Journal of Psychopathology and Behavioral Assessment (2022) 44:663–678
https://doi.org/10.1007/s10862-021-09925-9

Harmonized Phenotypes for Anxiety, Depression, and Attention-Deficit Hyperactivity Disorder (ADHD)

Miljan Jović[1] · Kratika Agarwal[1] · Andrew Whitehouse[2] · Stéphanie M. van den Berg[1]

Accepted: 5 August 2021 / Published online: 23 February 2022

Abstract
In multi-cohort consortia, the problem often arises that a phenotype is measured using different questionnaires. This study aimed to harmonize scores based on the Child Behaviour Check List (CBCL) and the Strength and Difficulties Questionnaire (SDQ) for anxiety/depression and ADHD. To link the scales, we used parent reports on 1330 children aged 10–11.5 years from the Raine study on both SDQ and CBCL. Harmonization was done based on Item Response Theory. We started from existing CBCL and SDQ scales related to anxiety/depression and ADHD (theoretical approach). Next, we conducted a data-driven approach using factor analysis to validate the theoretical approach. Both approaches yielded similar scales, validating the combination of existing scales. In addition, we studied the impact of harmonized (IRT-based) scores on the statistical power of the results in meta-analytic gene-finding studies. The results showed that the IRT-based harmonized scores increased the statistical power of the results compared to sum scores, even with an equal sample size. These findings can help future researchers to harmonize data from different samples and/or different questionnaires that measure anxiety, depression, and ADHD, in order to obtain the larger sample sizes, to compare research results across subpopulations or to increase generalizability, the validity or statistical power of research results. We recommend using our item parameters to estimate harmonized scores that represent commensurate phenotypes across cohorts, and we explained in detail how other researchers can use our results to harmonize data in their studies.

Keywords Data harmonization · Anxiety · Depression · ADHD · Raine study

Introduction

According to statistics by ROAMER (*Roadmap for Mental Health and Wellbeing Research in Europe*), mental health disorders form about 11% to 27% of all diseases (Haro et al., 2014). One of the most common mental health disorders is depression, which is a leading cause of disability worldwide with more than 322 million people all around the world suffering from it (WHO, 2017, 2020). Depression can lead to problems at work, at school, and in the family (WHO, 2020). The total estimated number of people living with depression increased by 18.4% between 2005 and 2015 (Vos et al., 2016). The other frequently common mental health problem besides depression is anxiety. The total estimated number of people living with anxiety disorders in the world is 264 million (WHO, 2017). There is an increase of 14.9% between 2005 and 2015 (Vos et al., 2016). Depression and anxiety affect all age groups, including children and adolescents. According to the World Health Organization, 10–20% of children and adolescents worldwide experience mental disorders (WHO, 1992), while 2–3% of children ages 6 to 12 and 6–8% of teens may have serious depression (ADAA, 2020). According to the Anxiety and Depression Association of America, about 80% of kids with anxiety, and 60% with depression are not getting treatment (ADAA, 2020).

The other common mental health problem among children is attention-deficit hyperactivity disorder (ADHD). Studies show that between 5 and 8.5 percent of children and 2.5 percent of adults have ADHD (Danielson et al., 2016; Dulcan, 1997; Polanczyk et al., 2007; Simon

Miljan Jović and Kratika Agarwal contributed equally to this work.

✉ Miljan Jović
m.jovic@utwente.nl

[1] Department of Learning, Data Analytics and Technology, University of Twente, 7500 Enschede, The Netherlands

[2] Telethon Kids Institute, University of Western Australia, Perth, Australia

et al., 2009), while symptoms may persist into adulthood in 50% of them (Karam et al., 2015; Schmitz et al., 2007).

Both genetic and non-genetic factors play important roles in mental health problems (e.g. Hettema et al., 2005; Kendler et al., 2011; Nadder et al., 1998; Silberg et al., 1999; Silberg et al., 2001; Silove et al., 1995; van den Berg et al., 2006). Therefore, it is important to study the causal mechanisms and the development of mental health symptoms across the entire lifespan and study prevention and the effect of interventions. This with the help of either genetically informative cohorts, with genotyping, or with a combination of both.

The EU-funded project CAPICE (*Childhood and Adolescence Psychopathology: unravelling the complex etiology by a large interdisciplinary collaboration in Europe*; Rajula et al., 2021) has objectives relevant to this. It focuses on the improvement of later outcomes of child and adolescent mental health problems related to anxiety, depression, and ADHD. The CAPICE consortium consists of several research groups with both phenotypic and genotypic data (Rajula et al., 2021). A common problem in research consortia is that the phenotypes are assessed with different questionnaires, resulting in difficulties combining the data from different studies (Luningham et al., 2019; van den Berg et al., 2014). This paper focuses on one objective of CAPICE: to construct a common metric for anxiety, depression, and ADHD phenotypes to harmonize them across research groups. At a later stage, these harmonized phenotypes can be used to make meaningful comparisons across countries, and generally boost statistical power in by increasing sample size, which is particularly important for genetic studies.

Two widely used screening instruments (or questionnaires) for psychopathology in children are *(a)* the Strengths and Difficulties Questionnaire (SDQ) and *(b)* the Child Behaviour Checklist (CBCL). These questionnaires have been used a lot by the consortium partners and it would help research tremendously if the data from the various partners could somehow be harmonized. Various studies have explored the development of internalizing and externalizing problem behaviour using the CBCL (Achenbach, 1991; Allen & Prior, 1995; Caspi et al., 1995) and SDQ (Muris et al., 2003; Ortuno-Sierra et al., 2015). The CBCL consists of 113 items and operationalizes childhood behaviour on eight subscales/dimensions (social withdrawal, somatic complaints, anxiety/depression, social problems, thought problems, attention problems, delinquent behaviour, and aggressive behaviour; Achenbach & Ruffle, 2000; Achenbach et al., 1991). The SDQ consists of 25 items equally divided across five scales, also called *dimensions* (Emotional, Conduct, Hyperactivity, Peer, and Prosocial problems; Goodman, 1997, 2001) and it is used for children aged 3–16 years.

