frontiers in Genetics

ORIGINAL RESEARCH
published: 10 December 2019
doi: 10.3389/fgene.2019.01227

Data Integration Methods for Phenotype Harmonization in Multi-Cohort Genome-Wide Association Studies With Behavioral Outcomes

Justin M. Luningham[1†], Daniel B. McArtor[1], Anne M. Hendriks[2,3], Catharina E. M. van Beijsterveldt[2,3], Paul Lichtenstein[4], Sebastian Lundström[5], Henrik Larsson[4,6], Meike Bartels[2,3,7], Dorret I. Boomsma[2,3,7*] and Gitta H. Lubke[1]*

[1] Department of Psychology, University of Notre Dame, Notre Dame, IN, United States, [2] Netherlands Twin Register, Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, [3] Faculty of Behavioural and Movement Sciences, Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, [4] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, [5] Gillberg Neuropsychiatry Centre, University of Gothenburg, Gothenburg, Sweden, [6] School of Medical Sciences, Örebro University, Örebro, Sweden, [7] Amsterdam Neuroscience, VU University Amsterdam, Amsterdam, Netherlands

OPEN ACCESS

Edited by:
Mariza De Andrade, Mayo Clinic, United States

Reviewed by:
Paola Sebastiani, Boston University, United States
Hongfang Liu, Mayo Clinic, United States

***Correspondence:**
Justin M. Luningham
jluningham@gsu.edu
Dorret I. Boomsma
di.boomsma@vu.nl

†Present Address:
Justin M. Luningham, School of Public Health, Georgia State University, Atlanta, GA, United States

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *04 October 2018*
Accepted: *05 November 2019*
Published: *10 December 2019*

Citation:
Luningham JM, McArtor DB, Hendriks AM, van Beijsterveldt CEM, Lichtenstein P, Lundström S, Larsson H, Bartels M, Boomsma DI and Lubke GH (2019) Data Integration Methods for Phenotype Harmonization in Multi-Cohort Genome-Wide Association Studies With Behavioral Outcomes. Front. Genet. 10:1227. doi: 10.3389/fgene.2019.01227

Parallel meta-analysis is a popular approach for increasing the power to detect genetic effects in genome-wide association studies across multiple cohorts. Consortia studying the genetics of behavioral phenotypes are oftentimes faced with systematic differences in phenotype measurement across cohorts, introducing heterogeneity into the meta-analysis and reducing statistical power. This study investigated integrative data analysis (IDA) as an approach for jointly modeling the phenotype across multiple datasets. We put forth a bi-factor integration model (BFIM) that provides a single common phenotype score and accounts for sources of study-specific variability in the phenotype. In order to capitalize on this modeling strategy, a phenotype reference panel was utilized as a supplemental sample with complete data on all behavioral measures. A simulation study showed that a mega-analysis of genetic variant effects in a BFIM were more powerful than meta-analysis of genetic effects on a cohort-specific sum score of items. Saving the factor scores from the BFIM and using those as the outcome in meta-analysis was also more powerful than the sum score in most simulation conditions, but a small degree of bias was introduced by this approach. The reference panel was necessary to realize these power gains. An empirical demonstration used the BFIM to harmonize aggression scores in 9-year old children across the Netherlands Twin Register and the Child and Adolescent Twin Study in Sweden, providing a template for application of the BFIM to a range of different phenotypes. A supplemental data collection in the Netherlands Twin Register served as a reference panel for phenotype modeling across both cohorts. Our results indicate that model-based harmonization for the study of complex traits is a useful step within genetic consortia.

Keywords: phenotype harmonization, genome-wide association studies, latent variable modeling, data integration, consortia

INTRODUCTION

Multi-study consortia and large-scale meta-analyses are the status quo for genome-wide analyses of complex traits (Evangelou and Ioannidis, 2013; Pedersen et al., 2013; Reitveld et al., 2014). Combining data from different studies presents an additional challenge when behavioral, psychological, or other complex phenotypes have been measured by different means across the studies. The most practical and widely used phenotype scoring approach is forming a sum or mean score of the available measures in each cohort (Wray et al., 2007; Bath et al., 2010; Pappa et al., 2016). However, sum scores overlook the systematic measurement differences that are brought about by different questionnaires. Sum scores of different item sets may not capture the same aspects of the behavior, so their use introduces phenotypic heterogeneity and reduces power in genome-wide association studies (GWAS; van den Berg et al., 2014). The current paper utilizes an integrative data analysis (IDA) framework for phenotype harmonization that can provide benefits for consortium-based GWAS meta-analyses by improving precision in phenotype measurement (Curran and Hussong, 2009). To quantify these benefits, we conduct a simulation study to assess the power to detect the effect of a genetic variant on a behavioral outcome that is modeled by IDA-based phenotype harmonization. In addition, we illustrate the IDA approach to harmonizing behavioral phenotypes.

IDA is a broad framework that holds great potential for improving the phenotype measure in GWAS meta-analyses because it is essentially *model-based* phenotype harmonization. The IDA framework allows researchers to adjust for measurement differences across studies, which is usually not possible when conducting meta-analyses of summary statistics. The common practice of forming sum scores of questionnaire scales is based on the often implicit assumption that the individual items available in each cohort measure the same phenotype across studies, which rarely holds in studies of complex behavioral outcomes. Different sets of items usually evaluate different aspects of a behavioral phenotype; and there are often measurement differences across countries or cultures, age groups, or different raters (Hudziak et al., 2003; Bartels et al., 2007; Jak, 2017). Typical approaches to phenotype harmonization, such as collapsing or rescaling response categories, are not sufficient when there are substantial differences underlying phenotype measurement (Gatz et al., 2015).

Phenotype precision has been demonstrated to improve statistical power and precision for genetic association tests. For example, removing poor measurement items can reduce heterogeneity in the phenotype, thereby increasing the signal associated with genetic variants (Laurin et al., 2015). In a different study, Xu et al. (2015) showed that fitting a complex psychometric model to mental health data led to larger single-nucleotide polymorphism (SNP) effects than performing a GWAS on a sum of mental health items. The study suggested that psychometric models more accurately reflect complex traits than the sum score, which ignores possible multidimensional subtypes of a trait. Indeed, simulations have shown that accounting for multidimensionality of a behavioral outcome with latent trait models can increase power in a GWAS compared to a sum score (van der Sluis et al., 2010). An additional advantage of psychometrically harmonizing behavioral phenotypes lies in the fact that, for many behaviors, subtypes of a given trait can have different levels of heritability (Ligthart et al., 2005; Yeh et al., 2010). This indicates that different SNPs may be acting on the different trait subtypes. Therefore, a single phenotype score ignoring dimensionality muddies strong genetic signals with weak or non-existent ones, resulting in less overall power than would be present if a subtype were accurately scored.

In IDA, item- or subscale-level data from different consortium partners are concatenated into one dataset. Psychometric modeling (Cattell, 1952; Lawley and Maxwell, 1963) allows items from different cohorts to contribute differentially to the scoring of the underlying trait which represents the phenotype and has the same metric across cohorts. The advantage of IDA is the flexibility to adjust for measurement differences in the measurement model specification such as rater, sex, and/or cohort differences. An inherent challenge of IDA is that combining the item-level data from different cohorts usually introduces a large amount of missing data due to the fact that not all cohorts use the same questionnaires. To illustrate, suppose cohort A uses questionnaire X, whereas cohort B uses questionnaire Y. Responses of questionnaire Y would be missing in cohort A and the reverse would be true for cohort B. IDA measurement models require the presence of overlapping items to adequately link data across all participants (Hussong et al., 2013). Collecting a supplemental sample with complete data on all items can help alleviate this problem (Hussong et al., 2013; Gatz et al., 2015). In this paper, the supplemental sample is called a phenotypic *reference panel.*

IDA has been used previously to combine different versions of cognitive batteries, personality measures, and alcohol use data across multiple studies (van den Berg et al., 2014; Xu et al., 2015; Marcoulides and Grimm, 2017). Many IDA approaches used a Rasch item response theory (IRT) model, a latent trait model that requires simplifying assumptions and specifies equivalent measurement across study (McArdle et al., 2009; Gatz et al., 2015; Marcoulides and Grimm, 2017). Other IDA models directly evaluate the differences of measurement properties of the pooled items. Curran and Hussong (2009) and Curran et al. (2014) proposed a moderated non-linear factor analysis (MNLFA) model that allows all of the item measurement parameters to vary across a set of covariates, such as study membership and country of origin. In a similar approach, Bauer et al. (2013) integrated data across multiple raters (e.g., mother and father ratings). This model extracts a single phenotype score while filtering out rater influences. The model is based on an adaptation of the bi-factor model, a classic psychometric model in clinical research (Holzinger and Swineford, 1937). The IDA model proposed in this paper is built upon the bi-factor model where a general factor represents the sought after phenotype that is common to multiple studies. Differences across cohorts are modeled in a set of specific factors. In addition to separating the common phenotype from cohort-specific influences, the bi-factor integration model also eliminates measurement error from the phenotype factor score.

This paper is structured as follows. A brief review of factor analysis and structural equation modeling (SEM) is presented. Practical issues for meta-analysis in GWAS consortia and the

potential advantages of a phenotype reference panel are discussed. A model specifically suited for IDA in multi-study GWAS, termed the bi-factor integration model (BFIM), is then presented. A series of simulations demonstrate how IDA can increase power to detect a genetic effect when phenotype reference panel data are available. A demonstration of the bi-factor integration model for two large datasets of aggressive behaviors in 9-year-old children is also presented. The implications of these results for behavior genetic consortia are discussed, as well as limitations of this approach and future directions.

METHODS

Factor Analysis Models

Behavioral and mental health phenotypes such as intelligence or depression are not directly observable but instead are measured with multiple questionnaire items. The observed items are individual indicators of an underlying construct. The observed indicators are designed to capture the different aspects of the construct, and item responses are considered as manifestations of the true trait. Latent variable models capture the information that is common across multivariate outcomes (i.e., shared variance), considered the underlying latent trait or factor (Bollen, 1989). For example, aggression is a trait that is not measured directly, but researchers administer multiple questions that pertain to different aspects of aggressive behaviors or attitudes. Individuals with higher levels of true aggression are expected to score higher on the items. Additionally, the underlying factor fully accounts for the covariance of the items; once the aggression factor is accounted for, the items are conditionally independent from each other.

