Behav Genet
DOI 10.1007/s10519-016-9783-5

ORIGINAL RESEARCH

Dependence of Gene-by-Environment Interactions (GxE) on Scaling: Comparing the Use of Sum Scores, Transformed Sum Scores and IRT Scores for the Phenotype in Tests of GxE

Aja Louise Murray[1,2,3] · Dylan Molenaar[4] · Wendy Johnson[1,2] · Robert F. Krueger[5]

Received: 6 May 2015 / Accepted: 18 January 2016

Abstract Estimates of gene–environment interactions (GxE) in behavior genetic models depend on how a phenotype is scaled. Inappropriately scaled phenotypes result in biased estimates of GxE and can sometimes even suggest GxE in the direction opposite to its true direction. Previously proposed solutions are mathematically complex, computationally demanding and may prove impractical for the substantive researcher. We, therefore, evaluated two simple-to-use alternatives: (1) straightforward non-linear transformation of sum scores and (2) factor scores from an appropriate item response theory (IRT) model. Within Purcell's (2002) GxM framework, both alternatives provided less biased parameter estimates, and improved false and true positive rates than using a raw sum score. These approaches are, therefore, recommended over using raw sum scores in tests of GxE. Circumstances under which IRT factor scores versus transformed sum scores should be preferred are discussed.

Keywords Gene–environment interaction · Item response theory · Transformation · Scaling · Skewness

Introduction

Increasingly, theoretical perspectives on phenotypic development and expression are recognising that genes and environments transact in dynamic ways. Many posit some kind of gene–environment interaction (GxE) where GxE is defined as a differential response to environmental circumstances depending on genotype, or, a differential genetic expression depending on environment (Boomsma and Martin 2002; Eaves et al. 1977). GxE plays a central role in major theoretical models such as the diathesis-stress model, the differential susceptibility model, the vantage sensitivity model, and the bioecological model (Bronfenbrenner and Ceci 1994; Pluess and Belsky 2013; Reiss et al. 2013; Rende and Plomin 1992). The diathesis-stress model, for example, predicts that the genetic variance in a psychopathological trait is greater in more adverse environments whereas the bioecological model predicts that the genetic potential for a positive trait is realised to a greater extent in more stimulating, higher-quality environments (Asbury et al. 2005; Rende and Plomin 1992). GxEs are also cited as mechanisms by which social factors regulate behavior, for example, in the idea that genetic influences on certain phenotypes are prevented from being expressed when there are stronger social norms or explicit prohibitions relating to those phenotypes (Shanahan and Hofer 2005).

To keep pace with these theoretical developments, it has been necessary to develop statistical methodologies capable of modelling the more complex forms of interplay that they imply (e.g. Purcell 2002). Despite the promise and

Edited by Stacey Cherny.

✉ Aja Louise Murray
am2367@cam.ac.uk

[1] Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK

[2] Department of Psychology, University of Edinburgh, Edinburgh, UK

[3] Violence Research Centre, Institute of Criminology, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA, UK

[4] Psychological Methods, University of Amsterdam, Amsterdam, The Netherlands

[5] Department of Psychology, University of Minnesota, Twin Cities, MN, USA

Published online: 01 February 2016

widespread uptake of these methodologies, the ability to test theoretically implied GxE interactions is affected in practice by dependency of tests of interactions on the observed distributions or scales of the phenotypes (Eaves et al. 1977, 2002; Eaves 2006; Mather and Jinks 1971; Purcell 2002; Schwabe and van den Berg 2014).

The problem of dependency of GxE on phenotype scaling has been known since the time of R.A. Fisher, who noted that GxE interactions could be manipulated by rescaling the variables involved. In fact, he went far as to advocate 'transformations of scale' to eliminate what he perceived to be nuisance non-additivity (Tabery 2008). This suggestion was controversial because he was recommending purging the same non-additivity that was and still is viewed by many substantive researchers as a meaningful clue as to the causal processes underlying phenotypic development. Since then, numerous methodological studies have further discussed and provided demonstrations of dependency of appearance of presence of GxE on scaling (Eaves et al. 1977; Martin 2000; Molenaar et al. 2012; Purcell 2002; Tucker-Drob et al. 2009; van der Sluis et al. 2006). In the section that follows we summarise and extend the key arguments of these authors.

The primary challenge in dependency of GxE on scaling concerns the multiplicity of possible causal structures that could underlie the same sample phenotypic distribution. Consider the case where the observed distribution of the phenotype is non-normal: a common occurrence in behavior genetic research, as well as psychological research in general (Beasley et al. 2009; Micceri 1989). The problem is that when an observed phenotypic distribution is non-normal, this non-normality could reflect the presence of GxE, or it could simply be that the measurement instrument used has been unable to capture the full range of variation in the trait, leading to a skewed score distribution. A statistical test of GxE will not be able to distinguish among these possibilities easily.

The challenge of choosing between a 'scaling' and 'GxE' explanation for an apparent moderation effect is just one example of the broader challenge of selecting the correct model when a range of causal generating mechanisms could produce similar patterns in the observed data. For example, non-normality could arise for a number of methodological reasons aside from improper scaling e.g., failing to adequately sample individuals with the lowest or highest trait levels from the population. In terms of theoretically important processes, GxE is also difficult to distinguish statistically from non-linear main effects of a moderator on a phenotype or from non-linear genetic or environmental influences on a phenotype (e.g. Rathouz et al. 2008; Zheng and Rathouz 2015). However, there are good reasons to begin by attempting to rule out scaling as the alternative explanation for GxE effects. First, if improper scaling can account for apparent moderation effects, there is no need to posit complex interactions between the etiological influences on a phenotype, whether this is GxE or some other form of interplay. At a scientific level, incorrectly accepting a 'complex interplay' explanation can lead to theories which lack parsimony and which when further pursued may lead to wasted research efforts. At a more practical level, falsely selecting a 'GxE' explanation may foster the mis-impression that a candidate moderator is an important factor with respect to understanding variation in some phenotype, able to constrain or promote the expression of genetic liability, when in fact it is merely correlated with that phenotype.

Second, there is evidence that many phenotypic measures suffer from sub-optimal scaling. Cases in point are measures of psychopathological constructs. These very commonly yield observed non-normal (positively skewed) distributions because majorities of participants score close to the low (non-pathological) ends of the measurement scales. It is often argued that these observed distributions are not necessarily appropriate representations of the population distributions of the phenotypes but arise as a result of the scales being developed with focus on the upper extremes of the traits (van den Oord et al. 2000, 2003). This argument is based on various pieces of evidence, including the apparent highly polygenic nature of many common psychopathological disorders (e.g. Wray et al. 2014); on the observed normal distributions obtained when special care is taken to measure 'non-clinical' levels of psychopathological traits (e.g. Baron-Cohen et al. 2001); and on statistical comparisons of models positing categorical versus dimensional models of psychopathological traits (e.g. Walton et al. 2011). None of these is definitive evidence that psychopathological traits are normally distributed in the population but together they suggest that this may be closer to the truth than the classical categorical models in which meaningful variation in psychopathological traits is restricted to a narrow, clinical range of trait values. Under this dimensional view, failure to observe a normal distribution for a trait may be a result of failing to measure that trait with items that have an appropriate range of difficulties to provide reliable coverage of the whole trait distribution.

Within an item response theory (IRT) framework, such a failure will be manifested as item difficulties that are tightly clustered at the high end of the range; a phenomenon observed in many psychometric studies of commonly used inventories of psychopathologies (Meijer and Egberink 2012; Reise and Waller 2009; Thomas 2011). These scales have high discrimination in and around clinical cut-off points but poor discrimination in the healthy ranges. Thus, in a population-representative sample that would include predominantly subjects considered healthy, most participants completing such a test will endorse the lowest response options for most items, leading to a positively skewed score distribution and an apparent lack of individual differences at low levels of the phenotype due to the absence of items tapping this level.

If raw scores such as sums of items from scales affected in this way are used to represent phenotypes, they are likely to provide biased tests of GxE (Molenaar and Dolan 2014; Schwabe and van den Berg 2014). This is because GxE estimates depend on the degrees of individual differences in a phenotype at different levels of the moderator. Use of a scale that fails to capture these adequately at lower levels will tend to falsely indicate less variation at lower levels, when in fact this is really a function of weaker measurement at lower levels. The direction of the resulting bias in GxE depends on both skewness of the score and extent of correlation with the moderator. Positive skew combined with a positive moderator-phenotype correlation is liable to produce a positive interaction parameter, while negative skew combined with a positive moderator-phenotype correlation is liable to produce a negative interaction parameter. Thus moderation effects can arise even when there are no causal processes corresponding to our conceptual models of GxE influencing phenotypic development.

In empirical studies a researcher is faced with the challenge of choosing the most appropriate scale for the measure used to capture the relevant phenotype. To the extent that any phenotype actually has a latent dimensional distribution, it can be thought of as having some correspondingly dimensional scale of measurement, but for psychological constructs, we have little or knowledge of what these scales might be. Still, there are more or less appropriate choices given what is known about the underlying etiology of a trait, its distribution in the population, and the research question of interest (e.g. see Falconer and Mackay 1996). The appropriate scale for a phenotypic measure cannot be selected based on its observed score distributions or other features of the data: it must be selected based on conceptual knowledge and assumptions regarding the phenotype underlying the measures. Deviations of phenotypic distributions from expectations derived from these assumptions should be cause for concern.

Compounding this challenge is the fact that most behavior genetic modelling approaches require assumptions of multivariate normality[1] and that violations of those assumptions can lead to incorrect inferences regarding the presence of GxE (Van Hulle and Rathouz 2015). With this in mind, researchers have tended to deal with non-normal score distributions by employing straightforward non-linear transformations intended to remove the non-normality. For positively skewed sum scores, the log-transformation is popular (e.g. Hicks et al. 2009b; Johnson et al. 2011) but the square root transformation is also sometimes used (e.g. Distel et al. 2011). Given that the same approach is recommended to remove GxE interactions that are artifacts of phenotypic scaling (e.g. see Falconer and Mackay 1996 ch. 17), one might conclude that this also represents a solution to the problem of dependency of GxE on scale. There are, however, at least two major reasons to question this. First, while there has been no systematic simulation study evaluating their effectiveness in mitigating bias due to sub-optimal scaling, Kang and Waller (2005) demonstrated that sum score transformations were only moderately successful in reducing the tendency towards spurious phenotypic interactions in the context of moderated multiple regression. Second, and more importantly: presence of GxE introduces non-normality into the phenotypic distribution because it is by definition a relative expansion or contraction of variance in the phenotype across levels of the moderator. This suggests that transforming a non-normal score to normality could ‘transform away’ the very interaction effect of potential interest.

As another possible solution, some authors have suggested separating out scaling and GxE sources of non-normality by modelling GxE using an explicit measurement model (the scaling part) in combination with a biometric model (the GxE part). Essentially, the proposal is to model the scaling properties of items to account for differences in informativeness of phenotypic estimates across levels of the moderator. For example, if a scale has items that have difficulties that are clustered towards one end of the scale, a psychometric model with potential to recognize this can be integrated into a broader biometric model so that these parameters can be freely estimated and reflected in the estimates of the biometric parameters. The particular choice of measurement model will vary from phenotype to phenotype and be dictated by expectations about the latent trait distribution and the item response format.