The two questionnaires largely assess the same dimensions of child psychopathology: *Emotional problems*, *Hyperactivity*, *Social problems*, and *Conduct problems*. When it comes to anxiety and depression, CBCL has Anxiety/depressed and Withdrawn/depressed subscales. Although the SDQ does not have separate scales for anxiety and depression, it has an Emotional problems scale that addresses these difficulties. On the other hand, when it comes to ADHD type of problems, SDQ has a hyperactivity-inattention scale that also includes items related to concentration problems, while the CBCL has an attention problems subscale that includes both hyperactivity and attention problems.

The main differences between the two questionnaires are when it comes to the phrasing of the individual questions and the response format. For example, the SDQ asks whether the child is easily distracted, and the CBCL asks whether the child can concentrate. The SDQ is rated with answer categories *not true*, *somewhat true,* and *certainly true*, whereas the CBCL is rated with answer categories *not true*, *somewhat/sometimes true*, and *very true/often true*. Given these differences in phrasing and response formats, item responses on the SDQ and CBCL or scale scores cannot be directly compared, as they contain slightly different sets of behaviour and different numbers of items. This means that if one child scores 4 points on the SDQ subscale for anxiety/depression (Emotional problems), and another child scores 4 points on the anxiety/depression subscale of the CBCL, it is usually not possible to say which child shows the most problematic behaviour and by how much they differ. We need to know how the items were phrased (do they address serious or less serious problems), how the items are scored with what response categories, and the number of items in order to compare SDQ and CBCL scores. Even if we know all the above-mentioned things, it is still hard to make conclusions regarding the quantitative differences between these two scores. Therefore, a common metric is needed to quantify individual differences on two subscales from different questionnaires. Once a common metric is found, scores from different questionnaires can be harmonized by transforming the original scores to new scores on the common metric.

Finding a common metric for comparable subscale scores can be achieved by different methodologies. A common choice is to use the methodology of test linking (Kolen & Brennan, 2014). There are multiple ways to carry out test linking and one of the widely used approaches is the framework of Item Response Theory (IRT; Embretson & Reise, 2000; Kolen & Brennan, 2014). The use of IRT is a common and flexible method for modeling the relationship between participants` trait levels and their responses to items (Park et al., 2019; van den Berg et al., 2007, 2014). In the IRT approach, a participant`s response to an item depends on both the participant`s trait level and the item parameters of that particular item (Embretson & Reise, 2000). In order to

construct a common metric, we need a sample from individuals with overlapping data on both questionnaires (Hussong et al., 2013; Luningham et al., 2019). By applying the IRT approach to such a sample with data on both questionnaires, we are able to define a common metric. As an example, in the Genetics of Personality Consortium, big-five personality phenotypes were harmonized across several personality questionnaires (Van den Berg et al., 2014). They used IRT models on the data sets where groups of participants responded to multiple questionnaires. Assuming that the level on the underlying trait does not change between filling in Questionnaire A and Questionnaire B, using IRT one can link the items from both questionnaires and define one common metric. Using the IRT approach, the data from more than 23 cohorts worldwide could be harmonized, resulting in genome-wide association (GWA) meta-analysis of neuroticism (De Moor et al., 2015) and extraversion (van den Berg et al., 2016) with large sample sizes. It not only increases sample size, but also overcomes the problem of comparing allelic effect sizes across studies (van den Berg & de Moor, 2020).

In this paper, we focus on harmonizing phenotypes for childhood psychopathology with a specific interest in the CAPICE phenotypes: anxiety, depression, and ADHD. We used the data from the Western Australian Pregnancy Cohort – Raine study (Newnham et al., 1993) to link the CBCL and SDQ data from children aged 10–11.5 years. In the Raine study, parents were asked to fill in the CBCL and SDQ at the same moment in time on their child between age 10–11.5 years. In this study, we have responses of 1330 participants on both CBCL and SDQ questionnaires and based on that we are able to link the items from both the questionnaires and to define one common metric. This provided a unique opportunity to see whether one can find anxiety, depression, and ADHD dimensions shared by both these questionnaires and to see whether common metrics can be found. The item-response theory models were applied to define common metrics.

The results of this study are useful for data harmonization of anxiety, depression, and ADHD in the CAPICE project as well as in other research consortia. Harmonization of phenotypes is especially important in large research consortia, but it can also be helpful in smaller research studies. Our findings can be applied in all situations where researchers need to harmonize data from two samples which have filled in two different questionnaires for measuring anxiety, depression, or ADHD (one sample with CBCL data, one sample with SDQ data) to be able to compare research results across countries or subpopulations or to increase the size of the sample (Hamilton et al., 2011; Smith-Warner et al., 2006; Thompson, 2009; van den Berg et al., 2014). Besides, it can also be used to analyse longitudinal data where CBCL data were gathered at age x and SDQ data were gathered at age y. Lastly, if a sample contains both CBCL and SDQ data at the same age from the same individual, our results can be used to increase measurement reliability or validity (Fortier et al., 2010, 2011).

It is especially important to mention the advantages of using harmonized scores in behaviour genetic studies. Most importantly, the use of harmonized scores leads to an increase in sample size and/or an increase in measurement precision of the phenotypes, and that in turn increases the statistical power of the results (van den Berg et al., 2014). We will introduce the advantages of using harmonized scores in behaviour genetic studies in more detail below.

