

Problems and Pit-Falls in Testing for $G \times E$ and Epistasis in Candidate Gene Studies of Human Behavior

Lindon Eaves · Brad Verhulst

Received: 5 February 2014 / Accepted: 20 August 2014 / Published online: 7 September 2014
© Springer Science+Business Media New York 2014

Abstract Conclusions about the genetic architecture of a phenotype relating to the contributions of genetic additivity, dominance, epistasis or genotype \times environment interaction, depend upon the statistical and distributional properties of the measured trait. This dependence is frequently ignored in contemporary genetic studies and can radically change the conclusions that may be drawn from the data. The interdependence of the conclusions about genetic architecture and instruments used for behavioral measurement is explored by simulated studies of the interaction between candidate genes and measured environment in psychiatric genetics. Trait values are simulated ($N = 100,000$) under several commonly encountered scenarios and subjected to two simulated 20-item psychological tests each comprising items with different patterns of difficulty and sensitivity to variation (discriminating power) in the latent trait. Test scores are generated for each test by summing the binary responses across all items. The full model for digenic additive and non-additive genetic effects and $G \times E$ is fitted to the trait values and test scores under a range of different simulated genetic architectures. Untransformed test scores show complex patterns of epistasis and $G \times E$ even when the underlying effects of genes and environment are purely additive and the transformation of symptom counts does not fully recover the simulated underlying genetic architecture. Accordingly, failing to allow for the theory of measurement when analyzing details of genetic architecture may frequently lead to

replicable over-reporting of interactions and mislead potential investigators and funding agencies.

Keywords $G \times E$ interaction · Epistasis · Candidate genes · Genetic architecture · Simulation · Item-response theory · Psychometrics

Introduction

It is hard to imagine that I first met John Loehlin when I was a first-year PhD student in Birmingham. It was July 1969, while Neil Armstrong and Buzz Aldrin were walking on the moon. The date coincided a side-trip to the Shakespeare Memorial Theatre from a “NATO Advanced Studies Institute for Psychogenetics” engineered by John Jinks and Peter Broadhurst partly to promote their emerging application of “Biometrical Genetics” to human and animal behavior. That night, on the bus to Stratford, the conferees listened with bated breath to the final moments of countdown as the astronauts prepared to blast off from the moon on the start of their return to earth.

Looking back, it was an incredible opportunity for a not-yet PhD to put faces to some of the names he had seen in print. Among those names and faces, was a younger John Loehlin, even then wearing his trade-mark black shoes and white socks. The tone of John’s thought was known already from his note modestly but concisely correcting some conceptual errors in Raymond Cattell’s Multiple Abstract Variance Analysis (MAVA: Loehlin 1965). I was not smart enough to understand either but I knew their importance from my teachers and idols, the late John Jinks and David Fulker. I remember being a fly on the pub wall as John Loehlin, Louis Guttman and David Fulker pored over a notepad in the bar discussing earnestly whether putting

L. Eaves (✉) · B. Verhulst
Virginia Institute for Psychiatric and Behavioral Genetics,
Virginia Commonwealth University School of Medicine,
Richmond, VA, USA
e-mail: eaves.lindon@gmail.com

heritability estimates down the diagonal of a correlation matrix would solve the “communality problem” in factor analysis. I didn’t have a clue what they were talking about.

A burning topic in those days, as it remains today, was genotype \times environment interaction ($G \times E$). Although still unpublished at that point, John (Jinks) and David had shared a pre-print of their seminal 1970 paper on the application of Biometrical Genetics to human behavior (Jinks and Fulker 1970). Among other significant issues they addressed was that of $G \times E$ and, in particular, the possibility of examining the regression of absolute intrapair differences for monozygotic (MZ) twins on pair means as a key to characterizing the relationship between sensitivity to random environmental influences (intrapair differences in MZ twins) and average genetic liability (measured by pair means).

Fired with enthusiasm for this insight, and challenged by David who had thought a lot about $G \times E$ and risk to psychopathology, we embarked on an exploration of $G \times E$ for personality applying the “Jinks and Fulker” approach to some early “EPQ” data on twins that Hans Eysenck had generously shared. Very soon I had generated some pretty diagrams and David had written the first draft of a joint paper showing significant, complex, non-linear, $G \times E$ for personality test scores. Recollection is hazy, but I think David sent an early draft to John Loehlin who suggested that we should check whether the interaction was “really” $G \times E$ or whether it was just a function of variation in measurement error over the range of test scores. “Goodbye” to a good paper, part of my doctoral dissertation and, for all intents and purposes, to a promising method since very few applications of the approach have been published in the 40 years since it first appeared.

Sadly, the fact that the approach, and John Loehlin’s early critical insight, lie all but forgotten by the contemporary literature means a fundamental lesson from quantitative genetics appears not to have been internalized by the new generation of behavioral researchers competing for the prestige and funding that goes with the pursuit of $G \times E$ in psychiatric genetics. The apparently forgotten lesson from those early efforts is quite simple. *You can generate almost any interaction you want by changing the scale of measurement.* The implication is equally simple: *Don’t make a career out of your interaction until you have excluded simpler psychometric considerations that owe nothing to the subtleties of the underlying genetic and environmental causes of human variation.*

The seminal contribution of Fisher, Immer and Tedin (1932) notwithstanding, geneticists have remained cautious about the using the properties of observed phenotypic distributions to infer subtleties of the genetic architecture of complex traits. This caution stems from the observation that a variety of more or less arbitrary factors, having little

or nothing to do with genetics, can affect the more subtle features of trait distributions. Paramount among such factors are those arising from the fact that the scales used to measure variation have an ill-defined, relationship to underlying biological differences. Hence, changes in the units or method of measurement can lead to drastically different conclusions about the genetic architecture of the underlying biological system. Mather and Jinks (1982) offer a classical statement of the interdependence of measurement and genetic inference:

“The scale on which the measurements are expressed for the purposes of genetical analysis must therefore be reached by empirical means. Obviously it should be one which facilitates both the analysis of the data and the interpretation and use of the resulting statistics...*The scale should preferably be one on which...the interactions among the genes and between genotype and environment are absent, or at any rate as small as they can reasonably be made.*” (p. 64, our italics). Lack of careful attention to this goal leaves in question the heuristic value of claims to find $G \times E$ in psychiatric data.

With respect to behavior, measurement often boils down to decisions about which constellations of items, combined in which way, best characterize the salient latent behavioral outcomes and psychosocial risk factors. The relationship between the numbers generated by a test and the way genes and environment work is tenuous and theory-dependent. There is an intimate connection between the choice of measure, and conclusions drawn about the relative importance of genes, environment and the various possible interactions between them. Elegant pictures of the role of $G \times E$ interaction may be no more robust than the items selected to measure the hypothesized latent variable, the rule used to combine them, or how the scores are scaled after they have been combined.

Mather and Jinks recommend that “So far as possible the non-allelic genes and non-heritable agents should all be additive in action” but also caution that such scales may be hard to find since “Each gene and each non-heritable agent may be acting on its own scale” and the elegance of a parsimonious additive model may be elusive. The problem is that psychiatric geneticists seldom bother to look. We are not blessed with decisions as simple as whether to measure body-weight in kilograms or log-kilograms, though even here the choice of scale will not be neutral with respect to conclusions about the contributions of additive and non-additive effects.