The factor analysis model is a direct implementation of this line of thought. Let y_{ij} represent a response to item i (i from 1, 2, …, p) for person j ($j$from1,2,…,N) Assuming a single latent variable underlies a set of observed, continuous item responses, the model can be written as

$$y_{ij} = \nu_i + \lambda_i \eta_j + \varepsilon_{ij} \tag{1}$$

where ν_i represents an item intercept, λ_i is an item loading (or slope) parameter, η_j represents the latent factor score for person j, and ε_{ij} is an error term. The latent factor is assumed to be normally distributed as $N(\alpha,\psi)$ and the error is normally distributed as $\varepsilon_{ij} \sim N(0,\sigma^2)$. Individuals are assumed to be independent from each other, and the items are conditionally independent given η_j. To identify the model, one item intercept and loading must be fixed to zero and one, respectively, or the factor variance must be fixed at one (Lawley and Maxwell, 1963; Bollen, 1989).

The Bi-Factor Integration Model

Several models developed for IDA can be used to test measurement differences in the item parameters one at a time for multiple items over multiple covariates (Curran et al., 2014). In the GWAS integration scenario, however, the only interest is in reducing the noise in the phenotype score introduced by differences in measurement across cohorts. If one can reasonably assume that the available questionnaire items are all indicators of the same phenotype, then the target trait of interest is simply a single common factor underlying the full item set. If the items used by the different cohorts tap into similar aspects of the phenotype and have similar measurement properties, then the sum score model is expected to work reasonably well. However, a simple unidimensional factor analysis model may not fit well if items used in the different cohorts measure more or less severe aspects of the phenotype, if cohorts differ with respect to raters, or if items have different meanings across cultures.

In this paper we propose a bi-factor model integration model (BFIM) that increases precision in the estimated target trait by modeling additional information specific to different questionnaires or cohorts. The BFIM is a special case of the factor model in equations 1 and 2 with multiple factors, which can be written as

$$y_{ij} = \nu_i + \lambda_{ig}\eta_{jg} + \sum_{k=1}^{K} \lambda_{ik}\eta_{jk} + \varepsilon_{ij} \tag{2}$$

where η_{jg} represents a *general* factor for person j, labeled T for the target trait, and there is an associated factor loading for all items on T, λ_{ij}. η_{jk} represents a *specific* factor k (k=1,2,…,K) that only subsets of items load onto, with λ_{ik} corresponding to item loadings on the k^{th} specific factor.

Similar to the general factor analysis model, the underlying factors are assumed to be normally distributed as $\eta_{jk} \sim N(\alpha_k,\psi_k)$, and the error is normally distributed as $\varepsilon_{ij} \sim N(0,\sigma^2)$. The bi-factor model specification also requires constraints to identify the model. The bi-factor model is identified by specifying that the factors follow a standard normal distribution (Gibbons and Hedeker, 1992). Further, each item loads onto the general factor and only one additional specific factor; otherwise, the model is not identifiable (Gibbons and Hedeker, 1992). Consider a case with eight total items and four items loading onto each of two specific factors. The matrix of factor loadings and vector of latent factors are then (3)

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{1g} & \lambda_{11} & 0 \\ \lambda_{2g} & \lambda_{21} & 0 \\ \lambda_{3g} & \lambda_{31} & 0 \\ \lambda_{4g} & \lambda_{41} & 0 \\ \lambda_{5g} & 0 & \lambda_{52} \\ \lambda_{6g} & 0 & \lambda_{62} \\ \lambda_{7g} & 0 & \lambda_{72} \\ \lambda_{8g} & 0 & \lambda_{82} \end{bmatrix}; \boldsymbol{\eta} = \begin{bmatrix} \eta_{jg} \\ \eta_{j1} \\ \eta_{j2} \end{bmatrix} \tag{3}$$

Orthogonality is imposed on the factors such that the specific factors are completely independent of the general factor and of the other specific factors. Due to this independence

specification, the general factor captures the target trait of interest, and the specific factors pull out the residual covariance among item *subsets* that is not captured in the common factor. For IDA across multiple studies, modeling the specific factors can follow known differences across the studies. For example, specific factors may be modeled from items that originate from the same questionnaire, the same study, or the same country of origin. By explicitly modeling these sources of heterogeneity, the precision of the general factor is increased. Improved phenotypic measurement, in turn, is likely to increase the precision in the SNP association coefficients. **Figure 1** depicts a bi-factor integration model for two different questionnaires with a general factor, two specific factors, and a SNP effect. In **Figure 1**, the specific factors are labeled Q_1 and Q_2 corresponding to questionnaire 1 and questionnaire 2, respectively.

The target factor scores can be computed from the bi-factor integration model and then used by partnering studies in a consortium to conduct a parallel meta-analysis. Using computed factor scores as outcomes in association tests should result in more precision in each study, therefore increasing precision in the average effect size. Reducing the standard error of the estimated $\bar{\beta}$ increases the power of the hypothesis test that the effect is significantly different from zero. Factor scores are not estimated in the latent variable model, but computed *post hoc* using fixed model parameters. There are multiple approaches for computing factor scores, such as regression scores, Bartlett scores, likelihood-based expected a-posteriori scores, and Bayesian plausible values (Muthén and Muthén, 1998-2017). Certain types of factor scores used as dependent variables can lead to bias in the regression coefficients (Grice, 2001; Skrondal and Laake, 2001). Devlieger et al. (2016) demonstrated that different methods of factor score calculations have similar rates of statistical power, so regression factor scores were calculated for this study.

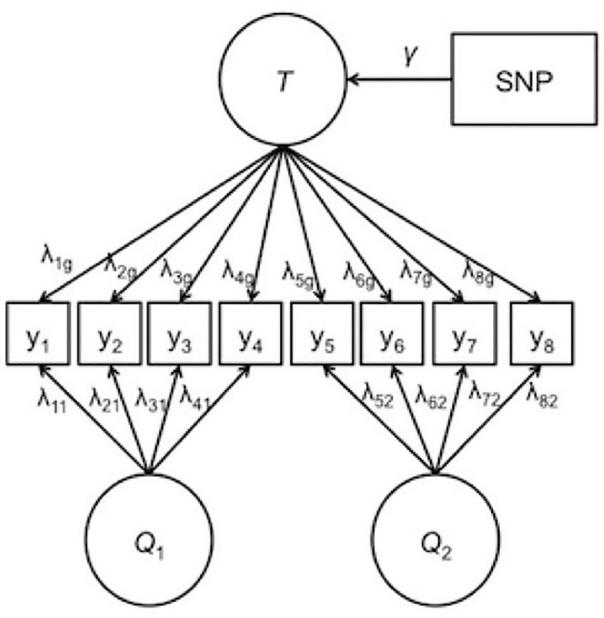

FIGURE 1 | Example of the bi-factor integration model. *T* represents the general trait. Q_1 and Q_2 are the specific factors corresponding to questionnaire 1 and questionnaire 2, respectively. λ represents the factor loading, γ represents the effect of the single-nucleotide polymorphism on the general trait. Thresholds and error terms are not depicted.

Combining data across studies with disjoint measurement variables creates a dataset with systematic missing data. Modeling an underlying factor in combined data is therefore also a missing data problem. The three most commonly assumed missing data mechanisms are missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR; Enders, 2010; Little and Rubin, 2002). When data are MCAR or MAR, full information maximum likelihood estimation is known to provide correct inferences (Rubin, 1976). The MAR mechanism states that missing values of *Y* are independent of the observed values of *Y* after taking other variables in the analysis into account. In the IDA case, the missing data are accounted for by study or cohort membership, and missingness is independent of scores on the trait of interest. Maximum likelihood estimation is well developed for SEM and confirmatory factor analysis approaches (Allison, 2003; Asparouhov and Muthén, 2010; Enders, 2010).

Genome-Wide Association Studies Meta-Analysis

Common practice for GWAS in a consortium is for each consortium partner to conduct association tests on a sum of all available items. The resulting regression coefficients are then meta-analyzed. In this case, the phenotype score is the sum of the items, which can also be written in the context of the factor model:

$$SS_j = \sum_{i=1}^{p} y_{ij} = \sum_{i=1}^{p} (\nu_i + \lambda_i \eta_j + \varepsilon_{ij}) \tag{4}$$

The general association test (prediction model) for the sum score is simply

$$SS_j = \beta_0 + \sum_{q=1}^{Q} \beta_q x_{qj} + \epsilon_j \tag{5}$$

where x_{qj} is the genotype available for person j at locus q, and β_q is the regression coefficient for SNP q. The primary interest is in the p-value and statistical significance of β coefficients corresponding to SNPs. In a meta-analysis, these coefficients are combined across the different individual analyses and their significance is re-evaluated. Consortia commonly use fixed-effects meta-analysis with inverse-variance weighting or the *Z*-score method (Evangelou and Ioannidis, 2013). Fixed effects meta-analysis requires the assumption that there is one true population value for the effect of interest. Given this, if the sum scores SS_j are calculated from different items across study, additional error will be introduced into the regression weight of a SNP on the heterogeneous sum scores.

Sum scores are straightforward, easy to compute, and easy to interpret. They do, however, come with drawbacks. With sum scores, the item uniquenesses and measurement errors are summed up along with the portions of the item relating to the true trait. This will increase the variance of the outcome relative to the underlying structural variance of the trait, leading to inflated standard errors for $\hat{\beta}_q$ in each study. Additionally, estimating the factor loadings in a measurement model permits each item to contribute to the latent factor with different weights, reflecting the fact that different questionnaire items are often not equally good indicators of an underlying trait. A sum score implicitly assigns equal weight to all items. Regressing a harmonized trait score on a set of SNPs should result in a meta-analysis of regression coefficients that are more directly comparable.

Genome-Wide Association Studies Mega-Analysis

In general, it is preferable to fit the integrated measurement model and simultaneously conduct genetic association tests. The ideal scenario is to carry out a mega-analysis with SEM in which the measurement model for η and the regression of η on x is estimated simultaneously (Bollen, 1989). While this may be computationally difficult for a full genome-wide search with millions of SNPs, it is certainly possible for cases in which a few hundred or even thousand candidate SNPs are identified (e.g., Xu et al., 2015). Furthermore, recent methodological advances have increased the computational feasibility of SEMs in GWAS, such as with genome-wide structural equation modeling (GW-SEM) (Verhulst et al., 2017). In the context of the bi-factor integration model, the covariate effect is truly expressed on the target trait factor T_j rather than the indicators themselves. This is specified as the structural portion of the SEM model. For observed covariates x_{qj} (q from 1, 2, …,Q), the prediction model is written

$$T_j = \sum_{q=1}^{Q} \gamma_q x_{qj} + \zeta_j \tag{6}$$

where T_j the target (general) trait score for person j, γ_q is the regression coefficient for the q^{th} covariate, and ζ_j is a residual disturbance term. In a GWAS, the γ_q of primary interest is the one associated with a SNP, but controlling covariates such as age, gender, and genetic relatedness principal components may also be included.