Increasing Statistical Power of the Results Using Harmonized Scores in Behaviour Genetic Studies

The effect sizes in the case of complex human traits are often small and that is one of the main reasons why meta-analytic studies (e.g. genome-wide association (GWA) studies) are required in the field of behaviour genetics (van den Berg et al., 2014; Zeggini & Ioannidis, 2009). One of the biggest problems in the meta-analysis of behavioural measures is caused by the use of different measurement instruments across studies for assessing a particular phenotype (van den Berg et al., 2014). In most meta-analytic studies, one is not able to meaningfully compare effect sizes across studies (van den Berg & de Moor, 2020). The coefficient of regression of phenotype on the number of alleles for a single-nucleotide polymorphism (SNP) gets a different meaning every time that the scale of the phenotypic measure changes. Instead, one could compare p-values, but that leads to the difficulty that we lose information about the direction of the effect (Begum et al., 2012; Sullivan & Feinn, 2012; Zeggini & Ioannidis, 2009). Instead of p-values, one could look at standardized regression coefficients, that yield information about the direction of the effect. But unfortunately, they do not give information about the absolute size of the effect, as unstandardized regression coefficients do. If the phenotype is assessed by the same instrument in all studies, the unstandardized regression coefficient tells us how many units on the scale of the common instrument we gain for every additional allele (in case of an additive inheritance pattern). It yields information on both the direction and size of the effect.

Thanks to harmonized scores, researchers can use larger samples in the meta-analytic studies, leading to the higher statistical power of the results (Sullivan et al., 2012). For example, van den Berg et al. (2014) showed that the statistical power to detect a single nucleotide polymorphism (SNP) at the genome-wide significance level that explains 0.1% of the true phenotypic variance with an allele frequency of 0.5, substantially increased from 18 to 44% after harmonization.

Besides that, the use of harmonized scores also leads to an increase in measurement precision, and higher measurement precision leads to an increase in the statistical power of the results (van den Berg et al., 2014). In other words, the use of harmonized scores can also lead to higher statistical power than the use of non-harmonized scores (i.e. sum scores) even if the size of the sample is the same in both of the cases. This happens for example when you include items from other scales into the measurement model.

In this study, after the scale linking, we will conduct a simulation analysis in order to compare the statistical power of the results in meta-analyses based on sum scores and harmonized scores. In both cases, we will use the same sample size and the same number of items.

Methods

Material

The Raine Study is a prospective cohort of children begun in 1989 and consisting of 2900 randomly assigned pregnant women (Chivers et al., 2010; Howard et al., 2011; McKnight et al., 2012; Newnham et al., 1993). Pregnant women who attended the public antenatal clinic at King Edward Memorial Hospital (KEMH; Perth, Western Australia) and nearby private clinics between May 1989 and November 1991 were enrolled in the Raine study (Newnham et al., 1993). It was required that women have sufficient English-language skills to give informed consent, an expectation to deliver at KEMH, and an intention to reside in Western Australia to enable future follow-ups of their child (Howard et al., 2011). Those women completed questionnaires at 18 and 34 weeks gestation, while further follow-up investigations took place at birth, and at 1, 2, 3, 5, 8, 10, 14, 17, 18, and 20 years (Howard et al., 2011; McKnight et al., 2012). The study had two main aims: to investigate the hypothesis that complications of pregnancy might be prevented by frequent ultrasound scans and to develop a long-term cohort to study the role that early life events have on later health (McKnight et al., 2012). The data are collected through questionnaires, physical measurements, and biological samples and consists of information about growth, cardiovascular, respiratory, immunological, musculoskeletal, nutritional, psychiatric, neurocognitive, and ophthalmic health (McKnight et al., 2012). The subset of the Raine dataset that we used for this study consists of both the CBCL and SDQ parent-filled questionnaires of 2861 children ('Generation 2') aged between 10 and 11.5 years (1417 girls, 1444 boys). Here, Raine used the 1991 Aseba version for the CBCL (age 4–18) by Achenbach (1991) and the 1997 SDQ version by Goodman (1997). The item scores consisting of either 0, 1, 2 which represented not true, somewhat/sometimes true, very true/often true, and 7, 8, or 9 which represented not done, not stated, and not applicable, respectively.

Data Analysis

Psychometric Analysis

Two different approaches were taken in the data analysis: a top-down (theoretical) approach and a bottom-up (data-driven) approach. In the top-down approach, we started from the existing CBCL and SDQ subscales related to anxiety/depression and ADHD, while in the bottom-up approach we conducted an exploratory factor analysis to identify anxiety/depression and ADHD scales. We used exploratory factor analysis to check whether the factor structure is in accordance with the theoretically expected structure of the scales. We used both approaches because we wanted to establish whether starting from theory and obtaining results that way was in any way supported by what the data would tell us without any preconceptions.

We used the R programming language and environment to carry out the factor and IRT analyses. First, we preprocessed the data of 2861 children aged between 10 and 11.5 years (1417 girls and 1444 boys) where we omitted the incomplete records from the dataset. The remaining 1330 complete cases (653 girls and 677 boys) were used for the analyses. We carried out an IRT analysis to investigate the psychometric quality of the scales, using the Multidimensional Item Response Theory (mirt) package (Chalmers, 2012). We used the Generalized Partial Credit Model to obtain the item parameters. This model contains discrimination and threshold item parameters (Embretson & Reise, 2000). The discrimination parameter is a measure of the capability of an item to differentiate respondents with different trait levels (Embretson & Reise, 2000). The discrimination parameter refers to the strength of the relationship between trait level and participants` responses on the item (Embretson & Reise, 2000). It can be compared to a factor loading in factor analysis (van den Berg et al., 2007). The threshold parameter is defined as the point on the latent trait continuum where the response probability for two adjacent response categories is equal (Wetzel & Carstensen, 2014). Accordingly, for a 3-point scale, we have two threshold parameters, between categories 1 and 2 and between categories 2 and 3 (Uto & Ueno, 2018). The chi-square item-fit approach was used for assessing the item fit. We used Bonferroni correction to correct the significance level of the chi-square test for multiple testing, limiting the risk of type I errors (Andrich & Marais, 2019). For each scale that we analysed, we multiplied the p-values by the number of items in that scale to get Bonferroni corrected p-values. We used an alpha level of 5% (after Bonferroni correction).