For continuous indicators, Molenaar et al. (2012) demonstrated the feasibility of this approach in a GxE model in which GxE was operationalised as heteroscedastic E or C variance across levels of A. They showed that when differences in item residual variances across phenotypic level were incorporated into a measurement model and combined with a test of GxE, biasing effects of poor scaling were substantially mitigated. Similarly, Tucker-Drob et al. (2009) suggested a procedure in which a factor model with quadratic factor loadings was estimated in one stage and then, in a second stage, the same

[1] Purcell’s GxM approach assumes a normal distribution for the phenotype conditional on the moderator; however, the presence of moderation will result in a skewed marginal distribution for the phenotype.

measurement model (with parameters fixed to the values estimated from the first stage) was combined with Purcell's GxE model. Quadratic factor loadings allow for the relation between the items and latent phenotype to vary across levels of the phenotype: an effect that could otherwise be mis-attributed to GxE. However, truly continuous indicators are rare; therefore, Molenaar and Dolan (2014) and Schwabe and van den Berg (2014) proposed models for (ordered) categorical data that could be combined with a test of GxE. Again, using these models there was evidence of substantial reduction of bias in tests of GxE compared to using biometric models that did not explicitly model the scaling properties of the items used to measure the phenotype.

In spite of the potential utility of incorporating explicit measurement models for the phenotype into tests of GxE when an assumption about the underlying distribution of the genetic and environmental influences on the phenotype can be made, there have been very few studies taking this approach. One reason may be that the approach is mathematically complex and thus somewhat inaccessible for non-methodologists. There may also be a misconception that, because scores from these models will be highly correlated with sum scores, there would be essentially no benefit from using such models. It is not valid, however, to conclude that highly correlated measures will have the same properties in moderated models such as those that test for GxE. This is because correlations are sensitive mainly to rank orders, which can be highly preserved even when distributional properties differ markedly. Distributional properties are particularly important in any situation involving any kind of nonlinearity such as that involved in interactions.

Misconceptions aside, there are practical limitations to the various approaches discussed above, and it is not clear what the best approach might be. For example, the Schwabe and van den Berg (2014) approach requires assumption that IRT parameters are known, the Molenaar and Dolan (2014) approach is computationally intensive, and the approaches of Molenaar et al. (2012) and Tucker-Drob et al. (2009) require continuous indicators. Further, all were applied within the context of specific GxE models, potentially limiting their general applicability in practice.

Given these potential practical limitations, another possibility is to use a two-step approach to estimating GxE. In this approach, an appropriate measurement model for the phenotype is estimated, factor scores are obtained from this model, and then in a separate stage, these factor scores are submitted to a biometric model to test GxE. The 'two steps' refer to the use of two separate models, and the approximation involved in using explicitly calculated factor scores to measure a variable conceptualized as latent. This is in contrast to the one-step approach described above in which the biometric and psychometric model are estimated together, in a single step.

Although there has been no systematic study of this approach in GxE models, simulation studies have shown that a two-step approach works well in reducing bias due to scaling in phenotypic-level interactions in moderated multiple regression and factorial ANOVA (Embretson 1996; Kang and Waller 2005; Morse et al. 2012). For example, Kang and Waller (2005) showed that the tendency for spurious interactions to result from poor item scaling was substantially mitigated when IRT scores from a 2-parameter logistic model were utilised in place of sum scores. This strategy also proved more effective than a simple non-linear transformation of the score. Therefore, it is possible that a two-step approach could provide a compromise between the greater conceptual and computational simplicity of using a sum score and the effectiveness of IRT-based latent trait estimates in accounting for the scaling properties of items.

Based on the preceding argument, we compared a two-step approach to the currently most commonly used methods for handling observed non-normal phenotypes, that is, the raw sum scores and the transformed sum scores. We compared these three approaches using a statistical simulation study complemented by a real data example.

Modelling approach

We based our analyses on the Purcellian GxM interaction (where the 'M' stands for measured environment) framework initially introduced by Purcell (2002) and subsequently extended and evaluated by others (Rathouz et al. 2008; van Hulle et al. 2013; Van Hulle and Rathouz 2015; Zheng and Rathouz 2015; Zheng et al. 2015). This framework is arguably the foremost in assessing theoretical hypotheses which predict moderation of genetic influences on a specific phenotype by a specific moderator because in addition to accommodating both gene–environment interaction and gene–environment correlation, it can also be used to evaluate a range of other forms of phenotype-moderator transactions (see Zheng and Rathouz 2013). Uptake of the GxM modelling approach has been extensive; it has been employed to assess substantive hypotheses relating to a diversity of phenotypes including cognitive ability (Harden et al. 2007), physical health (Johnson and Krueger 2005), health behaviors (Timberlake et al. 2006), social relationships (South et al. 2008), and psychopathological traits (South and Krueger 2011). The popularity and influence of the approach is indicated by the fact that, at time of writing, the Purcell (2002) article has been cited almost 500 times.

We focussed on a form of the model that can be used to assess gene-by-measured environment interaction. The moderator (M) is modelled as:

$$M = a_M A_M + c_M C_M + e_M E_M \tag{1}$$

and the phenotype (P) as:

$$\begin{aligned} P = {} & (a_C + \alpha_C M)A_M + (c_C + \gamma_C M)C_M + (e_C + \varepsilon_C M)E_M \\ & + (a_U + \alpha_U M)A_U + (c_U + \gamma_U M)C_U + (e_U + \varepsilon_U M)E_U \end{aligned} \tag{2}$$

where A, C and E refer to mutually uncorrelated multivariate normally distributed (each with mean = 0, variance = 1) latent additive genetic, shared environmental and unshared environmental influences respectively, α, γ and ε are moderation parameters that capture the moderation of A, C and E influences by M, with the subscripts $_C$ and $_U$ denoting 'common' (to P and M) and 'unique' (to P).

The parameter of interest is α_U which captures moderation of the genetic influences on the phenotype that are not shared with the moderator. When this parameter is positive, genetic influences unique to the phenotype increase with the moderator and when it is negative, they decrease with the moderator.

Simulation study

We evaluated the effect of poor scaling on estimates of α_U using Eqs. 1 and 2 as our population biometric model, simulating poor scaling of the phenotype (explained below), and then estimating the model in Eqs. 1 and 2 using this poorly scaled phenotype. For our population biometric model, we used the following parameter magnitudes: For the moderator and phenotypic means we set $\mu_M = \mu_P = 0$; for the latent genetic and environmental influences on the moderator and phenotype we set $a_U = \sqrt{0.2}$, $a_C = \sqrt{0.3}$, $a_M = \sqrt{0.3}$; $c_U = \sqrt{0.1}$, $c_C = \sqrt{0.1}$, $c_M = \sqrt{0.2}$; $e_U = \sqrt{0.2}$, $e_C = \sqrt{0.1}$, $e_M = \sqrt{0.5}$; and for the moderation parameters we set $\alpha_C = \gamma_C = \varepsilon_C = 0$ and varied the magnitude of α_U, γ_U and ε_U across conditions. To explore how bias in α_U was affected by direction of the skew of the observed score distribution and direction of the population interaction, we varied $\alpha_U =$ to be −.15, 0, and .15 across conditions. In addition, as resolvability of the α_U, γ_U, and ε_U parameters is often imperfect, we explored how the bias in α_U is affected by whether γ_U and ε_U represented interactions in the same versus the opposite direction to that of α_U. We did this by including a subset of conditions in which γ_U and ε_U were specified to have the same sign as α_U and a subset of conditions in which they were specified to have the opposite sign to α_U. In both cases the absolute magnitudes of γ_U and ε_U were specified to be .20 and .08 respectively while α_U was held constant at −.15. We chose these sets of conditions and corresponding parameter values with the goal of selecting realistic values based on our own experiences of working with empirical twin data and on other published studies. Because we could expect results to be broadly symmetrical for positive and negative skews and negative and positive interaction parameters, we did not implement a fully crossed simulation design, but focussed on models that were realistic and which covered key combinations of variables.

Together, this combination of population parameters resulted in a total of four population models, summarised in Tables 2 and 3. In each replication, we generated data for either 500 MZ and 500 DZ or 1000 MZ and 1000 DZ twins according to these models. To keep the model focussed on the question at hand, we did not consider sex differences.

Observed item-level data generation

We generated item level data for twin 1 and twin 2 phenotypes using two different models that reflect common scaling practices. We did not manipulate the scaling of the moderator because—as in moderated multiple regression—the scaling of the predictor is far less critical with respect to the accuracy of estimates of interactions (e.g. Van Hulle and Rathouz 2015). First, we used a graded response model (GRM; Samejima 1969) as the basis for linking the latent trait values for the phenotype (P) to observed item responses to give a set of conditions in which the scaling issues could be considered mild. These latent trait values were determined according to the GxM population models described in the previous section. We simulated these data using the catIrt package in R statistical software (Nydick 2014; R Core Team 2014). Here, the items are considered in dichotomous steps, each characterised by a 2-parameter logistic model but with discriminations constrained equal within items. Specifically, probability of a respondent i with level of the latent trait θ_i having a response x_{ij} that falls at or above a given category ($k = 1 \ldots m_j$) is specified as:

$$P^*_{ijk} = P\left(x_{ij} \geq k \middle| \theta_i, a_j, \beta_{jk}\right) = \frac{1}{1 + \exp\left[-a_j\left(\theta_i - \beta_{jk}\right)\right]} \tag{3}$$

where a_j is the discrimination parameter of item j and β_{jk} is the category difficulty parameter of category k in item j. Note that θ_i is identical to P in Eq. 2.

We generated data for 20 items with a_j and β_{jk} parameters provided in Table 1. This gave items with five ordinal levels. The β_{jk} parameters were chosen to yield positively skewed item and sum score distributions that mimicked

Table 1 Parameter values for IRT models used to simulate item responses

Item	α	Polytomous item parameters (GRM)				Binary item parameters (2PL)
		β_1	β_2	β_3	β_4	β
1	1.94	−0.27	0.84	2.23	2.74	1.29
2	1.93	−0.21	1.46	2.01	2.73	2.23
3	1.96	−0.11	1.50	2.38	2.82	0.67
4	2.13	−0.36	1.29	2.07	2.65	1.22
5	1.09	0.34	1.16	2.07	2.73	−0.03
6	1.13	−0.15	1.34	2.00	2.78	0.99
7	0.87	0.34	0.99	2.34	2.64	1.11
8	0.99	0.23	0.68	2.33	2.62	0.88
9	1.63	0.43	0.98	2.22	2.83	1.94
10	1.01	0.04	1.22	2.39	2.73	0.12
11	1.75	0.10	0.93	2.27	2.63	−0.33
12	0.80	0.01	0.67	2.20	2.75	0.89
13	0.67	0.37	1.49	2.42	2.67	0.45
14	1.91	0.13	0.89	2.29	2.92	1.01
15	1.06	0	1.29	2.09	2.96	2.20
16	0.55	0.50	0.76	2.32	2.81	2.03
17	1.88	−0.24	1.02	2.07	2.74	0.65
18	2.44	−0.40	0.80	2.09	2.86	1.00
19	0.90	−0.11	1.27	2.27	2.73	1.45
20	1.15	−0.24	0.65	2.17	2.73	1.20

α is an item discrimination parameter, β_1–β_4 and β are threshold parameters. The same α values were used in both the GRM- and 2PL-generated item responses

GRM graded response model, *2PL* 2-parameter logistic model

those commonly found in empirical research (e.g. Kang and Waller 2005). To do this, we selected β_{jk} for successive response categories so that a disproportionate number of responses would fall into the first and second response categories. We also specified the β_{jk} parameters for a given category to show variability across the 20 items within our simulated test which is more realistic than setting them all equal. Discrimination parameters, a_j, were selected by randomly sampling from a uniform distribution with $\min = 0.5$ and $\max = 2.5$.

