Behav Genet (2014) 44:295–313
DOI 10.1007/s10519-014-9654-x

ORIGINAL RESEARCH

Harmonization of Neuroticism and Extraversion phenotypes across inventories and cohorts in the Genetics of Personality Consortium: an application of Item Response Theory

Stéphanie M. van den Berg · Marleen H. M. de Moor · Matt McGue · Erik Pettersson · Antonio Terracciano · Karin J. H. Verweij · Najaf Amin · Jaime Derringer · Tõnu Esko · Gerard van Grootheest · Narelle K. Hansell · Jennifer Huffman · Bettina Konte · Jari Lahti · Michelle Luciano · Lindsay K. Matteson · Alexander Viktorin · Jasper Wouda · Arpana Agrawal · Jüri Allik · Laura Bierut · Ulla Broms · Harry Campbell · George Davey Smith · Johan G. Eriksson · Luigi Ferrucci · Barbera Franke · Jean-Paul Fox · Eco J. C. de Geus · Ina Giegling · Alan J. Gow · Richard Grucza · Annette M. Hartmann · Andrew C. Heath · Kauko Heikkilä · William G. Iacono · Joost Janzing · Markus Jokela · Lambertus Kiemeney · Terho Lehtimäki · Pamela A. F. Madden · Patrik K. E. Magnusson · Kate Northstone · Teresa Nutile · Klaasjan G. Ouwens · Aarno Palotie · Alison Pattie · Anu-Katriina Pesonen · Ozren Polasek · Lea Pulkkinen · Laura Pulkki-Råback · Olli T. Raitakari · Anu Realo · Richard J. Rose · Daniela Ruggiero · Ilkka Seppälä · Wendy S. Slutske · David C. Smyth · Rossella Sorice · John M. Starr · Angelina R. Sutin · Toshiko Tanaka · Josine Verhagen · Sita Vermeulen · Eero Vuoksimaa · Elisabeth Widen · Gonneke Willemsen · Margaret J. Wright · Lina Zgaga · Dan Rujescu · Andres Metspalu · James F. Wilson · Marina Ciullo · Caroline Hayward · Igor Rudan · Ian J. Deary · Katri Räikkönen · Alejandro Arias Vasquez · Paul T. Costa · Liisa Keltikangas-Järvinen · Cornelia M. van Duijn · Brenda W. J. H. Penninx · Robert F. Krueger · David M. Evans · Jaakko Kaprio · Nancy L. Pedersen · Nicholas G. Martin · Dorret I. Boomsma

Received: 21 October 2013 / Accepted: 20 March 2014 / Published online: 15 May 2014

Abstract Mega- or meta-analytic studies (e.g. genome-wide association studies) are increasingly used in behavior genetics. An issue in such studies is that phenotypes are often measured by different instruments across study cohorts, requiring harmonization of measures so that more powerful fixed effect meta-analyses can be employed. Within the Genetics of Personality Consortium, we demonstrate for two clinically relevant personality traits, Neuroticism and Extraversion, how Item-Response Theory (IRT) can be applied to map item data from different inventories to the same underlying constructs. Personality item data were analyzed in >160,000 individuals from 23

Edited by Kristen Jacobson.

Stéphanie M. van den Berg and Marleen H. M. de Moor are the co-first authors.

Electronic supplementary material The online version of this article (doi:10.1007/s10519-014-9654-x) contains supplementary material, which is available to authorized users.

S. M. van den Berg · J.-P. Fox
Department of Research Methodology, Measurement and Data-Analysis, University of Twente, Enschede, The Netherlands

S. M. van den Berg (✉)
Department of Behavioural Sciences, OMD, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands
e-mail: stephanie.vandenberg@utwente.nl

M. H. M. de Moor · J. Wouda · E. J. C. de Geus · K. G. Ouwens · G. Willemsen · D. I. Boomsma
Department of Biological Psychology, VU University, Amsterdam, The Netherlands

M. McGue · L. K. Matteson · W. G. Iacono · R. F. Krueger
Department of Psychology, University of Minnesota, Elliott Hall, Minneapolis, MN, USA

M. McGue
Institute of Public Health, University of Southern Denmark, Odense, Denmark

E. Pettersson · A. Viktorin · P. K. E. Magnusson · N. L. Pedersen
Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

cohorts across Europe, USA and Australia in which Neuroticism and Extraversion were assessed by nine different personality inventories. Results showed that harmonization was very successful for most personality inventories and moderately successful for some. Neuroticism and Extraversion inventories were largely measurement invariant across cohorts, in particular when comparing cohorts from countries where the same language is spoken. The IRT-based scores for Neuroticism and Extraversion were heritable (48 and 49 %, respectively, based on a meta-analysis of six twin cohorts, total N = 29,496 and 29,501 twin pairs, respectively) with a significant part of the heritability due to non-additive genetic factors. For Extraversion, these genetic factors qualitatively differ across sexes. We showed that our IRT method can lead to a large increase in sample size and therefore statistical power. The IRT approach may be applied to any mega- or meta-analytic study in which item-based behavioral measures need to be harmonized.

Keywords Personality · Item-Response Theory · Measurement · Genome-wide association studies · Consortium · Meta-analysis

Introduction

Mega- or meta-analytic studies (e.g. genome-wide association (GWA) studies) are increasingly used in behavior genetics. Because phenotypes have not always been assessed similarly across cohorts (and sometimes not even within cohorts), measures need to be harmonized, that is, phenotypic scores need to be made comparable such that data from individuals who were assessed by different inventories can be compared meaningfully. Such harmonization then enables fixed effect meta-analytic analyses (Hedges and Vevea 1998). Meta-analytic studies are required when effect sizes are small such as for complex human traits. For example, GWA studies for psychiatric disorders have led to important discoveries, but for many disorders, individual variants typically explain less than 1 % of the heritability, although in unison they can explain quite a large proportion of phenotypic variation (Craddock et al. 2008; Lee et al. 2013; Ripke et al. 2013; Sullivan et al. 2012). Sample size determines the number of significant loci discovered (Sullivan et al. 2012), so that meta-analysis of results is the gold standard. Consortium GWA studies for traits such as height and body-mass index now report sample sizes of >100,000 (Berndt et al. 2013; Lango Allen and et al. 2010; Speliotes et al. 2010). Consortia for psychiatric disorders and behavioral traits have also been formed, with sample sizes increasing rapidly to hundreds of thousands (Rietveld et al. 2012; Ripke et al. 2011; Wray et al. 2012), leading to the discovery of novel loci for psychiatric disorders and educational attainment. Thus, large sample sizes are essential for behavioral phenotypes.

A meta-analysis of behavioral measures will have most power if the same reliable and valid measurement instrument is administered in all cohorts. In practice, however, different instruments are often used, and, even when the instrument is the same, translations into different languages may cause problems. To tackle the problem that different inventories

A. Terracciano · L. Ferrucci · A. R. Sutin · T. Tanaka
National Institute on Aging, NIH, Baltimore, MD, USA

A. Terracciano · A. R. Sutin
College of Medicine, Florida State University, Tallahassee, FL, USA

K. J. H. Verweij · N. K. Hansell · D. C. Smyth · M. J. Wright · N. G. Martin
QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia

K. J. H. Verweij
Department of Developmental Psychology and EMGO Institute for Health and Care Research, VU University Amsterdam, Amsterdam, The Netherlands

N. Amin · C. M. van Duijn
Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands

J. Derringer
Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL, USA

T. Esko · A. Metspalu
Estonian Genome Center, University of Tartu, Tartu, Estonia

G. van Grootheest · B. W. J. H. Penninx
Department of Psychiatry, EMGO+ Institute, Neuroscience Campus Amsterdam, VU University Medical Center Amsterdam, Amsterdam, The Netherlands

J. Huffman · C. Hayward
MRC Human Genetics, MRC IGMM, Western General Hospital, University of Edinburgh, Edinburgh, Scotland, UK

B. Konte · I. Giegling · A. M. Hartmann · D. Rujescu
Department of Psychiatry, University of Halle, Halle, Germany

J. Lahti · M. Jokela · A.-K. Pesonen · L. Pulkki-Råback · K. Räikkönen · L. Keltikangas-Järvinen
Institute of Behavioural Sciences, University of Helsinki, Helsinki, Finland

J. Lahti · J. G. Eriksson · K. Räikkönen
Folkhälsan Research Center, Helsinki, Finland

M. Luciano · A. Pattie · I. J. Deary
Department of Psychology, University of Edinburgh, Edinburgh, UK

M. Luciano · A. Pattie · J. M. Starr · I. J. Deary
Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK

may not assess the same phenotype, we demonstrate how Item-Response Theory (IRT) test linking can be applied to map item data from different inventories to a common metric. We conduct such an analysis for Neuroticism and Extraversion personality traits, based on data from the Genetics of Personality Consortium (GPC). If different inventories indeed measure the same phenotype, the only requirement for this approach is that multiple inventories have been administered in at least a subset of individuals. That is, in order to be able to harmonize across different inventories, some participants must have filled in multiple inventories so that they can function as a "bridge" between inventories. This can be done if we assume that the true phenotype (personality) does not change between the multiple assessments. If this can be assumed, then for all individuals in the different (sub-)cohorts, a score on the latent construct can be estimated based on all available item data for that person. The IRT-based score estimates for Neuroticism and Extraversion can subsequently be meta-analyzed to assess heritability, or can be used as phenotypes in GWA or brain-imaging studies.

This IRT approach has multiple advantages. First, within each cohort there is increased measurement reliability, because when multiple inventories have been administered to the same individual, scores can be estimated using the items from all relevant inventories. In addition, items can be differentially and optimally weighted if necessary, and items that do not fit the measurement model can be identified and omitted, thereby increasing power. Subgroups of individuals that were assessed with only a subset of items can now also be included in the study. Moreover, the IRT approach can statistically evaluate the extent to which different inventories actually measure the same construct. Lastly, IRT enables researchers to determine the extent of measurement invariance across cohorts: can scores across cohorts be quantitatively compared and therefore pooled and meaningfully used in a meta-analysis?

Applying the IRT method to Neuroticism and Extraversion is especially relevant for the field of behavior genetics, as these personality traits are correlated with numerous other traits and disorders, not only phenotypically but also genetically (Heath et al. 1994; Hopwood et al. 2011; Klein et al. 2011; Markon et al. 2005; Samuel and Widiger 2008). For example, Neuroticism is highly related to a variety of psychiatric disorders, including major depression and borderline personality disorder (Distel et al. 2009; Kendler and Myers 2009), and Extraversion is associated with alcohol use (Dick et al. 2013). Earlier GWA studies of personality (De Moor et al. 2010; Service et al. 2012; Shifman et al. 2008; Terracciano et al. 2010; van den Oord et al. 2008) focused on single inventories, hence hampering sample size, and few, if any, genome-wide significant loci were detected. Large sample sizes are needed, which can be achieved by pooling results from multiple inventories.

This study included data obtained from 160,958 individuals from 23 cohorts, of which 6 were twin cohorts. Neuroticism and Extraversion were assessed by 9 different personality inventories; 7 cohorts assessed more than one inventory. The first objective was to determine the feasibility of the IRT approach in linking Neuroticism and Extraversion item data from different inventories: to what extent do the different inventories measure the same constructs? For instance, Harm Avoidance correlates moder-

A. Agrawal · L. Bierut · R. Grucza · A. C. Heath · P. A. F. Madden
Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, USA

J. Allik · A. Realo
Department of Psychology, University of Tartu, Tartu, Estonia

J. Allik · A. Metspalu
Estonian Academy of Sciences, Tallinn, Estonia

U. Broms · K. Heikkilä · E. Vuoksimaa · J. Kaprio
Department of Public Health, Hjelt Institute, University of Helsinki, Helsinki, Finland

U. Broms · J. G. Eriksson · K. Räikkönen · J. Kaprio
National Institute for Health and Welfare (THL), Helsinki, Finland

H. Campbell · L. Zgaga · J. F. Wilson · I. Rudan
Centre for Population Health Sciences, Medical School, University of Edinburgh, Edinburgh, UK

G. D. Smith · K. Northstone · D. M. Evans
MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol, UK

J. G. Eriksson
Department of General Practice and Primary Health Care, University of Helsinki, Helsinki, Finland

J. G. Eriksson
Unit of General Practice, Helsinki University Central Hospital, Helsinki, Finland

J. G. Eriksson
Vasa Central Hospital, Vaasa, Finland

B. Franke · A. Arias Vasquez
Donders Institute for Cognitive Neuroscience, Radboud University Nijmegen, Nijmegen, The Netherlands

B. Franke · J. Janzing · A. Arias Vasquez
Department of Psychiatry, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands

B. Franke · S. Vermeulen · A. Arias Vasquez
Department of Human Genetics, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands

A. J. Gow
Department of Psychology, School of Life Sciences, Heriot-Watt University, Edinburgh, UK

ately high with Neuroticism (r = 0.5–0.6) (De Fruyt et al. 2000). Therefore, we expect that mapping item data from Harm Avoidance with Neuroticism will be less perfect than mapping Neuroticism item data from other personality inventories (e.g. EPQ versus NEO neuroticism). We expect that this is even more the case for mapping Reward Dependence with Extraversion. Here we determine to what extent cross-inventory mapping is feasible, for the purpose of a GWAS meta-analysis in mind. The second objective was to test for measurement invariance across cohorts, and the third objective was to establish the heritability of the harmonized Neuroticism and Extraversion scores in the six participating twin cohorts. Sex differences in the genetic background of Neuroticism and Extraversion were studied, as well as the contribution of non-additive genetic factors. The contribution of non-additive genetic factors to variation in personality traits has been extensively discussed in the literature (Keller et al. 2005), but their assessment requires a large sample (Posthuma and Boomsma 2000). Lastly, we studied the theoretical increase in power of finding a quantitative trait locus due to the harmonization of phenotypes in two large cohorts.