Top-Down Approach

In the top-down (theoretical) approach, we started from what is known about the CBCL and SDQ scales and their internal structure. We departed from the theory and existing CBCL (Achenbach, 1991) and SDQ (Goodman, 1997) scales related to anxiety, depression, and ADHD.

For the SDQ, we selected all 5 items from the scale of the Emotional problem to represent anxiety and depression combined, while 5 items from the hyperactivity scale were selected to represent ADHD type of problems (Table 1; Goodman, 1997). Note that there are not enough SDQ items to identify anxiety and depression separately. Also, note that the SDQ hyperactivity scale also includes items related to concentration problems.

The CBCL has two separate subscales for depressive behaviour (Anxious/Depressed and Withdrawn/Depressed), of which only one is related to anxiety (Achenbach, 1991). Besides that, there is also the Somatic Complaints subscale which can be related to anxiety and depression (e.g. Löwe et al., 2008), since together with Anxious/Depressed and Withdrawn/Depressed it forms the internalizing problem behaviour subscale. We therefore decided to select the internalizing problem behaviour subscale into representing anxiety/depression. This CBCL anxiety/depression scale consists of 31 items (Table 1). To represent ADHD, we used the CBCL attention problems subscale as it contained the items which could be interpreted as ADHD type of problems. This CBCL ADHD scale consists of 11 items (Table 1).

First, we carried out IRT analyses on the anxiety/depression CBCL and SDQ items separately to investigate the psychometric quality of the scales on their own. Next, we ascertained whether it is possible to measure anxiety and depression using the CBCL anxiety/depression scale and the SDQ items combined. If we would find a good quality scale when these items are combined, harmonization is possible. Then the CBCL anxiety/depression scale can be directly linked to the SDQ anxiety/depression scale (van den Berg et al., 2014).

Before combining the scales, we applied Natural Language Processing (NLP) using the Applied Latent Semantic Analysis Functions package of R (LSAfun; Guenther et al., 2015) to examine the semantic similarity between items from the CBCL and SDQ scales. Redundant (almost identical or similarly worded) items in CBCL and SDQ scales are a problem because they can undermine the structure of the scale. Highly similar items are expected to correlate much more than less similar items so that the scale is likely to be dominated (i.e. high discrimination parameters) by similar items. This would severely bias the scale toward the content of these similar items, which would reduce the content validity but also affect the fit of a (unidimensional) model. Accordingly, we must exclude one item in the case when there are two identically or similarly worded items.

Table 1 Original anxiety/depression and ADHD scales consisting of the CBCL and SDQ items

Anxiety/depression items	ADHD items
CBCL lonely	CBCL acts too young
CBCL cries a lot	CBCL can't concentrate
CBCL fears might do bad	CBCL restless, hyperactive
CBCL fears has to be perfect	CBCL confused
CBCL feels unloved	CBCL day dreams
CBCL feels others out to get	CBCL impulsive
CBCL feels worthless	CBCL nervous or tense
CBCL loner	CBCL nervous movements
CBCL nervous or tense	CBCL poor school work
CBCL too fearful or anxious	CBCL poorly coordinated
CBCL feels dizzy	CBCL stares blankly
CBCL feels too guilty	SDQ restless
CBCL overtired	SDQ constantly fidgeting
CBCL no cause aches/pains	SDQ easily distracted
CBCL no cause headaches	SDQ thinks before acting
CBCL no cause nausea	SDQ good attention span
CBCL no cause eye problems	
CBCL no cause skin problems	
CBCL no cause stomachaches	
CBCL no cause vomiting	
CBCL refuses to talk	
CBCL secretive	
CBCL self-conscious	
CBCL shy or timid	
CBCL stares blankly	
CBCL sulks a lot	
CBCL suspicious	
CBCL lacks energy/slow	
CBCL unhappy depressed	
CBCL withdrawn	
CBCL worries	
SDQ complains of illness	
SDQ often seems worried	
SDQ unhappy/tearful	
SDQ nervous or clingy	
SDQ easily scared	

After that, we carried out an IRT analysis on the scale consisting of both the CBCL and SDQ anxiety/depression items combined. The same approach was taken for ADHD.

Bottom-Up Approach

In the bottom-up (data-driven) approach, we conducted an exploratory factor analysis to identify scales that measure anxiety, depression, and ADHD by looking at the correlational structure when all CBCL and SDQ items are combined. After identifying a suitable candidate scale for anxiety

and/or depression, we carried out IRT analyses on the CBCL and SDQ items separately to investigate the psychometric quality of the anxiety/depression scale(s) consisting only of the CBCL items and only of the SDQ items. To prevent psychometric problems, we performed a semantic similarity analysis to identify items that are similarly worded. After that, we carried out an IRT analysis on the combined anxiety/depression scale consisting of the CBCL and SDQ items both. The same procedure was used for ADHD (Table 2).

Table 2 Bottom-up Anxiety/depression and ADHD scales consisting of both CBCL and SDQ items

Anxiety/depression items	ADHD items
CBCL obsessions	CBCL acts too young
CBCL too dependent	CBCL can't concentrate
CBCL lonely	CBCL restless, hyperactive
CBCL confused	CBCL day dreams
CBCL cries a lot	CBCL impulsive
CBCL easily jealous	CBCL poor school work
CBCL fears animals, situations	CBCL poorly coordinated
CBCL fears going to school	CBCL speech problem
CBCL fears might do bad	SDQ restless
CBCL fears has to be perfect	SDQ constantly fidgeting
CBCL feels others out to get	SDQ easily distracted
CBCL feels worthless	SDQ thinks before acting
CBCL loner	SDQ good attention span
CBCL nervous or tense	
CBCL nightmares	
CBCL constipated	
CBCL too fearful or anxious	
CBCL feels dizzy	
CBCL feels too guilty	
CBCL overtired	
CBCL no cause aches/pains	
CBCL no cause nausea	
CBCL no cause stomachaches	
CBCL secretive	
CBCL self-conscious	
CBCL shy or timid	
CBCL sudden change in mood	
CBCL sulks a lot	
CBCL overly neat/clean	
CBCL lacks energy/slow	
CBCL unhappy/depressed	
CBCL withdrawn	
CBCL worries	
SDQ complains of illness	
SDQ often seems worried	
SDQ often unhappy/tearful	
SDQ nervous or clingy	
SDQ easily scared	

Evaluation of the Scales

We started this part with a qualitative comparison of the content of the items from the top-down and bottom-up approach to choose which versions of the scales we should use (Table 3).