Although the point had been made on several occasions (see e.g. Eaves and Eysenck 1977; Purcell 2002), a recent paper (Eaves 2014) reiterated the implications of common problems of measurement in psychiatric genetics for the

detection of interaction between measured environmental covariates and random genetic effects in twin studies. In particular, it was demonstrated that the use of symptom counts, characteristic of attempts to quantify clinical outcomes, would almost certainly generate statistical evidence for $G \times E$ when the underlying genetic and environmental causes of variation in liability were purely additive. Furthermore, because such interactions depend purely on the units of measurement rather than biology, they are almost certain to replicate, a sine qua non for publication.

Studies of $G \times E$ in humans are not confined to the study of multifactorial liability and the structural modeling of the patterns of covariance between relatives but extend to the detection and analysis of interaction between candidate genes and environmental covariates. Such studies enjoy a high profile and the generated enthusiasm influences the direction and funding of subsequent research in psychiatric and behavioral genetics. With this in mind, and knowing the inherent problems of interpreting interactions in twin and family studies, this paper explores the extent to which the same uncertainties attend apparent demonstrations of interaction between candidate genes and covariates in psychiatric genetics.

Approach

The problem is addressed by simulating the effects of two candidate loci and environment on liability to a psychiatric disorder. A general biometrical-genetic model for the additive, dominance and epistatic effects of the two loci (c.f. Mather and Jinks 1982) characterizes the main effects of the genes on liability and the (linear) response of genotypes to a continuously variable environmental covariate ($G \times E$ interaction). The model has been widely used in experimental organisms, including plants and fruitflies and has the advantage of capturing classical patterns of non-allelic interaction (epistasis) as special cases of the general model.

Genotypes, environments and liabilities were simulated for a large number of independent subjects ($N = 100,000$) under a variety of configurations for the additive and non-additive effects of the loci and covariate. Simulated subjects were scored using simulated responses to dichotomous items ($k = 20$) of checklists using two types of test analogous to those frequently encountered in behavioral measurement. The first, resembling a typical checklist of relatively infrequent symptoms, comprises items with equal low endorsement frequency (difficulty) and the same discriminating power. The second, more characteristic of tests used to assess abilities, comprises items with a wide range of difficulty and variable discriminating power.

Table 1 Contributions of two-locus homozygous, heterozygous and epistatic effects to expected genotypic values (after Mather and Jinks 1982, p.83, Table 19)

Locus	Genotype	A/a		
		AA	Aa	aa
B/b	BB	$d_a + d_b + i_{ab}$	$h_a + d_b + j_{ba}$	$-d_a + d_b - i_{ab}$
	Bb	$d_a + h_b + j_{ab}$	$h_a + h_b + l_{ab}$	$-d_a + h_b - j_{ab}$
	bb	$d_a - d_b - i_{ab}$	$h_a - d_b - j_{ba}$	$-d_a - d_b + i_{ab}$

See text and Table 2 for explanation of parameters

The general linear model for gene effects and $G \times E$ is fitted to the liabilities and test scores derived from the item responses to test the main effects and interactions of the candidate genes and environments simulated under various “true” configurations of their effects on liability. Parameter estimates and their sampling errors are recovered and t tests compared for the various types of measurement to assess the impact of different approaches to measurement on the outcome of tests for non-additive effects of genes and environment (epistasis and $G \times E$) under various combinations of “true” genetic model and mode of assessment.

For various configurations of genetic effects at the two candidate loci it is shown how estimates and significance levels of non-additive genetic effects and $G \times E$ are critically dependent on the items and rules for combining them chosen to measure a psychiatric outcome.

Genetic model

Table 1 presents the general model for the main effects and interactions of two diallelic loci on a continuous trait outlined by Mather and Jinks (1982, c.f. Van der Veen 1959).

Various notations and parameterizations may be found in the literature. The notation used here has enjoyed widespread application for the analysis of digenic effects on the means and variances of generations derived from crosses between inbred lines of diploid species and for specifying the components of genetic variance in randomly mating populations (Mather 1974). The model specifies the homozygous (additive effects) and heterozygous effects (dominance effects) of the two loci, d_a , d_b , h_a and h_b respectively, and the four possible types of epistatic interaction between them: between homozygotes, i_{ab} ; between homozygotes at the A/a locus (AA versus aa) and heterozygote at B/b, j_{ab} ; between heterozygote at the A/a locus and homozygote at B/b, j_{ba} ; between both heterozygous effects, l_{ab} . While the notation appears cumbersome, it has the advantage of generality in capturing characteristic patterns of classical epistatic segregation in Mendelian

Table 2 Definition of model parameters

Symbol	Definition
p_{aA}, p_b	Frequencies of increasing alleles at candidate loci A and B
m	Origin of main effects (“constant”)
d_a, d_b	Homozygous (“additive”) deviations at A and B
h_a, h_b	Heterozygous (“dominance”) effects at A and B
i_{ab}	Interaction between homozygous effects at A/a and B/b (“additive × additive”)
j_{ab}	Interaction between additive effect at A/a and dominance effect at B/b
j_{ba}	Interaction between dominance effect at A/a and additive effect at B/b
l_{ab}	Interaction between dominance effects at A/a and B/b (“dominance × dominance”)
β_m	Origin of (linear) response to covariate (“main effect of environment”)
β_{d_a}, β_{d_b}	Homozygous effects of A/a and B/b on linear response to environment (“additive genetic effects on $G \times E$ ”)
β_{h_a}, β_{h_b}	Heterozygous effects of A/a and B/b on linear response environment (“dominant genetic effects on $G \times E$ ”)
$\beta_{i_{ab}}$	Additive × additive epistatic genetic effects on response to environment ($G \times E$)
$\beta_{j_{ab}}$	Additive × dominant epistatic genetic effects on response to environment ($G \times E$)
$\beta_{i_{ba}}$	Dominant × additive epistatic genetic effects on response to environment ($G \times E$)
$\beta_{l_{ab}}$	Dominant × dominant epistatic genetic effects on response to environment ($G \times E$)
m_E	Mean of measured environment
σ_E	Standard deviation of measured environment
σ_δ	Residual standard deviation

dihybrid crosses while not being restricted by them. The classical patterns of epistasis were described in the first decade of the 20th century (see e.g. Miko 2008, for a recent didactic summary of the classical ratios). Thus, the 9:7 F_2 segregation characteristic of complementary gene interaction is realized when, inter alia, $d_a = d_b = h_a = h_b = i_{ab} = j_{ab} = j_{ba} = l_{ab}$ in Table 1. In contrast the 15:1 F_2 segregation characteristic of duplicate gene interaction arises, for example, when $d_a = d_b = h_a = h_b = -i_{ab} = -j_{ab} = -j_{ba} = -l_{ab}$. “Complementary” epistasis arises when genes form a series in a biological pathway such that failure of either component leads to failure of the pathway. “Duplicate” epistasis corresponds to systems that are buffered by redundant parallel pathways so that failure of both components is required for system failure and has commonly been associated with a strong linear component of the relationship between phenotype and fitness (see e.g. Mather 1966).

The model for the additive, dominant and epistatic effects of the locus pair may be extended to include their effects on the response to an environmental covariate

($G \times E$ interaction). Following the approach of, e.g. Bucio Alanis and Hill (1966) and developed by Jinks and his coworkers (see Mather and Jinks 1982) genotypes differ in their regression on the environmental covariate. Just as differences in the main effects of the gene pair may be represented by the parameters $d_a, d_b, h_a, h_b, i_{ab}, j_{ab}, j_{ba}$ and l_{ab} , so an analogous parameterization may be used to account for genotypic differences in the (e.g. linear) regression of phenotype on measured environment. For example, the regression of the AAbb genotype on environment is $\beta_m + \beta_{d_a} - \beta_{d_b}$ where β_m is the regression of the mid-homozygote on the environment, β_{d_a} the homozygous effect of locus A/a on regression and β_{d_b} the homozygous effect of the B/b locus on response to the environment.