Phenotypic Reference Panel

Retrospectively combining independent studies with different instruments often results in a sparse dataset with a high degree of missing data. This can lead to sets of subjects with no common items, resulting in latent variable models that often do not converge using modern estimation approaches for handling missing data. Lack of convergence leads to flawed estimators, if the model is able to provide estimates at all. Typical harmonization treats items that are similarly worded in the different questionnaires as the same item in the combined dataset, thus creating item overlap. This can also destabilize the model, however, if the items are not truly interchangeable.

A better strategy for understanding the relationships among all items is to collect a reference panel with complete data. The reference panel provides information about the association between items not jointly observed within cohorts when different surveys are used. This supplemental sample is critical for providing a link across cohorts and offers a potential gateway for psychometric harmonization through IDA. Similar approaches have been applied for multiple imputation integration, in which measurement models are not used (Carrig et al., 2015; Siddique et al., 2015). When there is available research and theory about a psychological phenotype, exploring a small number of measurement models can offer more precision for subsequent analyses than approaches making no assumptions about structure in the data (Collins et al., 2001). Others have collected a reference panel-type sample and analyzed it separately to evaluate the performance of more conventional phenotype harmonization approaches (Gatz et al., 2015). The ACTION Consortium (Aggression in Children: Unraveling gene-environment interplay to inform Treatment and InterventiON strategies) is actively collecting a *post hoc* phenotype reference panel in order to facilitate the multi-study integration of complex models of childhood aggressive behaviors (Boomsma, 2015).

The collection of the reference panel is imperative for the BFIM (and any measurement model) in the case of insufficient item overlap. In the next section, we present a Monte-Carlo simulation study evaluating the use of the bi-factor integration model compared to sum scores in hypothetical SNP association tests. These simulations provide insight into whether the collection of the reference panel and the extra effort in phenotype modeling are worth the costs.

SIMULATION

We conducted a simulation study with the goal of comparing the power of sum score meta-analysis, factor score meta-analysis, and integrated mega-analysis (full data integration model with SNP effect). A multiple imputation procedure was also carried out as an alternative method for handling missing items in the cohorts. The simulation was set up to represent a scenario in which two different studies used two different questionnaires to measure the same trait. Each cohort only had item responses on one questionnaire, with missing data on the items used in the other study. A small reference panel dataset was also included, in which subjects had responses on all items across the two questionnaires.

Data were simulated under four different data-generating models with five different sample size conditions, resulting in 20 total simulation conditions. To evaluate the necessity of the reference panel, data were also generated without a reference panel (with additional subjects added to one or both cohorts). **Table 1** lists the different models and sample size conditions utilized in the simulation study. In the first three sample size conditions, the reference panel makes up ~4, ~2.5, and ~7.5% of the total dataset, with equal sample sizes in the two cohorts. In

TABLE 1 | Various data-generating models and sample sizes used in simulations, resulting in 20 simulation conditions.

	Sample sizes	Data-generating models
Varying simulation conditions	N1: cohort 1 = 5,000, cohort 2 = 5,000, Ref= 400	Model 1: same measurement across item sets
	N2: cohort 1 = 2,500, cohort 2 = 2,500, Ref= 400	Model 2: different levels of item set reliability
	N3: cohort 1 = 7,500, cohort 2 = 7,500, Ref= 400	Model 3: mean and variance differences in cohort-specific factor
	N4: cohort 1 = 2,500, cohort 2 = 7,500, Ref= 400	Model 4: true model is a higher-order model (bi-factor model is misspecified)
	N5: cohort 1 = 4,500, cohort 2 = 4,500, Ref= 1,000	

the fourth condition, unequal sample sizes are introduced into the cohorts. In the fifth condition, the reference panel is increased to make up 10% of the total sample, maintaining similar sizes to condition 1 and 4. For all 20 combinations of model and sample size, 1,000 repetitions were executed in the simulation.

For all data-generating models, the underlying factors all followed a (marginally) standard normal distribution, and factor loadings were invariant across the reference group and cohort item sets. A SNP covariate accounted for 0.1% of the variance in the general factor, and a second covariate representing biological sex accounted for 20% of the variance in the general factor. Note that this assumes a single population-level effect size of the SNP, the same assumption made in fixed effects meta-analysis (Evangelou and Ioannidis, 2013). It was also assumed that the multiple cohorts originated from comparable populations, meaning that the minor allele frequency of the SNP was the same across groups. In practice, quality control checks with a reliable genotype reference should be conducted for either meta-analysis or mega-analysis approaches, and population stratification should be controlled for.

The genotypes for the SNP were generated with a minor allele frequency of 0.5, such that the data-generating equation for the target trait was

$$T_j = 0.0447 * SNP_j + 0.8944 * X_{2j} + \zeta_j, \zeta_j \sim N\left(0, 0.799\right) \quad (7)$$

leading to a marginal variance of 1 for T_j, and the specific factors were orthogonal to T_j and each other and were standard normal. Item-level data were then generated from the bi-factor model: (8)

$$y_{ij} = \nu_i + \lambda_{ig} T_j + \lambda_{ik} \eta_{jk} + \varepsilon_{ij} \quad (8)$$

where factor loadings λ_{ig} ranged from 0.3 to 0.6, depending on data-generating condition (specific simulation parameters are detailed in **Appendix I**). The loadings for the general and specific factors were controlled such that the general factor and specific factor collectively accounted for either 60 or 45% of each items' variance. For example, for factors with marginal unit variance, the explained variance of the factors is $\lambda_{ig}^2 + \lambda_{ik}^2$. There were eight items total: four items for each cohort, representing a brief item set tapping into a sub-domain of a behavior. Two items across cohorts were also generated with an additional residual correlation, reflecting items across questionnaires that were similar but not *exactly* the same, such as items that might be harmonized based on similar wording in the prompts. The residual correlation for these items was set at 0.6.

Data-Generating Models

Data were generated for the two cohort-specific item sets under a series of different measurement models These measurement differences were quite mild compared to the potential level of heterogeneity that is often encountered in practice, but this allowed us to examine the effect of increased measurement precision with even a small amount of measurement heterogeneity. All data were generated in R (R Core Team, 2018).

Model 1: ideal measurement. The first model, depicted in **Figure 2A**, represented identical measurement across the two cohorts. This model reflects ideal measurement conditions where the item sets have equal reliabilities. In other words, the factors account for the same amount of variance in the item responses. The factors collectively account for 60% of the variance in the item responses. The general factor accounts for 25.5% of the item response variability, and the specific factors account for 34.5% in the item responses. This contribution breakdown is equal across the two questionnaires.

Model 2: reliability differences. The second data-generating model is depicted in **Figure 2B**. In the second model, the variance explained by the second set of items was lower than the first set. This reflects an applied scenario in which less reliability in the second measure. To simulate this, the general factor and specific factor accounted for only 45% of the variance in the observed item responses of the second item set, rather than 60% of the variance in the more reliable set. Consequently, one set of four items has residual variance of 40%, but the other item set has residual variance of 55%. This weaker contribution is split between the general and specific factors, resulting in a slightly weaker contribution of the covariate effect to the manifest variables.

Model 3: mean and variance differences. The third data-generating model is described in **Figure 2C**. This model featured a larger mean and variance in the specific factor that contributes to the less reliable items of model 2. Rather than having mean of zero and unit variance, the specific factor had a mean of 1 and variance of 3. At the population level, the effect size of the SNP on the general factor remained the same. With increased mean and variance in the specific factor, the proportional contribution of the specific factor on the manifest variables also increased.

Model 4: higher-order data-generating model. Model 4 is depicted in **Figure 2D**. The fourth model reflected a scenario where the bi-factor model is somewhat misspecified. The data-generating model was re-formulated as a higher-order factor

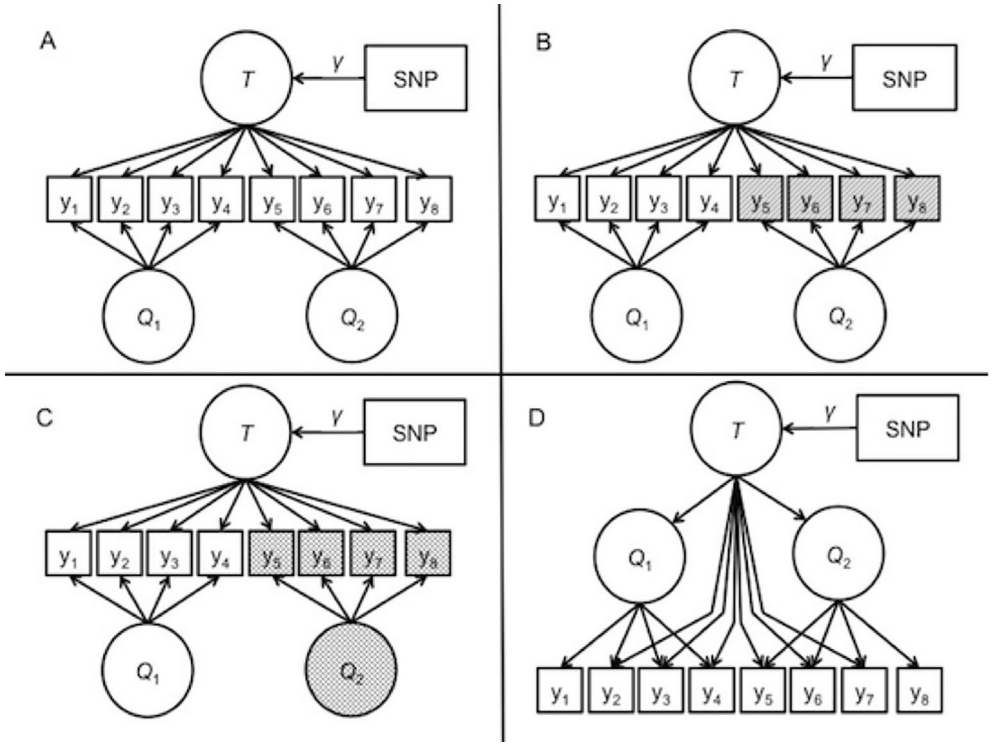

FIGURE 2 | Path diagrams of the data-generating models for the simulation. *T* represents the general trait. Q_1 and Q_2 are the specific factors corresponding to questionnaire 1 and questionnaire 2, respectively. **(A)** depicts model 1, representing the ideal measurement case with equal reliabilities for both questionnaires. **(B)** depicts model 2, in which the items on the second questionnaire have lower reliabilities than the first questionnaire, represented by shading. **(C)** depicts model 3, in which the Q_2 factor is shaded, representing an increased mean and variance along with reliability differences. **(D)** depicts model 4, a second-order model where the general factor summarizes the covariance among two questionnaire-specific factors, and the general factor additional direct effects on select items.

model in which the covariates influenced a target trait factor, which then influenced two separate cohort factors that finally determined the item responses in each of the cohorts. This model included direct effects from the true factor to some of the item responses. Chen et al., (2006) demonstrated that the higher-order correlated factors model is a constrained version of the bi-factor model. In fact, the correlated factors model with direct effects of the higher-order factor on all items is equivalent to the bi-factor model. In our simulation, some of the items did not include a direct effect from the trait factor to the item, meaning that the BFIM was slightly misspecified to the generated data.