Next, we performed additional analyses to further evaluate the resulting scales and to provide additional information about the practicalities of using them. One of the problems of combining the items from the two questionnaires is that often there are items in both scales that are almost identically worded. To avoid psychometric problems, one of the similarly worded items has to be deleted from the combined scale. But using fewer items in the test linking will lead to loss of information (reliability). To illustrate this, suppose that item 1 from the CBCL and item 2 from the SDQ are similarly worded, which causes a model fit problem due to their high correlation. In that case, we need to exclude one of the two similarly worded items. Further suppose that for the test linking, item 1 from the CBCL is therefore not included in the combined scale. As a result, we will have item parameters for all items on the combined scale, except for item 1 from the CBCL. Therefore, researchers who want to use our item parameters in their studies (for example, to harmonize their CBCL data to make them comparable with SDQ data from a different cohort) will not have item parameters for item 1 from the CBCL scale. Then this item cannot be used to estimate latent trait values and that leads to loss of information and consequently loss of reliability and validity. Note that exclusion of similarly worded items is necessary in the IRT test linking stage, but that is not the case in the situation when researchers want to use our item parameters in their studies.

We explored different solutions to this problem based on a set of analyses. First, we plotted test information curves of different versions of scales in order to compare them based on their informative value (i.e., reliability). This visualizes the extent to which we lose information if excluded items are not used. Second, we plotted scatter plots and regression lines to describe the relationship between item parameters from one scale before and after combining them with items from the other scale. Based on obtained regression coefficients (intercept and slope), we conducted regression analyses to examine whether it is possible to find a useful set of item parameters based on a linear transformation from one scale to the other. More precisely, we used regression analysis to predict item parameters of items that are excluded from the combined scale due to similarity. For example, the initial CBCL anxiety/depression scale consists of 31 items, but suppose that we need to exclude some of them from the combined scale due to similarity with some SDQ items. In that case, we do not have item parameters of items that are excluded from the combined scale, but we do have their item parameters from the initial scale (scale consisting only of the CBCL items) and we have regression coefficients that

Table 3 Anxiety/depression – Comparisons between scales from different approaches (Top-down and Bottom-up)

Top-down	Bottom-up
CBCL lonely*	CBCL obsessions
CBCL cries a lot*	CBCL too dependent
CBCL fears might do bad*	CBCL lonely*
CBCL fears has to be perfect*	CBCL confused
CBCL feels unloved	CBCL cries a lot*
CBCL feels others out to get*	CBCL easily jealous
CBCL feels worthless*	CBCL fears animals, situations
CBCL loner*	CBCL fears going to school
CBCL nervous or tense*	CBCL fears might do bad*
CBCL too fearful or anxious*	CBCL fears has to be perfect*
CBCL feels dizzy*	CBCL feels others out to get*
CBCL feels too guilty*	CBCL feels worthless*
CBCL overtired*	CBCL loner*
CBCL no cause aches/pains*	CBCL nervous or tense*
CBCL no cause headaches	CBCL nightmares
CBCL no cause nausea*	CBCL constipated
CBCL no cause eye problems	CBCL too fearful or anxious*
CBCL no cause skin problems	CBCL feels dizzy*
CBCL no cause stomachaches*	CBCL feels too guilty*
CBCL no cause vomiting	CBCL overtired*
CBCL refuses to talk	CBCL no cause aches/pains*
CBCL secretive*	CBCL no cause nausea*
CBCL self-conscious*	CBCL no cause stomachaches*
CBCL shy or timid*	CBCL secretive*
CBCL stares blankly	CBCL self-conscious*
CBCL sulks a lot*	CBCL shy or timid*
CBCL suspicious	CBCL sudden change in mood
CBCL lacks energy/slow*	CBCL sulks a lot*
CBCL unhappy depressed*	CBCL overly neat/clean
CBCL withdrawn*	CBCL lacks energy/slow*
CBCL worries*	CBCL unhappy/depressed*
	CBCL withdrawn*
	CBCL worries*

Note: * – Common items

describe the relationship between item parameters from the initial scale and item parameters from the combined scale, based on the items that are retained and a regression analysis. If a linear relationship exists between item parameters before and after combining them, this is a very easy method.

Increasing Statistical Power of the Results Using Harmonized Scores in Behaviour Genetic Studies

We computed the statistical power for an SNP explaining 1% of the true phenotypic variance (at the latent trait level) with an allele frequency 0.5, as previously applied in van den Berg et al. (2014). The simulation of item data was based on parameter estimates obtained from the empirical data of this study. We simulated 10,000 datasets, each with a sample size of 1330 participants.

In simulation scenario I, we obtained meta-analysis results based on the sum scores of only CBCL or only SDQ items. We are then not able to meaningfully compare unstandardized regression coefficients because they are obtained by using different measurement instruments. Instead, we need to calculate standardized regression coefficients. We conducted a fixed-effects meta-analysis to combine two standardized regression coefficients (one based on SDQ items and one based on CBCL items) in order to get one standardized regression coefficient and its *p*-value. We used the slope coefficient as a measure of effect size (Hedges & Vevea, 1998). When it comes to the *p*-value, we used a *p*-value smaller than 10^{-8} as a threshold for genome-wide significance (van den Berg et al., 2014; Zhang, 2016), and we calculated the proportion of statistically significant *p*-values as a measure of the statistical power of the results.