Second, we generated item-level data designed to be less favorable with respect to its scaling properties. Specifically, we used the same discrimination values but instead of using five ordinal levels, we used a 2PL model with only 2 ordinal levels (i.e., binary items), again selecting difficulty parameters such that disproportionate numbers of responses fell into the response category indicating a lower trait level. This gave us a set of conditions in which the scaling issues could be considered more serious than the polytomous case. Here, the model linking latent trait values to observed item level responses was:

$$P^{*}_{ijk}\left(x_{ij} = 1 | a_j, \beta_j\right) = \frac{1}{1 + \exp\left[-a_j\left(\theta_i - \beta_j\right)\right]} \tag{4}$$

The a_j and β_{jk} parameters used are provided in Table 1.

True score

As a control condition, we generated scores for the phenotype according to Eqs. 1 and 2 for without introducing any scaling issues. These scores can therefore be considered ‘true’ phenotypic scores. We considered these true phenotypic scores in order to provide a baseline against which we could compare the results. This is necessary because even in the absence of any scaling problems, it is likely that the GxM model will not perfectly recover all moderation parameters and because moderation parameters may be difficult to resolve from one another. For example, moderation of shared environmental influence may be to some extent mis-attributed to moderation of genetic influences.

Sum score

We created a sum score for the phenotype summing the scores from the 20 item responses generated as described above by Eqs. 1, 2, and 3 for the GRM and by summing the 20 item responses generated as described by Eqs. 1, 2 and

4 for the 2PL. Examples of the resulting sum score distributions are shown in Fig. 1 (polytomous) and Fig. 2 (binary). These sum score distributions exhibited positive skew, similar to that observed in many measures of psychopathological traits. In the binary case, this would correspond to the kind of summed 'presence versus absence' symptom scores found in diagnostic data. Skew also depended on the direction of interaction in the population model, with positive interactions making score distributions more positively skewed and negative interactions making score distributions more negatively skewed. However, these effects were relatively minor in comparison to the effect of scaling on the phenotypic distribution.

Transformed sum score

We created transformed sum scores by $\log_{10}$ transformations of the sum scores generated as described in the previous section. The $\log_{10}$ transformation, the natural log transformation, and other similar kinds of transformation of the phenotype are commonly used in GxE models when the phenotype has a positively skewed distribution (e.g. Button et al. 2010; Hicks et al. 2009a; Hicks et al. 2009a, b; Johnson et al. 2011; Silventoinen et al. 2009; Tuvblad et al. 2006). Transforming the sum scores gave rise to approximately normal distributions (see Figs. 3 and 4).

IRT scores

We obtained factor scores by fitting an IRT model to the item data and using the resulting item parameters to estimate IRT-based individual phenotype scores, usually referred to as 'factor scores' (Chalmers 2012). To estimate item parameters for the polytomous items we fit graded response models and to estimate item parameters for the binary items we fit 2PL models. As we originally generated the data according to these models, we knew that these were the appropriate measurement models, however, in real applications this choice should be based on considerations of the response format of items and the likely form of relations between item responses and the latent phenotype. We then computed IRT-based estimates of the phenotypic level for each individual in the sample by combining information from their patterns of item scores with the estimated item parameters from fitting the graded response model. We used Expected a Posteriori (EAP) scores: a Bayesian approach based on finding the mean of a posterior distribution representing the likelihood of phenotypic scores given a response pattern (Embretson and Reise 2000). The posterior distribution is computed by multiplying the prior distribution (likelihoods of phenotypic levels

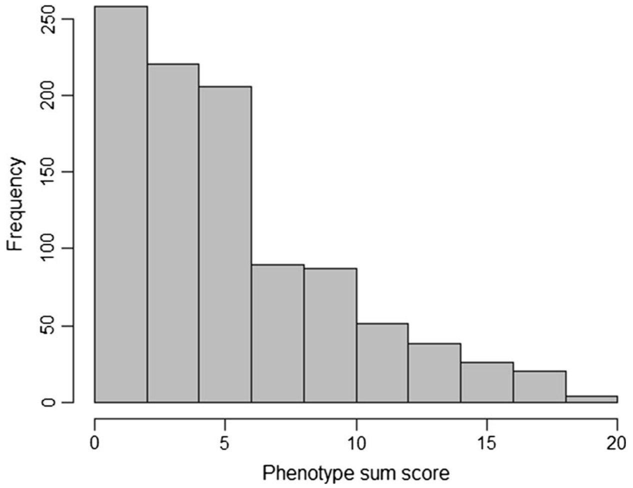


Fig. 2 Histogram showing the distribution of the sum score derived from generating item level data according to Eq. 4 with parameters in Table 1 (binary)

Fig. 1 Histogram showing the distribution of the sum score derived from generating item level data according to Eq. 3 with parameters in Table 1 (polytomous)

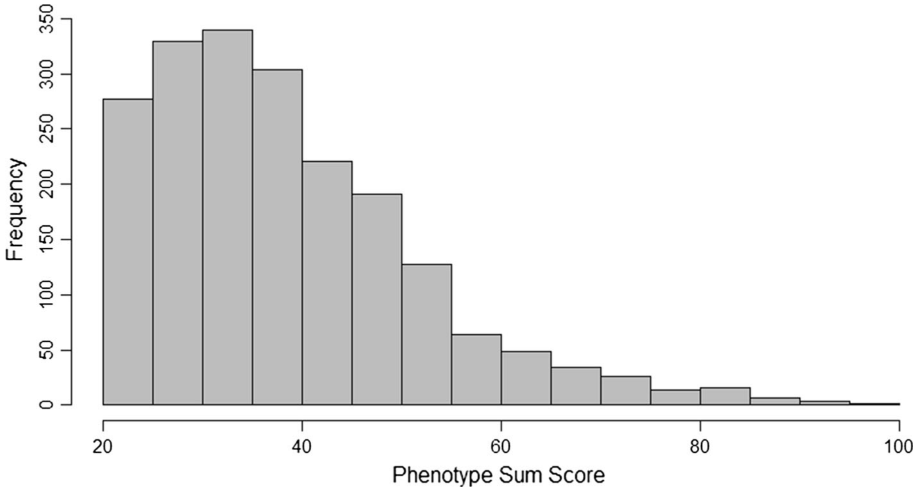
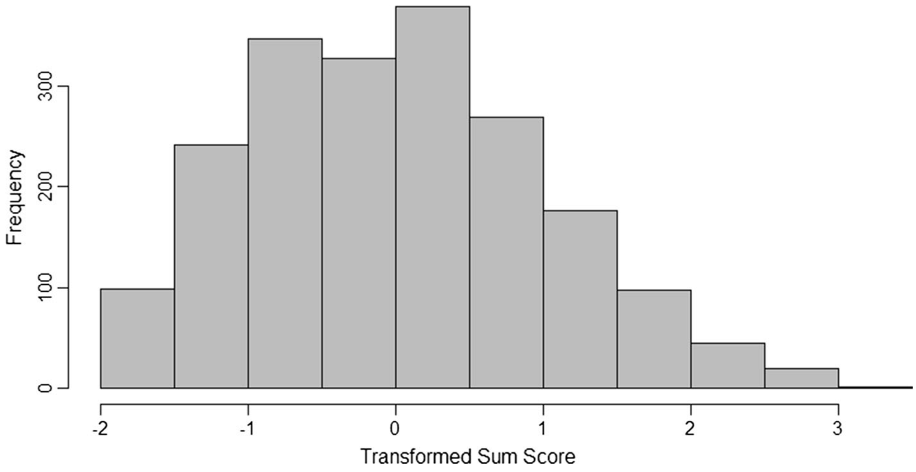


Fig. 3 Histogram showing the distribution of the transformed sum score derived from generating item level data according to Eq. 3 with parameters in Table 1 (polytomous) and then applying a $\log_{10}$ transformation



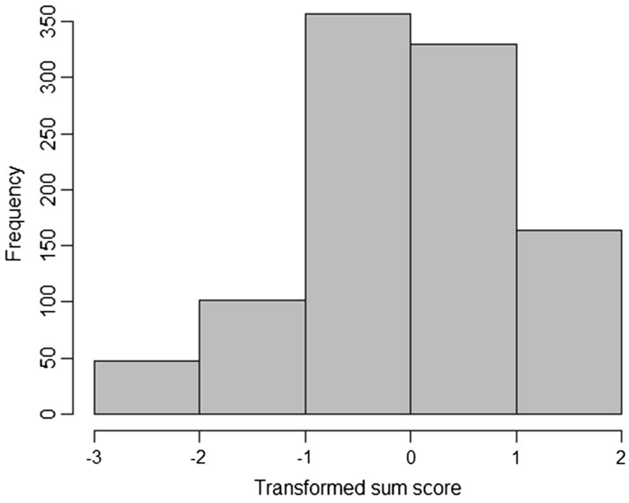


Fig. 4 Histogram showing the distribution of the transformed sum score derived from generating item level data according to Eq. 4 with parameters in Table 1 (binary) and then applying a $\log_{10}$ transformation

occurring in the population) by the likelihood of the observed response pattern given the phenotypic level (Embretson and Reise 2000). This method was selected among available factor score estimation approaches because it is easy to implement and available in most IRT software packages. In the context of the models used here in which the trait of interest was uni-dimensional and the sample size large, other commonly used scoring methods such as maximum a posteriori (MAP) scoring or maximum likelihood estimates (ML) should perform similarly to EAP. Unlike using sum scores as a proxy for the phenotype, this method takes into account the scaling properties of the items. For example, in an IRT model in which items differ in discrimination, each item's contribution to the sum score will depend on its discrimination. Estimating factor scores in this way gave phenotypic scores with an approximately normal distribution (see Figs. 5 and 6).

Summary of simulation conditions

The combination of GxE interaction parameters ($\alpha_U, = -.15$ vs 0 vs .15), other interaction parameters ($\gamma_U = .20$ and $\varepsilon_U = .08$ vs $\gamma_U = -.20$ and $\varepsilon_U = -.08$), item response model (GRM vs 2PL) and score type (true, sum, transformed, IRT) resulted in 28 simulation conditions. These are outlined in Tables 2 and 3. We generated 100 datasets for each condition to give 100 replications per condition.