Materials and methods

Cohorts

Twenty-three cohorts of the GPC were included in this study (for detailed descriptions, see Supplementary Materials Online). Seventeen cohorts originated from Europe, 4 cohorts were from the USA and 2 cohorts from Australia. Most cohorts are large epidemiological studies. Some of the cohorts focused on specific birth cohorts and/or recruited individuals of specific regions in the country (e.g. ERF, VIS, KORCULA, NBS, LBC1921, LBC1936 and HBCS), or targeted twins and their family members (QIMR cohorts, NTR, MCTFR, STR, Finnish Twin Cohort). Three cohorts were designed to include cases and controls for Nicotine dependence, Alcoholism or Mood and Anxiety disorders (respectively, COGEND, SAGE-COGA and NESDA). The data collection in some of the cohorts is longitudinal in nature.

Personality assessment

Supplementary Table 1 and Supplementary Fig. 3 give an overview of the personality inventories administered in each cohort. The Supplementary Materials Online describes these inventories in detail. For the Neuroticism analysis, we included all Neuroticism items from the NEO, the International Personality Item Pool (IPIP) and Eysenck (EPQ, EPI, ABV) inventories, the Harm Avoidance (HA) items from the Temperament and Character Inventory (TCI), and the Negative Emotionality (NEM) items (excluding the aggression items) from the Multidimensional Personality Questionnaire (MPQ). The Neuroticism scales of the NEO, IPIP and Eysenck inventories consist of different items, but there is strong overlap in item content and the sum scores correlate highly across inventories (Aluja et al. 2004; Draycott and Kline 1995; Larstone et al. 2002). HA correlates most strongly with Neuroticism (as assessed with the NEO-PI-R or EPQ-R) (De Fruyt et al. 2000; Gillespie et al. 2001). NEM corresponds most closely to Neuroticism, although NEM is a broader concept because it also includes items about aggressive behavior.

M. Jokela · T. Lehtimäki · I. Seppälä
Department of Clinical Chemistry, Fimlab Laboratories and School of Medicine, University of Tampere, Tampere, Finland

L. Kiemeney · S. Vermeulen
Department of Health Evidence, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands

L. Kiemeney
Department of Urology, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands

T. Nutile · D. Ruggiero · R. Sorice · M. Ciullo
Institute of Genetics and Biophysics "A. Buzzati-Traverso" – CNR, Naples, Italy

A. Palotie
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

A. Palotie · E. Widen · J. Kaprio
Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

O. Polasek
Department of Public Health, Faculty of Medicine, University of Split, Split, Croatia

L. Pulkkinen
Department of Psychology, University of Jyväskylä, Jyväskylä, Finland

O. T. Raitakari
Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku, Finland

O. T. Raitakari
Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku, Finland

R. J. Rose
Department of Psychological & Brain Sciences, Indiana University, Bloomington, IN, USA

W. S. Slutske
Department of Psychological Sciences and Missouri Alcoholism Research Center, University of Missouri, Columbia, MO, USA

For the Extraversion analysis, all Extraversion items from the NEO, IPIP and Eysenck inventories were analyzed, a selection of Reward Dependence (RD) items from the TCI, and the Positive Emotionality (PEM) items from the MPQ. Extraversion sum scores derived from the NEO, IPIP and Eysenck inventories correlate highly across inventories (Aluja et al. 2004; Draycott and Kline 1995; Larstone et al. 2002). The relationship between Extraversion and the temperament traits is less clear, but Extraversion correlates strongest with RD (De Fruyt et al. 2000; Gillespie et al. 2001). Based on the item correlations among the RD items with the Extraversion items from the NEO-PI-R and EPQ in the HBCS, PAGES and QIMR adults cohorts, we decided to include a subset of RD items that correlated strongest with the Extraversion items (see Supplementary Fig. 3 for number of items included and Supplementary Table 2 for overview of the items).

Estimating Neuroticism and Extraversion scores

The harmonization goal is to estimate personality scores that are not biased by the number of items and the specific inventory used. In the field of IRT, such harmonization is termed 'test linking'. By fitting IRT models (Lord 1980) to item data, personality scores can be estimated conditional on the observed items and their respective item parameters. This leads to personality scores for individuals that are comparable irrespective of what items were assessed in a particular individual. For example, image an intelligence assessment: If we know that items 1–10 are very easy test items, and items 11–20 are very difficult, we are pretty confident that a person that scores 1 on the items 1–10 is less bright than a person that scores 9 on items 11–20. The exact knowledge of the difficulties of the 20 items allows us to estimate the difference in intelligence.

A basic IRT model assumes a one-dimensional latent variable representing the trait that predicts the probability of a certain response on a particular item: the higher the latent trait value, the higher the probability of a high score on the item. Item parameters determine the exact relationship between the latent trait and the probability of the response to a particular item. The so-called difficulty parameter provides information about the general probability of a positive response to a particular item, and is very similar to the threshold parameter in liability models. The discrimination parameter value of an item indicates how strong the relationship is between the latent trait and the item response variable, and is therefore similar to a factor loading. Because latent scores are estimated conditional on the item parameters for the administered items, the scoring process becomes independent of the particular items in the test. For example, this allows the comparison of a child's achievement on a test with easy questions with the achievement of another child on a test with difficult questions. IRT test linking was applied in each cohort separately and used to link all data from one cohort to one common metric for Neuroticism and one common metric for Extraversion. For more details, see Supplementary Materials Online.

Appropriateness of Item Response Theory to harmonize Neuroticism and Extraversion scores

We assessed whether the IRT Neuroticism and Extraversion scores in the 23 cohorts were truly independent of the specific inventory used. First, the appropriateness of linking tests *within cohorts* was investigated by testing basic assumptions of IRT models: the idea that scoring is independent of the specific item set that was administered (local independence), and unidimensionality. For every cohort and every inventory separately, item parameters were estimated based on data from individuals without missing data. Such a set of parameter values for a particular sample of items assessed in a particular sample is termed a calibration. Calibrations were also obtained for *combinations* of item sets from various inventories, if there was a subsample of individuals that was assessed with those inventories. Based on these calibrations, (i.e., sets of item parameter values), latent scores can be estimated for those individuals for which one has either complete data or data with some missing values, assuming these are missing at random. In order to investigate local independence, latent scores for a particular item set (say, item scores for NEO-PI-R) were estimated and compared based on different calibrations: one based on the calibration of several inventories combined (e.g., NEO-PI-R and EPQ-R Neuroticism) and one based on only one inventory (NEO-PI-R items). The resulting scores were then correlated. A

J. Verhagen
Department of Psychological Methods, University of Amsterdam, Amsterdam, The Netherlands

E. Vuoksimaa
Department of Psychiatry, University of California, La Jolla, CA, USA

L. Zgaga
Department of Public Health and Primary Care, Trinity College Dublin, Dublin, Ireland

A. Arias Vasquez
Department of Cognitive Neuroscience, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands

P. T. Costa
Behavioral Medicine Research Center, Duke University School of Medicine, Durham, NC, USA

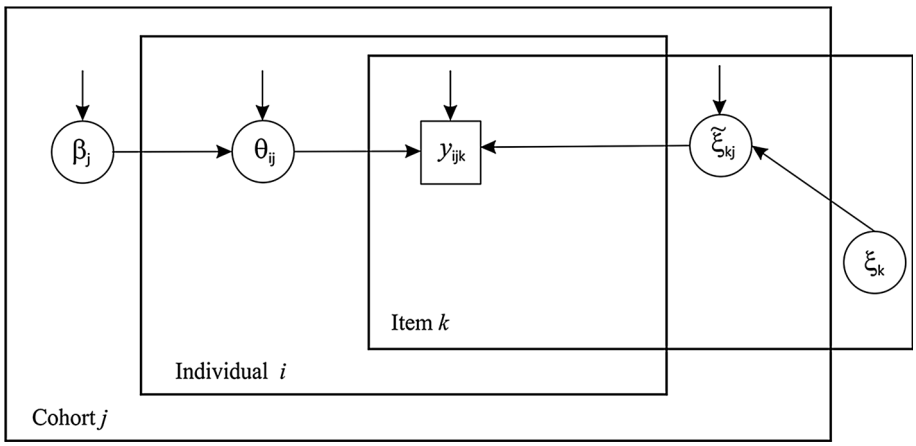

Fig. 1 A graph representation of the hierarchical model for measurement variance. Item parameters ξ (thresholds and discrimination parameter) are allowed to vary across cohorts, but person parameters are allowed to vary both across cohorts and within cohorts. Observed response Y_{ijk} from person i in cohort j to item k is predicted by a latent score θ_{ij} for that person and item parameters ξ_{kj} for item k that is specific for cohort j

correlation of 1 indicates that the estimated scores are completely independent of what inventory was used for assessment (see also Supplementary Materials Online).

Unidimensionality was assessed by plotting the test information curves (TICs) (Lord 1980; van den Berg and Service 2012) for inventories separately and with two or more inventories combined. If two tests measure the same underlying construct, the TIC of the tests combined should be the sum of the TICs of the two separate tests. These curves also show the increase in measurement precision for those individuals that were administered multiple inventories.

The choice for the above approach to assessing model fit, which is a bit unconventional, was motivated by the fact that the personality inventories are well-developed and validated instruments. Also, from previous research we know that two-parameter models generally are more appropriate for personality data than one- and three-parameter models (Chernyshenko et al. 2001; Reise and Waller 1990). As one aim is to use as much information as possible from the personality inventories, to establish a linear relationship between personality scales and an external variable, such as a SNP, we chose to retain all items in the analyses.

The above analysis determines whether within cohorts, items from inventories can be combined, that is, whether different inventories can be used to measure the same trait. In addition, it is important to assess whether *across cohorts*, the same trait is being measured. If Neuroticism and Extraversion were very differently expressed across cohorts, a meta-analysis is rather meaningless. Due to a host of reasons (culture, language, sample selection criteria, etc.), the same test items might have different parameters across cohorts. Ignoring these differences results in systematic bias when comparing individual sum scores from different cohorts. The assumption of equal item parameters across groups is usually termed measurement invariance (Meredith 1993). If one item has different parameter values across groups, this is called differential item functioning (DIF) (Glas 1998, 2001; Speliotes et al. 2010). There are two ways of dealing with DIF, either (1) omitting the item entirely in estimating individual scores, or (2) allowing for different item parameters for that particular DIF item across groups (Weisscher et al. 2010). The first approach leads to loss of information, so that the second is generally more attractive.

A new alternative Bayesian method for modeling measurement non-invariance (Verhagen and Fox 2013a, b) was applied to assess variance of item parameters across cohorts and that identifies true differences in means and variances of Neuroticism and Extraversion across cohorts, while controlling for any measurement non-invariance. The Bayesian approach allows for estimating complicated models in a straightforward way, and through hierarchical modeling one borrows statistical strength for small cohorts from information in larger cohorts. The Bayesian hierarchical approach assumes there is at least some violation of measurement invariance, and quantifies its extent. Since there are some important differences across cohorts in terms of population and language, we expect there will be at least some difference in item parameters across cohorts.

In the Bayesian hierarchical approach, item and person parameters are estimated using a Markov Chain Monte Carlo procedure, in which cohort-specific item parameters are considered level-1 parameters randomly distributed around overall mean item parameters at level 2. See Fig. 1 for a graph representation of the hierarchical structure of both item and person parameters across cohorts. As the identification constraint, the average difficulty of the items is assumed equal across cohorts. That is, cohorts may differ in mean and variance of the latent trait, and particular item parameters might be different across cohorts, but the *average* difficulty of items is the same (for example, in case of an IQ test for males and females: the assumption is that overall the test has the same difficulty, although it can be the case that some items are relatively more difficult for

males, and other items are relatively more difficult for females). In addition, to identify the variance of the scale the product of the discrimination parameters was fixed at 1. Allowing for such random fluctuations in difficulty and discrimination across cohorts is also referred to as the assumption of approximate measurement invariance. This Bayesian method was only applied to NEO-FFI and EPQ-R test items, as for those tests, the numbers of cohorts were sufficiently large. We randomly selected 1,000 individuals from each cohort (or all individuals if sample size was smaller) and determined which items showed considerable DIF across cohorts by computing Bayes factors (Verhagen and Fox 2013a, b). When testing invariance hypotheses, an advantage of the Bayes factor is that you can gather evidence in favor of the (null) hypothesis of invariance. A Bayes factor smaller than 0.3 was regarded as clear evidence of DIF. A Bayes factor larger than 3 was regarded as evidence of measurement invariance (i.e., no DIF). Taking into account possible DIF, all individuals with either NEO or EPQ data were mapped to a common scale for Neuroticism and Extraversion and mean Neuroticism and Extraversion scores and variances were estimated for each cohort.