The power of the results based on the sum scores was compared with the statistical power of the results based on harmonized scores (simulation scenario II). The harmonization was done by using the item parameters from the empirical part of this paper and estimating expected a posteriori (EAP) estimates for each individual with either SDQ or CBCL data. The availability of harmonized scores allows us to combine *un*standardized regression coefficients (slopes) in a fixed-effect meta-analysis (Hedges & Vevea, 1998) and to calculate the power of the results. We used the "rma.uni" function from the "metafor" package in R (Viechtbauer, 2010), in order to get effect size based on combining two unstandardized regression coefficients. After obtaining the effect size, we calculated its *p*-value. We used the proportion of statistically significant *p*-values (smaller than 10^{-8}) as a measure of the statistical power of the results.

Results

Psychometric Analysis

Top-Down Approach

Anxiety/Depression Problems There were in total 36 items in the anxiety/depression scale (31 CBCL and 5 SDQ items; Table 1). We carried out an IRT analysis on the 31 CBCL items to investigate the psychometric quality of the original anxiety/depression scale consisting only of the CBCL items. Supplementary Table 1 shows the item parameters. We looked at the chi-square statistics for item fit and observed there are no items with statistically significant *p*-values

(Supplementary Table 1). Based on that, we can conclude that the original anxiety/depression scale consisting only of the CBCL items is a good scale for this cohort of children.

Second, we carried out an IRT analysis on the anxiety/depression scale consisting only of the 5 SDQ items (Supplementary Table 2). We looked at the chi-square statistics and observed that there are no items with statistically significant *p*-values (Supplementary Table 2). We can conclude that it is a good scale for this cohort of children.

Therefore, we can take these 31 CBCL items and 5 SDQ items as a combined subset and investigate the psychometric quality of the combined scale using IRT. Before doing that, we used NLP to identify items that are similarly worded in both scales. The semantic similarity analysis showed that three items refer to the same feelings or behaviours (worries, nervousness, and unhappiness) in both the CBCL and SDQ scales. In all of the three cases, the degree of the similarity was higher than 0.90 (on a scale from 0 to 1). In the case when we have two almost identical or similarly worded items in both scales, we need to exclude one of them to prevent psychometric problems. We decided to exclude the CBCL items in the case of similarity because there are only 5 SDQ items in the combined scale, and their exclusion will lead to loss of a large amount of the information.

After excluding 3 CBCL items due to similarity with SDQ items, we carried out a psychometric analysis with 28 CBCL and 5 SDQ items combined. When it comes to the chi-square statistics, the results showed that there are no items with statistically significant *p*-values (Table 5). We concluded that the combined anxiety/depression scale consisting of both CBCL and SDQ items is a good scale for this cohort of children.

Table 4 ADHD – Comparisons between scales from different approaches (Top-down and Bottom-up)

Top-down	Bottom-up
CBCL acts too young*	CBCL acts too young*
CBCL can't concentrate*	CBCL can`t concentrate*
CBCL restless, hyperactive*	CBCL restless, hyperactive*
CBCL confused	CBCL impulsive*
CBCL day dreams	CBCL poor school work*
CBCL impulsive*	CBCL poorly coordinated*
CBCL nervous or tense	CBCL speech problem
CBCL nervous movements	
CBCL poor school work*	
CBCL poorly coordinated*	
CBCL stares blankly	

Note: * – Common items

ADHD There were in total of 16 items in the combined ADHD scale (11 CBCL and 5 SDQ items; Table 1). First, we carried out the IRT analysis on the 11 CBCL items. We looked at the chi-square statistics and observed that there is only one item with a statistically significant *p*-value (Supplementary Table 3). We plotted and analysed the item fit curve (Supplementary Fig. 1) to investigate the problem for this particular item. It shows the relationship between expected and observed category counts (proportions) of the item. The item fit curve showed that observed responses follow expected patterns and there is no unusual behaviour seen in them. We can conclude that the ADHD scale consisting only of the CBCL items is a good scale.

Second, we conducted a psychometric analysis of the 5 SDQ items. The results of the chi-square statistics showed that there is only 1 item with a statistically significant *p*-value (Supplementary Table 4). We plotted and analysed the item fit curve (Supplementary Fig. 2) for this item. It showed that observed responses show only minor deviations from expected proportions (i.e., less than 0.1 proportion points). Accordingly, we can conclude that the ADHD scale consisting only of the SDQ items is a good scale.

We used NLP to examine the semantic similarity between the CBCL and SDQ items. The results showed that two items refer to restless behaviour, one in each scale, with the level of semantic similarity higher than 0.90. There are 11 CBCL items and only 5 SDQ items, so we decided to exclude the CBCL item.

After that, we carried out the analysis on the 10 CBCL and 5 SDQ items. The chi-square statistics showed one item with a statistically significant result ("SDQ restless"; Table 6). We plotted and analysed the item fit curve, and the results showed that observed responses follow expected patterns and there is no unusual behaviour seen in them (Supplementary Fig. 3). We concluded that the combined ADHD scale consisting of the CBCL and SDQ items is a good scale for this cohort of children.