Model fitting

To the 100 simulated datasets for each simulation condition (see Tables 2 and 3), we fit the GxM model described in Eqs. 1, 2. We fit the models in Mx (Neale et al. 2006) using maximum likelihood estimation, making use of the script accompanying Purcell (2002) which the author made available on his website. All latent A,C and E variances and covariances were freely estimated, α_C, γ_C, and ε_C were fixed to zero, and α_U, γ_U and ε_U were freely estimated. In other words, the model we fit to each dataset was consistent with the true model. The main parameter of interest was α_U, which captures the moderation of the additive genetic variance unique to the phenotype by M. Parameter bias was the difference between the population magnitude and the mean estimated value across the 100 replications within a condition. In addition, we conducted a likelihood ratio test (comparing a model in which α_U was freely estimated to one in which it was constrained to zero) for each replication to evaluate the statistical significance (using alpha = .05) of the α_U, parameter. Based on these, we computed false positive and false negative rates across the 100 replications. False negative rate was defined as the proportion of replications in which α_U, was non-significant in the presence of a non-zero population parameter. False positive rate was defined as the proportion of replications

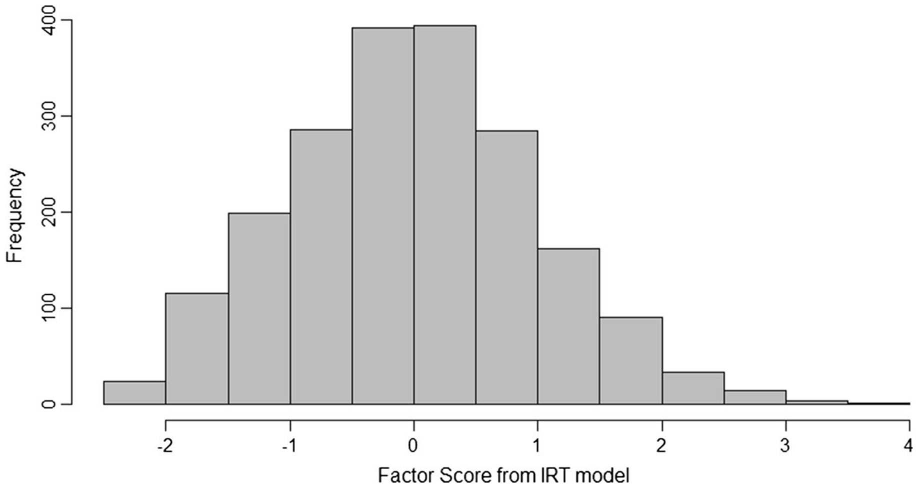


Fig. 5 Histogram showing the distribution of factor scores derived from generating item level data according to Eq. 3 with parameters in Table 1 (polytomous), fitting a graded response model, and then obtaining factor scores based on this model

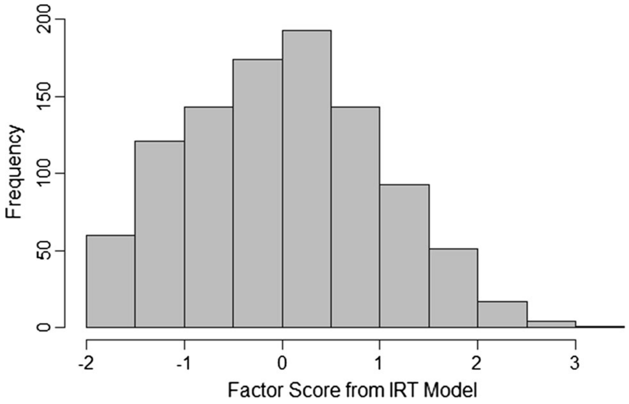


Fig. 6 Histogram showing the distribution of factor scores derived from generating item level data according to Eq. 4 with parameters in Table 1 (binary), fitting a 2PL, and then obtaining factor scores based on this model

in which α_U was significant in the presence of a null population parameter or where α_U was statistically significant but its value was in the opposite direction to its population value (e.g. negative sample value with a positive population value).

Simulation study results

Simulation study results are provided in Tables 2 and 3. There was only one convergence failure across all the models fit; therefore, scaling of the phenotype did not seem to have a strong influence on model convergence. Both transforming to normality and using IRT scores provided overall improvement over using raw sum scores. Whether transformed or IRT scores performed better depended on the number of response options: IRT scores were superior for polytomous items but transformations to normality were superior for binary items. More specific results are discussed below.

Control conditions

Results for the control conditions are provided in the 'true score' rows of Table 2. In these conditions, the α_U parameters were generally recovered well. There was a slight positive bias when the α_U parameter was in the opposite direction to the other moderation parameters. This bias appeared to reflect the imperfect resolvability of α_U from $\boldsymbol{\gamma}_{\mathbf{U}}$ and ε_U because it was accompanied by a negative bias in these two parameters. Power to detect moderation of the genetic influences unique to the phenotype was also generally good, as indicated by the true positive rates of 75 % and above. It was lowest in the condition in which α_U was in the opposite direction to the other moderation parameters. The type 1 error rates fell short of nominal levels (i.e. 5 %), staying at 0 % across all population models at both sample sizes.

Sum scores conditions

Results using a poorly scaled sum score are provided in the 'sum score' rows of Tables 2 and 3. In all of these conditions there was positive bias in the α_U parameter. These biases are in the positive direction because the IRT parameters used to generate the data produced positively skewed sum scores when the true scores were approximately normally distributed. Had item parameters been selected to produce negatively skewed sum scores, negative biases would have occurred.

Positive α_U bias was largest in conditions in which the true moderation parameter was in the opposite direction to the direction of skew (i.e. a negative or null population

Table 2 Performance of sum score, transformed score and IRT score with polytomous items

Score type	Population GxM values						N = 1000 twin pairs				N = 2000 twin pairs			
	a_c	c_c	e_c	α_U	γ_U	ε_U	Average α_U (SD)	α_U Bias	α_U true positive rate (%)	α_U false positive rate[a] (%)	Average α_U (SD)	α_U Bias	α_U true positive rate (%)	α_U false positive rate[a] (%)
True	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	.15	.20	.08	.15 (.04)	.00	98	0	.15 (.03)	+.00	100	0
True	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	−.15	.20	.08	−.12 (.05)	+.03	75	0	−.14 (.03)	+.01	97	0
True	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	0	.20	.08	.00 (.03)	.00	N/A	0	.00 (.02)	+.00	N/A	0
True	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	−.15	−.20	−.08	−.15 (.05)	.00	96	0	−.15 (.04)	.00	96	0
Sum	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	.15	.20	.08	.22 (.05)	+.07	94	0	.23 (.04)	+.08	98	0
Sum	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	−.15	.20	.08	.03 (.08)	+.18	1	8	.02 (.05)	+.17	2	8
Sum	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	0	.20	.08	.14 (.07)	+.14	N/A	54	.13(.05)	+.13	N/A	87
Sum	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	−.15	−.20	−.08	−.06 (.05)	+.09	15	0	−.05 (.03)	+.10	23	0
Transformed	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	.15	.20	.08	.16 (.03)	+.01	73	0	.16 (.03)	+.01	98	0
Transformed	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	−.15	.20	.08	−.02 (.05)	+.13	4	0	−.02 (.03)	+.13	8	1
Transformed	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	0	.20	.08	.08 (.04)	+.08	N/A	23	.08 (.03)	+.08	N/A	63
Transformed	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	−.15	−.20	−.08	−.11 (.03)	+.04	68	0	−.11 (.02)	+.04	97	0
IRT	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	.15	.20	.08	.16 (.04)	+.01	80	0	.16 (.03)	+.01	98	0
IRT	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	−.15	.20	.08	−.06 (.05)	+.09	13	0	−.07 (.03)	+.08	50	0
IRT	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	0	.20	.08	.06 (.05)	+.06	N/A	16	.05 (.03)	+.05	N/A	26
IRT	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	−.15	−.20	−.08	−.13 (.03)	+.02	79	0	−.12 (.02)	+.03	98	0

[a] False positive defined as significant effect in opposite direction to population parameter or significant effect in any direction when population parameter is zero. True positive defined as significant effect in the correct direction

moderation parameter with a positively skewed score) and the other moderation parameters. Here the biasing effects of scaling and imperfect resolvability of the α_U and $\boldsymbol{\gamma}_{\mathbf{U}}$ parameters combined to give a larger overall positive bias. Bias was slightly worse when using binary rather than polytomous items.

Both false and true positive rates varied considerably depending on the combination of skew and moderation direction. Power was lower when using binary items than when using ordered-categorical items and when analysing 1000 rather than 2000 twin pairs. Power was also, with the exception of the condition in which the scaling enhanced a positive moderation effect, quite poor.

False positive rates were also unacceptably high and far above nominal levels. For example, in the conditions in which there was no moderation effect; significant moderation was detected 54 and 46 % of the time using polytomous and binary items respectively. One notable result was that when α_U was negative and γ_U and ε_U were positive., detection of moderation using sum scores derived from summing binary items occurred *only* in the wrong direction. That is, while there were 13 % false positives, there were no true positives at all. Collectively, these results suggest that moderation detected using sum scores suspected to depart from the distribution of the underlying phenotype should not be relied upon.

Transformed sum scores conditions

Overall, the effect of transforming sum scores to normality was to reduce the bias in the GxE estimates. The effectiveness of the transformation varied considerably and for the most part some positive bias remained. The exception was that in the conditions in which a sum score was formed from binary items and in which α_U was in the same direction as the other moderation parameters, the transformation over-corrected the scaling problems, leading to a negative bias in α_U.

In the conditions in which α_U was negative, transforming sum scores improved but did not universally

Table 3 Performance of sum score, transformed score and IRT score with binary items

Score type	Population GxM values						N = 1000 twin pairs				N = 2000 twin pairs			
	a_c	c_c	e_c	α_U	γ_U	ε_U	Average α_U (SD)	α_U Bias	α_U true positive rate (%)	α_U false positive rate[a] (%)	Average α_U (SD)	α_U Bias	α_U true positive rate (%)	α_U false positive rate[a] (%)
Sum	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	.15	.20	.08	.23 (.05)	+.08	81	0	.22 (.04)	+.07	97	0
Sum	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	−.15	.20	.08	.05 (.09)	+.20	0	13	.03 (.05)	+.18	0	11
Sum	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	0	.20	.08	.14 (.07)	+.14	N/A	46	.14 (.05)	+.14	N/A	79
Sum	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	−.15	−.20	−.08	−.04 (.05)	+.11	15	0	−.04 (.04)	+.11	15	0
Transformed	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	.15	.20	.08	.09 (.04)	−.06	34	0	.10 (.03)	−.05	67	0
Transformed	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	−.15	.20	.08	−.04 (.05)	+.11	4	0	−.05 (.03)	+.10	13	0
Transformed	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	0	.20	.08	.03 (.03)	+.03	N/A	0	.03 (.02)	+.03	N/A	4
Transformed	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	−.15	−.20	−.08	−.13 (.04)	+.02	49	0	−.14 (.03)	+.01	88	0
IRT	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	.15	.20	.08	.17 (.03)	+.02	79	0	.16 (.03)	+.01	98	0
IRT	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	−.15	.20	.08	.01 (.06)	+.16	0	2	.00 (.04)	+.15	1	3
IRT	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	0	.20	.08	.09 (.04)	+.09	N/A	32	.09 (.03)	+.09	N/A	59
IRT	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	−.15	−.20	−.08	−.08 (.03)	+.07	24	0	−.08 (.02)	+.07	67	0

[a] False positive defined as significant effect in opposite direction to population parameter or significant effect in any direction when population parameter is zero. True positive defined as significant effect in the correct direction. Refer to Table 2 for results of control conditions

successfully recover all the statistical power lost by using inappropriately scaled sum scores. Again the conditions most affected were those in which α_U was in the opposite direction to the other moderation effects. For example, the true positive rate dropped from 75 % for the true scores to only 4 % for the transformed sum scores when using either binary or polytomous items. However, transforming the sum scores to normality had the benefit of producing marked reductions in false positive rates. For example, when the population parameter was zero and N = 2000 twin pairs, the false positive rate was only 23 % when using a transformed sum score obtained from polytomous items, compared with 54 % when using a raw sum score. The corresponding drop for the sum scores obtained from binary items was 46 to 0 %.