Table 2 summarizes the parameters of the full model for the effects of a pair of diallelic loci on a quantitative phenotype. Additional parameters specify the allele frequencies, the mean and variance of the hypothesized environmental covariate and the variance of residual effects.

Simulations

The large number of parameters in the digenic model for epistasis and $G \times E$ precludes consideration of any but a small fraction of the set of possible genetic systems. Arbitrariness of the units used to measure behavior introduces an additional dimension to explore the impact of test construction and scoring for the detection of epistasis and $G \times E$.

Five two-locus models were chosen for simulation:

1. The two locus model with a main effect of environment (β_m) with no dominance, epistasis or $G \times E$.
2. Model 1 with the addition of heterozygous effects, h_a and h_b without epistasis or $G \times E$.
3. Model 2 plus complementary gene interaction, without $G \times E$.
4. Model 2 plus duplicate gene interaction, without $G \times E$.
5. Model 1, with the addition of homozygous effects, β_{d_a} and β_{d_b} , on linear response to the measured environment ($G \times E$).

The parameter values employed to simulate the genotypes and individual continuous phenotypes are summarized in Table 3. Each simulation assumed further that the origin for the genetic main effects (m) was 10. The measured environment was assumed to be distributed normally ($\mu = 5, \sigma = 1$) and residual effects of unmeasured genes and environment to be distributed normally ($\mu = 0, \sigma = 1$).

Table 3 Parameter values employed in candidate gene models for epistatic and $G \times E$ interactions

Model	Simulated parameter values																
	Genetic main effects (G)								Environmental effects (E) and $G \times E$								
	Homozygous		Heterozygous		Epistatic				Homozygous			Heterozygous		Epistatic			
	d_a	d_b	h_a	h_b	i_{ab}	j_{ab}	j_{ba}	l_{ab}	β_m	βd_a	βd_b	βh_a	βh_b	βi_{ab}	βj_{ab}	βj_{ba}	βl_{ab}
1	1	1	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0.5	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	0.5	0	0	0	0	0	0	0	0
4	1	1	1	1	-1	-1	-1	-1	0.5	0	0	0	0	0	0	0	0
5	1	1	0	0	0	0	0	0	0.5	0.25	0.25	0	0	0	0	0	0

Table 4 Item-response parameters (a = difficulty, s = sensitivity) used in second simulated test (Test 1 has $a = 1$ and $s = 1$ for all items)

Item	1	2	3	4	5	6	7	8	9	10
Difficulty (a)	0.361	-0.564	-1.968	-0.708	-1.136	0.119	-1.005	1.200	0.287	1.146
Sensitivity (s)	0.859	1.015	0.703	1.372	0.834	0.809	1.347	0.727	1.430	0.618
Item	11	12	13	14	15	16	17	18	19	20
Difficulty (a)	1.616	0.653	-1.640	0.041	-0.520	-1.826	0.276	-1.480	1.602	1.269
Sensitivity (s)	1.296	1.476	0.837	1.182	0.550	1.288	0.603	1.203	1.193	1.388

Parameters assume normal ogive item characteristic curves with latent trait scaled to unit variance

The traits simulated under each model, standardized to zero mean and unit variance, were then “administered” two simulated tests comprising 20 binary items. Item parameters were chosen to reflect two different extreme measurement models. Raw test scores were generated by summing the 0/1 item responses across items. Both tests assumed normal ogive item characteristic curves for each item. The items of the first test were assumed to have unit thresholds (item difficulties) and sensitivities (discriminating powers). Item difficulties of the second test were assumed to be distributed uniformly (ranging from -2 to 2) with discrimination parameters distributed uniformly (ranging from 0.5 to 1.5). Thus, the first test generated symptom counts with a J-shaped distribution characteristic of those often encountered in psychiatric assessment. The second test, with item difficulties distributed uniformly across most of the range of simulated trait values generated scores more symmetrically distributed around an intermediate mode. Table 4 shows the specific item parameters simulated for the second test. In addition, the raw trait values and test scores were dichotomized to generate outcome (“disease” phenotypes) at thresholds giving the closest to 20 % prevalence in the population. The raw sum scores for the first test were also subjected to a square root transformation to minimize the effects of heteroscedasticity on the subsequent regression analysis of the raw symptom counts (c.f. Bartlett 1947) and spurious non-additive

genetic effects (c.f. Eaves and Eysenck 1977). It will be seen that simple transformation does not always have the desired result.

100,000 independent observations were simulated under each of the five genetic models. The sample size was chosen to give estimates that were stable enough to allow relatively reliable inferences about the power and biases implicit in the detection of the main effects and interactions of the pair of candidate loci and the environment but not so large as to overwhelm a typical laptop computer. Simulations and regression analyses were conducted in R 2.13.2.

Statistical analysis of simulated data

The full linear regression model, allowing for additive, dominance and epistatic effects of the two loci on the average phenotype and linear response to the covariate ($G \times E$) was fitted to the data generated under each of the five models for genetic and environmental effects (above). The raw trait values, sum scores for the two simulated tests and transformed scores for the first test were all analyzed on the assumption of normal errors. The dichotomous disease phenotypes were analyzed by logistic regression assuming binomial errors. In addition to the full model, the “true” model, assumed in generating each data set, was

Table 5 Estimates recovered fitting candidate gene models to simulated trait values and derived variables under Model 1 (no interactions)