Bottom-Up Approach

Factor Analysis We performed an exploratory factor analysis using all 145 CBCL and SDQ items in the dataset. We computed the eigenvalues and made a scree plot (Cattell & Vogelmann, 1977). Based on the scree plot, we chose a 2-factor solution (see Supplementary Fig. 4). We then performed a factor analysis with 2 factors. There are a lot of studies that show there is a statistically significant correlation between ADHD and anxiety/depression (e.g., Faraone et al., 2019; Tenenbaum et al., 2019). An oblique rotation should be used in the case when a correlation is expected between factors. We therefore used the oblique Promax rotation method. Factor 1 showed 53 items with factor loadings higher than 0.3 or lower than -0.3 (Supplementary Table 5):

Table 5 Original anxiety/depression scale consisting of the CBCL and SDQ items (after excluding similar items)

Item	Discrimination parameter	Threshold parameter Category 1	Threshold parameter Category 2	Chi-square	Degrees of Freedom	Corrected p-value
CBCL lonely	1.28	-2.22	-6.27	24.23	24	1.00
CBCL cries a lot	1.21	-2.56	-6.19	21.45	24	1.00
CBCL fears might do bad	1.22	-2.79	-7.84	32.10	25	1.00
CBCL fears has to be perfect	0.84	-1.26	-3.89	28.92	43	1.00
CBCL feels unloved	1.12	-1.66	-5.39	36.49	28	1.00
CBCL feels others out to get	1.57	-2.77	-7.46	24.04	24	1.00
CBCL feels worthless	1.87	-2.78	-8.20	26.24	24	1.00
CBCL loner	1.02	-2.15	-5.60	24.59	25	1.00
CBCL nervous or tense*	1.47	-2.11	-6.08			
CBCL too fearful or anxious	1.78	-2.89	-7.79	35.82	25	1.00
CBCL feels dizzy	1.36	-3.75	-9.28	19.46	26	1.00
CBCL feels too guilty	1.84	-4.16	-10.22	12.44	24	1.00
CBCL overtired	1.27	-2.15	-6.16	33.51	24	1.00
CBCL no cause aches/pains	0.85	-1.31	-4.61	42.58	34	1.00
CBCL no cause headaches	0.78	-1.07	-4.06	38.26	38	1.00
CBCL no cause nausea	1.19	-2.14	-6.55	21.54	24	1.00
CBCL no cause eye problems	0.62	-3.84	-5.78	14.59	19	1.00
CBCL no cause skin problems	0.64	-2.25	-5.08	13.03	26	1.00
CBCL no cause stomachaches	0.91	-1.27	-4.83	25.53	28	1.00
CBCL no cause vomiting	0.92	-3.24	-7.44	21.45	24	1.00
CBCL refuses to talk	1.33	-3.23	-7.70	13.16	25	1.00
CBCL secretive	1.20	-1.74	-5.95	21.63	23	1.00
CBCL self-conscious	1.26	-0.28	-3.63	37.65	41	1.00
CBCL shy or timid	1.00	-0.95	-4.27	43.61	37	1.00
CBCL stares blankly	1.29	-3.14	-8.15	29.92	25	1.00
CBCL sulks a lot	1.18	-1.73	-5.13	35.00	35	1.00
CBCL suspicious	1.64	-3.98	-9.51	32.47	24	1.00
CBCL lacks energy/slow	1.25	-3.18	-6.70	40.98	25	0.76
CBCL unhappy/depressed*	2.15	-3.13	-9.93			
CBCL withdrawn	1.86	-3.91	-10.46	43.26	24	0.31
CBCL worries*	1.84	-1.26	-6.16			
SDQ complains of illness	0.91	-0.84	-3.21	54.20	46	1.00
SDQ often seems worried	1.42	-1.14	-4.36	62.95	39	0.29
SDQ often unhappy/tearful	1.33	-2.09	-5.92	35.98	30	1.00
SDQ nervous or clingy	0.98	-0.63	-3.30	47.67	45	1.00
SDQ easily scared	1.17	-1.70	-4.80	43.82	42	1.00

Note: *- Items excluded from the combined scale due to similarity with SDQ items to prevent psychometric problems. Their item parameters are predicted through linear regression to keep the CBCL scale intact and to avoid information loss in the data harmonization process

38 CBCL items and 15 SDQ items. Factor 2 had 39 items with factor loadings higher than 0.3 or lower than -0.3 (Supplementary Table 5): 33 CBCL items and 6 SDQ items. The sets of items showed no overlap. To be more precise, there are no items that have factor loadings higher than 0.3 or lower than -0.3 on both factors. If we carefully interpret these factors, we can see that they can be interpreted as an Externalizing problem behaviour (Factor 1) and Internalizing problems behaviour (Factor 2) dimensions.

The results did not yield an ADHD type of dimension. Therefore, the number of factors was doubled in a second Promax factor analysis. Again, the same selection was made based on the factor loadings. Two of these four scales were identified as ADHD and anxiety/depression.

Factor 1 (anxiety/depression) showed 38 items with factor loadings higher than 0.3 or lower than -0.3 (Table 2; for details see Supplementary Table 6), consisting of 33 CBCL items and 5 SDQ items. Factor 4 (ADHD) showed

Table 6 Original ADHD scale consisting of both CBCL and SDQ items (excluded similar items)

Item	Discrimination parameter	Threshold parameter Category 1	Threshold parameter Category 2	Chi-square	Degrees of Freedom	Corrected p-value
CBCL acts too young	1.28	-1.82	-5.39	15.03	28	1.00
CBCL can't concentrate	3.71	-2.07	-8.22	16.59	19	1.00
CBCL restless, hyperactive*	1.31	-1.51	-4.80			
CBCL confused	1.71	-3.59	-8.63	13.20	17	1.00
CBCL day dreams	1.11	-1.15	-3.96	34.28	32	1.00
CBCL impulsive	1.58	-0.88	-4.45	35.25	28	1.00
CBCL nervous or tense	0.65	-1.69	-4.26	35.12	24	1.00
CBCL nervous movements	0.83	-3.53	-5.36	27.03	23	1.00
CBCL poor school work	1.84	-2.49	-6.52	19.30	27	1.00
CBCL poorly coordinated	1.19	-2.77	-6.27	26.34	23	1.00
CBCL stares blankly	1.19	-3.08	-8.05	36.53	18	0.09
SDQ restless	1.41	-0.89	-3.72	54.64	29	0.04
SDQ constantly fidgeting	1.34	-1.16	-3.92	34.07	30	1.00
SDQ easily distracted	3.24	0.15	-4.02	13.07	19	1.00
SDQ thinks before acting	1.44	1.58	-1.28	45.80	25	0.10
SDQ good attention span	1.99	0.70	-2.37	23.35	23	1.00

Note: *- Item excluded from the combined scale due to similarity with SDQ item to prevent psychometric problems. Its item parameters are predicted through linear regression to keep the CBCL scale intact and to avoid information loss in the data harmonization process

13 items with factor loadings higher than 0.3 or lower than -0.3 (Table 2; for details see Supplementary Table 6). This factor consisting of 8 CBCL items and 5 SDQ items. We observed that two items had negative factor loadings: "SDQ thinks before acting" and "SDQ good attention span". We observe that both are formulated such that a lower score on the item can be interpreted as a higher trait score of the ADHD. Based on that, we recoded these items such that a higher score on the scale reflects a higher trait score for ADHD.