IRT scores conditions

Results using factor scores derived from the relevant IRT model are provided in the 'IRT' rows of Tables 2 and 3. Like the transformed sum scores, these gave consistently less biased α_U parameter estimates than the raw sum scores. However, some positive bias remained in all cases, ranging from very mild (+.01) to substantial (+.16) and was again most pronounced when α_U was in the opposite direction to the other moderation parameters. When considering smaller sample sizes, the IRT scores yielded less biased α_U estimates than transformed sum scores for polytomous items; however, the opposite was true for binary items.

Similar to transformed sum scores, IRT scores recovered some but not all of the statistical power lost by inappropriate scaling. Whether it yielded superior power to transforming sum scores depended on the directions of moderation parameters and whether binary or polytomous items were used. In general, IRT scores provided greater power when items were polytomous but transformed sum scores were superior in this respect with binary items. This suggests that IRT scores are advantageous primarily when trait-level indicators are rated at greater levels of detail.

IRT scores did not prevent scaling-related false positives and although they did bring the false positive rates down, these rates remained above nominal levels. Using polytomous items, IRT scores were more effective in reducing the false positive rates than transforming sum scores; however, transforming was more effective when using binary items.

Real data example

Participants

We used data from the Minnesota Twin Registry (MTR), a comprehensive description of which can be found in

Krueger and Johnson (2002). The full MTR includes data from twin pairs born in Minnesota in one of 3 year ranges. It includes 4307 twin pairs born between 1936 and 1955, 901 twin pairs born between 1904 and 1943, and 391 male twin pairs born between 1961 and 1964. Eligible participants were identified from birth records, located, and invited to participate via mail. Additional incentives and invitations to participate were offered to those who did not initially respond. Zygosity determination was by self-reported similarity in eye colour, hair colour, overall appearance, and the difficulties others had in distinguishing two members of a pair. Analysis of a sub-sample of 74 twin pairs who underwent zygosity determination by serological analysis suggested that the self-report method had an estimated accuracy of 96 %.

Different subsets of the total MTR received different sets of measures. Data used in the current study were from 528 monozygotic twin pairs and 411 dizygotic twin pairs comprising 614 males and 1264 females who had completed measures of both personality and leisure time interests. The mean age of the sample was 37.11 (SD = 7.8).

Measures

Moderator

As our moderator we used a composite of items from the Minnesota Leisure Time Interest Test (Lykken et al. 1990). The scale asks participants to rate the extent to which they would be interested in pursuing a given activity assuming no time, health, or financial constraints. Participants rated their interest on a 5-point scale from 1 = 'No interest at all' to 5 = 'I would certainly do this'. In total, 120 activities were rated, but we selected 6 items to form an 'Intellectual Interests' scale. Selected items refer to the following activities: reading current non-fiction, taking a college course, reading literary classics, visiting galleries/museums/exhibitions, reading books/magazines or watching TV programs on science, and reading history/philosophy/biography. We checked that these items formed a reasonable uni-dimensional scale by fitting a single factor confirmatory factor model to the data from twin 1 of each twin pair. We used the Weighted Least Squares Means and Variances (WLSMV) in estimator in *Mplus 7.0* (Muthén and Muthén 2010) to account for the categorical item response format. The 6 items all showed standardised loadings of .50 or greater and yielded a good-fitting single factor model (RMSEA = .05, CFI = .99, TLI = .99, WRMR = 0.56). We therefore used the unweighted sum score of these six items as our moderator variable. Cronbach's alpha of the scale was .63.

Phenotype

As our phenotypes we used personality scales from the 300-item Multidimensional Personality Questionnaire (MPQ; Tellegen and Waller 2008). Participants were administered a version of the MPQ using a 2-point response scale. Items are phrased as statements to which participants answer 'True' or 'False' depending on whether they believe the statement describes their attitudes, opinions, interests or other characteristics.

We selected two scales that yielded oppositely skewed scores. First, we used the negatively skewed 'Well-being' scale comprising 18 items. High scores on this scale are presumed to be indicative of a cheerful and happy disposition, feeling good about oneself, being optimistic, and enjoying an interesting and exciting life. Second, we used the positively skewed 'Aggression' scale comprising 18 items. High scores on this scale are presumed to be indicative of physical aggression, enjoyment of scenes of violence or upsetting or frightening others, victimisation of others for personal advantage, and vindictive and retaliatory tendencies.

We varied how each phenotype was operationalised across conditions to mirror our simulation conditions. First, we used the raw sum score from each scale. Second, we used a transformation of the sum score that yielded an approximately normal distribution. Third, we used an IRT score for each scale. For this, we used a 2-parameter logistic model with a procedure otherwise identical to that described in the simulation study to estimate factor scores.

Model fitting

Model fitting broadly followed the procedure outlined in the simulation. However, because we were working with real data we did not know the true model and, therefore, relied on model fit comparisons to guide model selection. We first assessed whether it was possible to constrain moderation of the influences common to moderator and phenotype to zero without significant decrease in fit. We then attended to moderation of the influences unique to the phenotype. We present the parameter estimates from best-fitting model(s). In all cases, all latent A, C, and E variances and covariances were freely estimated.

Real data example results

Descriptive statistics

Descriptive statistics for the moderator and phenotypes are provided in Table 4. For the phenotypes, descriptive statistics are provided for sum scores, transformed sum scores and IRT scores. The Well-being sum score showed negative skew which was reduced considerably by a normalising transformation. The IRT factor scores for this phenotype showed a level of non-normality similar to the transformed sum score but slightly more negative. The correlation between Well-being and Intellectual interests was around $r = .18$ and practically unaffected by which phenotypic proxy was used. The correlations between the three kinds of scores derived from the Well-being items were all $>.97$.

The Aggression sum score showed positive skewness. The transformation to normality produced scores with a near-normal distribution. The IRT factor scores for this phenotype also substantially reduced non-normality but these scores were more positively skewed than the transformed sum scores. The correlation between Aggression and Intellectual interests was around $r = -.12$ and practically identical across the three different kinds of phenotypic proxy. The correlations between the three kinds of scores derived from the Aggression items were also all $>.97$.

GxM model fitting

Fits for selected models for each phenotype and type of phenotypic score are provided in Tables 5 and 6. The parameter estimates from the best-fitting model for each phenotype across the three different phenotypic proxies (sum score, transformed sum score, and IRT score) are provided in Table 7.

Well-being

In the GxE models for Well-being, it was possible to constrain moderation of the common influences to zero without significant decrease in fit irrespective of whether a sum score, transformed sum score, or IRT score represented the phenotype. Therefore, this became the baseline model for all further model comparisons.

Using sum scores, model comparisons supported moderation of the genetic influences unique to the phenotype fairly unequivocally. Constraining this parameter to zero produced significant decreases in fit irrespective of whether moderation of the unique C and E influences on the phenotype were freely estimated or fixed to zero. Model fit comparisons suggested the latter model provided the best overall representation of the data: a conclusion on which there was agreement across all the information theoretical criteria examined. Thus, results suggested that the genetic influences unique to Well-being were smaller at higher levels of intellectual interests.

Using transformed sum scores, model fit comparisons suggested some moderation of unique influences for which moderation of the A influences unique to the phenotype best accounted. However, this result was not completely unequivocal: it was possible to constrain moderation of the A influences unique to the phenotype to zero without significant decrease in fit when moderation of the C and E influences were freely estimated but not when they were both fixed to zero. This further illustrates the lack of resolvability of α_U and γ_U effects noted in the simulation study. The fact that GxE evidence was more marginal here was also reflected in the information theoretic fit criteria; for example, AIC was more negative for a model including α_U while BIC was more negative for the nested model excluding this parameter. This is consistent with BIC having a larger parsimony penalty for these models. For these sets of comparisons, results suggested that the genetic variance unique to Well-being may be higher at higher levels of intellectual interests.

When using IRT scores, results were highly similar to those for the transformed sum score in terms of fit differences and parameter magnitudes (α_U was 0.04 when freely

Table 4 Descriptive statistics for well-being, aggression and intellectual interests phenotypes

Phenotypic proxy	N MZ pairs	N DZ pairs	Mean (SD)	Skew	Kurtosis	Correlation with moderator
Intellectual Interests sum score	528	411	13.32 (3.75)	0.13	-0.27	N/A
Well-being sum score	525	406[a]	11.15 (2.21)	-1.06	0.71	.18
Well-being sum score transformed	525	406[a]	0 (1)	-0.36	-0.90	.19
Well-being IRT score	528	411	0 (0.89)	-0.42	-0.32	.18
Aggression sum score	525	411	3.66 (3.21)	1.12	1.09	−.12
Aggression sum score transformed	525	411	0 (1)	0.23	-0.79	−.12
Aggression IRT score	528	411	-0.04 (0.86)	0.46	-0.40	−.13

[a] There were an additional 4 incomplete twin pairs for these measures which were included in the analysis

Table 5 GxM model fits for Well-being phenotype

Model (freely estimated parameters)	−2LL	df	BIC	AIC	saBIC	DIC
Sum score						
$a_C, c_C, e_C, \alpha_C, \gamma_C, \varepsilon_C, \alpha_U, \gamma_U, \varepsilon_U$	10,204.50	3727	−7653.07	2750.50	−1734.73	−4228.18
$a_C, c_C, e_C, \alpha_U, \gamma_U, \varepsilon_U$	10,204.78	3730	−7663.18	2744.80	−1740.08	−4235.54
$a_C, c_C, e_C, \gamma_U, \varepsilon_U$	10,209.33	3731	−7664.34	2747.33	−1739.65	−4335.78
a_C, c_C, e_C, α_U	10,206.10	3732	−7669.38	2742.09	−1743.11	−4239.91
a_C, c_C, e_C	10,222.75	3733	−7664.47	2756.75	−1736.61	−4234.08
Transformed sum score						
$a_C, c_C, e_C, \alpha_C, \gamma_C, \varepsilon_C, \alpha_U, \gamma_U, \varepsilon_U$	10,214.25	3727	−7648.19	2760.25	−1729.85	−4223.30
$a_C, c_C, e_C, \alpha_U, \gamma_U, \varepsilon_U$	10,214.92	3730	−7658.12	2754.92	−1735.02	−4230.48
$a_C, c_C, e_C, \gamma_U, \varepsilon_U$	10,216.73	3731	−7660.64	2754.73	−1735.95	−4332.08
a_C, c_C, e_C, α_U	10,215.09	3732	−7664.88	2751.09	−1738.60	−4235.40
a_C, c_C, e_C	10,219.96	3733	−7665.87	2753.96	−1738.00	−4235.47
IRT score						
$a_C, c_C, e_C, \alpha_C, \gamma_C, \varepsilon_C, \alpha_U, \gamma_U, \varepsilon_U$	9806.21	3739	−7893.28	2328.21	−1955.88	−4457.37
$a_C, c_C, e_C, \alpha_U, \gamma_U, \varepsilon_U$	9806.89	3742	−7903.21	2322.89	−1961.05	−4464.54
$a_C, c_C, e_C, \gamma_U, \varepsilon_U$	9808.22	3743	−7905.96	2322.22	−1962.22	−4466.38
a_C, c_C, e_C, α_U	9807.08	3744	−7909.96	2319.09	−1964.62	−4469.45
a_C, c_C, e_C	9810.82	3745	−7911.51	2320.82	−1964.59	−4470.08