Each model was used in combination with the five different sample size conditions, and for each model and sample size, 1,000 repetitions were generated to conduct analyses. The codes used to conduct the simulation are attached as a supplementary downloadable folder.

Analyses

Four different types of analyses were carried out to evaluate the effect of the genetic variant on a trait score across the simulated cohorts. The analyses were designed to compare sum scores in each cohort with the measurement model integration approach utilizing the reference panel. The different analysis procedures are listed below:

1) **Sum score meta-analysis:** The mean score of available items in each cohort and in the reference panel was computed. A meta-analysis of the SNP effect on the mean score in the two cohorts and reference panel was conducted. Meta-analyses were conducted directly in R (R Core Team, 2018).
2) **Factor score meta-analysis:** A BFIM was fitted to the overlapping phenotype information with no genetic variant included. The BFIM models were fitted using Mplus 7.11 (Muthén and Muthén, 1998-2017), and estimation was carried out using maximum likelihood estimation with numeric integration. Numeric integration was needed because of the low rates of coverage for some items. Regression-type factor scores were saved and used as the outcome in association tests within the two cohorts and reference panel and subsequently meta-analyzed. Meta-analyses were computed in R (R Core Team, 2018).
3) **BFIM mega-analysis:** The BFIM was fitted to all of the items to model a harmonized phenotype, and the general factor of the BFIM was regressed on the SNP simultaneously in the structural part of the model. The full SEM mega-analysis models were fitted using Mplus 7.11 (Muthén and Muthén, 1998-2017), and estimation was carried out using maximum likelihood estimation with numeric integration.
4) **Multiple imputation:** Multiple imputation of the individual items is an alternative approach to addressing missingness in

the combined data. For a benchmark comparison, we used multivariate imputation by chained equations (MICE, van Buuren and Groothuis-Oudshoorn, 2011) to impute each item from all other items and the covariates. We then summed the items for an imputed sum score as the outcome in mega-analysis association tests. Predictive mean matching, which is the default imputation model of the "mice" package in R[1], was specified as the imputation model. The default value of 5 iterations was used for each imputed dataset, and 50 datasets were imputed for each to account for the large proportion of missing data. Results from the regression analyses were pooled according to Rubin's rules to obtain correct standard errors and degrees of freedom (see Rubin, 1987; van Buuren, 2018; for details on pooling, predictive mean matching, and multiple imputation).

A bi-factor SEM was also carried out with complete data (i.e., no missingness) for all participants as a benchmark for the maximum power that could be achieved. To evaluate the utility of the reference panel, the factor score meta-analysis and SEM mega-analysis were also conducted without a reference panel included. In order to fit the models, some overlapping item information must be present. Therefore, the items with large residual correlation across questionnaires were treated as the same item. This is consistent with the practical scenario in which similarly worded items are recoded as equivalent items in the harmonization process.

The primary outcome of interest was the empirical power to detect the SNP effect (i.e., proportion of significant findings). Raw empirical power and empirical power relative to the maximum power under the complete data condition were computed. Type I errors, relative bias, and 95% coverage rates were also recorded. Relative bias is the difference between the true parameter and the average estimate across repetitions divided by the true value.

RESULTS

The overall results indicated a small-to-moderate advantage to detect the SNP effect for the data integration approach, in general. Under the fourth model data-generating model, in which the bi-factor model is misspecified, the sum score meta-analysis outperforms the factor score meta-analysis, but the BFIM mega-analysis still provides the best overall result.

Power and Type I Error

Figure 3 presents the different rates of power of the sum score meta-analysis, factor score meta-analysis, and BFIM mega-analysis relative to the power obtained with no missing data. Power is plotted as a function of different data-generating model across different panels representing the various sample size conditions. This was included because the raw power is not truly generalizable, as some data-generating models by design had less power to detect the genetic effect even with complete data than others. Therefore, the empirical power relative to the maximal power is more comparable across conditions. **Table 2** presents the raw power rates of each of the methods, as well as type I error rates.

As seen in **Figure 3**, the BFIM mega-analysis resulted in the most statistical power to detect the variant effect across all conditions. The advantage in relative power for the BFIM mega-analysis over the sum score ranged from about 4% (model 1, N4 condition) to about 19% (model 3, N1 condition). The BFIM mega-analysis also displayed larger power rates than the meta-analysis using harmonized factor scores. This is true even in the fourth data-generating model in which the BFIM is slightly misspecified. However, in this fourth model condition, the advantage of the BFIM mega-analysis over the sum score is generally at its smallest, compared to the other conditions.

The meta-analysis of factor scores also resulted in more statistical power than the meta-analysis of sum scores in 13 of the 20 different conditions. The power rates across these two methods were essentially equivalent in three conditions, and the sum score approach was more powerful in four conditions. The factor score meta-analysis resulted in more power for the first three data-generating models. However, with smaller sample sizes or unbalanced sizes across cohorts, the advantage of the factor score meta-analysis over the sum score is fairly small. In the fourth data-generating model, where the BFIM is slightly misspecified, the meta-analysis of factor scores is generally less powerful than sum score meta-analysis.

The multiple imputation approach used in these simulations resulted in the least power across nearly all conditions. This is especially true under data-generating model 3, when there are measurement differences across cohort, and in the fourth sample size condition, when there is significant imbalance in cohort size. For balanced sample sizes and no measurement differences across cohorts, the imputation approach performed similarly to the meta-analysis of sum scores. As seen in **Table 2**, all methods displayed acceptable type I error rates (between 2.5 and 7.5%).

Bias and Coverage

Figure 4 depicts the relative bias of the different methods across models and sample size conditions. Across all conditions, the BFIM mega-analysis estimates were within acceptable levels (±0.05) of relative bias. The bias in sum score meta-analysis fell within the acceptable range in 18 of the 20 conditions, and only presented problematic bias with unbalanced sample sizes. The bias resulting from factor score meta-analysis was between 5 and 10% in 8 of the 20 conditions, and was within acceptable levels in 12 of the 20 conditions. The imputation approach resulted in negative bias that was greater than 5% when the data-generating model was model 3. The sum score meta-analysis and the BFIM mega-analysis had excellent coverage rates, and the factor score meta-analysis had good coverage rates in 18 of the 20 conditions, as seen in **Table 3**.

[1]We note that multiple imputation can utilize a large number of different prediction models, and the number of iterations and imputed datasets can be tuned and optimized across different settings. As this paper focuses on the explicit psychometric modeling and not imputation, we used the default model and a single set of imputation settings.

TABLE 2 | Empirical power and type I error results with different analysis methods under the four data-generating models and five sample size conditions.

Model	N	Complete data power	Mega SEM	FS meta	SS meta	Impute power	Full T1	Mega T1	FS T1	SS T1	Impute T1
Model1	N1	0.760	0.567	0.557	0.519	0.496	0.046	0.046	0.044	0.040	0.037
Model2	N1	0.710	0.517	0.481	0.445	0.446	0.051	0.044	0.043	0.055	0.039
Model3	N1	0.678	0.479	0.399	0.356	0.275	0.045	0.047	0.071	0.067	0.037
Model4	N1	0.760	0.549	0.486	0.531	0.505	0.049	0.047	0.047	0.059	0.047
Model1	N2	0.480	0.358	0.329	0.323	0.295	0.046	0.040	0.057	0.051	0.033
Model2	N2	0.444	0.305	0.277	0.280	0.266	0.045	0.057	0.064	0.057	0.047
Model3	N2	0.384	0.241	0.219	0.213	0.151	0.048	0.045	0.045	0.044	0.043
Model4	N2	0.466	0.327	0.264	0.285	0.287	0.047	0.051	0.050	0.044	0.051
Model1	N3	0.878	0.743	0.723	0.666	0.637	0.055	0.054	0.042	0.043	0.034
Model2	N3	0.858	0.675	0.658	0.619	0.596	0.045	0.049	0.054	0.048	0.044
Model3	N3	0.816	0.596	0.554	0.526	0.424	0.049	0.046	0.049	0.047	0.032
Model4	N3	0.886	0.719	0.665	0.681	0.665	0.048	0.061	0.054	0.066	0.042
Model1	N4	0.772	0.576	0.553	0.551	0.410	0.045	0.046	0.055	0.058	0.049
Model2	N4	0.719	0.549	0.508	0.497	0.380	0.062	0.047	0.064	0.061	0.053
Model3	N4	0.653	0.509	0.487	0.457	0.203	0.048	0.041	0.057	0.052	0.029
Model4	N4	0.738	0.549	0.498	0.513	0.420	0.051	0.050	0.049	0.050	0.042
Model1	N5	0.747	0.588	0.572	0.535	0.512	0.036	0.044	0.044	0.039	0.039
Model2	N5	0.711	0.535	0.515	0.488	0.478	0.064	0.066	0.052	0.046	0.028
Model3	N5	0.626	0.450	0.398	0.399	0.318	0.051	0.049	0.055	0.058	0.040
Model4	N5	0.713	0.525	0.504	0.494	0.497	0.048	0.049	0.056	0.052	0.048

FS, factor score meta-analysis; Mega, BFIM SEM mega-analysis; SS, mean score meta-analysis; T1, type I error rate.