Significant DIF does not imply that its effects are dramatic. To assess the extent to which DIF results in different scoring, depending on what calibration is used, Neuroticism and Extraversion scores were estimated using different cohort-specific calibrations and these were compared. For example, how much would the estimated scores for individuals in the Dutch NTR sample differ if instead of using the NTR calibration (i.e., using item parameters as estimated using NTR data), the Finnish HBCS calibration were used? If measurement invariance holds perfectly, the correlation between the different score estimates should be very close to 1. These correlations were computed for NEO-FFI, NEO-PI-R and EPQ inventories in the appropriate cohorts.

Meta-analysis of heritability

In each of the 6 cohorts with twin data separately, twin correlations for the IRT latent trait scores were estimated using the structural equation modeling package OpenMx within the statistical software program R (Boker et al. 2011). This was done by fitting a fully saturated model using full information likelihood to the data of twins in five sex-by-zygosity groups: monozygotic male twin pairs (MZM), dizygotic male twin pairs (DZM), monozygotic female twin pairs (MZM), dizygotic female twin pairs (DZM) and dizygotic twin pairs of opposite sex (DOS; if available in the particular cohort). Twin pairs in which Neuroticism and Extraversion scores were available for both twins were included, as well as twin pairs for which information was available for only one of the twins. In each cohort including a DOS group, 16 parameters were estimated: 5 means (5 sex by zygosity groups), 1 regression parameter for the effect of age on the means, 5 variances (5 sex by zygosity groups) and 5 covariances (for 5 sex by zygosity groups). In the cohorts without a DOS group, 4 means, 1 regression parameter for age, 4 variances and 4 covariances were estimated (13 parameters in total). The 4 or 5 covariances were standardized in each sex-by-zygosity group in order to obtain 4 or 5 twin correlations in each cohort. In addition, the 95 % confidence intervals for the twin correlations were computed. It was further tested whether the twin correlations could be constrained to be equal across sex (MZM = MZF and DZM = DZF = DOS).

Under the classical twin model assumptions, the expected MZ twin correlation is a function of the proportions of variance in a trait explained by additive (h^2) and non-additive (d^2) genetic effects: $r(MZ) = h^2 + d^2$. The expected DZ twin correlation is a different function of these two types of effects: $r(DZ) = \frac{1}{2}h^2 + \frac{1}{4}d^2$. IRT-score-based twin correlations (Table 1) were used as the basis to assess both qualitative and quantitative sex effects. This was done by fitting the same model to data from all six cohorts simultaneously allowing for different estimates of h^2 and d^2 in each sex, and allowing the opposite-sex twin correlation to be different from its expectation, $\frac{1}{2}h_m h_f + \frac{1}{4}d_m d_f$. The estimates of parameters (h^2, e^2 and d^2 by sex) thus were constrained to be the same across cohorts. First it was tested whether the correlation in opposite-sex twins could be equated to the expectation above (i.e. testing for qualitative sex effects). Next, it was tested whether the relative sizes of the genetic components could be equated across sexes, that is, whether $h_m^2 = h_f^2$ and $d_m^2 = d_f^2$. Lastly, it was tested whether non-additive genetic effects were present, by comparing the fit of the model with a model in which $d^2 = 0$.

Power study

For the NTR and the QIMR-adult cohorts, the increase in statistical power for a GWAS on Neuroticism was determined that results from the increase in sample size and measurement precision due to the IRT test linking. A baseline condition of using 12 NEO-FFI items as in a previous meta-analysis (De Moor et al. 2010) was compared with using all available data from NEO-PI-R and other available inventories. We assumed that genotype data was non-missing for all phenotypes. Power was computed for a single nucleotide polymorphism (SNP) explaining 0.1 % of true phenotypic variance (latent trait) with allele

Table 1 Twin correlations for the IRT-based Neuroticism and Extraversion scores

Cohort	Twin pairs	Trait	r_{MZ}	N	95 % CI	r_{DZ}	N	95 % CI
7. FINNISH TWINS	M–M	Neuroticism	0.43	1998	0.39–0.47	0.20	4862	0.16–0.23
		Extraversion	0.44	1999	0.40–0.48	0.14	4861	0.11–0.17
	F–F	Neuroticism	0.48	2226	0.45–0.52	0.19	4658	0.16–0.22
		Extraversion	0.52	2227	0.49–0.55	0.15	4663	0.12–0.18
	All	Neuroticism	0.46	4224	0.43–0.48	0.19	9520	0.17–0.21
		Extraversion	0.48	4226	0.46–0.51	0.14	9524	0.12–0.17
12. MCTFR	M–M	Neuroticism	0.53	922	0.47–0.60	0.17	506	0.05–0.28
		Extraversion	0.52	922	0.45–0.58	0.23	506	0.11–0.34
	F–F	Neuroticism	0.45	1054	0.38–0.52	0.26	580	0.15–0.37
		Extraversion	0.51	1054	0.45–0.57	0.13	580	0.02–0.25
	All	Neuroticism	0.48	1976	0.44–0.53	0.22	1086	0.14–0.30
		Extraversion	0.52	1976	0.47–0.56	0.17	1086	0.09–0.25
15. NTR	M–M	Neuroticism	0.45	1124	0.40–0.50	0.22	855	0.14–0.29
		Extraversion	0.47	1123	0.42–0.52	0.13	855	0.06–0.21
	F–F	Neuroticism	0.51	2249	0.47–0.54	0.23	1391	0.17–0.28
		Extraversion	0.49	2248	0.46–0.52	0.20	1392	0.14–0.26
	M–F	Neuroticism	–	–	–	0.21	2044	0.16–0.26
		Extraversion	–	–	–	0.14	2044	0.09–0.19
	All	Neuroticism	0.49	3373	0.46–0.52	0.22	4290	0.18–0.25
		Extraversion	0.48	3371	0.46–0.51	0.16	4291	0.13–0.19
18. QIMR adolescents	M–M	Neuroticism	0.51	304	0.42–0.59	0.27	252	0.15–0.38
		Extraversion	0.49	304	0.40–0.57	0.18	252	0.06–0.30
	F–F	Neuroticism	0.39	329	0.29–0.48	0.19	268	0.07–0.30
		Extraversion	0.45	329	0.36–0.53	0.19	268	0.07–0.31
	M–F	Neuroticism	–	–	–	0.21	463	0.13–0.30
		Extraversion	–	–	–	0.12	463	0.03–0.21
	All	Neuroticism	0.44	633	0.38–0.50	0.22	983	0.16–0.28
		Extraversion	0.47	633	0.40–0.53	0.16	983	0.09–0.22
19. QIMR adults	M–M	Neuroticism	0.45	1182	0.40–0.50	0.11	889	0.04–0.19
		Extraversion	0.48	1182	0.43–0.53	0.19	889	0.11–0.26
	F–F	Neuroticism	0.48	2075	0.45–0.52	0.22	1435	0.17–0.28
		Extraversion	0.48	2075	0.44–0.51	0.16	1435	0.11–0.21
	M–F	Neuroticism	–	–	–	0.13	1827	0.08–0.18
		Extraversion	–	–	–	0.14	1827	0.09–0.19
	All	Neuroticism	0.47	3257	0.44–0.50	0.16	4151	0.13–0.19
		Extraversion	0.48	3257	0.45–0.51	0.16	4151	0.12–0.19
21. STR	M–M	Neuroticism	0.54	3188	0.51–0.56	0.18	4841	0.15–0.21
		Extraversion	0.54	3188	0.51–0.56	0.25	4841	0.22–0.28
	F–F	Neuroticism	0.45	2830	0.42–0.49	0.16	4625	0.13–0.19
		Extraversion	0.44	2830	0.41–0.48	0.20	4625	0.17–0.23
	All	Neuroticism	0.51	6018	0.49–0.53	0.19	9466	0.17–0.21
		Extraversion	0.52	6018	0.50–0.54	0.26	9466	0.23–0.28

r_{MZ} correlation in monozygotic twin pairs, r_{DZ} correlation in dizygotic twin pairs, *N* number of twin pairs (pairs are included with personality data for both twins and with data for one twin), *95 % CI* 95 % confidence interval, *M–M* male–male twin pairs, *F–F* female–female twin pairs, *M–F* male–female twin pairs, *All* twin pairs combined across gender

frequency 0.5. Item data were simulated with parameter settings equal to the observed parameter estimates in the empirical data. Sample sizes were also the same as in the empirical data. For each power estimate, 100 data sets were simulated and analyzed, and the proportion of *p*-values smaller than 10^{-8} was calculated.

Table 2 Correlations between the IRT-based Neuroticism and Extraversion scores and the personality inventory-based sum scores

	Neuroticism		Extraversion	
Cohort	**N**	**r**	**N**	**r**
1. ALSPAC	6,068	0.98 (IPIP)	6,072	0.97 (IPIP)
2. BLSA	1,917	0.96 (NEO-PI-R)	1,917	0.97 (NEO-PI-R)
3. CILENTO	800	0.97 (NEO-PI-R)	800	0.98 (NEO-PI-R)
4. COGEND	2,712	0.98 (NEO-FFI)	2,712	0.98 (NEO-FFI)
5. EGCUT	1,730	0.98 (NEO-PI-3)	1,730	0.98 (NEO-PI-3)
6. ERF	2,474	0.93 (NEO-FFI)	2,479	0.87 (NEO-FFI)
7. FINNISH TWINS	30,073	0.96 (NEO-FFI)	30,120	0.94 (NEO-FFI)
		0.98 (EPI)		0.97 (EPI)
8. HBCS	1,698	0.91 (NEO-PI-R)	1,698	0.92 (NEO-PI-R)
		0.85 (TCI)		0.63 (TCI)
9. KORCULA	810	0.97 (EPQ)	809	0.79 (EPQ)
10. LBC1921	478	0.96 (IPIP)	478	0.98 (IPIP)
11. LBC1936	1,032	0.92 (NEO-FFI)	1,032	0.85 (NEO-FFI)
		0.92 (IPIP)		0.93 (IPIP)
12. MCTFR	9,063	0.97 (MPQ)	9,063	0.96 (MPQ)
13. NBS	1,818	0.96 (EPQ)	1,821	0.96 (EPQ)
14. NESDA	2,961	0.99 (NEO-FFI)	2,961	0.96 (NEO-FFI)
15. NTR	31,299	0.91 (NEO-FFI)	31,294	0.85 (NEO-FFI)
		0.89 (ABV)		0.86 (ABV)
16. ORCADES	602	0.98 (EPQ)	602	0.88 (EPQ)
17. PAGES	476	0.95 (NEO-PI-R)	476	0.93 (NEO-PI-R)
		0.73 (TCI)		0.60 (TCI)
18. QIMR-adolescents	4,100	0.93 (NEO-PI-R)	4,100	0.88 (NEO-PI-R)
		0.94 (NEO-FFI)		0.77 (NEO-FFI)
		0.86 (JEPQ)		0.81 (JEPQ)
19. QIMR-adults	26,681	0.94 (NEO-PI-R)	26,681	0.90 (NEO-PI-R)
		0.92 (NEO-FFI)		0.89 (NEO-FFI)
		0.86 (EPQ)		0.94 (EPQ)
		0.88 (TCI)		0.64 (TCI)
		0.87 (MPQ)		0.85 (MPQ)
20. SAGE-COGA	649	0.97 (TCI)	649	0.89 (TCI)
21. STR	30,264	0.96 (EPI)	30,253	0.97 (EPI)
22. VIS	909	0.98 (EPQ)	909	0.75 (EPQ)
23. YOUNG FINNS	2,057	0.97 (NEO-FFI)	2,057	0.96 (NEO-FFI)
TOTAL	160,671		160,713	

Results

Estimating Neuroticism and Extraversion scores

Personality scores were estimated for 160,671 (Neuroticism) and 160,713 individuals (Extraversion). Correlations between estimated latent scores and sum scores were high for Neuroticism (79 % of the correlations >0.90, and 50 % >0.95; lowest correlation 0.73) and moderately high for Extraversion (82 % of the correlations >0.80, and 48 % >0.90; lowest correlation 0.60) (Table 2). Correlations were highest with NEO, EPQ and IPIP-based sum scores, and lowest with TCI-based sum scores.

Appropriateness of Item Response Theory to harmonize Neuroticism and Extraversion scores

To assess whether test linking was successful within the seven cohorts that assessed more than one personality inventory, latent scores were computed based on different calibrations. In the majority of cohorts, the correlations among estimated scores were very high for most of the

inventories ($r > 0.96$). Only for TCI Neuroticism in the HBCS cohort, was the correlation lower ($r = 0.87$). Thus, the latent scores are largely independent of the inventories included. TICs for these cohorts are presented in Supplementary Figs. 4–27. Supplementary Figs. 11, 14, 18, and 20 thru 23 show that combining tests always leads to higher information content, and therefore more measurement precision for those individuals with multiple-inventory data. However, the TICs of the combined tests are not a simple sum of the TICs of the individual tests, showing that the personality inventories largely, but not completely, measure the same phenotypes.