Table 6 GxM model fits for Aggression phenotype

Model (freely estimated parameters)	−2LL	df	BIC	AIC	saBIC	DIC
Sum score						
$a_C, c_C, e_C, \alpha_C, \gamma_C, \varepsilon_C, \alpha_U, \gamma_U, \varepsilon_U$	10,218.91	3732	−7662.97	2754.91	−1736.69	−4233.49
$a_C, c_C, e_C, \alpha_U, \gamma_U, \varepsilon_U$	10,222.38	3735	−7671.51	2752.38	−1740.46	−4239.27
$a_C, c_C, e_C, \alpha_U, \gamma_U,$	10,232.63	3736	−7669.80	2760.63	−1737.17	−4236.65
$a_C, c_C, e_C, \varepsilon_U$	10,224.28	3737	−7677.40	2750.28	−1743.18	−4243.33
a_C, c_C, e_C	10,240.40	3738	−7672.76	2764.40	−1736.96	−4237.77
Transformed sum score						
$a_C, c_C, e_C, \alpha_C, \gamma_C, \varepsilon_C, \alpha_U, \gamma_U, \varepsilon_U$	10,228.85	3732	−7658.00	2764.85	−1731.72	−4228.52
$a_C, c_C, e_C, \alpha_U, \gamma_U, \varepsilon_U$	10,232.34	3735	−7666.52	2762.34	−1735.48	−4234.29
$a_C, c_C, e_C, \alpha_U, \gamma_U,$	10,234.20	3736	−7669.01	2762.20	−1736.38	−4235.86
$a_C, c_C, e_C, \varepsilon_U$	10,234.73	3737	−7672.17	2760.73	−1737.96	−4238.10
a_C, c_C, e_C	10,238.00	3738	−7673.96	2762.00	−1738.16	−4238.97
IRT score						
$a_C, c_C, e_C, \alpha_C, \gamma_C, \varepsilon_C, \alpha_U, \gamma_U, \varepsilon_U$	9676.16	3739	−7958.30	2198.16	−2020.91	−4522.39
$a_C, c_C, e_C, \alpha_U, \gamma_U, \varepsilon_U$	9679.97	3742	−7966.67	2195.97	−2024.51	−4528.00
$a_C, c_C, e_C, \alpha_U, \gamma_U,$	9682.29	3743	−7968.93	2196.29	−2025.18	−4529.34
$a_C, c_C, e_C, \varepsilon_U$	9682.21	3744	−7972.39	2194.21	−2027.06	−4531.88
a_C, c_C, e_C	9687.08	3745	−7973.38	2197.08	−2026.46	−4531.95

estimated but the other moderation parameters were fixed to zero). However, the difference in fit between the model in which moderation of all the unique A,C and E influences on the phenotype was fixed to zero and the model in which moderation of the unique A influences was freely estimated happened to fall just short of statistical significance. In addition, with the exception of saBIC, all information theoretic criteria were more positive for the model with α_U than in the nested model excluding it. Therefore, there was essentially no statistical evidence for GxE when using the IRT factor score, suggesting that the genetic influences unique to Well-being did not depend on level of intellectual Interests.

To summarise results from the Well-being scale, based on a naïve interpretation, all favoured different conclusions regarding the presence of GxE: GxE was in evidence using a sum score, was somewhat in evidence using a transformed sum score, and was not in evidence using an IRT

Table 7 Parameter estimates from best-fitting models for Well-being and Aggression phenotypes

Phenotype				GxM parameter estimates			
Phenotypic proxy	Correlation with moderator	α_C	α_U	γ_C	γ_U	ε_C	ε_U
Well-being							
Sum score	.18	0 (fixed)	−.11	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)
Transformed sum score	.19	0 (fixed)	−.06	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)
IRT factor score	.18	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)
Aggression							
Sum score	−.12	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)	-0.07
Transformed sum score	−.12	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)
IRT factor score	−.13	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)	-0.03

score. While the results in the latter two conditions were in actuality very similar, the fact that the statistical evidence lay on opposite sides of a statistical significance threshold and a naïve interpretation could lead to very different substantive conclusions in practice. Only the sum score condition appeared to show unambiguous support for GxE. This is consistent with the simulation conditions in which the presence of non-normality resulted in detection of GxE, irrespective of whether this non-normality was a result of moderation or poor scaling. The moderation observed using the sum score was in the direction expected for a negatively skewed sum score even when there was no true moderation. Thus, there would be reason to question the validity of the evidence for GxE observed in this real data example.

Aggression

In all conditions, it was possible to constrain moderation of the influences common to moderator and phenotype to zero without significant drop in fit. From here, the best-fitting model using sum scores was one in which there was moderation of the unshared environmental influences on the phenotype captured by the ε_U parameter. Fixing ε_U to zero resulted in a significant deterioration in fit both when α_U and γ_U, were freely estimated and when they fixed to zero. Information theoretical criteria also unanimously supported the inclusion of ε_U. However, when this parameter was freely estimated, constraining moderation of neither shared environmental influences nor genetic influences on the phenotype resulted in statistically significant decrease in fit. Thus, using a sum score, there was evidence that only the unshared environmental influences unique to Aggression decreased with increasing Intellectual Interests. The direction of this moderation was in the opposite direction to the direction of the skew of the sum score. Given that the phenotype and moderator were negatively correlated, the moderation was in the direction consistent with the skew of the sum score.

Using transformed sum scores, after constraining moderation of the influences common to moderator and phenotype to zero, the best-fitting model involved no moderation of the influences unique the phenotype. These could all be individually constrained to zero without significant decrease in fit, irrespective of whether moderation parameters for the other unique influences were also constrained or freely estimated. Based on information theoretic criteria, model fit was close between models including and excluding ε_U, but was—except according to AIC—better when it was excluded. Thus, on balance there was technically no evidence that the genetic or environmental influences on Aggression depended on level of Intellectual Interests.

Using IRT scores, after constraining moderation of the influences common to the moderator and phenotype to zero, there was some very weak support for moderation of the unshared environmental influences unique to the phenotype. Specifically, fixing moderation of unshared environmental influences unique to the phenotype to zero resulted in significant decrease in fit when all other moderation parameters were fixed to zero; however, the decrease in fit on constraining this parameter to zero was not statistically significant when moderation of the shared environmental and genetic influences unique to the phenotype was freely estimated. The best-fitting model according to BIC included no moderation, albeit by a small margin compared with one in which the moderation of the unshared environmental influences unique to the phenotype was freely estimated (ΔBIC = 0.99). However, AIC and saBIC favoured the model with moderation (DIC differed only in the 2nd decimal place between the two models). Considering these results together, the IRT factor score condition showed only very weak evidence for moderation intermediate between the results for the sum score (which showed evidence for moderation) and the transformed sum score (which showed no evidence for moderation). Again, the direction of moderation suggested smaller unshared environmental influences unique to Aggression at higher levels of Intellectual Interests.

Discussion

It is well known that poorly scaled sum scores as phenotypic proxies in GxE tests can seriously bias tests of GxE. For example, using sets of items where the difficulty or location parameters are clustered near the high end of the phenotypic continuum can lead to positively skewed sum scores and, in turn, positively biased tests of GxE. In a simulation study, we assessed the extent to which this bias was mitigated by transforming non-normal sum scores to normality. We compared this to estimating phenotypic scores from an IRT model: a method that explicitly takes account of the scaling properties of items. Our results suggest that using IRT methods to provide formal models for the phenotype or appropriately transforming score distributions can provide much more accurate detection and quantification of GxE effects. Transformation may be preferred where there is insufficient information in the data (e.g. small sample size, small number of items, binary item response format) to provide good IRT latent trait estimates.

Based on our analyses, we can extend the arguments set out in the introduction in the following ways. First, we confirmed that biases in estimates of GxE can be introduced by phenotypic scaling that results in a sum score that fails to reflect the underlying distribution of the target phenotype. The nature of this bias is predictable: sum scores that are negatively skewed relative to their underlying phenotypic distribution will tend to produce negatively biased moderation parameters and sum scores that are positively skewed relative to their underlying phenotypic distribution will tend to produce positively biased moderation parameters. When there is no true moderation effect, this will often lead to unacceptably high false positive rates.

These effects occur because non-normality due to poor scaling is not completely statistically distinguishable from non-normality due to presence of interaction. Where there is non-normality, the model is liable to attribute this to interaction; however, only when the observed phenotypic distribution reflects its population distribution will this estimate provide accurate quantification of GxE. Measuring the phenotype and capturing its population distribution as accurately as possible is, therefore, important in ensuring accurate assessment of GxE. When the raw score from an inventory fails to do this, there may be options for recovering this distribution via post hoc manipulations of its measurement scale.

Our results showed, in particular, that transforming a score or using an IRT score in place of a non-normal sum score can be used to reduce in bias. We studied the case in which the *latent* genetic and environmental influences on the phenotype, absent the influence of the moderator could be assumed normally distributed in the population. This is a reasonable assumption in cases where there are a large number of small, independent effects on the phenotype. Here, a normal distribution of the joint effects of etiological contributors is predicted based on the central limit theorem. Under these conditions, using either a simple transformation or IRT scores reduced bias in GxE because they led to score distributions that better approximated the population distribution of the phenotype.

In cases where there is no true moderation effect, using a phenotypic proxy that better reflects its population distribution than a sum score reduces false positive rates substantially. When the direction of the moderation is consistent with the direction of skew, either transforming to normality or using an IRT score will give close to unbiased parameter estimates and result in good power to detect the effect. However, in cases where moderation and skew are in opposite directions, these methods will under-estimate the effect and reduce power to detect GxE relative to situations in which the phenotype is not subject to scaling problems.

We also provided a real data example from the Minnesota Twin Registry using two phenotypes with non-normal sum scores. Analysing the Well-being phenotype using (negatively skewed) sum scores yielded statistically and practically significant GxE whereas using IRT scores suggested no significant GxE. The transformed sum scores yielded evidence intermediate between these two outcomes. The direction of the GxE using sum scores was consistent with the direction of the skewness of the sum score. This suggests that the observed effect could be due to item scaling. Moreover, based on these results, researchers using sum scores rather than IRT scores could easily have been led to opposite substantive conclusions despite the high correlations between the raw and IRT scores.