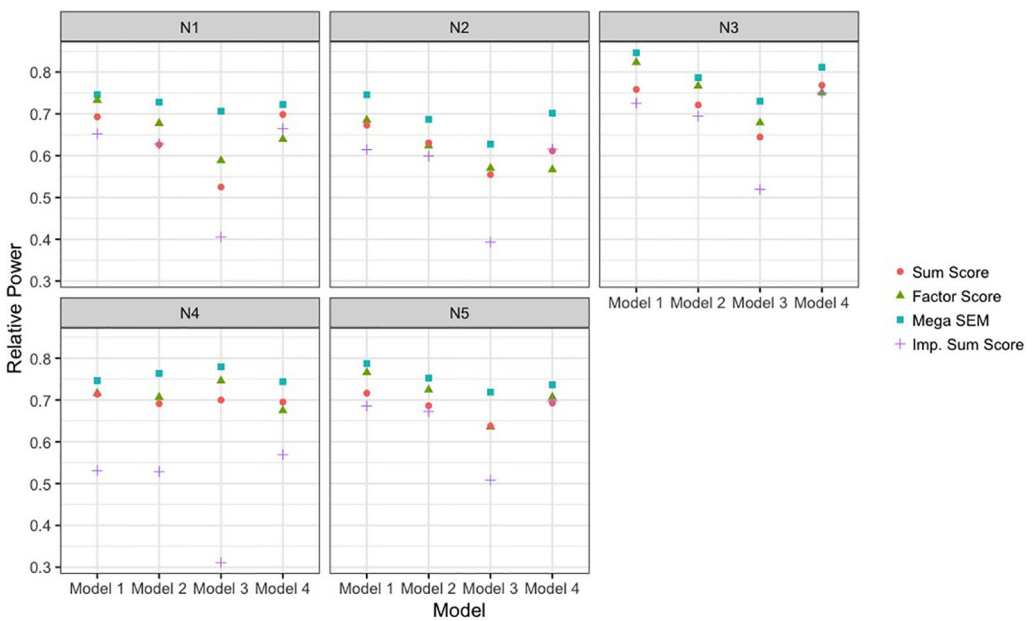

FIGURE 3 | Power to detect single-nucleotide polymorphism effect of each analysis approach relative to complete data power. Results are presented across the different data-generating models and the different sample size conditions detailed in **Table 1**. The analysis approaches were sum score meta-analysis, factor score meta-analysis, mega-analysis with the bi-factor integration mode, and multiple imputation of the missing items. SEM, structural equation modeling; Imp., imputation.

Differences Across Simulation Conditions

The BFIM mega-analysis was most powerful and unbiased across all conditions. The pattern of results was also the same across the conditions, except with unbiased sample sizes across cohorts (N4). In this condition, the relative power of the BFIM mega-analysis increased as the models became more complicated, and the relative bias departed from zero more in this sample size condition.

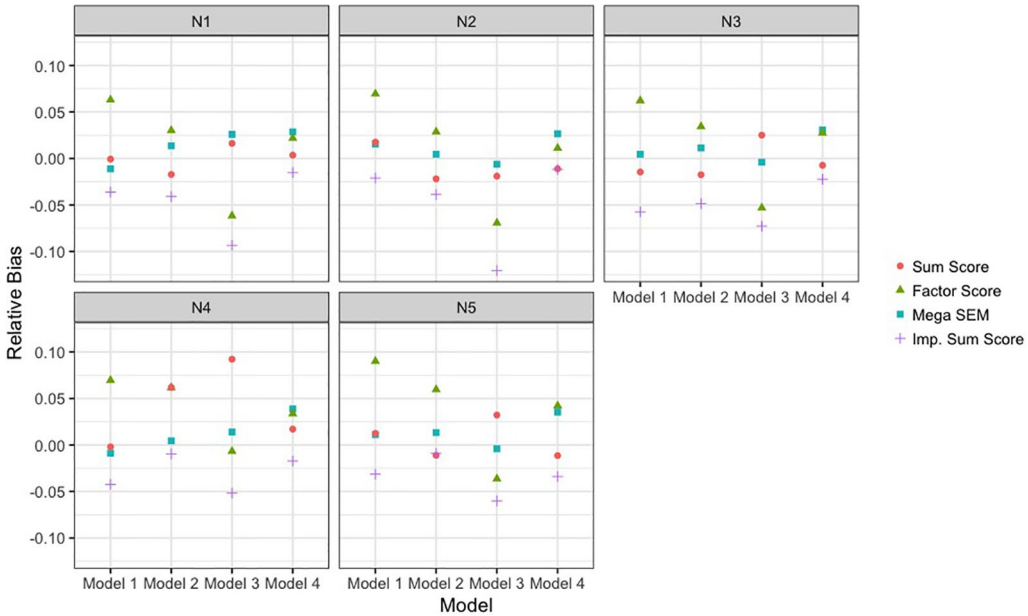

FIGURE 4 | Empirical bias of single-nucleotide polymorphism effect of each analysis approach relative to true effect size. Results are presented across the different data-generating models and the different sample size conditions detailed in **Table 1**. The analysis approaches were sum score meta-analysis, factor score meta-analysis, mega-analysis with the bi-factor integration mode, and multiple imputation of the missing items. SEM, structural equation modeling; Imp., imputation.

TABLE 3 | Relative bias, coverage rates, type I error rates, and standard errors computed with different analysis methods under the four data-generating models and five sample size conditions.

Model	N	Mega bias	FS bias	SS bias	Impute bias	Mega coverage	FS coverage	SS coverage	Impute coverage
Model1	N1	−0.011	0.063	−0.001	−0.036	0.944	0.939	0.962	0.972
Model2	N1	0.014	0.030	−0.017	−0.041	0.958	0.940	0.944	0.950
Model3	N1	0.026	−0.062	0.016	−0.093	0.959	0.903	0.948	0.966
Model4	N1	0.029	0.022	0.004	−0.015	0.949	0.941	0.956	0.955
Model1	N2	0.016	0.070	0.018	−0.021	0.941	0.926	0.937	0.947
Model2	N2	0.005	0.029	−0.022	−0.039	0.941	0.934	0.934	0.949
Model3	N2	−0.006	−0.069	-0.019	−0.121	0.941	0.910	0.944	0.963
Model4	N2	0.027	0.011	−0.011	−0.012	0.948	0.933	0.943	0.942
Model1	N3	0.005	0.062	−0.015	−0.057	0.952	0.929	0.949	0.962
Model2	N3	0.011	0.034	−0.017	−0.049	0.944	0.944	0.953	0.952
Model3	N3	−0.004	−0.053	0.025	−0.073	0.941	0.883	0.951	0.968
Model4	N3	0.031	0.027	−0.007	−0.022	0.958	0.944	0.950	0.966
Model1	N4	−0.009	0.070	−0.002	−0.042	0.954	0.934	0.956	0.966
Model2	N4	0.004	0.062	0.062	−0.010	0.957	0.938	0.949	0.960
Model3	N4	0.014	−0.007	0.092	−0.052	0.946	0.917	0.945	0.965
Model4	N4	0.039	0.034	0.017	−0.017	0.945	0.921	0.945	0.947
Model1	N5	0.011	0.090	0.012	−0.031	0.954	0.917	0.953	0.952
Model2	N5	0.013	0.060	−0.011	−0.009	0.955	0.929	0.950	0.958
Model3	N5	−0.004	−0.036	0.032	−0.060	0.938	0.917	0.936	0.961
Model4	N5	0.035	0.042	−0.011	−0.034	0.945	0.938	0.934	0.956

FS, factor score meta-analysis; Mega, BFIM SEM mega-analysis; SS, mean score meta-analysis; Impute, multiple imputation.

The meta-analysis of factor scores was least effective under data-generating model 4. Under this model, the factor score meta-analysis was slightly less powerful than the sum score-meta analysis. In the other three models, the factor-score meta-analysis was generally more powerful, but the degree of advantage was largest with larger sample sizes.

Advantage of the Reference Panel

The necessity of the reference panel was apparent in the simulations due to the low rates of convergence for both the BFIM mega-analysis and the BFIM factor analysis models. Without any overlapping items, the covariance matrix of the combined data had completely missing cells across the

cohorts. When two items were treated as the same item, reflecting the real-world scenario of recoding items based on similar face validity, the BFIM models often did not converge. When estimates were obtained, the power of the BFIM mega-analysis was similar to the sum score meta-analysis, and the estimates were downwardly biased. The meta-analysis of factor scores resulted in very low rates of power, due in part to non-convergence across repetitions and to unstable estimates with large standard errors. **Table 4** provides details across the four data-generating models with the N1 condition when no reference panel was included.

Results Summary

In conclusion, the BFIM mega-analysis approach employing the bi-factor integration model provided meaningful power increases, very low bias, and appropriate coverage. The factor score meta-analysis also resulted in power gains compared to sum score meta-analysis when the BFIM is correctly specified, although there was a small amount of bias in the estimates. Additionally, the use of the reference panel was crucial for the BFIM models. The measurement models were completely unstable when there was no item overlap, and harmonization carried out on two non-identical items caused issues for model estimation. The integration approach will be problematic when there is sparse item overlap, as would happen in consortia using different instruments across studies. The reference panel overcame this limitation and resulted in more power gain that the same amount of additional subjects added through one of the partners. Multiple imputation of the items using default settings followed by mega-analysis of sum scores resulted in the lowest power rates in all conditions.

APPLICATION

The BFIM approach was demonstrated on data from two cohorts participating in the ACTION Consortium (http://www.action-euproject.eu/; Boomsma, 2015; Bartels et al., 2018). The main objective of ACTION is to improve understanding of the sources of individual differences in aggression among children to better inform treatment strategies. The ACTION consortium is unique because several participating cohorts used distinct questionnaires to measure aggression in children. The ACTION Consortium also collected reference panel data: parents (fathers and mothers) of young twins from the Netherlands Twin Register (van Beijsterveldt et al. (2013)) were contacted when their children were around 9 years old to complete supplemental questionnaires that were also administered among other partnering studies. The reference panel data facilitate harmonization using the BFIM approach. We demonstrate the data management and analysis plan for saving harmonized factor scores. These scores can then but used as the phenotype in any genetic analysis.

TABLE 4 | Results for sample size condition 1 when no reference panel data were included.

Model	N	Mega rel. power	FS rel. power	Mega bias	FS bias
mdl1	N1	0.690	0.465	–0.378	0.012
mdl2	N1	0.620	0.449	–0.426	0.158
mdl3	N1	0.531	0.410	–0.497	0.212
mdl4	N1	0.662	0.483	–0.228	0.189

Participants

For this application, data were analyzed from mother report for twins around age 9. In all cases, subjects were retained for analysis if they had less than 30% missing values on the aggression items administered within their cohort. Details of samples and measures used are below.