Data	Method	Estimated parameter values														σ	r^2								
		Genetic main effects (G)							Environmental effects (E) and G × E																
		Homozygous		Heterozygous		Epistatic			Homozygous		Heterozygous		Epistatic												
		d_A	d_B	h_A	h_B	i_{AB}	j_{AB}	j_{BA}	i_{AB}	i_{BA}	i_{AB}	i_{BA}	β_{d_A}	β_{d_B}	β_{h_A}	β_{h_B}	$\beta_{i_{AB}}$	$\beta_{i_{BA}}$	$\beta_{j_{AB}}$	$\beta_{j_{BA}}$	$\beta_{i_{AB}}$	$\beta_{i_{BA}}$	$\beta_{j_{AB}}$	$\beta_{j_{BA}}$	
		1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		β_n																							
		.5																							
True latent trait	Normal errors	0	0.989	0.959	0.023	0.028	0.017	0.024	0.019	0.009	0.009	0.009	0.501	0.002	0.008	-0.004	-0.006	-0.004	-0.003	-0.003	-0.003	-0.003	-0.002	1.006	0.554
		t	30.39	29.47	0.50	0.61	0.52	0.53	0.42	0.13	0.13	0.13	78.63	0.36	1.22	-0.43	-0.67	-0.55	-0.28	-0.28	-0.29	-0.18	-	-	-
Normal errors		0	0.981	-	-0.010	-	-	-	-	-	-	-	0.494	0.006	-	0.004	-	-	-	-	-	-	1.230	0.332	
		t	89.48	-	-0.25	-	-	-	-	-	-	-	89.48	1.11	-	0.47	-	-	-	-	-	-	-	-	-
Normal errors		0	1.006	1.001	-	-	-	-	-	-	-	-	0.496	-	-	-	-	-	-	-	-	-	-	1.006	0.554
		t	223.0	222.8	-	-	-	-	-	-	-	-	156.3	-	-	-	-	-	-	-	-	-	-	-	-
Logistic regression		0	2.558	2.785	-0.403	-0.329	-0.449	1.251	0.896	1.084	1.084	1.068	-0.094	-0.140	-0.140	0.060	0.019	0.037	-0.155	-0.118	-0.099	-	-	-	-
		t	3.15	3.43	-0.43	-0.34	-0.55	1.30	0.96	1.00	1.00	8.46	-0.74	-1.11	-1.11	0.42	0.13	0.29	-1.03	-0.82	-0.59	-	-	-	-
Sum score		0	0.730	0.634	-0.611	-0.646	1.864	-0.563	-0.606	-0.365	-0.365	1.401	0.423	0.442	0.442	0.043	0.049	-0.164	0.160	0.164	0.021	3.658	0.464	-	
		t	6.16	5.36	-3.68	-3.86	15.74	-3.36	-3.64	-1.55	-1.55	60.39	18.22	19.07	19.07	1.34	1.49	-7.05	4.87	5.03	0.46	-	-	-	-
Square root transformation		0	0.583	0.532	-0.150	-0.179	0.483	0.059	0.091	-0.087	-0.087	0.352	0.031	0.041	0.025	0.030	0.030	-0.081	0.008	0.001	0.014	0.933	0.470	-	
		t	19.31	17.60	-3.53	-4.20	15.97	1.38	2.14	-1.45	-1.45	59.39	5.25	6.92	3.05	3.55	-13.73	0.96	0.06	0.88	-	-	-	-	
Logistic regression		0	2.462	2.466	0.371	0.615	-0.741	0.079	0.176	-0.151	-0.151	0.961	-0.112	-0.112	-0.054	-0.084	0.082	-0.029	-0.020	0.039	-	-	-	-	
		t	6.16	6.17	0.82	1.38	-1.85	0.18	0.17	-0.30	-0.30	15.11	-1.76	-1.77	-0.74	-1.17	1.28	-0.40	-0.28	0.47	-	-	-	-	
Sum score		0	2.771	2.801	-0.246	-0.195	0.847	0.147	0.040	-0.082	-0.082	1.309	-0.011	-0.019	0.057	0.042	-0.175	-0.002	0.023	0.011	3.206	0.484	-	-	
		t	26.70	26.98	-1.69	-0.13	8.16	1.00	0.27	-0.40	-0.40	64.37	-0.53	-0.95	2.00	1.47	-8.63	-0.06	0.81	0.27	-	-	-	-	
Square root transformation		0	0.653	0.655	0.034	-0.000	-0.020	0.003	-0.002	-0.008	-0.008	0.230	-0.037	-0.038	0.002	-0.000	-0.011	0.000	0.002	0.001	0.561	0.470	-	-	
		t	35.93	36.07	1.32	-0.08	-1.11	0.10	-0.09	-0.22	-0.22	64.66	-10.29	-10.54	0.37	-0.08	-3.23	0.06	0.34	0.10	-	-	-	-	
Logistic regression		0	2.919	2.857	0.622	0.418	-0.946	-0.002	-0.264	-0.240	-0.240	0.964	-0.185	-0.173	-0.078	-0.046	0.091	-0.014	0.027	0.049	-	-	-	-	
		t	5.29	5.18	1.03	0.69	-0.17	-0.00	-0.44	-0.36	-0.36	11.28	-2.17	-2.02	-0.83	-0.49	1.05	-0.15	0.28	0.46	-	-	-	-	

The sample size for each model is 100,000 independent observations. For the analytical method, “Normal errors” are true latent trait regression with normal errors, “Logistic Regression” is logistic regression of dichotomized outcome with binomial errors, “Sum Scores” are the regression of sum score with normal errors and “Square-Root Transformation” is the regression of square-root transformed sum score with normal errors

Table 6 Estimates recovered fitting candidate gene models to simulated trait values and derived variables under Model 2 (no epistasis or $G \times E$)

Data	Method	Estimated parameter values																σ	r^2				
		Genetic main effects (G)																					
		Homozygous				Heterozygous				Epistatic				Environmental effects (E) and $G \times E$									
		d_A	d_B	h_A	h_B	i_{AB}	j_{AB}	i_{BA}	j_{BA}	i_{AB}	i_{BA}	β_{mE}	β_{dA}	β_{dB}	β_{hA}	β_{hB}	β_{iAB}			β_{iBA}	β_{jAB}	β_{jBA}	
True latent trait	Normal errors	0	0.989	1.023	1.028	0.017	0.024	0.019	0.009	0.501	0.002	0.008	-0.004	-0.005	-0.004	-0.003	-0.003	-0.003	-0.003	-0.002	1.006	0.634	
	Normal errors	t	30.39	29.47	22.34	0.52	0.53	0.42	0.13	78.63	0.36	1.22	-0.43	-0.67	-0.55	-0.28	-0.30	-0.18	-	-	-	-	
	Normal errors	0	0.997	1.001	-	-	-	-	-	0.492	0.003	-	0.001	-	-	-	-	-	-	-	1.327	0.353	
Equal item parameters	Normal errors	t	32.81	23.43	-	-	-	-	-	82.61	0.54	-	0.12	-	-	-	-	-	-	-	-	-	
	Normal errors	0	1.006	1.003	0.997	-	-	-	-	0.496	-	-	-	-	-	-	-	-	-	-	1.006	0.634	
	Logistic regression	t	223.0	222.8	157.6	156.7	-	-	-	156.3	-	-	-	-	-	-	-	-	-	-	-	-	-
Variable item parameters	Logistic regression	0	1.600	1.766	1.840	1.102	1.449	1.191	1.238	0.948	0.084	0.060	0.060	0.036	-0.221	-0.266	-0.229	-0.231	-0.231	-	-	-	-
	Logistic regression	t	1.11	1.23	1.16	1.26	0.77	0.99	0.82	0.83	3.76	0.34	0.24	0.24	0.14	-0.88	-1.04	-0.89	-0.89	-	-	-	-
	Sum score	0	-0.508	-0.481	-0.160	-0.240	0.442	0.655	0.602	0.478	0.930	0.450	0.454	0.401	0.416	0.107	0.076	0.084	0.084	0.084	3.530	0.462	
Sum score	Sum score	t	-4.44	-4.20	-1.00	-1.51	3.86	4.06	3.76	2.10	41.56	20.51	20.28	12.78	13.14	4.76	2.41	2.63	1.90	-	-	-	-
	Square root transformation	0	0.201	0.177	0.281	0.241	0.388	0.401	0.425	0.388	0.289	0.078	0.083	0.065	0.071	-0.042	-0.042	-0.049	-0.044	-0.044	0.868	0.530	
	Logistic regression	t	7.15	6.30	7.11	6.06	13.81	10.10	10.78	6.95	52.57	14.16	15.14	8.44	9.17	-7.62	-5.39	-6.34	-4.06	-	-	-	-
Sum score	Logistic regression	0	1.532	2.004	1.869	2.262	0.852	0.922	0.650	0.546	0.252	0.089	0.012	0.031	-0.032	-0.202	-0.210	-0.166	-0.155	-	-	-	-
	Logistic regression	t	1.07	1.29	1.40	1.56	0.59	0.64	0.45	0.37	3.51	0.35	0.05	0.12	-0.13	-0.80	-0.83	-0.65	-0.60	-	-	-	-
	Sum score	0	1.941	2.067	1.925	2.141	0.855	0.737	0.661	0.611	1.188	0.094	0.072	0.010	0.056	-0.135	-0.105	-0.093	-0.088	-0.088	2.977	0.555	
Sum score	Sum score	t	20.14	15.28	19.97	15.72	8.88	5.41	4.89	3.19	62.95	4.95	2.73	5.28	2.07	-7.16	-3.94	-3.51	-2.35	-	-	-	-
	Square root transformation	0	0.585	0.589	0.607	0.618	0.021	-0.002	-0.013	-0.016	0.236	-0.021	-0.022	-0.025	-0.027	-0.015	-0.009	-0.008	-0.007	-0.007	0.520	0.568	
	Logistic regression	t	34.71	34.98	25.68	28.87	1.26	-0.09	-0.57	-0.49	71.69	-6.50	-6.57	-5.47	-5.87	-4.47	-1.98	-1.65	-1.13	-	-	-	-
Logistic regression	Logistic regression	0	3.304	2.690	3.588	3.351	0.099	-0.260	-0.109	-0.894	1.186	-0.264	-0.319	-0.165	-0.277	-0.018	0.034	0.030	0.145	-	-	-	-
	Logistic regression	t	3.31	2.70	3.47	3.24	0.10	-0.25	-0.11	-0.78	7.92	-1.76	-2.33	-1.10	-1.77	-0.12	0.28	0.19	0.89	-	-	-	-