The Aggression phenotype did not yield evidence of GxE irrespective of whether (positively skewed) sum scores, transformed sum scores, or IRT scores were used. This shows that non-normal trait distributions will not automatically result in the appearance of GxE and that altering phenotypic distributions will not necessarily affect the GxE parameter. However, there was evidence for dependence of another moderation parameter on scaling: using sum scores and an IRT scores, negative moderation of the unshared environmental influences unique to the phenotype (captured by the ε_U parameter) was detected. There was no such evidence using a transformed sum score. Taking into account the fact that the phenotype and moderator were negatively correlated, the ε_U parameter was proportional to and in the direction consistent with the skew of the phenotypic proxy. That is, the parameter was most negative when the phenotypic proxy was strongly skewed (sum score), less negative when the phenotypic proxy was moderately positively skewed (IRT score) and effectively zero when the phenotypic proxy was only slightly positively skewed (transformed sum score). Thus, although we have focussed on the α_U parameter because it is most often used to operationalise theoretical hypotheses, this example highlights the fact that the effects of scaling on GxE models are not confined to that one parameter.

Our results reinforce the message that poorly scaled sum scores should be avoided in tests of GxE. Poorly scaled sum scores, in addition to producing high false positive rates, can yield results that suggest significant moderation in the opposite direction to the true moderation effect. Demonstrating that sum scores are highly correlated with transformed sum scores or IRT scores for the same phenotype is thus not sufficient justification for using them in place of these better-performing methods. Because correlation coefficients are relatively unaffected by rank-preserving transformations, sum and functionally transformed scores will show very high correlations, even when their distributions are markedly different. IRT scoring basically differentially weights the items or response options rather than weighting each one equivalently as does sum scoring, thus very closely preserving rank ordering. This was illustrated in our real data examples where, in spite of leading to diverging conclusions about the presence and strength of moderation effects, the three types of score were correlated with one another at >.97.

The strategies of transforming sum scores to normality or using an IRT score did not suffer the limitations of poorly scaled sum scores to anywhere near the same extent; however, both resulted in tests that lacked statistical power when the moderation was in the opposite direction to skew and failed to control the type 1 error rate completely when GxE was not present. Overall, transforming non-normal sum scores to normality or using IRT scores will in many cases fail to address the biasing effects of poor scaling on GxE tests, especially when there is non-genetic moderation in the opposite direction to the genetic moderation. Therefore, evidence of GxE (or lack thereof) should be considered tentative even when obtained using transformed or IRT scores.

Although using IRT scores is more time consuming and technically demanding than transformations to normality, it may be worth the additional effort, especially when the raw scale items were rated using multiple response options. IRT scores can be estimated reasonably easily in a range of freely available software packages and have several practical and theoretical advantages over transformed sum scores. First, they are easily estimable in the presence of missing item data, or when respondents did not complete an identical set of items (Embretson and Reise 2000). Second, the diversity of available IRT models means that many kinds of response formats, scale structures, or theories about how the latent trait relates to item responses can be accommodated. For example, a bi-factor model could be fit when it is desirable to partition general and specific trait variance captured by a set of items (Cai et al. 2011); if a scale has a categorical response format, a nominal response model could be fit (Bock 1972); or if items follow an ideal point process an unfolding model can be fit (e.g. Chernyshenko et al. 2007). All of these and other features can be easily dealt with in an IRT framework while posing significant problems or being simply impossible to take account of when using sum scores, both raw and transformed to normality. Furthermore, while an IRT model can be chosen based on theoretical considerations, the choice of a transformation is somewhat arbitrary and usually driven by pragmatic considerations. The choice of an IRT model can be evaluated both overall and with respect to individual items using well-studied goodness-of-fit statistics and graphical checks. A beneficial side effect of this is that the process of fitting and evaluating IRT model(s) is likely to encourage explicit consideration of the assumptions that underpin the phenotypic proxy used. However, no analogous tests exist for transformations. More importantly, from a conceptual perspective, if the genetic and environmental influences on the phenotype in the absence of the influence of the moderator are normally distributed and there is true GxE in the population then the phenotype *should* show a non-normal distribution because GxE involves an expansion (or contraction) of the variance in a phenotype according to the levels of moderator. This expansion (or contraction) of variance shows up in the marginal distribution of the phenotype as non-normality that is commensurate with the GxE effect. Using a transformation to normality is, therefore, directly at odds with theoretical expectations when GxE is hypothesised. In IRT models, this is also a problem to some extent; however, the assumption of a normal latent distribution is not a necessity; where appropriate alternative prior distributions can be specified in a manner that is far more flexible than attempting to obtain that distribution through transformation of observed scores.

The primary disadvantage of IRT scoring is practical: to be effective requires large sample sizes and ideally a large number of items with polytomous response formats. Where any of these is lacking, transformed sum scores may be more effective than IRT scores. This underlines the importance of assessing the empirical reliability of factor scores from IRT models, as one would for sum scores (see

Culpepper 2013). Unreliable IRT scores will not only be ineffective in addressing bias in GxE; they will also result in attenuated estimates of twin correlations and bias other model parameters (van den Berg et al. 2007). Similarly, as the extent to which the accuracy of the scores as measures of the intended underlying dimension depends on the appropriateness of the IRT model, its specification should be carefully considered and its fit assessed empirically (see Embretson and Reise 2000).

Where both approaches are limited is that the underlying liability distribution absent the influence of the moderator could be non-normal due to other moderators or the effects of rare but highly influential etiological factors that engender extreme effects. Analogous to the problem of distinguishing non-normality due to moderation versus poor scaling, it is not easy to disentangle non-normality due to the effect of a moderator of interest and non-normality due to other etiological factors without detailed a priori knowledge.

Further, the performance of the IRT scores in the simulation study should be interpreted in light of the fact that they were estimated under idealised conditions. In practice their use is more complicated and may be less effective. For example, we fit graded response and 2-parameter logistic models to our polytomoyus and binary data respectively because we knew that these models had been used to generate the item responses. Thus, there was no risk of seriously mis-specifying the psychometric model. In reality, the appropriate model for the items will not be known in advance- it will have to be chosen on the basis of the item format and a hypothesis about how the latent trait is related to item responding and then tested for appropriateness. The lack of a priori knowledge about the appropriate IRT model for a given set of items increases the risk that the chosen model will be mis-specified in some important way. Further, parametric IRT models are also often poor fits to the very same kinds of data that prove problematic in GxE tests, such as those concerning psychopathological phenotypes. Less restrictive non-parametric IRT models are sometimes recommended as alternatives (Meijer and Baneke 2004) but these methods do not allow estimation of factor scores for use in GxE tests. Finally, at a very pragmatic level, IRT models are only useful when item-level data are available, which is not always the case.

In practice, it is worthwhile to compare results obtained using IRT scores with those obtained using raw and transformed sum scores. Comparison can highlight how sensitive results are to phenotypic scaling. Under some conditions, e.g. when the phenotype and moderator do not have strong association or the phenotypic distribution departs only slightly from its population distribution, scaling of the phenotype may make little difference to results. In addition, in rare cases where the phenotypic distribution is mis-specified in the IRT model used to estimate the scores but well approximated by the sum scores, the sum scores could, in principle, produce less biased results than the IRT scores. Even when the phenotypic distribution is correctly assumed to be normal, no non-linear transformation or IRT score estimation method guarantees a perfect reconstruction of the phenotypic distribution as it exists in the population. In fact, as argued above, the scores produced by a transformation to normality could be ‘too normal’ in the sense that in the presence of GxE non-normality of the phenotype would usually be expected. This is exactly what occurred in, for example, the condition of the simulation study in which all moderation parameters were positive in the population and in which a sum score from binary items was transformed to normality. Transforming to normality yielded a parameter estimate that was almost as negatively biased as the original estimate from using the sum score was positively biased. Moreover, the true positive rate dropped from 81 to 34 % suggesting a significant drop in the power to detect GxE.

This result underscores the fact that near-normal observed score distributions should not always be the goal. Non-normal latent distributions would be expected when, for example, a phenotype is influenced by GxE processes (perhaps not related to the moderator of interest), when it is influenced by some genetic (or environmental) variants of disproportionately large effect, or when phenotypic expression is subject to a liability threshold. Without some knowledge of the etiology of the trait, the appropriate distribution to which to transform or to assume in an IRT model will not be obvious. For example, although empirical methods exist that attempt to determine a latent trait distribution and IRT parameters simultaneously (e.g. Woods 2006), in practice the same patterns of item responses may be represented equally well by a range of combinations of distributions and IRT parameters (e.g. Pilkonis et al. 2011). There remains an important role of theoretical knowledge in determining which of these combinations is the most biologically plausible. We believe that the continuing advances in characterising the etiologies of complex traits will increasingly serve to inform the reasonableness of distributional assumptions and measurement models for phenotypes in testing GxE. Although it was once necessary (at least in practical terms) to assume multivariate normality for parameter estimation, recent and continuing developments in statistical methodology mean that this is no longer the case. Rather, the primary limiting factor at present is the theoretical knowledge to guide the specification of an appropriate (implicit or explicit) measurement model, rather than the statistical models to operationalise it.

Finally, our results highlight some challenges with testing GxE even under optimal scaling conditions. In our control conditions, there was a slight negative bias in GxE estimates when this effect was in the opposite direction to moderation of shared and unshared environmental influences. In addition, although power to detect GxE was under optimal scaling, type 1 error rates were below nominal levels. This has also been observed in previous studies of the GxM model (van Hulle et al. 2013) and suggests that nested model comparisons for the GxE provide conservative tests.

Limitations

A limitation of the current study is that we did not directly compare the two-step IRT approach with a one-step approach presented here. A one-step approach has yet to be developed for testing of GxE within the Purcellian framework; however, it is possible to anticipate some of its disadvantages and advantages. First, the approach would share the limitation of the two-step approach that the true phenotypic distribution would not be known but assumed. Assuming a normal distribution for the phenotype when the true distribution is non-normal could, in principle, result in biased GxE tests in a similar way to using a poorly scaled sum score. It would also share the necessity to select an appropriate IRT model and freely estimate its parameters in a finite sample. A further disadvantage would be its statistical and computational complexity as compared to a two-step approach. However, an important advantage would be that the error-free latent trait could be decomposed directly and this is likely to result in less biased GxE tests. It would have the related advantage that the IRT parameters would not have to be taken as given as they are in the second step of the two-step approach. Therefore, the imprecision in these parameter estimates could be appropriately taken account of. Further, and perhaps most importantly, a one-step approach is more appropriate from a conceptual perspective because it provides a much more direct operationalization of GxE hypotheses. In the two-step approach, a distribution for the phenotype is assumed in the first step; however, in tests of GxE it is important to distinguish between assumptions about the marginal distribution of the phenotype and the distribution of the underlying genetic and environmental influences absent the influence of the moderator. While the former would be expected to be non-normal because being subject to moderation skews the phenotypic distribution, the latter can usually be assumed normal. The two-step approach unfortunately conflates these distinct contributions because it specifies a distribution only for the latent phenotype. In addition, although we designed our simulation conditions to be as realistic as possible, we covered only a limited range of the possible conditions that could occur in the real world. Although the principles discussed are likely general, we conducted our analyses within specific GxE and IRT frameworks and used a limited range of parameter values. Similarly, while inclusion of a real data example is important to test conclusions from simulation studies in a more ecologically valid context, these too are limited by their specificity.