To assess whether personality scores could be compared *across* cohorts, latent scores in each cohort were estimated several times based on different values for the item parameters coming from different cohorts (different calibrations). That is, a certain pattern of item responses was used to estimate the latent trait based on the item parameters as calibrated in one cohort, and this was repeated but then using item parameters as calibrated in another cohort. The correlations (see Supplementary Tables 4 and 5) are generally very high (most >0.95; only 3 out of the $84 < 0.90$, with the lowest correlation 0.81). Thus, ranking is not much affected by the particular cohort that individuals were in.

Figures 2 and 3 display item parameter values for the NEO-FFI and EPQ-R Neuroticism and Extraversion items for all cohorts in which these inventories were assessed. These parameters are based on a Bayesian hierarchical analysis (Verhagen and Fox 2013a, b) which takes into account any potential mean and variance differences across cohorts. All Bayes factors were smaller than 0.3. However, the item parameters were largely the same across cohorts for most items, with few striking differences. Item parameters tend to be more similar when cohorts have the same language. An example is NEO-FFI Neuroticism item 9 ('At times I have been so ashamed I just wanted to hide') for which the two Finnish cohorts show somewhat different item parameter values compared to the other cohorts. Examples from the NEO-FFI Extraversion scale are items 10 ('I don't consider myself especially "light-hearted" (R)) and 11 ('I am a cheerful, high-spirited person') that show differences across English speaking (red lines) and Dutch speaking cohorts (green lines). Similarly for the EPQ-R items, where item parameters for the Croatian cohorts (black lines) are very similar, as are the parameters for the English-speaking cohorts (green lines), with clear differences between the two language groups. This suggests some evidence for measurement variance across cohorts, which could be due to slightly different item content after translation.

Allowing for these significant deviations from measurement invariance across cohorts by applying the Bayesian model, Tables 3 and 4 show uncorrected means and variances per cohort, as measured by the NEO-FFI and EPQ-R items. Note that we included all cohorts with NEO data (NEO-PI-R or NEO-FFI), but using only the 12 items that are part of both the NEO-PI-R and the NEO-FFI. NESDA shows the highest mean Neuroticism score (which is expected given that it concerns a sample selected for depression and anxiety) and PAGES the lowest mean for NEO data. For NEO Extraversion, the QIMR adolescents show the highest mean (as expected based on their age), and CILENTO the lowest mean. Based on the EPQ data, the Croatian samples have the highest Neuroticism and Extraversion scores, and ORCADES the lowest. Some variance differences across cohorts are also observed, which can partly be explained by differences in age distribution, birth cohort and inclusion criteria. Note that for the NEO, the variances for Neuroticism are larger than for Extraversion, which is explained by the higher reliability of the Neuroticism scale. This is because in the hierarchical modeling, in order to identify scale, the product of the discrimination parameters was fixed at 1, both for Neuroticism and for Extraversion. Larger variance of the latent trait implies that in case the latent variance was fixed to a constant instead of the discrimination parameters, the discrimination parameters would be higher for Neuroticism than for Extraversion. As these discrimination parameters are used for computing test information (Lord 1980), and therefore reliability, we can conclude that Neuroticism is more reliably assessed than Extraversion.

Meta-analysis of heritability

MZ twin correlations for the estimated Neuroticism and Extraversion scores ranged between 0.39 and 0.54 (Table 1). DZ correlations were typically smaller than half the MZ correlations, suggesting non-additive genetic effects on variation in Neuroticism and Extraversion. Significant sex differences in same-sex twin correlations (p value <0.01) were found in the MCTFR, Finnish Twin and STR cohorts, but not in the NTR and two QIMR cohorts. The NTR and QIMR cohorts included opposite-sex twins. Table 1 shows that in the NTR and in the QIMR adolescent cohorts, the opposite-sex twin correlations are not significantly different from the same-sex DZ twin correlations, nor are the male same-sex DZ twin correlations different from the female same-sex DZ twin correlations. Only in the QIMR-adult cohort, there is some evidence of a larger same-sex DZ correlation for Neuroticism in females compared to males.

In the meta-analysis of the 27 twin correlations in Table 1, the base model for Neuroticism with 5 parameters (h_m, h_f, d_m, d_f, and one for allowing the opposite-sex twin correlation to differ from its expectation under the

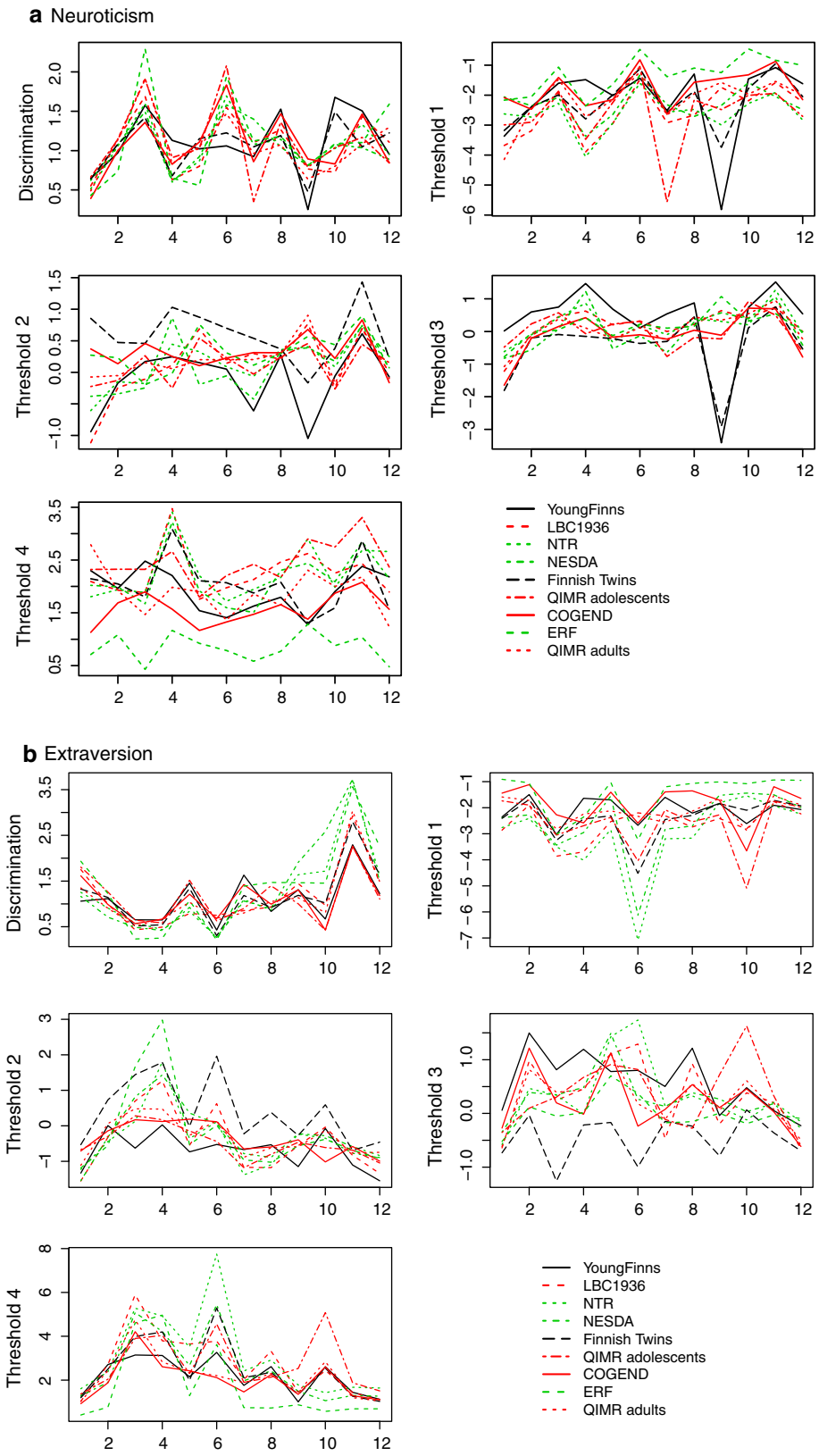

Fig. 2 Parameter estimates (thresholds and discrimination parameters) for 12 items (*x*-axis) from the NEO-FFI personality inventory for different cohorts, separately for Neuroticism and Extraversion. In *black*, the item parameter values for Finnish language cohorts, in *green* for Dutch language cohorts, and in *red* for English language cohorts (Color figure online)

hypothesis of no qualitative sex differences) did not show a better fit than one where the opposite-sex twin correlation was equated to its expected value (total N = 29,496 pairs). The base model χ^2 was 88.33, and the restricted model χ^2 was 88.89, a non-significant change with 1 degree of freedom. Next, this restricted model with qualitatively the

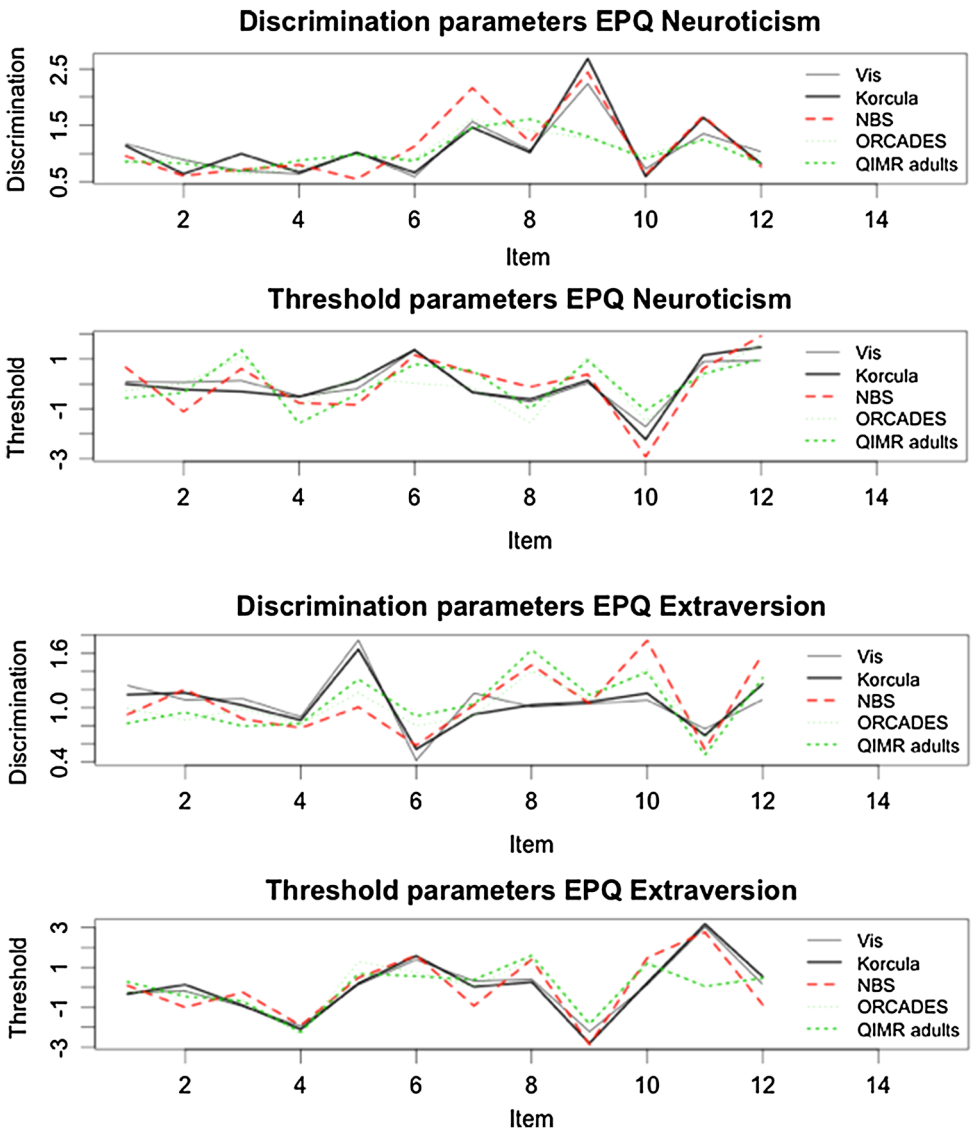

Fig. 3 Parameter estimates (thresholds and discrimination parameters) for 12 items (*x*-axis) from the EPQ-R personality inventory for different cohorts, separately for Neuroticism and Extraversion. In *black*, the item parameter values for Croatian cohorts, in *green* for English language cohorts, and in *red* for a Dutch language cohort (Color figure online)

same additive and non-additive genetic effects for males and females was compared with a model that specified that the proportions additive and non-additive genetic variance were equal across sexes. This model had a χ^2 statistic of 91.63, a non-significant increase of the χ^2 statistic by 2.74 for 2 degrees of freedom. Next, it was tested whether the non-additive genetic effects could be dropped from the model. The χ^2 statistic increased to 170.39, which is highly significant. Thus, for Neuroticism, both additive and non-additive genetic effects seem to be operating, which seem to be the same in males and females, and of equal importance in males and females. Proportions of additive and non-additive genetic variance were estimated at 27 and 21 %, respectively.