Netherlands Twin Register

For this study, 22,772 mother reports for the Netherlands Twin Register (NTR) were used from collections when children were approximately 9 (mean = 9.94, SD = 0.51). The sample was 50.4% female. For details on data collection in the NTR see e.g., van Beijsterveldt et al. (2013).

Child and Adolescent Twin Study in Sweden

Parents from the Swedish Twin Register were interviewed *via* telephone on the 9th birthday of their children. Mother report data were available for 18,278 children at age 9. The sample was approximately 49.4% female at age 9. For details on Child and Adolescent Twin Study in Sweden (CATSS) see Anckarsäter et al. (2011).

The Phenotypic Reference Panel

The reference panel is a supplemental collection of participants from the NTR with additional questionnaires collected. Throughout 2017, the complete survey items were collected. Questionnaires were mailed to families with children around age 9 (mean = 9.42, SD = 0.78). The current study utilized mother report data on 2,205 children. The reference panel is 51.5% female.

Measures

Child Behavior Checklist

The Child Behavior Checklist (CBCL) 6–18 (Achenbach and Rescorla, 2001) was used by the NTR and the reference panel. The CBCL 6–18 consists of 120 items which are rated on a three-point scale ranging from "not true = 0," "somewhat or sometimes true = 1," to "very true or often true = 2." In the CBCL 6–18 aggressive symptom subscale, we identified 8 items that pertain directly to an overt/physical subtype of aggression for this analysis (see Lubke et al., 2018). These items are listed in **Table 5**.

Autism-Tics, Attention-Deficit Hyperactivity Disorder, and Other Comorbidities Inventory

The Autism-Tics, Attention-Deficit Hyperactivity Disorder, and Other Comorbidities Inventory (ATAC) (Larson et al., 2010) was administered in CATSS and the reference panel.

TABLE 5 | Overt/physical aggression items in Aggression in Children: Unraveling gene-environment interplay to inform Treatment and InterventiON strategies.

Item code	Item
ATAC63	Has there ever been a time when he/she would be angry to the extent that he/she cannot be reached?
ATAC65	Does he/she often tease others by deliberately doing things that are perceived as provocative?
ATAC70	Has he/she ever been deliberately been physical cruel to anybody?
ATAC71	Does he/she often get into fights?
CBCL016	Cruelty, bullying, or meanness to others
CBCL020	Destroys his/her own things
CBCL021	Destroys things belonging to his/her family or others
CBCL023	Disobedient at school
CBCL037	Gets in many fights
CBCL057	Physically attacks people
CBCL094	Teases a lot
CBCL095	Temper tantrums or hot temper

The ATAC is a comprehensive screening interview for autism spectrum disorders, attention deficit/hyperactivity disorder, tic disorders, developmental coordination disorder, learning disorders, and other childhood mental disorders. The ATAC included four items related to overt/physical subtype of aggression, and responses were scored on a three-point scale (response options "yes," "yes, to some extent," and "no"). These items are listed in **Table 5**.

Analysis Plan

The NTR, CATSS, and reference panel data were concatenated into the same dataset for analysis. A BFIM was constructed in which all items were modeled by a general factor, representing the target trait of overt/physical type aggression. Specific factors were used to model ATAC-specific and CBCL-specific item subsets. The factors were specified to be standard normal, with all factor loadings freely estimated, and the factors were all uncorrelated with each other.

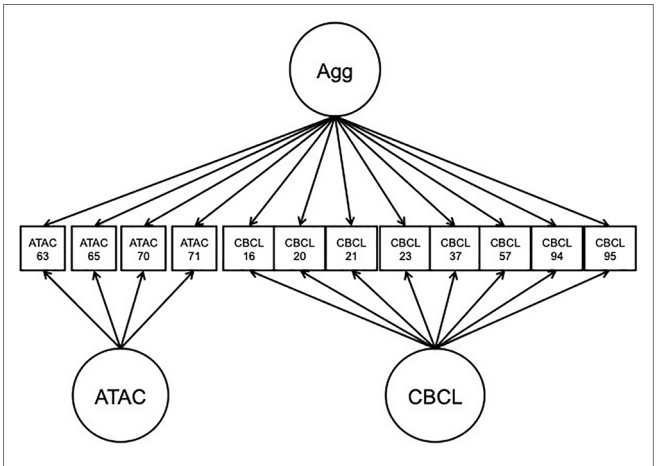

FIGURE 5 | Path diagram of bi-factor integration model application to Aggression in Children: Unraveling gene-environment interplay to inform Treatment and InterventiON strategies data. Agg, overt aggression factor; CBCL, child behavior check-list; ATAC, autism-tics, attention-deficit hyperactivity disorder, and other comorbidities scale.

The model is presented in **Figure 5**. Because subjects were nearly always twin pairings, sandwich-type robust standard errors were used for twins clustered within the same family. Analyses were carried out in Mplus 7.11 (Muthén and Muthén, 1998-2017).

RESULTS

The BFIM model overall displayed excellent fit to the data $\chi^2(42) = 418.98$, $p < 0.001$, $RMSEA = 0.014$ [0.013, 0.016], $CFI = 0.992$, $TLI = 0.988$).While the sum of a limited number of categorical items provide a small number of possible observed scores, the factor scores provided more nuance based on the relationship among all items. The factor scores and the sum scores were correlated at 0.91. Factor scores were computed as maximum a posteriori estimates of the factor scores (the only option for categorical variables in Mplus). The harmonized scores could be returned to individual cohorts for genetic analyses prior to meta-analysis. If genetic data is shared among cohorts, the integration model could be used as part of a larger mega-analysis.

DISCUSSION

The current paper presented IDA as a phenotype scoring framework for combining data across multiple independent studies. A bi-factor model for data integration was proposed that was designed specifically to adjust for measurement differences across multiple cohorts such as the use of different questionnaires. A series of simulation studies compared the BFIM to the standard approach of using sum scores of available items in each cohort and demonstrated the benefits of IDA in terms of increased power to detect a SNP effect. The BFIM was applied to two partnering cohorts in the ACTION Consortium using the collection of a reference panel dataset with responses to the questionnaires from both cohorts.

The IDA approach presented here has implications for joint gene association analyses carried out in genetic consortia. Several reviews have emphasized the need to improve phenotype measurement and consistent phenotype definition across the individual studies participating in GWAS meta-analyses (Evangelou and Ioannidis, 2013; Robinson et al, 2014). Psychometric measurement models have been promoted in other areas of behavior genetics research as well, such as twin and family studies (van den Berg et al., 2007; van der Sluis et al., 2010; Schwabe and van den Berg, 2014; Luningham et al., 2017). Researchers have suggested item response models for use in multi-cohort studies of personality traits (van den Berg et al., 2014). The BFIM proposed in this paper is designed specifically for harmonization in association tests within a consortium of disparate studies.

Importantly, this study revealed that a factor score meta-analysis provided a gain in power over separate studies using sum score scores that were not directly comparable, provided the bi-factor model was adequately specified. This reflects the likely scenario where phenotypic data can be shared and jointly modeled in a consortium, but a full genome-wide search with SEM is not tenable. While the power gains found in this study were small in an absolute sense, the only difference in obtaining empirical power was in the

method used. For GWAS analyses, where power is at a premium, a 4% gain in power simply through modeling the phenotype more precisely is a meaningful advantage. However, as advances such as GW-SEM (Verhulst et al., 2017) and genomic SEM (Grotzinger et al., 2019) make GWAS with multivariate outcomes and SEMs more feasible, our results suggest that the increase in power simply through IDA phenotypic modeling could be much greater.

A crucial aspect of the proposed IDA approach is the collection and use of a phenotypic reference panel. The concept of a phenotype reference panel to facilitate phenotype modeling was essential under our simulation scenarios. An important finding in the current paper is that the reference panel need only be a proportionately small increase in overall sample size to stabilize the integration model. It is more important to obtain a representative sample with complete phenotype information than to obtain new data as large as many of the participating partners. Collecting a small set of individuals that can bridge the gap in measurement items used across studies provide more benefit than increasing an individual cohort by the same amount of subjects.

The advantages of using a latent factor model to define a phenotype rather than a sum or mean score are reflected in the results of this paper. These advantages stem from using the full set of all available items and from accounting for different sources of heterogeneity in the observed score. The BFIM allows for modeling shared information in item subsets that does not pertain to the phenotype of interest, but to other sources of shared variance specific to certain studies. In our simulation, the residual covariance unrelated to the target trait was represented by different questionnaires. Additional sources of covariance among items can be incorporated into integrated measurement models, such as specific factors for different raters (e.g., mother and father) and residual covariance among similarly worded items. By modeling these as separate sources of commonality, the general factor becomes more precise. Covariate effects can also be included, such that item parameters differ as a function of age or gender (Curran et al., 2014).

More generally, the BFIM is not limited to GWAS, but can be applied in any joint analysis effort across multiple studies. IDA was proposed in psychological literature as a way to promote cumulative science, increase replicability of results, obtain broader psychometric assessments of constructs, and increase power (Curran and Hussong, 2009; Hussong et al., 2013). The bi-factor integration model presented in this paper is straightforward and has potential wide-ranging uses for detecting meaningful covariate effects on an integrated outcome. The bi-factor integration approach represents a potentially more powerful alternative to meta-analysis when phenotypic heterogeneity across studies needs to be taken into account.

The current study utilizes simulated and applied items that were already identified as pertaining to a unidimensional trait, in this case, overt aggression. An unexplored potential advantage of the IDA harmonization framework is the possibility to fit more complex models to larger item sets. Previous research has demonstrated that removing unreliable items in a questionnaire of a phenotype increases the power in a GWAS of that phenotype (Laurin et al., 2015). Further, conducting an association test of a sum score of items that actually originate from multiple subtypes of a trait can reduce GWAS power substantially compared to appropriately modeling the separate, but correlated, traits (van der Sluis et al., 2010). Applications of BFIM are contingent upon investigating the psychometric properties of the available questionnaires, and sum scores including irrelevant or unrelated items would likely lead to increasingly less power than an IDA approach. Future investigations can consider more complicated integration models.

In practice, the BFIM will need to be adapted to the particular data available. For example, one could utilize the bi-factor integration model and the multi-rater integration model of Bauer et al. (2013) to combine mother and father data across different studies in one analysis. Additionally, one could incorporate a limited number of covariate effects on item parameters as in the MNLFA model (Curran and Hussong, 2009; Curran et al., 2014). Consider a data integration scenario in which there were slight measurement differences across males and females. Rather than fitting separate models for each gender, one could use the bi-factor integration model and adjust for gender differences on individual items. Though real-world applications require careful application of complex measurement models, our study indicates that better phenotypic modeling with a reference panel can increase power at less cost than simply increasing sample size.