Conclusions

Tests of GxE can be biased by inappropriate scaling of a phenotype, and reliance on raw scores that are suspected to mis-represent the underlying distribution of the target phenotype. Two potentially useful solutions are to transform sum scores to normality or to estimate IRT scores based on an appropriate model. Although these strategies will suffer low statistical power, they reduce the rate of spurious GxE detection and recover the correct direction of effects. Therefore, researchers can be more confident about the presence and direction of GxE when it is identified using one of these strategies than when using a raw sum score.

References

Asbury K, Wachs TD, Plomin R (2005) Environmental moderators of genetic influence on verbal and nonverbal abilities in early childhood. Intelligence 33(6):643–661

Baron-Cohen S, Wheelwright S, Skinner R, Martin J, Clubley E (2001) The autism-spectrum quotient (AQ): evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. J Autism Dev Disord 31(1):5–17

Beasley TM, Erickson S, Allison DB (2009) Rank-based inverse normal transformations are increasingly used, but are they merited? Behav Genet 39(5):580–595

Bock RD (1972) Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika 37(1):29–51

Boomsma DI, Martin NG (2002) Gene–environment interactions. In: D'haenen H, den Boer JA, Willner P (eds) Biological psychiatry. Wiley, New York, pp 181–187

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

Button TM, Hewitt JK, Rhee SH, Corley RP, Stallings MC (2010) The moderating effect of religiosity on the genetic variance of problem alcohol use. Alcohol Clin Exp Res 34(9):1619–1624

Cai L, Yang JS, Hansen M (2011) Generalized full-information item bifactor analysis. Psychol Methods 16(3):221–248

Chalmers RP (2012) mirt: a multidimensional item response theory package for the R environment. J Stat Softw 48(60):1–29

Chernyshenko OS, Stark S, Drasgow F, Roberts BW (2007) Constructing personality scales under the assumptions of an ideal point response process: toward increasing the flexibility of personality measures. Psychol Assess 19(1):88–106

Culpepper SA (2013) The reliability and precision of total scores and IRT estimates as a function of polytomous IRT parameters and latent trait distribution. Appl Psych Meas 37(3):201–225

Distel MA, Middeldorp CM, Trull TJ, Derom CA, Willemsen G, Boomsma DI (2011) Life events and borderline personality features: the influence of gene–environment interaction and gene–environment correlation. Psychol Med 41(4):849–860

Eaves LJ (2006) Genotype × environment interaction in psychopathology: fact or artifact? Twin Res 9(1):1–8

Eaves LJ, Last K, Martin NG, Jinks JL (1977) A progressive approach to non-additivity and genotype-environmental covariance in the analysis of human differences. Br J Math Stat Psychol 30(1):1–42

Embretson SE (1996) Item response theory models and spurious interaction effects in factorial ANOVA designs. Appl Psychol Meas 20(3):201–212

Embretson SE, Reise SP (2000) Item response theory for psychologists. Psychology Press, Hove

Falconer DS, Mackay TF (1996) Introduction to quantitative genetics. Harlow, Longman

Harden KP, Turkheimer E, Loehlin JC (2007) Genotype by environment interaction in adolescents' cognitive aptitude. Behav Genet 37(2):273–283

Hicks BM, South SC, DiRago AC, Iacono WG, McGue M (2009a) Environmental adversity and increasing genetic risk for externalizing disorders. Arch Gen Psychiatry 66(6):640–648

Hicks BM, DiRago AC, Iacono WG, McGue M (2009b) Gene–environment interplay in internalizing disorders: consistent findings across six environmental risk factors. J Child Psychol Psychiatry 50(10):1309–1317

Johnson W, Krueger RF (2005) Genetic effects on physical health: lower at higher income levels. Behav Genet 35(5):579–590

Johnson W, Kyvik KO, Mortensen EL, Skytthe A, Batty GD, Deary IJ (2011) Does education confer a culture of healthy behavior? Smoking and drinking patterns in Danish twins. Am J Epidemiol 173(1):55–63

Kang SM, Waller NG (2005) Moderated multiple regression, spurious interaction effects, and IRT. Appl Psychol Meas 29(2):87–105

Krueger RF, Johnson W (2002) The Minnesota twin registry: current status and future directions. Twin Res Hum Genet 5(5):488–492

Lykken DT, Bouchard TJ, McGue M, Tellegen A (1990) The Minnesota twin family registry: some initial findings. Acta Genet Med Gemellol 39(1):35–70

Martin N (2000) Gene–environment interaction and twin studies. In: Spector T, Sneider H, MacGregor A (eds) Advances in twin and sib-pair analysis. London, Greenwich Medical Media

Mather K, Jinks JL (1971) Biometrical genetics. Biometrical genetics, 2nd edn. London, Chapman and Hall

Meijer RR, Baneke JJ (2004) Analyzing psychopathology items: a case for nonparametric item response theory modeling. Psychol Methods 9(3):354–368

Meijer RR, Egberink IJ (2012) Investigating invariant item ordering in personality and clinical scales some empirical findings and a discussion. Educ Psychol Meas 72(4):589–607

Micceri T (1989) The unicorn, the normal curve, and other improbable creatures. Psychol Bull 105(1):156–166

Molenaar D, Dolan CV (2014) Testing systematic genotype by environment interactions using item level data. Behav Genet 44(3):212–231

Molenaar D, van der Sluis S, Boomsma DI, Dolan CV (2012) Detecting specific genotype by environment interactions using marginal maximum likelihood estimation in the classical twin design. Behav Genet 42(3):483–499

Morse BJ, Johanson GA, Griffeth RW (2012) Using the graded response model to control spurious interactions in moderated multiple regression. Appl Psych Meas 36(2):122–146

Muthén LK, Muthén BO (2010). Mplus user's guide: statistical analysis with latent variables

Neale MC, Boker SM, Xie G, Maes HH (2006) Mx: statistical modeling, 7th edn. VCU Department of Psychiatry, Richmond

Nydick SW (2014) catIrt: An R Package for Simulating IRT-Based Computerized Adaptive Tests. R package version 0.5-0. http://CRAN.R-project.org/package=catIrt

Pilkonis PA, Choi SW, Reise SP, Stover AM, Cella D (2011) Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS$^{®}$): depression, anxiety, and anger. Assessment 18(3):263–283

Pluess M, Belsky J (2013) Vantage sensitivity: individual differences in response to positive experiences. Psychol Bull 139(4):901–916

Purcell S (2002) Variance components models for gene–environment interaction in twin analysis. Twin Res 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(3):301–315

Reise SP, Waller NG (2009) Item response theory and clinical measurement. Annu Rev Clin Psychol 5:27–48

Reiss D, Leve LD, Neiderhiser JM (2013) How genes and the social environment moderate each other. Am J Public Health 103(S1):S111–S121

Rende R, Plomin R (1992) Diathesis-stress models of psychopathology: a quantitative genetic perspective. Appl Prev Psychol 1(4):177–182

Samejima F (1969) Estimation of latent ability using a response pattern of graded scores. Psychometrika Monogr Suppl 34(4):100

Schwabe I, van den Berg SM (2014) Assessing genotype by environment interaction in case of heterogeneous measurement error. Behav Genet 44(4):394–406

Shanahan MJ, Hofer SM (2005) Social context in gene–environment interactions: retrospect and prospect. J Gerontol B 60(1):65–76

Silventoinen K, Hasselbalch AL, Lallukka T, Bogl L, Pietiläinen KH, Heitmann BL et al (2009) Modification effects of physical activity and protein intake on heritability of body size and composition. Am J Clin Nutr 90(4):1096–1103

South SC, Krueger RF (2011) Genetic and environmental influences on internalizing psychopathology vary as a function of economic status. Psychol Med 41(1):107–117

South SC, Krueger RF, Johnson W, Iacono WG (2008) Adolescent personality moderates genetic and environmental influences on relationships with parents. J Pers Soc Psychol 94(5):899–912

Tabery J (2008) RA Fisher, Lancelot Hogben, and the origin (s) of genotype–environment interaction. J Hist Biol 41(4):717–761

R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Tellegen A, Waller NG (2008) Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In: Boyle G, Matthews G, Saklofske DH (eds) The SAGE handbook of personality theory and assessment, volume 2: personality measurement and testing. Thousand Oaks, Sage Publications

Thomas ML (2011) The value of item response theory in clinical assessment: a review. Assessment 18(3):291–307

Timberlake DS, Rhee SH, Haberstick BC, Hopfer C, Ehringer M, Lessem JM et al (2006) The moderating effects of religiosity on the genetic and environmental determinants of smoking initiation. Nicotine Tob Res 8(1):123–133

Tucker-Drob EM, Harden KP, Turkheimer E (2009) Combining nonlinear biometric and psychometric models of cognitive abilities. Behav Genet 39(5):461–471

Tuvblad C, Grann M, Lichtenstein P (2006) Heritability for adolescent antisocial behavior differs with socioeconomic status: gene–environment interaction. J Child Psychol Psychiatry 47(7):734–743

van den Berg SM, Glas CA, Boomsma DI (2007) Variance decomposition using an IRT measurement model. Behav Genet 37(4):604–616

van den Oord EJ, Simonoff E, Eaves LJ, Pickles A, Silberg J, Maes H (2000) An evaluation of different approaches for behavior genetic analyses with psychiatric symptom scores. Behav Genet 30(1):1–18

van den Oord EJ, Pickles A, Waldman ID (2003) Normal variation and abnormality: an empirical study of the liability distributions underlying depression and delinquency. J Child Psychol Psychiatry 44(2):180–192

van der Sluis S, Dolan CV, Neale MC, Boomsma DI, Posthuma D (2006) Detecting genotype– environment interaction in monozygotic twin data: comparing the Jinks and Fulker test and a new test based on marginal maximum likelihood estimation. Twin Res Hum Genet 9(3):377–392

Van Hulle CA, Rathouz PJ (2015) Operating characteristics of statistical methods for detecting gene-by- measured environment interaction in the presence of gene–environment correlation under violations of distributional assumptions. Twin Res Hum Genet 18(1):19–27

van Hulle CA, Lahey BB, Rathouz PJ (2013) Operating characteristics of alternative statistical methods for detecting gene-by-measured environment interaction in the presence of gene–environment correlation in twin and sibling studies. Behav Genet 43(1):71–84

Walton KE, Ormel J, Krueger RF (2011) The dimensional nature of externalizing behaviors in adolescence: evidence from a direct comparison of categorical, dimensional, and hybrid models. J Abnorm Child Psychol 39(4):553–561

Woods CM (2006) Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. Psychol Methods 11(3):253–273

Wray NR, Lee SH, Mehta D, Vinkhuyzen AA, Dudbridge F, Middeldorp CM (2014) Research Review: polygenic methods and their application to psychiatric traits. J Child Psychol Psychiatry 55(10):1068–1087

Zheng H, Rathouz P (2013) GxM: Maximum Likelihood Estimation for Gene-by-Measured Environment Interaction Models. R package version 1.0. http://CRAN.Rproject.org/package=GxM

Zheng H, Rathouz PJ (2015) Fitting procedures for novel gene-by-measured environment interaction models in behavior genetic designs. Behav Genet 45(4):467–479

Zheng H, Van Hulle CA, Rathouz PJ (2015) Comparing alternative biometric models with and without gene-by-measured environment interaction in behavior genetic designs: statistical operating characteristics. Behav Genet 45(4):480–491