For Extraversion (total N = 29,501 pairs), the base model had a χ^2 of 97.15. Restricting the opposite-sex twin correlation led to a χ^2 of 104.67, a difference of 7.54, which is significant at one degree of freedom. We therefore allowed for qualitative sex difference when testing for quantitative sex differences (equating h_m to h_f, and d_m, to d_f,). This restriction led to a χ^2 of 101.20, a non-significant change of 4.06 at 2 degrees of freedom, $p = 0.13$. Thus, there seem to be only qualitative differences in genetic variance components. Dropping non-additive genetic variance from the model resulted in a significantly higher χ^2 statistic, of 194.60, a difference of 93.40.

Thus, for Extraversion, there are qualitative sex differences in the additive and non-additive genetic effects, but the additive and non-additive genetic effects are of equal magnitude in males and females: 24 % and 25 %, respectively. The χ^2 statistic for these qualitative sex differences was relatively small given the large sample size, but nevertheless, the opposite sex twin correlation was a factor 0.76 smaller than expected under no qualitative differences (i.e., 0.14 instead of 0.18).

Table 3 Estimated means and variances of IRT-based Neuroticism and Extraversion latent scores based on NEO-FFI item data, after taking into account measurement non-invariance across cohorts

Cohort	Neuroticism		Extraversion	
	Mean (SE)	Variance	Mean (SE)	Variance
2. BLSA	−0.93 (0.04)	0.93	0.50 (0.03)	0.56
3. CILENTO	−0.14 (0.03)	0.43	−0.15 (0.04)	0.25
4. COGEND	−0.45 (0.03)	0.69	0.40 (0.03)	0.39
5. ERF	−0.28 (0.02)	0.38	0.06 (0.03)	0.23
6. EGCUT	−0.16 (0.03)	0.37	0.04 (0.04)	0.11
7. FINNISH TWINS	−0.41 (0.04)	0.74	0.34 (0.03)	0.41
8. HBCS	−0.59 (0.04)	0.65	0.13 (0.06)	0.37
11. LBC1936	−0.77 (0.04)	1.10	0.25 (0.03)	0.50
14. NESDA	0.05 (0.04)	1.12	0.03 (0.03)	0.62
15. NTR	−0.69 (0.04)	0.88	0.57 (0.03)	0.55
17. PAGES	−1.02 (0.05)	0.74	0.28 (0.07)	0.50
18. QIMR adolescents	−0.11 (0.03)	0.60	0.68 (0.03)	0.49
19. QIMR adults	−0.43 (0.03)	0.81	0.36 (0.03)	0.40
23. YOUNG FINNS	−0.73 (0.04)	1.24	0.50 (0.03)	0.61
Overall average	−0.47 (0.09)	0.12[a]	0.28 (0.07)	0.07[a]

[a] Between cohort variance

Table 4 Estimated means and variances of IRT-based Neuroticism and Extraversion latent scores based on EPQ-R item data, after taking into account measurement non-invariance across cohorts

Cohort	Neuroticism		Extraversion	
	Mean (SE)	Variance	Mean (SE)	Variance
9. Korcula	−0.55 (0.06)	2.28	1.41 (0.07)	2.10
13. NBS	−1.33 (0.07)	2.94	0.60 (0.07)	3.52
16. ORCADES	−1.47 (0.08)	2.56	0.36 (0.08)	3.10
19. QIMR adults	−0.72 (0.06)	2.35	0.76 (0.07)	4.12
22. VIS	−0.33 (0.06)	2.22	1.10 (0.06)	2.02
Overall average	−0.83 (0.23)	0.30[a]	0.82 (0.21)	0.23[a]

[a] Between cohort variance

Power study

For the NTR cohort, the statistical power to detect a SNP at the genome-wide significance level that explains 0.1 % of the true phenotypic variance (latent trait) with an allele frequency of 0.5 when using only the 12 NEO-FFI items was 18 % (N = 5,299 individuals with NEO-FFI data on Neuroticism) and increased to 44 % when using IRT scores based on both NEO-FFI and ABV data (N = 31,309 individuals with either NEO-FFI data, ABV data or both). In the QIMR-adult sample, the power with only 12 NEO-FFI items was 0 % (N = 3,712). Using all available data from all inventories and analyzing IRT scores yielded a power of 30 % (N = 26,692). Thus, the power in GWAS substantially increases if item data from multiple inventories are included, if available.

Discussion

This study examined for Neuroticism and Extraversion personality traits whether measures from different inventories could be harmonized using IRT test linking. The IRT analyses showed that the linked scores for Neuroticism and Extraversion that were estimated in >160,000 individuals from 23 cohorts were largely independent of the particular inventory. The success of this approach is demonstrated by the power study that showed a clear increase in statistical power to find a genetic variant associated with personality that is mainly the result of an increase in sample size.

The NEO, Eysenck and IPIP inventories were especially conducive to being linked. Linking was slightly less successful for TCI and MPQ with the NEO, Eysenck and IPIP inventories. The mapping of Harm Avoidance onto Neuroticism, despite theoretical differences between the concepts, was found to be relatively good. However, the mapping of Reward Dependence to Extraversion was less feasible, as was suspected. Such imperfect linking results in bias when individuals are ranked, which is very important in for example educational settings (e.g. pass/fail decisions on a test or determining the final class rank). However, when scientific interest is in population effects, like a correlation in twins or between the phenotype and a SNP, results are highly satisfactory. When dealing with non-identical but correlated traits, an alternative could be the use of multidimensional IRT models (van den Berg and Service 2012), because such models allow for relatively low correlations between multiple latent construct, but still enable borrowing statistical information from the respective sets of items, which leads to more precise estimation of latent scores.

Across cohorts, personality scores were largely comparable; that is, the extent of measurement variance was overall not large. We did, however, observe measurement variance for a few cohorts and for some items. Differences in item parameters across cohorts seem largest in cohorts with different spoken languages, suggesting cultural and/or language effects on some of the items. As a consequence, the estimated latent scores across cohorts are not based on completely identical scales. Again, for individual scoring this has consequences (e.g., a person's ranking within a population), but these imperfections have little effect on results for population effects, because the correlation of two scores based on different calibrations was generally very high. Overall, the conclusion is that data pooling within cohorts and subsequently pooling results across cohorts in a meta-analysis is meaningful for Neuroticism and Extraversion and these inventories. As the power study showed, such pooling of data within cohorts can lead to a potentially large increase in statistical power. Such increase in power is largely due to the increase in sample size, but also of using more phenotypic information per individual.

Note that IRT test linking is always *possible*: the only requirement is that there is either overlap in individuals that were administered several inventories, or overlap in items, when some items are present in multiple inventories. It remains however to be determined whether the linking leads to psychometrically sound re-scaled phenotypes in order to for the test linking to be meaningful and successful.

Based on six cohorts with twin data, the meta-analysis broad-sense heritability was 48 % for Neuroticism and 49 % for Extraversion (total N = 29,496 and 29,501 twin pairs, respectively). There was clear evidence of non-additive genetic variance for both traits. Although this finding could be partly due to a scale effect (the test information curves are slightly skewed, so therefore the distributions of sum scores *and* IRT score estimates are skewed as well, see (van den Berg et al. 2007; van den Berg and Service 2012), the relatively large size of the dominance genetic variance component suggests there is truly non-additive gene action. Sex differences in the kind of, and the relative size of, genetic factors on Neuroticism and Extraversion were suggested in only a subset of cohorts. The meta-analysis showed that qualitative sex effects were only significant for Extraversion. Proportions of additive and non-additive genetic variance were not significantly different across sexes.

We reported high correlations among the IRT-based scores and the sum scores for the specific personality inventories. One may argue that sum scores can serve just as well in analyses. There are several reasons however why the IRT approach is superior. First, the IRT approach leads to less biased estimates for Neuroticism and Extraversion if not exactly the same set of items is administered to all individuals, as was often the case in the cohorts because of missing data or because of assessing multiple inventories or versions. In addition, the IRT approach results in increased measurement precision for individuals who have been assessed using multiple inventories. Without fitting an IRT model, it is not clear how to weigh items from different inventories. Moreover, by using IRT, groups of individuals within cohorts with different item sets can be compared since all individuals are scored on one common metric, once linking is possible. Lastly, and most importantly, the IRT approach enables one to make explicit the extent to which item data from multiple inventories can be combined, both within and across cohorts. When simply using sum scores for different inventories separately and pooling results, it remains unknown whether this is actually appropriate.

When estimating latent trait scores, we preferred linking inventories within cohorts, but not across cohorts. Arguably, linking across cohorts would be even better, scaling all individuals from all cohorts to one common metric. Although theoretically possible, it can be infeasible in practice. In our study, it would require analyzing hundreds of items in over 160,000 individuals in one analysis, which is computationally infeasible. This approach would also only be possible if all inventories could in fact be linked to one another. In our study, this was not the case; for instance, different versions with different answer categories of the same inventory were used in different cohorts.

Limitations of the current study are that we did not include all items in cases of repeated measures, item data were assumed to be missing at random (Little and Rubin 1989), and we preselected items to belong to Neuroticism or Extraversion, rather than making this choice data-driven. Future extensions of the IRT linking approach may address these issues. Also note that our method deals with harmonization of continuously distributed data. Generally, harmonization of case–control status requires a different approach, but in cases where diagnosis is based on cut-off scores on continuous measures (e.g. a symptom count), the application of IRT models could prove helpful; IRT models are also used to compare pass/fail decisions in educational measurement where students are differentially assessed.

To conclude, the IRT results show that the Neuroticism and Extraversion item data from different inventories in different cohorts can be harmonized (for general recommendations and an example R analysis script, see Supplementary Materials Online). The harmonized phenotypes can now also be confidently correlated with brain measures or used in a GWA study. The IRT analysis is not only useful for harmonizing phenotypes, it is also informative regarding the power to find significant genetic variants of various allele frequencies. The TICs show where in the distribution of Neuroticism and Extraversion scores there is most phenotypic information. Relating these TICs to the power a phenotypic test might give in a GWAS (van den

Berg and Service 2012), we conclude that there generally is more power to detect low frequency genetic variants associated with scoring at the low end of the Extraversion distribution than towards the high end of the distribution. Similarly, there is more power to detect low-frequency genetic variants associated with scoring above-average on Neuroticism, compared to scoring below-average. Overall, the phenotypic information content is higher for Neuroticism than for Extraversion in most cohorts, suggesting more power to find loci for Neuroticism than for Extraversion. Combined with the finding of more additive genetic variance in Neuroticism than in Extraversion, we expect that Neuroticism loci will be easier to find than Extraversion loci.

Acknowledgments Phenotype harmonization was financially supported by the European Network of Genomic and Genetic Epidemiology (ENGAGE).

ALSPAC thanks all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The UK Medical Research Council and the Wellcome Trust (Grant ref: 092731) and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors and DME, KN and GDS will serve as guarantors for the contents of this paper.

The BLSA was supported by the Intramural Research Program of the National Institutes of Health, National Institute on Aging.

Cilento give special thanks to the Cilento populations for their participation in the study, acknowledge Dr. Maria Enza Amendola for the test administration and thank the personnel working in the organization of the study in the villages. This work was supported by grants from the Fondazione CON IL SUD (2011-PDR-13) and the Fondazione Banco di Napoli to MC.

Funding support for the Study of Addiction: Genetics and Environment (SAGE) was provided through the NIH Genes, Environment and Health Initiative [GEI] (U01 HG004422). SAGE is one of the GWA studies funded as part of the Gene Environment Association Studies (GENEVA) under GEI. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01 HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Support for collection of datasets and samples was provided by the Collaborative Study on the Genetics of Alcoholism (COGA; U10 AA008401) and the Collaborative Genetic Study of Nicotine Dependence (COGEND; P01 CA089392).

The Collaborative Study on the Genetics of Alcoholism (COGA), Principal Investigators B. Porjesz, V. Hesselbrock, H. Edenberg, L. Bierut, includes ten different centers: University of Connecticut (V.Hesselbrock); Indiana University (H.J. Edenberg, J. Nurnberger Jr., T. Foroud); University of Iowa (S. Kuperman, J. Kramer); SUNY Downstate (B. Porjesz); Washington University in St. Louis (L.Bierut, A. Goate, J. Rice, K. Bucholz); University of California at San Diego (M. Schuckit); Rutgers University (J. Tischfield); Texas Biomedical Research Institute (L. Almasy), Howard University (R.Taylor) and Virginia Commonwealth University (D. Dick). Other COGA collaborators include: L. Bauer (University of Connecticut); D. Koller, S. O'Connor, L. Wetherill, X. Xuei (Indiana University); Grace Chan (University of Iowa); S. Kang, N. Manz, M. Rangaswamy (SUNY Downstate); J. Rohrbaugh, J-C Wang (Washington University in St. Louis); A. Brooks (Rutgers University); and F. Aliev (Virginia Commonwealth University). A. Parsian and M. Reilly are the NIAAA Staff Collaborators. This national collaborative study is supported by NIH Grant U10AA008401 from the National Institute on Alcohol Abuse and Alcoholism (NIAAA) and the National Institute on Drug Abuse (NIDA).