Limitations

Our simulation design represents a fairly simplified scenario compared to the complexity of research designs in applied data. In real data applications, on the other hand, it is difficult to calculate exact power gains in a specific study because the true data-generating process in the measurement model is usually unknown. Furthermore, results from simulation studies are specific to the chosen conditions and do not necessarily generalize to all possible scenarios. The pattern of power gains found in this study is expected to hold whenever phenotype precision is improved. The multiple imputation procedure included for comparison in our simulation was also limited to using the default imputation model and limited settings. In practice, multiple imputation can employ a wide range of prediction models, and the procedure can be optimized by increasing iterations and/or the number of datasets. Our results should not be seen as an indictment on imputation itself but on the shortcomings of using only default settings. However, the BFIM is better suited to explicitly adjust for differences in item sets as they pertain to the true underlying phenotype score, compared to a composite score of imputed items.

Data integration approaches also face challenges and potential pitfalls. Data integration requires extensive data sharing efforts among collaborators. While data sharing is often quite streamlined in genetic consortia, such agreements are not the norm for joint research ventures. The social sciences in general may benefit by working more collaboratively across existing studies. Data integration requires extremely careful planning, and joint model-fitting efforts can be very complex. Furthermore, model fitting with

a combined dataset that has a high degree of missingness requires computationally intensive estimation algorithms. On the other hand, computational resources are increasingly affordable, and models can be fitted with the help of distributed cluster computing and cloud storage. Finally, IDA is only effective if models are properly specified. Substantive experts and data analysts must work together closely to ensure that integrated phenotype models are theoretically sound.

Data integration is not a cure-all procedure to improve SNP detection, but it is a reasonable and additional step that can be taken in genetic collaborations to improve power in GWAS of complex phenotypes. With all of the effort in genetic research projects to collect genetic data, impute gene SNP information, control for genetic relatedness, and collaborate internationally, the additional effort in phenotype modeling is certainly a small price to pay for meaningful gains in power.

DATA AVAILABILITY STATEMENT

All datasets and scripts for the simulation study are included in the article/**Supplementary Material**. The NTR and CATSS datasets are not publicly available to protect sensitive phenotype information for participating children. The NTR and CATSS datasets are available by submitting a data request.

ETHICS STATEMENT

Data were previously collected under approval of the participating cohorts' original governing boards. All data used in the current analyses were collected under protocols that have been approved by the appropriate ethics committees, and studies were performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments.

AUTHOR CONTRIBUTIONS

JL, DM, and GL devised the bi-factor approach for integration. JL designed and conducted simulation studies. DM and GL advised simulation design. JL and GL drafted the first manuscript. DB, MB and CB supervised NTR data collections and set up the reference panel. AH facilitated the merging of aggression data across multiple ACTION partners. PL, HL, and SL supervised CATSS data collection and its partnership in ACTION. All authors edited the manuscript.

FUNDING

This work was supported by FP7-602768 "ACTION: Aggression in Children: Unraveling gene-environment interplay to inform Treatment and InterventiON strategies" from the European Commission/European Union Seventh Framework Program. GL was in addition supported by DA-018673 awarded by the National Institutes of Health: The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01227/full#supplementary-material

REFERENCES

Achenbach, T. M., and Rescorla, L. A. (2001). Manual for the ASEBA school-age forms and profiles (child behavior checklist for ages 6–18). *ASEBA, Burlington, Vermont*.

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of abnormal psychology,* 112 (4), 545

Anckarsäter, H., Lundström, S., Kollberg, L., Kerekes, N., Palm, C., Carlström, E., et al. (2011). The Child and Adolescent Twin Study in Sweden (CATSS). *Twin Res. Hum. Genet.* 14 (6), 495–508. doi: 10.1375/twin.14.6.495

Asparouhov, T., and Muthén, B. (2010). Weighted least squares estimation with missing data. *Mplus Technical Appendix,* 2010, 1–10.

Bartels, M., Boomsma, D. I., Hudziak, J. J., van Beijsterveldt, C. E. M., and van den Oord, E. J. C. G. (2007). Twins and the study of rater (dis)agreement. *Psychol. Methods* 12, 451–466. doi: 10.1037/1082-989X.12.4.451

Bartels, M., Hendriks, A., Mauri, M., Krapohl, E., Whipp, A., Bolhuis, K., et al. (2018). Childhood aggression and the co-occurrence of behavioural and emotional problems: results across ages 3-16 years from multiple raters in six cohorts in the EU-ACTION project. *Eur. Child Adolesc. Psychiatry* 9, 1105–1121. doi: 10.1007/s00787-018-1169-1

Bath, P. A., Deeg, D., and Poppelaars, J. (2010). The harmonisation of longitudinal data: a case study using data from cohort studies in The Netherlands and the United Kingdom. *Ageing Soc.* 30, 1419–1437. doi: 10.1017/S0144686X1000070X

Bauer, D. J., Howard, A. L., Baldasaro, R. E., Curran, P. J., Hussong, A. M., Chassin, L., et al. (2013). A trifactor model for integrating ratings across multiple informants. *Psychol. Methods* 18, 475–493. doi: 10.1037/a0032475

Bennett, S. N., Caporaso, N., Fitzpatrick, A. L., Agrawal, A., Barnes, K., Boyd, H. A., et al. (2011). Phenotype harmonization and cross-study collaboration in GWAS consortia: the GENEVA experience. *Genet. Epidemiol.* 35, 159–173. doi: 10.1002/gepi.20564

Bollen, K. A. (1989). *Structural equations with latent variables* (New York: Wiley and Sons). doi: 10.1002/9781118619179

Boomsma, D. I. (2015). Aggression in children: Unraveling the interplay of genes and environment through (epi)genetics and metabolomics. *J. Pediatr. Neonatal Individualized Med.* 4, e040251.

Cattell, R. B. (1952). *Factor analysis: an introduction and manual for the psychologist and social scientist*. New York: Harper.

Carrig, M. M., Manrique-Vallier, D., Randby, K. W., Reiter, J. P., and Hoyle, R. K. (2015). A nonparametric, multiple imputation-based method for the retrospective integration of data sets. *Multivariate Behav. Res.* 50, 383–397. doi: 10.1080/00273171.2015.1022641

Chen, F. F., West, S. G., and Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behav. Res.* 41, 189–225. doi: 10.1207/s15327906mbr4102_5

Collins, L. M., Schafer, J. L., and Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol. Methods* 6, 440–451. doi: 10.1037/1082-989X.6.4.330

Curran, P. J., and Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychol. Methods* 14, 81–100. doi: 10.1037/a0015914

Curran, P. J., Mcginley, J. S., Bauer, D. J., Hussong, A. M., Burns, A., Chassin, L., et al. (2014). A moderated nonlinear factor model for the development of commensurate measures in integrative data analysis. *Multivariate Behav. Res.* 49, 214–231. doi: 10.1080/00273171.2014.889594

Devlieger, I., Mayer, A., and Rosseel, Y. (2016). Hypothesis Testing Using Factor Score Regression: A Comparison of Four Methods. *Educational and Psychological Measurement,* 76(5), 741–770. doi: 10.1177/0013164415607618

Enders, C. K. (2010). *Applied missing data analysis.* New York: Guilford press.

Evangelou, E., and Ioannidis, J. P. A. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* 14, 379–389. doi: 10.1038/nrg3472

Gatz, M., Reynolds, C. A., Finkel, D., Hahn, C. J., Zhou, Y., and Zavala, C. (2015). Data harmonization in aging research: Not so fast. *Exp. Aging Res.* 41, 475–495. doi: 10.1080/0361073X.2015.1085748

Gibbons, R. D., and Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika* 57, 423–436. doi: 10.1007/BF02295430

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychol. Methods* 6, 430–450. doi: 10.1037/1082-989X.6.4.430

Grotzinger, A. D., Rhemtulla, M., de Vlaming, R., Ritchie, S. J., Mallard, T. T., Hill, W. D., et al. (2019). Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Human Behav.* 3 (5), 513. doi:10.1038/s41562-019-0566-x

Holzinger, K., and Swineford, F. (1937). The bi-factor method. *Psychometrika* 2, 41–54. doi: 10.1007/BF02287965

Hudziak, J. J., van Beijsterveldt, C. E. M., Bartels, M., Reitveld, M., Rettew, D., Derks, E., et al. (2003). Individual differences in aggression: Genetic analyses by age, gender, and informant in 3-, 7-, and 10-year-old Dutch twins. *Behav. Genet.* 5, 575–589. doi: 10.1023/A:1025782918793

Hussong, A. M., Curran, P. J., and Bauer, D. J. (2013). Integrative data analysis in clinical psychology research. *Annu. Rev. Clin. Psychol.* 9, 61–89. doi: 10.1146/annurev-clinpsy-050212-185522

Jak, S. (2017). Testing and explaining differences in common and residual factors across many countries. *J. Cross-Cult. Psychol.* 48, 75–92. doi: 10.1177/0022022116674599

Larson, T., Anckarsäter, H., Gillberg, C., Ståhlberg, O., Carlström, E., and Kadesjö, B. (2010). The autism–tics, AD/HD and other comorbidities inventory (A-TAC): further validation of a telephone interview for epidemiological research. *BMC Psychiatry* 10, 1. doi: 10.1186/1471-244X-10-1

Laurin, C. A., Hottenga, J. J., Willemsen, G., Boomsma, D. I., and Lubke, G. H. (2015). Genetic analyses benefit from using less heterogeneous phenotypes: an illustration with the hospital anxiety and depression scale (HADS). *Genet. Epidemiol.* 39, 317–324. doi: 10.1002/gepi.21897

Lawley, D. N., and Maxwell, A. E. (1963). *Factor analysis as a statistical method* (London: Buttersworth).

Ligthart, L., Bartels, M., Hoekstra, R. A., Hudziak, J. J., and Boomsma, D. I. (2005). Genetic contributions to subtypes of aggression. *Twin Res. Hum. Genet.* 8, 483–491. doi: 10.1375/twin.8.5.483

Little, R. J. A., and Rubin, D. B. (2002). *Statistical analysis with missing data. (2nd ed)* (New York: Wiley). doi: 10.1002/9781119013563

Lord, F. M., and Novick, M. R. (1968). *Statistical theories of mental test scores* (Reading, Mass: Addison-Wesley).