Table 7 Estimates recovered fitting full model to simulated trait values and derived variables under Model 3 (Complementary epistasis, no G × E)

Data	Method	Estimated parameter values																		σ	r ²					
		Genetic main effects (G)									Environmental effects (E) and G × E															
		Homozygous			Heterozygous			Epistatic			E			Homozygous			Heterozygous					Epistatic				
		d _A	d _B	b _A	b _B	i _{AB}	j _{AB}	j _{BA}	i _{AB}	β _{in}	β _{d_A}	β _{d_B}	β _{d_{AB}}	β _{h_A}	β _{h_B}	β _{h_{AB}}	β _{i_{AB}}	β _{i_{BA}}	β _{i_{AB}}	β _{i_{BA}}	β _{i_{AB}}	β _{i_{BA}}				
		1	1	1	1	1	1	1	1	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0			
True latent trait	Normal errors	θ	0.989	0.959	1.023	1.028	1.017	1.024	1.019	1.009	0.501	0.002	0.008	0.008	-0.004	-0.006	-0.004	-0.004	-0.003	-0.003	-0.003	-0.003	-0.003	-0.002	1.006	0.806
		t	30.39	29.47	22.39	22.34	31.24	22.26	22.30	15.58	78.63	0.36	1.22	1.22	-0.43	-0.67	-0.43	-0.55	-0.28	-0.30	-0.30	-0.30	-0.30	-0.18	-	-
	Normal errors	θ	1.526	-	1.509	-	-	-	-	-	0.494	-0.003	-	-	-0.002	-	-	-	-	-	-	-	-	1.807	0.372	
		t	36.91	-	25.95	-	-	-	-	-	60.97	-0.40	-	-	-0.17	-	-	-	-	-	-	-	-	-	-	-
Equal item parameters	Normal errors	θ	1.001	0.998	1.004	0.998	0.999	1.012	1.006	0.997	0.496	-	-	-	-	-	-	-	-	-	-	-	-	1.006	0.806	
		t	156.6	156.3	111.7	110.6	156.5	112.1	111.9	78.4	156.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
		θ	2.964	0.428	0.538	0.627	4.287	4.487	6.934	6.840	0.623	0.332	-0.189	-0.206	0.201	0.102	0.102	0.102	0.070	0.605	0.611	0.611	0.611	-	-	
Variable item parameters	Logistic regression	t	0.02	0.00	0.00	0.00	0.03	0.02	0.02	0.03	0.02	0.01	-0.01	-0.00	-0.00	-0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	-	-	
		θ	0.243	0.267	0.564	0.557	0.281	0.597	0.475	0.331	0.637	0.328	0.321	0.274	0.274	0.275	0.321	0.321	0.269	0.295	0.299	0.299	0.299	2.901	0.649	
	Square root transformation	t	2.59	2.85	4.28	4.20	3.00	4.51	3.60	1.78	34.60	17.80	17.44	17.44	17.80	17.44	17.44	17.45	10.35	11.44	8.17	8.17	8.17	-	-	
		θ	0.400	0.400	0.476	0.452	0.425	0.480	0.456	0.430	0.214	0.032	0.031	0.018	0.018	0.023	0.028	0.028	0.019	0.024	0.025	0.025	0.025	0.685	0.736	
Variable item parameters	Sum score	t	18.01	18.03	15.27	14.42	19.16	15.30	14.63	9.75	49.18	7.29	7.12	3.01	3.67	3.67	6.36	6.36	3.06	3.94	2.86	2.86	2.86	-	-	
		θ	1.860	1.911	1.847	1.946	1.977	2.001	1.908	1.934	0.900	0.015	0.003	0.017	-0.007	-0.014	-0.014	-0.017	0.008	-0.002	2.418	2.418	2.418	0.725		
	Square root transformation	t	23.75	24.11	16.81	17.59	25.25	18.15	17.37	12.43	56.68	0.99	0.18	0.81	0.81	-0.33	-0.93	-0.93	-0.77	0.37	-0.06	-0.06	-0.06	-	-	
		θ	0.394	0.411	0.394	0.409	0.420	0.427	0.405	0.416	0.175	-0.014	-0.017	-0.014	-0.018	-0.018	-0.020	-0.021	-0.016	-0.018	0.429	0.429	0.429	0.703		
Variable item parameters	Sum score	t	28.33	29.57	20.20	20.83	30.23	21.74	20.76	15.06	64.45	-5.13	-6.40	-3.74	-4.66	-7.44	-5.40	-4.18	-3.38	-	-	-	-	-	-	
		θ	1.860	1.911	1.847	1.946	1.977	2.001	1.908	1.934	0.900	0.015	0.003	0.017	-0.007	-0.014	-0.014	-0.017	0.008	-0.002	2.418	2.418	2.418	0.725		

Estimates from fitting logistic regression to dichotomized sum scores are uninformative and not presented for this model

Table 8 Estimates recovered fitting full model to simulated trait values and derived variables under Model 4 (Duplicate gene epistasis, no G × E)