The Collaborative Genetic Study of Nicotine Dependence (COGEND) project is a collaborative research group and part of the NIDA Genetics Consortium. Subject collection was supported by NIH grant P01 CA089392 (L.J. Bierut) from the National Cancer Institute. Phenotypic and genotypic data are stored in the NIDA Center for Genetic Studies (NCGS) at http://zork.wustl.edu/under NIDA Contract HHSN271200477451C (J. Tischfield and J. Rice).

EGCUT received targeted financing from Estonian Government SF0180142s08, Center of Excellence in Genomics (EXCEGEN) and University of Tartu (SP1GVARENG). We acknowledge EGCUT technical personnel.

The ERF study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)/grant agreement HEALTH-F4-2007-201413 by the European Commission under the programme "Quality of Life and Management of the Living Resources" of 5th Framework Programme (no. QLG2-CT-2002-01254). We are grateful to all study participants and their relatives, general practitioners and neurologists for their contributions and to P. Veraart for her help in genealogy, J. Vergeer for the supervision of the laboratory work and P. Snijders for his help in data collection.

The Finnish Twin Cohort acknowledges NIH grants DA12854 to P A F Madden; AA-12502, AA-00145, and AA-09203 to R J Rose; AA15416 to D M Dick; and Academy of Finland grants 118555, 141054, 265240, 263278, and 264146 to J Kaprio.

HBCS thanks all study participants as well as everybody involved in the Helsinki Birth Cohort Study. Helsinki Birth Cohort Study has been supported by grants from the Academy of Finland, the Finnish Diabetes Research Society, Folkhälsan Research Foundation, Novo Nordisk Foundation, Finska Läkaresällskapet, Signe and Ane Gyllenberg Foundation, University of Helsinki, Ministry of Education, Ahokas Foundation, Emil Aaltonen Foundation.

The CROATIA-Korcula study was funded by grants from the Medical Research Council (UK), European Commission Framework 6 project EUROSPAN (Contract No. LSHG-CT-2006-018947) and Republic of Croatia Ministry of Science, Education and Sports research grants to I.R. (108-1080315-0302). We would like to acknowledge the invaluable contributions of the recruitment team in Korcula, the administrative teams in Croatia and Edinburgh and the people of Korcula.

LBC1921 and LBC1936 phenotype collection was carried out within the Centre for Cognitive Ageing and Cognitive Epidemiology funded by a Lifelong Health and Wellbeing initiative (G0700704/84698). LBC1921 also acknowledge financial support from the Scottish Executive Chief Scientist Office (CZG/3/2/79) and BBSRC (15/SAG09977). LBC1936 also acknowledge financial support from Research into Ageing (Grant No. 251) and Age UK (Disconnected Mind grant).

MCTFR acknowledges support by the National Institutes of Health under award numbers R37DA005147, R01AA009367, R01AA011886, R01DA013240, and R01MH066140.

Principal investigators of the Nijmegen Biomedical Study (NBS) are L.A.L.M. Kiemeney, M. den Heijer, A.L.M. Verbeek, D.W. Swinkels and B. Franke.

NESDA acknowledges financial support from the Geestkracht program of the Netherlands Organisation for Health Research and Development (ZonMW, grant number 10-000-1002) and participating institutes (VU University Medical Center, GGZ inGeest, Leiden

University Medical Center, GGZ Rivierduinen, University Medical Center Groningen, Lentis, GGZ Friesland, GGZ Drenthe, Netherlands Institute of Mental Health and Addiction).

NTR acknowledges financial support from the Netherlands Organization for Scientific Research (NWO) Grants No. 575-25-006, 480-04-004, 904-61-090; 904-61-193, 400-05-717, 311-60008 and Spinozapremie SPI 56-464-14192 and the European Research Council (ERC 230374). MHMdeM is financially supported by NWO VENI Grant No. 016-115-035.

ORCADES was supported by the Chief Scientist Office of the Scottish Government, the Royal Society, the MRC Human Genetics Unit, Arthritis Research UK and the European Union framework program 6 EUROSPAN project (contract no. LSHG-CT-2006-018947). DNA extractions were performed at the Wellcome Trust Clinical Research Facility in Edinburgh. We would like to acknowledge the research nurses in Orkney, the administrative team in Edinburgh and the people of Orkney.

QIMR adolescents We acknowledge financial support from the Australian Research Council (A79600334, A79906588, A79801419, DP0212016, DP0343921, DP0664638, DP1093900), Beyond Blue, and the Borderline Personality Disorder Research Foundation.

QIMR adults We acknowledge financial support from NIH (DA12854, AA07728, AA10248, AA07580, AA11998, AA13320, AA13321, AA13326, DA019951, AA014041, AA07535, MH66206, AGO4954 and GM30250), the Australian National Health and Medical Research Council, Gemini Genomics Plc, the Borderline Personality Disorder Research Foundation, the Australian Associated Brewers, ADAMHA (AA06781 and MH40828), and the American Cancer Society (IRG-58-010-50).

Funding support for the Study of Addiction Genetics and Environment (SAGE) was provided through the NIH Genes, Environment and Health Initiative [GEI] (U01 HG004422). SAGE is one of the GWA studies funded as part of the Gene Environment Association Studies (GENEVA) under GEI. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01 HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Support for collection of datasets and samples was provided by the Collaborative Study on the Genetics of Alcoholism (COGA; U10 AA008401) and the Collaborative Genetic Study of Nicotine Dependence (COGEND; P01 CA089392).

The STR is financially supported by the Swedish Ministry for Higher Education For the various projects financial support has been provided by: TwinGene; the Swedish Research Council (M-2005-1112), GenomEUtwin (EU/QLRT-2001-01254; QLG2-CT-2002-01254), NIH DK U01-066134, The Swedish Foundation for Strategic Research (SSF), the Heart and Lung foundation no. 20070481. STOPPA; the Strategic Research Program in Epidemiology at Karolinska Institutet, the Swedish Research Council (grant number 2011-3060), the Swedish Asthma and Allergy Association and the regional agreement on medical training and clinical research (ALF) between Stockholm County Council and Karolinska Institutet. CATSS; support was provided by the Swedish Council for Working Life and Social Research, the Swedish Research Council, Systembolaget, the National Board of Forensic Medicine, the Swedish Prison and Probation Service, Bank of Sweden Tercentenary Foundation, the Söderström–Königska foundation, and the Karolinska Institutet Center of Neurodevelopmental Disorders (KIND). BIRTH; supported by the Swedish Council for Working Life and Social Research (2004-0174 and 2007-0231), the Swedish Research Council (K2006-71X-14676-04-2 and 2008-54 × 20638-01-3), the Swedish Cancer Society (4594-B01-01XAC and 4594-B04-04XAB), and the European Union-funded Network of Excellence Lifespan (FP6036894).

The CROATIA-Vis study was funded by grants from the Medical Research Council (UK) and Republic of Croatia Ministry of Science, Education and Sports research grants to I.R. (108-1080315-0302). We would like to acknowledge the staff of several institutions in Croatia that supported the fieldwork, including but not limited to The University of Split and Zagreb Medical Schools, the Institute for Anthropological Research in Zagreb and Croatian Institute for Public Health.

The Cardiovascular Risk in Young Finns Study acknowledges financial support from the Academy of Finland (grants 126925, 121584, 124282, 129378 [Salve], 117787 [Gendi], 265869 (MIND), 258711 and 41071 [Skidi]); the Social Insurance Institution of Finland; the Kuopio, Tampere, and Turku University Hospital Medical Funds (grant 9M048 and 9N035 for Dr. Lehtimäki); the Juho Vainio Foundation; the Paavo Nurmi Foundation; Signe and Ane Gyllenberg's Foundation, the Finnish Foundation of Cardiovascular Research; the Finnish Cultural Foundation; as well as the Tampere Tuberculosis Foundation and the Emil Aaltonen Foundation (Dr. Lehtimäki). The expert technical assistance in data management and statistical analyses by Irina Lisinen and Ville Aalto is gratefully acknowledged.

Conflict of Interest Paul T. Costa receives royalties from the NEO inventories.

Human and Animal Rights and Informed Consent The procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2000 and 2008. All participants provided informed consent.

References

Aluja A, Garcia O, Garcia LF (2004) Replicability of the three, four and five Zuckerman's personality super-factors: exploratory and confirmatory factor analysis of the EPQ-RS, ZKPQ and NEO-PI-R. Personal Individ Differ 36(5):1093–1108

Berndt SI, Gustafsson S, Magi R, Ganna A, Wheeler E, Feitosa MF, Justice AE, Monda KL, Croteau-Chonka DC, Day FR, Esko T, Fall T, Ferreira T, Gentilini D, Jackson AU, Luan JA, Randall JC, Vedantam S, Willer CJ, Winkler TW, Wood AR, Workalemahu T, Hu YJ, Lee SH, Liang LM, Lin DY, Min JL, Neale BM, Thorleifsson G, Yang J, Albrecht E, Amin N, Bragg-Gresham JL, Cadby G, den Heijer M, Eklund N, Fischer K, Goel A, Hottenga JJ, Huffman JE, Jarick I, Johansson A, Johnson T, Kanoni S, Kleber ME, Konig IR, Kristiansson K, Kutalik Z, Lamina C, Lecoeur C, Li G, Mangino M, McArdle WL, Medina-Gomez C, Muller-Nurasyid M, Ngwa JS, Nolte IM, Lavinia P, Pechlivanis S, Perola M, Peters MJ, Preuss M, Rose LM, Shi JX, Shungin D, Smith AV, Strawbridge RJ, Surakka I, Teumer A, Trip MD, Tyrer J, Van Vliet-Ostaptchouk JV, Vandenput L, Waite LL, Zhao JH, Absher D, Asselbergs FW, Atalay M, Attwood AP, Balmforth AJ, Basart H, Beilby J, Bonnycastle LL, Brambilla P, Bruinenberg M, Campbell H, Chasman DI, Chines PS, Collins FS, Connell JM, Cookson WO, de Faire U, de Vegt F, Dei M, Dimitriou M, Edkins S, Estrada K, Evans DM, Farrall M, Ferrario MM, Ferrieres J, Franke L, Frau F, Gejman PV, Grallert H, Gronberg H, Gudnason V, Hall AS, Hall P, Hartikainen AL, Hayward C, Heard-Costa NL, Heath AC, Hebebrand J, Homuth G, Hu FB, Hunt SE, Hypponen E, Iribarren C, Jacobs KB, Jansson JO, Jula A, Kahonen M, Kathiresan S, Kee F, Khaw KT, Kivimaki M, Koenig W, Kraja

AT, Kumari M, Kuulasmaa K, Kuusisto J, Laitinen JH, Lakka TA, Langenberg C, Launer LJ, Lind L, Lindstrom J, Liu JJ, Liuzzi A, Lokki ML, Lorentzon M, Madden PA, Magnusson PK, Manunta P, Marek D, Marz W, Leach IM, McKnight B, Medland SE, Mihailov E, Milani L, Montgomery GW, Mooser V, Muhleisen TW, Munroe PB, Musk AW, Narisu N, Navis G, Nicholson G, Nohr EA, Ong KK, Oostra BA, Palmer CNA, Palotie A, Peden JF, Pedersen N, Peters A, Polasek O, Pouta A, Pramstaller PP, Prokopenko I, Putter C, Radhakrishnan A, Raitakari O, Rendon A, Rivadeneira F, Rudan I, Saaristo TE, Sambrook JG, Sanders AR, Sanna S, Saramies J, Schipf S, Schreiber S, Schunkert H, Shin SY, Signorini S, Sinisalo J, Skrobek B, Soranzo N, Stancakova A, Stark K, Stephens JC, Stirrups K, Stolk RP, Stumvoll M, Swift AJ, Theodoraki EV, Thorand B, Tregouet DA, Tremoli E, Van der Klauw MM, van Meurs JBJ, Vermeulen SH, Viikari J, Virtamo J, Vitart V, Waeber G, Wang ZM, Widen E, Wild SH, Willemsen G, Winkelmann BR, Witteman JCM, Wolffenbuttel BHR, Wong A, Wright AF, Zillikens MC, Amouyel P, Boehm BO, Boerwinkle E, Boomsma DI, Caulfield MJ, Chanock SJ, Cupples LA, Cusi D, Dedoussis GV, Erdmann J, Eriksson JG, Franks PW, Froguel P, Gieger C, Gyllensten U, Hamsten A, Harris TB, Hengstenberg C, Hicks AA, Hingorani A, Hinney A, Hofman A, Hovingh KG, Hveem K, Illig T, Jarvelin MR, Jockel KH, Keinanen-Kiukaanniemi SM, Kiemeney LA, Kuh D, Laakso M, Lehtimaki T, Levinson DF, Martin NG, Metspalu A, Morris AD, Nieminen MS, Njolstad I, Ohlsson C, Oldehinkel AJ, Ouwehand WH, Palmer LJ, Penninx B, Power C, Province MA, Psaty BM, Qi L, Rauramaa R, Ridker PM, Ripatti S, Salomaa V, Samani NJ, Snieder H, Sorensen TIA, Spector TD, Stefansson K, Tonjes A, Tuomilehto J, Uitterlinden AG, Uusitupa M, van der Harst P, Vollenweider P, Wallaschofski H, Wareham NJ, Watkins H, Wichmann HE, Wilson JF, Abecasis GR, Assimes TL, Barroso I, Boehnke M, Borecki IB, Deloukas P, Fox CS, Frayling T, Groop LC, Haritunian T, Heid IM, Hunter D, Kaplan RC, Karpe F, Moffatt MF, Mohlke KL, O'Connell JR, Pawitan Y, Schadt EE, Schlessinger D, Steinthorsdottir V, Strachan DP, Thorsteinsdottir U, van Duijn CM, Visscher PM, Di Blasio AM, Hirschhorn JN, Lindgren CM, Morris AP, Meyre D, Scherag A, McCarthy MI, Speliotes EK, North KE, Loos RJF, Ingelsson E (2013) Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. Nat Genetics 45(5):501–569