Lubke, G. H., McArtor, D. B., Boomsma, D. I., and Bartels, M. (2018). Genetic and environmental contributions to the development of childhood aggression. *Developmental psychology,* 54 (1), 39.

Luningham, J. M., McArtor, D. B., Bartels, M., Boomsma, D. I., and Lubke, G. H. (2017). Sum scores in twin growth curve models: practicality versus bias. *Behav. Genet.* 47, 516–536. doi: 10.1007/s10519-017-9864-0

Marcoulides, K. M., and Grimm, K. J. (2017). Data integration approaches to longitudinal growth modeling. *Educ. Psychol. Measurement* 77, 971–989. doi: 10.1177/0013164416664117

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., and Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychol. Methods* 14, 126–149. doi: 10.1037/a0015857

Muthén, L. K., and Muthén, B. O. (1998-2017). *Mplus user's guide*. 8th ed. (Los Angeles: Muthén and Muthén).

Nugent, W. R. (2009). Construct validity invariance and discrepancies in meta-analytic effect sizes based on different measures: A simulation study. *Educ. Psychol. Measurement* 69, 62–78. doi: 10.1177/0013164408318762

Pappa, I., St. Pourcain, B., Benke, K., Cavadino, A., Hakulinen, C., Nivard, M. G., et al. (2016). A genome-wide approach to children's aggressive behavior: the EAGLE consortium. *Am. J. Med. Genet. Part B.: Neuropsychiatr. Genet.* 171, 562–572. doi: 10.1002/ajmg.b.32333

Pedersen, N. L., Christensen, K., Dahl, A. K., Finkel, D., Franz, C. E., and Gatz, M. (2013). IGEMS: the consortium on interplay of genes and environment across multiple studies. *Twin Res. Hum. Genet.* 16, 481–489. doi: 10.1017/thg.2012.110

R Core Team (2018). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. Available at: https://www.R-project.org/.

Rietveld, C. A., Conley, D., Eriksson, N., Ekso, T., Medland, S. E., Vinkhuyzen, A. A., et al. (2014). Replicability and robustness of genome-wide association studies for behavioral traits. *Psychol. Sci.* 25, 1975–1986. doi: 10.1177/0956797614545132

Robinson, M. R., Wray, N. R., and Visscher, P. M. (2014). Explaining additional genetic variation in complex traits. *Trends In Genet.* 30, 124–132. doi: 10.1016/j.tig.2014.02.003

Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63 (3), 581–592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (New York: Wiley). doi: 10.1002/9780470316696

Schwabe, I., and van den Berg, S. M. (2014). Assessing genotype by environment interaction in case of heterogeneous measurement error. *Behav. Genet.* 44 (4), 394–406. doi: 10.1007/s10519-014-9649-7

Siddique, J., Reiter, J. P., Brincks, A., Gibbons, R. D., Crespi, C. M., and Brown, C. H. (2015). Multiple imputation for harmonizing longitudinal non-commensurate measures in individual participant meta-analysis. *Stat In Med.* 34, 3399–3414. doi: 10.1002/sim.6562

Skrondal, A., and Laake, P. (2001). Regression among factor scores. *Psychometrika* 66, 563–575. doi: 10.1007/BF02296196

Van Beijsterveldt, C., Groen-Blokhuis, M., Hottenga, J., Franić, S., Hudziak, J., Lamb, D., Boomsma, D. (2013). The Young Netherlands Twin Register (YNTR): longitudinal twin and family studies in over 70,000 children. *Twin Research and Human Genetics* 16 (1), 252–267. doi: 10.1017/thg.2012.118

van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *J. Stat. Software* 45 (3), 1–67. doi: 10.18637/jss.v045.i03

van Buuren, S. (2018). *Flexible imputation of missing data*. 3rd Ed (Chapman and Hall/CRC Press). doi: 10.1201/b11826

van den Berg, S. M., Glas, C. A. W., and Boomsma, D. I. (2007). Variance decomposition using an IRT measurement model. *Behav. Genet.* 37, 604–616. doi: 10.1007/s10519-007-9156-1

van den Berg, S. M., de Moor, M. H. M., Mcgue, M., Pettersson, E., Terracciano, A., Verweij, K. J. H., et al. (2014). Harmonization of the neuroticism and extraversion phenotypes across inventories and cohorts in the Genetics of Personality Consortium: an application of item response theory. *Behav. Genet.* 44, 295–313. doi: 10.1007/s10519-014-9654-x

van der Sluis, S., Verhage, M., Posthuma, D., and Dolan, C. V. (2010). Phenotypic complexity, measurement bias, and poor phenotypic resolution contribute to the missing heritability problem in genetic association studies. *PloS One* 5, e13929. doi: 10.1371/journal.pone.0013929

Verhulst, B., Maes, H., and Neale, M. (2017). GW-SEM: a statistical package to conduct genome-wide structural equation modeling. *Behav. Genet.* 47, 345–359. doi: 10.1007/s10519-017-9842-6

Wray, N. R., Birley, A. J., Sullivan, P. F., Visscher, P. M., and Martin, N. G. (2007). Genetic and phenotypic stability of measures of neuroticism over 22 years. *Twin Res. Hum. Genet.* 10, 695–702. doi: 10.1375/twin.10.5.695

Xu, M. K., Gaysina, D., Barnett, J. H., Scoriels, L., van de Lagemaat, L. N., Wong, A., et al. (2015). Psychometric precision in phenotype definition is a useful step in molecular genetic investigation of psychiatric disorders. *Trans. Psychiatry* 8, 316–326. doi: 10.1038/tp.2015.86

Yeh, M. T., Coccaro, E. F., and Jacobson, K. C. (2010). Multivariate behavior genetic analyses of aggressive behavior subtypes. *Behav. Genet.* 40, 603–617. doi: 10.1007/s10519-010-9363-z

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX I: DATA-GENERATING MODEL PARAMETERS

Four data-generating models were used in the simulations—three using a bi-factor model, and a fourth using a higher-order latent variable model that resulted in the bi-factor model having slight misspecification to the generated data. The vector of latent variables was generated from the following model for an individual *j*

$$\boldsymbol{\eta}_j = \boldsymbol{\beta}\boldsymbol{x}_j + \boldsymbol{\zeta}_j \tag{9}$$

where η is a *k*×1 vector of *k* latent factors for person *j*, $\mathbf{x}_j$ is a $Q \times 1$ vector of *Q* predictors (the genetic variant and any covariates) for person *j*, β is a $k \times Q$ vector of effect sizes of each covariate on each latent factor, and ζ is a vector of residuals for the *k* factors. In these simulations, we generated a general factor that was predicted by a single nucleotide polymorphism (SNP) and a covariate term, along with two independent specific factors that did not have exogenous predictors:

$$\begin{pmatrix} \eta_{jg} \\ \eta_{j1} \\ \eta_{j2} \end{pmatrix} = \begin{pmatrix} 0.045 & 0.894 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} SNP_j & x_j \end{pmatrix} + \begin{pmatrix} \zeta_{jg} \\ \zeta_{j1} \\ \zeta_{j2} \end{pmatrix},$$

$$\zeta_j \sim MVN\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} .799 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right) \tag{10}$$

In this equation, the residual variance of the general factor is 0.799 because it is calculated as 1 minus the sum of the variance explained by the predictors.

The item-level data is then generated from the factors and the factor loading matrices:

$$\boldsymbol{y}_j = \boldsymbol{\Lambda}\boldsymbol{\eta}_j + \boldsymbol{\varepsilon}_j \tag{11}$$

where y_j is a *p*×1 vector of items for person *j*, Λ is a *p*×*k* matrix of factor loadings for each *p* items on the *k* factors, and ε_j is a *p*×1 vector of item residuals. To manipulate the conditions across the first three data-generating conditions, the loadings matrices and residual variance matrix was manipulated:

$$\boldsymbol{\Lambda_1} = \begin{bmatrix} 0.6 & 0.49 & 0 \\ 0.4 & 0.66 & 0 \\ 0.5 & 0.59 & 0 \\ 0.5 & 0.59 & 0 \\ 0.6 & 0 & 0.49 \\ 0.5 & 0 & 0.59 \\ 0.5 & 0 & 0.59 \\ 0.4 & 0 & 0.66 \end{bmatrix}; \boldsymbol{\Lambda_2} = \begin{bmatrix} 0.5 & 0.45 & 0 \\ 0.3 & 0.60 & 0 \\ 0.4 & 0.54 & 0 \\ 0.4 & 0.54 & 0 \\ 0.6 & 0 & 0.49 \\ 0.5 & 0 & 0.59 \\ 0.5 & 0 & 0.59 \\ 0.4 & 0 & 0.66 \end{bmatrix};$$

$$\boldsymbol{\Lambda_3} = \begin{bmatrix} 0.5 & 0.45 & 0 \\ 0.3 & 0.60 & 0 \\ 0.4 & 0.54 & 0 \\ 0.4 & 0.54 & 0 \\ 0.6 & 0 & 0.49 \\ 0.5 & 0 & 0.59 \\ 0.5 & 0 & 0.59 \\ 0.4 & 0 & 0.66 \end{bmatrix}; \tag{12}$$

The residual variance is calculated as:

$$Var\left(\varepsilon_j\right) \sim MVN\left(0, \boldsymbol{I_p} - diag\left(\boldsymbol{\Lambda'\Lambda}\right)\right) \tag{13}$$

Meaning that the variance is a diagonal matrix. In addition, a covariance is inserted in the variance matrix such that $\sigma_{38} = 0.6$.

For the fourth data-generating model, the general factor is directly predicted by the SNP and covariate as before, and then the two cohort-specific factors are directly caused by the general factor itself, such that it accounts for 60% of the variance in the two factors:

$$\begin{pmatrix} \eta_{j1} \\ \eta_{j2} \end{pmatrix} = \begin{pmatrix} 0.77 \\ 0.65 \end{pmatrix} \begin{pmatrix} \eta_{jg} \end{pmatrix} + \begin{pmatrix} \zeta_{j1} \\ \zeta_{j2} \end{pmatrix},$$

$$\zeta_j \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{I} - diag\left(\begin{pmatrix} 0.77 \\ 0.65 \end{pmatrix} \begin{pmatrix} 0.77 & 0.65 \end{pmatrix} \right) \right) \tag{14}$$

The item loading matrix is then *p*×2 instead of *p*×3 and the items also have a direct effect of the general factor. For graphical representations, see **Figure 2**. The codes used to conduct the simulation are attached as a supplementary downloadable folder.