Data	Method	Estimated parameter values																				σ	r^2										
		Genetic main effects (G)										Environmental effects (E) and G × E																					
		Homozygous					Heterozygous					Epistatic					Homozygous							Heterozygous					Epistatic				
		d_A	d_B	h_A	h_B	i_{AB}	j_{BA}	j_{AB}	i_{BA}	i_{AB}	j_{BA}	β_m	β_{dA}	β_{dB}	β_{hA}	β_{hB}	$\beta_{i_{AB}}$	$\beta_{i_{BA}}$	$\beta_{j_{AB}}$	$\beta_{j_{BA}}$	$\beta_{i_{AB}}$			$\beta_{i_{BA}}$	$\beta_{j_{AB}}$	$\beta_{j_{BA}}$							
True latent trait	Normal errors	0	0.989	0.959	1.023	1.028	1.028	-0.983	-0.976	-0.981	-0.991	0.501	0.002	0.008	-0.004	-0.006	-0.004	-0.003	-0.003	-0.003	-0.003	-0.002	1.006	0.541									
	t	30.39	29.47	22.39	22.34	22.34	-30.19	-21.20	-21.47	-15.32	78.63	0.36	1.22	-0.43	-0.67	-0.55	-0.28	-0.30	-0.18	-	-	-	-	-									
Normal errors	0	0.467	-	0.492	-	-	-	-	-	-	0.490	0.010	-	0.004	-	-	-	-	-	-	-	1.328	0.199										
	t	5.39	-	11.53	-	-	-	-	-	-	82.22	1.63	-	0.47	-	-	-	-	-	-	-	-	-	-									
Normal errors	0	.001	0.998	1.004	0.998	-1.001	-0.988	-0.994	-1.002	0.496	-	-	-	-	-	-	-	-	-	-	-	1.006	0.541										
	t	156.7	156.3	111.7	110.6	-156.7	-109.5	-110.6	-78.84	156.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-									
Logistic regression	0	2.912	2.988	3.106	3.149	-2.775	-2.787	-3.016	-3.260	0.674	0.218	0.210	0.210	0.183	0.176	-0.246	-0.236	-0.203	-0.156	-	-	-	-	-									
	t	0.05	0.05	0.05	0.05	-0.04	-0.04	-0.05	-0.05	0.05	0.02	0.02	0.02	0.02	0.01	-0.02	-0.02	-0.02	-0.01	-	-	-	-	-									
Equal item parameters	Sum score	0	-0.876	-0.985	-0.621	-0.765	0.082	1.021	0.920	0.880	1.331	0.425	0.450	0.379	0.402	-0.416	-0.415	-0.431	-0.433	3.918	0.218	-	-	-									
	t	-6.91	-7.77	-3.49	-4.27	6.47	5.69	5.17	3.49	53.56	17.10	18.13	10.87	10.87	11.45	-16.76	-12.66	-12.36	-8.75	-	-	-	-	-									
Square root transformation	0	-0.025	-0.086	0.029	-0.015	0.014	0.156	0.076	0.040	0.344	0.101	0.113	0.092	0.099	-0.099	-0.105	-0.110	-0.106	0.986	0.292	-	-	-										
	t	-0.79	-2.71	0.65	-0.33	0.42	1.23	1.69	0.63	54.98	16.21	18.09	10.45	10.45	11.21	-15.77	-11.89	-12.52	-8.49	-	-	-	-	-									
Variable item parameters	Sum score	0	1.268	1.242	1.523	1.532	-1.220	-1.222	-1.355	-1.524	1.271	0.218	0.224	0.173	0.164	-0.230	-0.221	-0.203	-0.172	3.277	0.404	-	-	-									
	t	11.95	11.71	10.22	10.21	-11.50	-8.15	-9.10	-7.23	61.17	10.49	10.79	5.93	5.93	5.58	-11.05	-7.50	-6.95	-4.16	-	-	-	-	-									
Square root transformation	0	0.595	0.586	0.623	0.623	-0.584	-0.588	-0.606	-0.621	0.240	-0.006	-0.005	-0.011	-0.013	0.004	0.006	0.008	0.011	0.555	0.534	-	-	-	-									
	t	33.12	32.66	24.71	24.69	-32.53	-23.16	-24.05	-17.40	68.17	-1.83	-1.34	-2.23	-2.62	1.18	1.21	1.68	1.56	1.56	-	-	-	-	-									

Estimates from fitting logistic regression to dichotomized sum scores are uninformative and not presented for this model

Table 9 Estimates recovered fitting full model to simulated trait values and derived variables under Model 5 (no dominance; homozygous G × E)

Simulated parameters		Estimated parameter values																				σ	r^2									
		Genetic main effects (G)										Environmental effects (E) and G × E																				
		Homozygous					Heterozygous					Epistatic					Homozygous							Heterozygous					Epistatic			
d_A	d_B	d_{AB}	h_A	h_B	h_{AB}	i_{AB}	j_{AB}	j_{BA}	i_{AB}	i_{BA}	i_{AB}	h_B	h_A	h_{AB}	β_{d_A}	β_{d_B}	$\beta_{d_{AB}}$	β_{h_A}	β_{h_B}	$\beta_{h_{AB}}$	$\beta_{i_{AB}}$	$\beta_{i_{BA}}$	$\beta_{j_{AB}}$	$\beta_{j_{BA}}$	$\beta_{i_{AB}}$	$\beta_{i_{BA}}$	$\beta_{j_{AB}}$	$\beta_{j_{BA}}$				
0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.25	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0		
Data																																
True latent trait	Normal errors	0	0.989	0.959	0.023	0.028	0.017	0.024	0.019	0.009	0.009	0.028	0.028	0.017	0.024	0.019	0.009	0.501	0.252	0.258	-0.004	-0.006	-0.004	-0.003	-0.003	-0.004	-0.003	-0.003	-0.003	-0.002	1.006	0.842
	t	30.39	26.47	0.50	0.61	0.52	0.53	0.42	0.42	0.13	0.13	0.61	0.61	0.52	0.53	0.42	0.42	78.69	35.26	40.42	-0.043	-0.67	-0.55	-0.30	-0.30	-0.55	-0.28	-0.30	-0.18	-	-	
Normal errors	0	1.001	0.969	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.496	0.251	0.256	-	-	-	-	-	-	-	-	-	1.006	0.842	
	t	43.52	42.43	-	-	-	-	-	-	-	-	-	-	-	-	-	-	156.4	55.65	57.32	-	-	-	-	-	-	-	-	-	-	-	
Normal errors	0	2.258	2.254	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.495	-	-	-	-	-	-	-	-	-	-	-	1.037	0.832	
	t	485.1	486.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	151.4	-	-	-	-	-	-	-	-	-	-	-	-	-	
Normal errors	0	2.256	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.497	-	0.443	-	-	-	-	-	-	-	-	-	1.030	0.835	
	t	488.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	153.0	-	491.2	-	-	-	-	-	-	-	-	-	-	-	
Logistic regression	0	3.447	3.736	-1.784	-1.609	-1.229	3.356	2.891	0.01	4.909	4.909	-1.609	-1.229	3.356	2.891	0.01	1.041	0.505	0.405	-0.397	-0.433	-0.091	0.198	0.198	-0.091	0.103	0.00	0.01	-	-	-	
	t	0.020	0.022	-0.01	-0.01	-0.01	0.01	0.01	0.01	0.01	0.01	-0.01	-0.01	0.01	0.01	0.01	0.031	0.031	0.02	0.01	-0.01	-0.01	-0.00	0.00	0.00	-0.00	0.00	0.00	0.01	-	-	
Equal item parameters	Sum score	0	1.045	1.007	-1.102	-1.114	1.656	-1.654	-1.692	0.630	0.630	-1.114	-1.114	1.656	-1.654	-1.692	1.011	0.538	0.543	0.112	0.112	0.064	0.440	0.440	0.064	0.430	0.430	0.440	-0.300	2.543	0.747	
	t	12.69	12.23	-9.53	-9.57	20.12	-14.21	-14.64	3.85	3.85	3.85	-9.57	-9.57	20.12	-14.21	-14.64	62.67	33.36	33.66	4.95	4.92	3.398	19.45	19.45	3.398	18.85	18.85	19.45	-9.36	-	-	
Square root transformation	0	0.576	0.535	-0.148	-0.170	0.465	-0.098	-0.081	-0.119	0.040	0.040	-0.170	0.465	-0.098	-0.081	0.214	0.073	0.080	0.017	0.021	0.021	-0.067	0.068	0.065	-0.067	0.068	0.065	0.026	0.675	0.728		
	t	26.35	24.50	-4.84	-5.52	21.29	-3.16	-2.64	-2.73	0.040	0.040	-5.52	21.29	-3.16	-2.64	49.92	17.02	18.68	2.80	3.40	3.40	-15.65	11.23	10.86	-15.65	11.23	10.86	3.00	-	-		
Variable item parameters	Sum score	0	2.164	2.175	-0.389	-0.291	0.924	-0.221	-0.270	0.040	0.040	-0.389	0.924	-0.221	-0.270	0.725	0.275	0.273	0.092	0.068	0.068	-0.203	0.118	0.129	-0.203	0.118	0.129	-0.013	2.299	0.736		
	t	29.08	29.23	-3.72	-2.77	12.42	-2.10	-2.59	0.27	0.040	0.040	-3.72	12.42	-2.10	-2.59	47.74	18.88	18.72	4.52	3.31	3.31	-13.90	5.72	6.32	-13.90	5.72	6.32	-0.53	-	-		
Square root transformation	0	0.444	0.445	-0.026	-0.010	0.069	-0.007	-0.011	0.015	0.015	0.015	-0.010	0.069	-0.007	-0.011	0.105	0.035	0.035	0.021	0.017	0.017	-0.042	0.004	0.005	-0.042	0.004	0.005	-0.004	0.408	0.716		
	t	33.63	33.70	-1.41	-0.56	5.23	-0.37	-0.61	0.56	0.015	0.015	-0.56	5.23	-0.37	-0.61	40.44	13.60	13.52	5.66	5.47	5.47	-16.44	1.09	1.37	-16.44	1.09	1.37	-0.76	-	-		