Boker S, Neale M, Maes H, Wilde M, Spiegel M, Brick T, Spies J, Estabrook R, Kenny S, Bates T, Mehta P, Fox J (2011) OpenMx: an open source extended structural equation modeling framework. Psychometrika 76(2):306–317

Chernyshenko OS, Stark S, Chan KY, Drasgow F, Williams B (2001) Fitting item response theory models to two personality inventories: issues and insights. Multivar Behav Res 36(4):523–562

Craddock N, O'Donovan MC, Owen MJ (2008) Genome-wide association studies in psychiatry: lessons from early studies of non-psychiatric and psychiatric phenotypes. Mol Psychiatr 13(7):649–653

De Fruyt F, Van de Wiele L, Van Heeringen C (2000) Cloninger's psychobiological model of temperament and character and the five-factor model of personality. Personal Individ Differ 29(3):441–452

De Moor MHM, Costa PT, Terracciano A, Krueger RF, de Geus EJC, Tanaka T, Penninx BWJH, Esko T, Madden PAF, Derringer J, Amin N, Willemsen G, Hottenga JJ, Distel MA, Uda M, Sanna S, Spinhoven P, Hartman CA, Sullivan P, Realo A, Allik J, Heath AC, Pergadia ML, Agrawal A, Lin P, Grucza RA, Widen E, Cousminer DL, Eriksson JG, Palotie A, Peltonen L, Luciano M, Tenesa A, Davies G, Houlihan LM, Hansell NK, Medland SE, Ferrucci L, Schlessinger D, Montgomery GW, Wright MJ, Aulchenko YS, Janssens ACJW, Oostra BA, Metspalu A, Abecasis GR, Deary IJ, Raikkonen K, Bierut LJ, Martin NG, van Duijn CM, Boomsma DI (2010) Meta-analysis of genome-wide association studies for personality. Mol Psychiatr 17(3):337–349

Dick DM, Aliev F, Latendresse SJ, Hickman M, Heron J, Macleod J, Joinson C, Maughan B, Lewis G, Kendler KS (2013) Adolescent Alcohol Use is Predicted by Childhood Temperament Factors Before Age 5, with Mediation Through Personality and Peers. Alcohol Clin Exp Res 37(12):2108–2117

Distel MA, Trull TJ, Willemsen G, Vink JM, Derom CA, Lynskey MT, Martin NG, Boomsma DI (2009) The Five Factor Model of personality and borderline personality disorder: a genetic analysis of comorbidity. Biol Psychiat 66:1131–1138

Draycott SG, Kline P (1995) The Big-3 Or the Big-5-the Epq-R vs the Neo-Pi-a research note, replication and elaboration. Personal Individ Differ 18(6):801–804

Gillespie NA, Johnstone SJ, Boyce P, Heath AC, Martin NG (2001) The genetic and environmental relationship between the interpersonal sensitivity measure (IPSM) and the personality dimensions of Eysenck and Cloninger. Personal Individ Differ 31(7):1039–1051

Glas CAW (1998) Detection of differential item functioning using Lagrange multiplier tests. Stat Sin 8(3):647–667

Glas CAW (2001) Differential item functioning depending on general covariates. Lect Notes Stat 157:131–148

Heath AC, Cloninger CR, Martin NG (1994) Testing a model for the genetic-structure of personality: a comparison of the personality Systems of Cloninger and Eysenck. J Pers Soc Psychol 66(4):762–775

Hedges LV, Vevea JL (1998) Fixed- and random-effects models in meta-analysis. Psychol Methods 3(4):486–504

Hopwood CJ, Wright AGC, Donnellan MB (2011) Evaluating the evidence for the general factor of personality across multiple inventories. J Res Pers 45(5):468–478

Keller M, Coventry W, Heath A, Martin N (2005) Widespread evidence for non-additive genetic variation in Cloninger's and Eysenck's Personality dimensions using a twin plus sibling design. Behav Genet 35(6):707–721

Kendler KS, Myers J (2009) The genetic and environmental relationship between major depression and the five-factor model of personality. Psychol Med 40(5):1–6

Klein DN, Kotov R, Bufferd SJ (2011) Personality and depression: explanatory models and review of the evidence. Annu Rev 7:269–295

Lango Allen H et al (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467(7317):832–838

Larstone RM, Jang KL, Livesley WJ, Vernon PA, Wolf H (2002) The relationship between Eysenck's P-E-N model of personality, the five-factor model of personality, and traits delineating personality dysfunction. Personal Individ Differ 33(1):25–37

Lee SH, Ripke S, Neale BM, Faraone SV, Purcell SM, Perlis RH, Mowry BJ, Thapar A, Goddard ME, Witte JS, Absher D, Agartz I, Akil H, Amin F, Andreassen OA, Anjorin A, Anney R, Anttila V, Arking DE, Asherson P, Azevedo MH, Backlund L, Badner JA, Bailey AJ, Banaschewski T, Barchas JD, Barnes MR, Barrett TB, Bass N, Battaglia A, Bauer M, Bayes M, Bellivier F, Bergen SE, Berrettini W, Betancur C, Bettecken T, Biederman J, Binder EB, Black DW, Blackwood DHR, Bloss CS, Boehnke M, Boomsma DI, Breen G, Breuer R, Bruggeman R, Cormican P, Buccola NG, Buitelaar JK, Bunney WE, Buxbaum JD, Byerley WF, Byrne EM, Caesar S, Cahn W, Cantor RM, Casas M, Chakravarti A, Chambert K, Choudhury K, Cichon S, Cloninger CR, Collier DA, Cook EH, Coon H, Cormand B, Corvin A, Coryell WH, Craig DW, Craig IW, Crosbie J, Cuccaro ML, Curtis D, Czamara D, Datta S, Dawson G, Day R, De Geus EJ, Degenhardt F, Djurovic S, Donohoe GJ, Doyle AE, Duan JB, Dudbridge F, Duketis E, Ebstein RP, Edenberg HJ,

Elia J, Ennis S, Etain B, Fanous A, Farmer AE, Ferrier IN, Flickinger M, Fombonne E, Foroud T, Frank J, Franke B, Fraser C, Freedman R, Freimer NB, Freitag CM, Friedl M, Frisen L, Gallagher L, Gejman PV, Georgieva L, Gershon ES, Geschwind DH, Giegling I, Gill M, Gordon SD, Gordon-Smith K, Green EK, Greenwood TA, Grice DE, Gross M, Grozeva D, Guan WH, Gurling H, De Haan L, Haines JL, Hakonarson H, Hallmayer J, Hamilton SP, Hamshere ML, Hansen TF, Hartmann AM, Hautzinger M, Heath AC, Henders AK, Herms S, Hickie IB, Hipolito M, Hoefels S, Holmans PA, Holsboer F, Hoogendijk WJ, Hottenga JJ, Hultman CM, Hus V, Ingason A, Ising M, Jamain S, Jones EG, Jones I, Jones L, Tzeng JY, Kahler AK, Kahn RS, Kandaswamy R, Keller MC, Kennedy JL, Kenny E, Kent L, Kim Y, Kirov GK, Klauck SM, Klei L, Knowles JA, Kohli MA, Koller DL, Konte B, Korszun A, Krabbendam L, Krasucki R, Kuntsi J, Kwan P, Landen M, Langstrom N, Lathrop M, Lawrence J, Lawson WB, Leboyer M, Ledbetter DH, Lee PH, Lencz T, Lesch KP, Levinson DF, Lewis CM, Li J, Lichtenstein P, Lieberman JA, Lin DY, Linszen DH, Liu CY, Lohoff FW, Loo SK, Lord C, Lowe JK, Lucae S, MacIntyre DJ, Madden PAF, Maestrini E, Magnusson PKE, Mahon PB, Maier W, Malhotra AK, Mane SM, Martin CL, Martin NG, Mattheisen M, Matthews K, Mattingsdal M, McCarroll SA, McGhee KA, McGough JJ, McGrath PJ, McGuffin P, McInnis MG, McIntosh A, McKinney R, McLean AW, McMahon FJ, McMahon WM, McQuillin A, Medeiros H, Medland SE, Meier S, Melle I, Meng F, Meyer J, Middeldorp CM, Middleton L, Milanova V, Miranda A, Monaco AP, Montgomery GW, Moran JL, Moreno-De-Luca D, Morken G, Morris DW, Morrow EM, Moskvina V, Muglia P, Muhleisen TW, Muir WJ, Muller-Myhsok B, Murtha M, Myers RM, Myin-Germeys I, Neale MC, Nelson SF, Nievergelt CM, Nikolov I, Nimgaonkar V, Nolen WA, Nothen MM, Nurnberger JI, Nwulia EA, Nyholt DR, O'Dushlaine C, Oades RD, Olincy A, Oliveira G, Olsen L, Ophoff RA, Osby U, Owen MJ, Palotie A, Parr JR, Paterson AD, Pato CN, Pato MT, Penninx BW, Pergadia ML, Pericak-Vance MA, Pickard BS, Pimm J, Piven J, Posthuma D, Potash JB, Poustka F, Propping P, Puri V, Quested DJ, Quinn EM, Ramos-Quiroga JA, Rasmussen HB, Raychaudhuri S, Rehnstrom K, Reif A, Ribases M, Rice JP, Rietschel M, Roeder K, Roeyers H, Rossin L, Rothenberger A, Rouleau G, Ruderfer D, Rujescu D, Sanders AR, Sanders SJ, Santangelo SL, Sergeant JA, Schachar R, Schalling M, Schatzberg AF, Scheftner WA, Schellenberg GD, Scherer SW, Schork NJ, Schulze TG, Schumacher J, Schwarz M, Scolnick E, Scott LJ, Shi JX, Shilling PD, Shyn SI, Silverman JM, Slager SL, Smalley SL, Smit JH, Smith EN, Sonuga-Barke EJS, St Clair D, State M, Steffens M, Steinhausen HC, Strauss JS, Strohmaier J, Stroup TS, Sutcliffe JS, Szatmari P, Szelinger S, Thirumalai S, Thompson RC, Todorov AA, Tozzi F, Treutlein J, Uhr M, van den Oord EJCG, Van Grootheest G, Van Os J, Vicente AM, Vieland VJ, Vincent JB, Visscher PM, Walsh CA, Wassink TH, Watson SJ, Weissman MM, Werge T, Wienker TF, Wijsman EM, Willemsen G, Williams N, Willsey AJ, Witt SH, Xu W, Young AH, Yu TW, Zammit S, Zandi PP, Zhang P, Zitman FG, Zollner S, Devlin B, Kelsoe JR, Sklar P, Daly MJ, O'Donovan MC, Craddock N, Sullivan PF, Smoller JW, Kendler KS, Wray NR, Genomi C-DGP, Genetic IIBD (2013) Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat Genetics 45(9):984

Little RJA, Rubin DB (1989) The analysis of social science data with missing values. Sociol Methods Res 18:292–326

Lord FM (1980) Applications of item response theory to practical testing problems. Erlbaum, Mahwah, NJ

Markon KE, Krueger RF, Watson D (2005) Delineating the structure of normal and abnormal personality: an integrative hierarchical approach. J Pers Soc Psychol 88(1):139–157

Meredith W (1993) Measurement invariance. Factor-analysis and factorial invariance. Psychometrika 58(4):525–543

Posthuma D, Boomsma DI (2000) A note on the statistical power in extended twin designs. Behav Genet 30(2):147–158

Reise SP, Waller NG (1990) Fitting the 2-parameter model to personality data. Appl Psychol Meas 14(1):45–58

Rietveld N, Medland S, Derringer J, Benjamin D, Cesarini D, Koellinger P, Ssgac S (2012) Meta-analysis of educational attainment GWAS: the Social Science Genetics Association Consortium. Behav Genetics 42(6):964–964

