.1007/s10519-011-9480-3