Estimates from fitting logistic regression to dichotomized sum scores are uninformative and not presented for this model

fitted and a variety of reduced models that were expected to illuminate errors of inference that might attend the unwary.

Results

The results of fitting regression models for candidate genes and environmental effects are summarized for each of the five simulated data sets in Tables 5, 6, 7, 8, 9. Parameter estimates and t-values are given for each model, test and simulated data set ($N = 100,000$). Residual standard errors and squared multiple correlations from regression models are also tabulated where appropriate.

Central to the current exercise, when the full model is estimated for the true latent parameters, with all the parameters for the main effects and interaction of genes and environment, the precise pattern of simulated values for all five cases is recovered. In each case, parameter estimates are very close to their simulated values and the values of non-zero parameters typically yield highly significant t-values with these very large samples. Thus, for data simulated under the digenic additive model with no non-additive genetic effects or $G \times E$ (Table 5), estimates of d_a and d_b are 0.989 and 0.959 respectively, the regression on phenotype on environment is 0.501 and the residual variance is 1.006 as expected. All other estimates are close to their zero expected values. The results for the other data sets (Tables 6, 7, 8, 9) also correspond to their expected values as long as the true latent phenotypes are measured directly.

In contrast with the regressions on the true latent trait, the picture changes markedly when analysis is based on test scores for the digenic additive model. For example, even when data are simulated under the simplest additive model (Table 5), regression analysis of the symptom counts, S , derived from a test with equal item parameters (“Test 1”) yields a much less parsimonious model, in which not only homozygous effects of both loci are significant (though less so) but there are also marked non-additive genetic effects, including some dominance and strong additive \times additive epistasis. Furthermore, raw symptom counts yield highly significant evidence of homozygous effects on sensitivity to the environment: $(\beta_{da}, \beta_{db}) = (0.423, 0.442)$ and even evidence of higher order interactions with apparent epistatic effects contributing to $G \times E$: $(\beta_{iab}, \beta_{jab}, \beta_{jab}) = (-0.164, 0.160, 0.164)$.

A square root transformation of the test scores improves the fit of the additive model but, with this large sample, still yields evidence of significant homozygote \times homozygote epistasis ($i_{ab} = 0.483$) and supports some epistatic effects on $G \times E$ ($\beta_{iab} = 0.483$). Although non-additive effects may not be statistically significant with smaller sample sizes, estimates will be biased in the direction of detecting

spurious $G \times E$ leading to inflated type I errors when the properties of measurement are ignored.

The situation is much improved when analysis is conducted on scores derived from items with difficulties distributed uniformly over the range of latent trait values. Only the homozygous main genetic effects and main effect of the measured environment are significant. There is no convincing evidence of dominance, epistasis or $G \times E$.

Fitting the full model to the dichotomized trait values of test scores yields the correct conclusion for data simulated under the additive genetic model (Table 5), showing little support for any but homozygous main effects of the two candidate loci. However, significance levels are much reduced under the full model, reflecting substantial loss of information when the continuous variables are dichotomized. Fitting a model that ignores all possible non-additive effects yield highly significant estimates of the additive main effects of both loci but the gain in significance presumes prior knowledge of the genetic architecture that might not be justified in practice (compare results for other simulated genetic models).

Taken overall, the results of testing candidate gene models for epistasis and $G \times E$ may be seriously misleading even under the simplest additive genetic model (Model 1) when investigators are forced to analyze test scores based on items with restricted range of difficulty. Dichotomizing scores and trait values avoids much of the potential bias but at the cost of dramatically reduced power in exploratory analysis. The problem is only partly resolved by a square root transformation of test scores but difficulties can be minimized if it is possible to construct a test in which the item parameters span the range of hypothesized trait values.

Results for other, more complex, genetic architectures (Models 2–5, Tables 6, 7, 8, 9) only get worse. In every case, fitting the full regression model to the simulated latent trait values with normal errors (N) yields unbiased estimates and most conclusions are qualitatively correct when models are fitted to scores on the second test with items spanning a wide range of difficulty. However, dichotomizing the trait or test scores, even with these large samples, leads to such marked loss of information that recovery of the true genetic architecture may be difficult or impossible given the range of possibilities a priori. In virtually every case, scores based on counts of relatively infrequent symptoms yield spurious results of remarkable complexity. The problem is not generally resolved by simple transformation.

When the “true” model involves only additive and completely dominant effects at the two candidate loci, the results for the untransformed symptom counts (Test with equal item parameters, Table 6) provide strong support for complex non-additive effects, especially epistatic

interactions and $G \times E$ interaction. Transformation makes matters worse by strengthening support for epistatic interaction between the candidates. Scores on a test with uniformly distributed difficulties (Test with variable item parameters, Table 6) also suggest some epistasis but provide no hint of $G \times E$. Dichotomizing the scale makes it virtually impossible to say anything certain about genetic architecture in the two-locus case.

When the true model involves complementary gene interaction (Table 7) analysis of the sum of the raw symptom counts shows striking evidence for all types of $G \times E$ at the two loci. Indeed, statistical support for $G \times E$ far outweighs that for the additive, dominant and epistatic main effects of the candidate loci. If the effect of one interacting candidate locus is removed from the model, the main effects of the other are grossly overestimated.

In this example of complementary gene interaction, transformation redresses the balance somewhat by reducing the support for $G \times E$ but still yields markedly inflated type I error rates. A test with variable item difficulties (Table 7) recovers the right qualitative answer for the genetic architecture. Again, dichotomizing any of the scales makes it all but impossible to estimate any parameters of the full model with sufficient precision to resolve individual components of the model (results not tabulated).

The qualitative results in the presence of duplicate gene interaction (Table 8) resemble those for complementary epistasis but the symptom counts show still far greater support for $G \times E$ and the effects are largely untouched by transformation. Attempts to resolve all parameters of the full two-locus model are completely frustrated by lack of information about the critical features of the model in the dichotomous case (estimates not tabulated). In contrast to the finding in the presence of complementary epistasis, when one of the interacting loci is omitted from the models for the trait with duplicate gene interaction, estimates of the effect of the other locus are too small and far less significant than expected under the correct model.