Ripke S, Sanders AR, Kendler KS, Levinson DF, Sklar P, Holmans PA, Lin DY, Duan J, Ophoff RA, Andreassen OA, Scolnick E, Cichon S, Clair DS, Corvin A, Gurling H, Werge T, Rujescu D, Blackwood DHR, Pato CN, Malhotra AK, Purcell S, Dudbridge F, Neale BM, Rossin L, Visscher PM, Posthuma D, Ruderfer DM, Fanous A, Stefansson H, Steinberg S, Mowry BJ, Golimbet V, De Hert M, Jonsson EG, Bitter I, Pietilainen OPH, Collier DA, Tosato S, Agartz I, Albus M, Alexander M, Amdur RL, Amin F, Bass N, Bergen SE, Black DW, Borglum AD, Brown MA, Bruggeman R, Buccola NG, Byerley WF, Cahn W, Cantor RM, Carr VJ, Catts SV, Choudhury K, Cloninger CR, Cormican P, Craddock N, Danoy PA, Datta S, De Haan L, Demontis D, Dikeos D, Djurovic S, Donnelly P, Donohoe G, Duong L, Dwyer S, Fink-Jensen A, Freedman R, Freimer NB, Friedl M, Georgieva L, Giegling I, Gill M, Glenthoj B, Godard S, Hamshere M, Hansen M, Hansen T, Hartmann AM, Henskens FA, Hougaard DM, Hultman CM, Ingason A, Jablensky AV, Jakobsen KD, Jay M, Jurgens G, Kahn R, Keller MC, Kenis G, Kenny E, Kim Y, Kirov GK, Konnerth H, Konte B, Krabbendam L, Krasucki R, Lasseter VK, Laurent C, Lawrence J, Lencz T, Lerer FB, Liang KY, Lichtenstein P, Lieberman JA, Linszen DH, Lonnqvist J, Loughland CM, Maclean AW, Maher BS, Maier W, Mallet J, Malloy P, Mattheisen M, Mattingsdal M, McGhee KA, McGrath JJ, McIntosh A, McLean DE, McQuillin A, Melle I, Michie PT, Milanova V, Morris DW, Mors O, Mortensen PB, Moskvina V, Muglia P, Myin-Germeys I, Nertney DA, Nestadt G, Nielsen J, Nikolov I, Nordentoft M, Norton N, Nothen MM, O'Dushlaine CT, Olincy A, Olsen L, O'Neill FA, Orntoft TF, Owen MJ, Pantelis C, Papadimitriou G, Pato MT, Peltonen L, Petursson H, Pickard B, Pimm J, Pulver AE, Puri V, Quested D, Quinn EM, Rasmussen HB, Rethelyi JM, Ribble R, Rietschel M, Riley BP, Ruggeri M, Schall U, Schulze TG, Schwab SG, Scott RJ, Shi JX, Sigurdsson E, Silverman JM, Spencer CCA, Stefansson K, Strange A, Strengman E, Stroup TS, Suvisaari J, Terenius L, Thirumalai S, Thygesen JH, Timm S, Toncheva D, van den Oord E, van Os J, van Winkel R, Veldink J, Walsh D, Wang AG, Wiersma D, Wildenauer DB, Williams HJ, Williams NM, Wormley B, Zammit S, Sullivan PF, O'Donovan MC, Daly MJ, Gejman PV (2011) Genome-wide association study identifies five new schizophrenia loci. Nat Genetics 43(10):969–977

Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kahler AK, Akterin S, Bergen SE, Collins AL, Crowley JJ, Fromer M, Kim Y, Lee SH, Magnusson PK, Sanchez N, Stahl EA, Williams S, Wray NR, Xia K, Bettella F, Borglum AD, Bulik-Sullivan BK, Cormican P, Craddock N, de Leeuw C, Durmishi N, Gill M, Golimbet V, Hamshere ML, Holmans P, Hougaard DM, Kendler KS, Lin K, Morris DW, Mors O, Mortensen PB, Neale BM, O'Neill FA, Owen MJ, Milovancevic MP, Posthuma D, Powell J, Richards AL, Riley BP, Ruderfer D, Rujescu D, Sigurdsson E, Silagadze T, Smit AB, Stefansson H, Steinberg S, Suvisaari J, Tosato S, Verhage M, Walters JT, Levinson DF, Gejman PV, Laurent C, Mowry BJ, O'Donovan MC, Pulver AE, Schwab SG, Wildenauer DB, Dudbridge F, Shi J, Albus M, Alexander M, Campion D, Cohen D, Dikeos D, Duan J, Eichhammer P, Godard S, Hansen M, Lerer FB, Liang KY, Maier W, Mallet J, Nertney

DA, Nestadt G, Norton N, Papadimitriou GN, Ribble R, Sanders AR, Silverman JM, Walsh D, Williams NM, Wormley B, Arranz MJ, Bakker S, Bender S, Bramon E, Collier D, Crespo-Facorro B, Hall J, Iyegbe C, Jablensky A, Kahn RS, Kalaydjieva L, Lawrie S, Lewis CM, Linszen DH, Mata I, McIntosh A, Murray RM, Ophoff RA, Van Os J, Walshe M, Weisbrod M, Wiersma D, Donnelly P, Barroso I, Blackwell JM, Brown MA, Casas JP, Corvin AP, Deloukas P, Duncanson A, Jankowski J, Markus HS, Mathew CG, Palmer CN, Plomin R, Rautanen A, Sawcer SJ, Trembath RC, Viswanathan AC, Wood NW, Spencer CC, Band G, Bellenguez C, Freeman C, Hellenthal G, Giannoulatou E, Pirinen M, Pearson RD, Strange A, Su Z, Vukcevic D, Langford C, Hunt SE, Edkins S, Gwilliam R, Blackburn H, Bumpstead SJ, Dronov S, Gillman M, Gray E, Hammond N, Jayakumar A, McCann OT, Liddle J, Potter SC, Ravindrarajah R, Ricketts M, Tashakkori-Ghanbaria A, Waller MJ, Weston P, Widaa S, Whittaker P, McCarthy MI, Stefansson K, Scolnick E, Purcell S, McCarroll SA, Sklar P, Hultman CM, Sullivan PF (2013) Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet 45(10):1150–1159

Samuel DB, Widiger TA (2008) A meta-analytic review of the relationships between the five-factor model and DSM-IV-TR personality disorders: a facet level analysis. Clin Psychol Rev 28(8):1326–1342

Service SK, Verweij KJ, Lahti J, Congdon E, Ekelund J, Hintsanen M, Raikkonen K, Lehtimaki T, Kahonen M, Widen E, Taanila A, Veijola J, Heath AC, Madden PA, Montgomery GW, Sabatti C, Jarvelin MR, Palotie A, Raitakari O, Viikari J, Martin NG, Eriksson JG, Keltikangas-Jarvinen L, Wray NR, Freimer NB (2012) A genome-wide meta-analysis of association studies of Cloninger's Temperament Scales. Transl Psychiat 2e116

Shifman S, Bhomra A, Smiley S, Wray NR, James MR, Martin NG, Hettema JM, An SS, Neale MC, Van den Oord EJCG, Kendler KS, Chen X, Boomsma DI, Middeldorp CM, Hottenga JJ, Slagboom PE, Flint J (2008) A whole genome association study of neuroticism using DNA pooling. Mol Psychiat 13(3):302–312

Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Allen HL, Lindgren CM, Luan J, Magi R, Randall JC, Vedantam S, Winkler TW, Qi L, Workalemahu T, Heid IM, Steinthorsdottir V, Stringham HM, Weedon MN, Wheeler E, Wood AR, Ferreira T, Weyant RJ, Segre AV, Estrada K, Liang LM, Nemesh J, Park JH, Gustafsson S, Kilpelanen TO, Yang JA, Bouatia-Naji N, Esko T, Feitosa MF, Kutalik Z, Mangino M, Raychaudhuri S, Scherag A, Smith AV, Welch R, Zhao JH, Aben KK, Absher DM, Amin N, Dixon AL, Fisher E, Glazer NL, Goddard ME, Heard-Costa NL, Hoesel V, Hottenga JJ, Johansson A, Johnson T, Ketkar S, Lamina C, Li SX, Moffatt MF, Myers RH, Narisu N, Perry JRB, Peters MJ, Preuss M, Ripatti S, Rivadeneira F, Sandholt C, Scott LJ, Timpson NJ, Tyrer JP, van Wingerden S, Watanabe RM, White CC, Wiklund F, Barlassina C, Chasman DI, Cooper MN, Jansson JO, Lawrence RW, Pellikka N, Prokopenko I, Shi JX, Thiering E, Alavere H, Alibrandi MTS, Almgren P, Arnold AM, Aspelund T, Atwood LD, Balkau B, Balmforth AJ, Bennett AJ, Ben-Shlomo Y, Bergman RN, Bergmann S, Biebermann H, Blakemore AIF, Boes T, Bonnycastle LL, Bornstein SR, Brown MJ, Buchanan TA, Busonero F, Campbell H, Cappuccio FP, Cavalcanti-Proenca C, Chen YDI, Chen CM, Chines PS, Clarke R, Coin L, Connell J, Day INM, den Heijer M, Duan JB, Ebrahim S, Elliott P, Elosua R, Eiriksdottir G, Erdos MR, Eriksson JG, Facheris MF, Felix SB, Fischer-Posovszky P, Folsom AR, Friedrich N, Freimer NB, Fu M, Gaget S, Gejman PV, Geus EJC, Gieger C, Gjesing AP, Goel A, Goyette P, Grallert H, Grassler J, Greenawalt DM, Groves CJ, Gudnason V, Guiducci C, Hartikainen AL, Hassanali N, Hall AS, Havulinna AS, Hayward C, Heath AC, Hengstenberg C, Hicks AA, Hinney A, Hofman A, Homuth G, Hui J, Igl W, Iribarren C, Isomaa B, Jacobs KB, Jarick I, Jewell E, John U, Jorgensen T, Jousilahti P, Jula A, Kaakinen M, Kajantie E, Kaplan LM, Kathiresan S, Kettunen J, Kinnunen L, Knowles JW, Kolcic I, Konig IR, Koskinen S, Kovacs P, Kuusisto J, Kraft P, Kvaloy K, Laitinen J, Lantieri O, Lanzani C, Launer LJ, Lecoeur C, Lehtimaki T, Lettre G, Liu JJ, Lokki ML, Lorentzon M, Luben RN, Ludwig B, Manunta P, Marek D, Marre M, Martin NG, McArdle WL, McCarthy A, McKnight B, Meitinger T, Melander O, Meyre D, Midthjell K, Montgomery GW, Morken MA, Morris AP, Mulic R, Ngwa JS, Nelis M, Neville MJ, Nyholt DR, O'Donnell CJ, O'Rahilly S, Ong KK, Oostra B, Pare G, Parker AN, Perola M, Pichler I, Pietilainen KH, Platou CGP, Polasek O, Pouta A, Rafelt S, Raitakari O, Rayner NW, Ridderstrale M, Rief W, Ruokonen A, Robertson NR, Rzehak P, Salomaa V, Sanders AR, Sandhu MS, Sanna S, Saramies J, Savolainen MJ, Scherag S, Schipf S, Schreiber S, Schunkert H, Silander K, Sinisalo J, Siscovick DS, Smit JH, Soranzo N, Sovio U, Stephens J, Surakka I, Swift AJ, Tammesoo ML, Tardif JC, Teder-Laving M, Teslovich TM, Thompson JR, Thomson B, Tonjes A, Tuomi T, van Meurs JBJ, van Ommen GJ (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genetics 42(11):937–953

Sullivan PF, Daly MJ, O'Donovan M (2012) Genetic architecture of psychiatric disorders: the emerging picture and its implications. Nat Rev Genetics 13:537–551

Terracciano A, Sanna S, Uda M, Deiana B, Usala G, Busonero F, Maschio A, Scally M, Patriciu N, Chen WM, Distel MA, Slagboom EP, Boomsma DI, Villafuerte S, Sliwerska E, Burmeister M, Amin N, Janssens ACJW, van Duijn CM, Schlessinger D, Abecasis GR, Costa PT (2010) Genome-wide association scan for five major dimensions of personality. Mol Psychiat 15(6):647–656

van den Berg SM, Service SK (2012) Power of IRT in GWAS: successful QTL mapping of sum score phenotypes depends on interplay between risk allele frequency, variance explained by the risk allele, and test characteristics. Genet Epidemiol 36(8):882–889

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

van den Berg SM, Paap MC, Derks EM, GROUP (2013) Using multidimensional modeling to combine self-report symptoms with clinical judgment of schizotypy. Psychiat Res 206(1):75–80

van den Oord EJ, Kuo PH, Hartmann AM, Webb BT, Moller HJ, Hettema JM, Giegling I, Bukszar J, Rujescu D (2008) Genome-wide association analysis followed by a replication study implicates a novel candidate gene for neuroticism. Arch Gen Psychiat 65(9):1062–1071

Verhagen AJ, Fox J-P (2013a) Bayesian tests of measurement invariance. Br J Math Stat Psychol 66(3):383–401

Verhagen AJ, Fox JP (2013b) Longitudinal measurement in health-related surveys. A Bayesian joint growth model for multivariate ordinal responses. Stat Med 32(17):2988–3005

Weisscher N, Glas CA, Vermeulen M, De Haan RJ (2010) The use of an item response theory-based disability item bank across diseases: accounting for differential item functioning. J Clin Epidemiol 63(5):543–549

Wray NR, Pergadia ML, Blackwood DHR, Penninx BWJH, Gordon SD, Nyholt DR, Ripke S, MacIntyre DJ, McGhee KA, Maclean AW, Smit JH, Hottenga JJ, Willemsen G, Middeldorp CM, de Geus EJC, Lewis CM, McGuffin P, Hickie IB, Van den Oord EJCG, Liu JZ, Macgregor S, McEvoy BP, Byrne EM, Medland SE, Statham DJ, Henders AK, Heath AC, Montgomery GW, Martin NG, Boomsma DI, Madden PAF, Sullivan PF (2012) Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. Mol Psychiat 17(1):36–48