All the above datasets were generated on the assumption of no $G \times E$ interaction in liability yet all provide strong evidence of non-additive effects when subjected to the vagaries of psychological testing. The final data set (Table 9) explores the consequences of simple digenic $G \times E$ in which the main effects of both loci are homozygous (only $d_a = d_b > 0$) and both loci show homozygous differences in their linear response to the environment ($\beta d_a = \beta d_b > 0$). If the true scores are known, the parameter estimates of the full model, including GE and epistasis, correspond to those of the underlying genetic architecture. Two further “wrong” models were fitted to the true scores to illustrate the possible biases that ensue from model misspecification. Omitting the two homozygous effects on $G \times E$ leads to grossly inflated estimates of

the main effects. Allowing one locus to affect the average response and the other to affect $G \times E$ ($d_a > 0$, $d_b = 0$, $\beta d_a = 0$, $\beta d_b > 0$) leads to biased estimates of both genetic parameters. As in other cases fitting the model to untransformed symptom counts (test with equal item parameters) produces substantially biased estimates and misleading conclusions supporting much more complicated models than necessary to account for variation in latent trait values. Consequences include spurious support for epistatic effects on average response and on response to the environment ($G \times E$). If anything, square root transformation only makes matters worse.

Can “Truth” be recovered?

The simulations presented are not intended to exhaust all the nuances of epistasis, $G \times E$ and measurement that might apply in any specific context but they certainly warn investigators not to oversell claims to seek or find $G \times E$ for measures of human behavior. Given that human behavioral and psychiatric genetics do not have access to *true* latent trait values or continuous measures of underlying biological processes, investigators have to rely on scores derived from clusters of indicators such as test items or symptoms. The simulations above confirm the intimate connection between the statistical conclusions drawn about the additive and non-additive contributions of candidate loci and the measured environment to behavioral traits. Even in the simplest case (Table 5) of a two-locus additive model (with no dominance, epistasis or $G \times E$), statistical analysis of counts based on many relatively infrequent symptoms biases results in the direction of detecting substantial epistatic and $G \times E$ effects. Indeed, *in this simple case, the effects of $G \times E$ and epistasis are expected to be more significant than the main effects of genes and environment.* A square-root transformation of the skewed symptom counts strengthens support for additive effects, but fails to remove the apparent contribution of epistasis and $G \times E$. In large samples, such as those simulated, the effects of $G \times E$ are expected to be statistically significant. With the smaller samples currently employed in psychiatric genetic epidemiology, significance of non-additive effects is comparable with that of the main effects pointing to a serious bias towards Type I Errors for the detection of epistasis or $G \times E$ even in transformed symptom counts.

Several possible solutions might be offered in the pursuit of unbiased truth. The symptom counts may be categorized (for example into affected and unaffected subjects) and models fitted by logistic regression. This approach may minimize spurious interaction in simple cases but usually leads to such a serious loss of power that choosing between models of different complexity will prove difficult if not

impossible with feasible sample sizes. It was difficult to find significant results with the large sample sizes used in the simulations.

A second approach is to design a better test, i.e. one in which item difficulties span a wide range of latent trait values, resembling the second simulated test in the examples above. In this case, regression analysis of a 20-item test recovers the “true” (additive) model with parameter sampling errors close to those that would be obtained if the true trait values were measured and little evidence for genetic effects on linear response to the environment. However, even a better test of this type is still affected by issues of scale.

We are thus led to the frustrating conclusion that anything we say about $G \times E$ in psychiatric genetics is critically dependent on the interface between biology and psychometrics to the point that analysis symptom counts and dichotomous outcomes is likely to be seriously misleading since estimates are biased and/or the type I error rates are higher than assumed. Patterns of main effect and interaction change as a function of the items chosen for measurement and the underlying truth about the genetic architecture of liability.

The ideal approach, suggested in a parallel set of simulations of $G \times E$ in twin data (Eaves 2014) is to integrate the model for genetic and environmental effects on liability with an item-response theory (IRT) model for the relationship between latent trait and test responses. If the IRT model is correctly specified, unbiased tests of the main effects may be recovered and some of the problems of misleading inference may be avoided. This approach has still to be tested fully in the candidate-gene context (though see Wray et al. 2008) but would seem to be a sine qua non for the development of a credible research program in the study of $G \times E$.

The last decade has witnessed unprecedented investment by researchers and funding agencies in the pursuit of $G \times E$ across many dimensions of human variation. Many of the models employed have been far simpler than some of those considered here and, once statistical significance has been achieved, publishable rationalization lurks close behind. Unfortunately, errors of the type described in this note are among the easiest to replicate and their uncritical dissemination risks distracting researchers from the more time-consuming task of “trying to get it right.”

Acknowledgments We are indebted to John Loehlin for his career of exemplary and critical scholarship that help generations of students realize that truth is not always what it seems. The age difference between the authors of this MS requires that the personal recollections in the first paragraphs are those of the first author alone. LJE is supported by grant R01AA5130267 from the National Institutes of Health (PI KS Kendler). We thank Nick Martin for suggesting this line of inquiry and Mike Neale for nuanced discussion of the implications of $G \times E$ and measurement for psychiatric genetics. Many of his thoughtful suggestions are not reflected in the current manuscript and warrant further inquiry.

Conflict of Interest There are no conflicting interests to declare.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

References

- Bartlett MS (1947) The use of transformations. *Biometrics* 3:39–52
- Bucio-Alanis L, Hill J (1966) Environmental and genotype-environmental components of variability II. Heterozygotes. *Heredity* 21:115–127
- Eaves LJ (2014) Genotype \times environment interaction in psychiatric genetics: deep truth or thin ice? *Psychol Med.* (in press)
- Eaves LJ, Eysenck HJ (1977) A genotype-environmental model for psychotocism. *Adv Behav Res Ther* 1:5–26
- Fisher RA, Immer FR, Tedin O (1932) The genetical interpretation of statistics of the third degree in the study of quantitative inheritance. *Genetics* 17:107–124
- Jinks JL, Fulker DW (1970) Comparison of the biometrical genetical, MAVA and classical approaches to the analysis of human behavior. *Psychol Bull* 73:311–349
- Loehlin JC (1965) Some methodological problems in Cattell’s multiple abstract variance analysis. *Psychol Rev* 72:156–161
- Mather K (1966) Variability and selection. *Proc Roy Soc Lond B* 164:328–340
- Mather K (1974) Non-allelic interaction in continuous variation of randomly breeding populations. *Heredity* 32:414–419
- Mather K, Jinks JL (1982) *Biometrical genetics, the study of continuous variation*, 3rd edn. Chapman and Hall, London
- Miko I (2008) Epistasis: gene interaction and phenotype effects. *Nat Educ* 1:197
- Purcell S (2002) Variance components models for gene–environment interaction in twin analysis. *Twin Res* 5:554–571
- Van der Veen JH (1959) Tests of non-allelic interaction and linkage for quantitative characters in generations derived from two diploid pure lines. *Genetica* 30:201–232
- Wray NR, Coventry WL, James MR, Montgomery GW, Eaves LJ, Martin NG (2008) Use of monozygotic twins to investigate the relationship between 5HTTLPR genotype, depression and stressful life events: an application of item response theory. *Novartis Found Symp* 293:48–59