Anxiety/Depression For the anxiety/depression factor, we observed that there were 38 items in total (33 CBCL items and 5 SDQ items) with factor loadings higher than 0.3 or lower than -0.3 (Table 2; for details see Supplementary Table 6).

First, we carried out an IRT analysis on the anxiety/depression scale consisting only of the CBCL items to investigate the psychometric quality of this scale. This scale consisting of 33 CBCL items. We looked at the chi-square statistics and observed that there are no items with statistically significant *p*-values (see Supplementary Table 7) and concluded that it is a good scale for this cohort of children.

Similarly, for the SDQ scale, we carried out the analysis for all the 5 items originally from the SDQ. We carried out the analysis and looked at the item fit and at the discrimination parameter value. We looked at the chi-square statistics and observed that there were no items with statistically significant *p*-values (see Supplementary Table 8). Based on that, we concluded that it is a good scale for this cohort of children. Before combining the CBCL and SDQ items, we conducted semantic similarity analysis and observed that three items refer to the same feelings or behaviours (worries, nervousness, and unhappiness) in both the CBCL and SDQ scales. In all of the three cases, the degree of the semantic similarity was higher than 0.90 (on a scale from 0 to 1). To prevent psychometric problems, we decided to exclude the CBCL items because of the already small number of SDQ items (there are 33 CBCL items and only 5 SDQ items).

After omitting these items, we conducted a psychometric analysis on the combined anxiety/depression scale consisting of the 35 items (30 CBCL items and 5 SDQ items). The chi-square fit statistics showed that there were no items with statistically significant *p*-values (Supplementary Table 9). We concluded that the combined anxiety/depression scale consisting of both CBCL and SDQ items is a good scale for this cohort of children.

ADHD For the ADHD factor, we observed that there were 13 items in total with factor loadings higher than 0.3 or lower than -0.3 (8 CBCL items and 5 SDQ items; Table 2; for more details see Supplementary Table 6).

We carried out the IRT analysis on the 8 CBCL items to investigate the psychometric quality of the ADHD scale consisting only of the CBCL items. We looked at the chi-square statistics and observed that the *p*-values of 2 out of 8 items were significant (Supplementary Table 10). The item fit curves showed that observed responses follow expected

patterns and there was no unusual behaviour seen for item "CBCL restless, hyperactive" (Supplementary Fig. 5), but there was for item "CBCL day dreams" (Supplementary Fig. 6). There we see that the differences between observed and expected proportions of responses are often clearly larger than 0.1. Accordingly, we decided to exclude this item from the scale. After excluding this item, we conducted a psychometric analysis on the ADHD scale consisting of the remaining 7 CBCL items. When it comes to the chi-squares, we observed that 1 out of 7 items had a statistically significant *p*-value (Supplementary Table 11). That is the item "CBCL restless, hyperactive". The item fit curve showed that observed responses follow expected patterns and there was no unusual behaviour seen in the item "CBCL restless, hyperactive" (Supplementary Fig. 7). Accordingly, we concluded that the ADHD scale consisting only of the CBCL items is a good scale for this cohort of children.

Similarly, for the SDQ scale, we carried out the analysis for all 5 SDQ items. We carried out the analysis and looked at the item fit and at the discrimination parameter value. The chi-square fit statistics showed that the *p*-value of 1 out of 5 items was statistically significant ("SDQ good attention span"; Supplementary Table 12). The item fit curve (Supplementary Fig. 8) showed that observed responses follow expected patterns and there was no unusual behaviour seen in the item "SDQ good attention span". We concluded that the ADHD scale consisting only of the SDQ items is a good scale for this cohort of children.

Further, we can combine the SDQ items with CBCL items to create a combined ADHD scale. Before combining, we conducted semantic similarity analysis and observed that the item "SDQ restless" is very similar to the item "CBCL restless, hyperactive", with the degree of similarity higher than 0.90. To prevent psychometric problems, we decided to exclude the CBCL item because of the smaller number of SDQ items in the combined scale (there are 7 CBCL items and 5 SDQ items in the combined scale).

After that, we conducted a psychometric analysis on the combined scale consisting of the 11 items (6 CBCL items and 5 SDQ items). We looked at the chi-square statistics and observed that there were no items with statistically significant *p*-values (Supplementary Table 13). Therefore, we concluded that the ADHD scale consisting of both the CBCL and SDQ items is a good scale for this cohort of children.

Evaluation of the Scales

The results showed that anxiety/depression and ADHD scales consisting only of CBCL or only of SDQ items are good scales, as well as combined scales (consisting of the CBCL and SDQ items), both in the top-down and bottom-up approaches. The data-driven (bottom-up) approach yielded scales very similar to those from the theoretically (top-down) approach. The SDQ anxiety/depression and ADHD scales from the top-down and bottom-up approaches are completely the same. When it comes to CBCL scales, content analysis of items showed that in the case of anxiety/depression both scales have a similar number of items (top-down – 31 items, bottom-up – 33 items) while most of them (23) are common items (Table 3). Regarding ADHD, analysis of the items showed that the bottom-up approach scale consists of 7 items, and 6 of them are also present in the top-down approach scale (Table 4). We can conclude that the bottom-up approach validates the theory-driven, top-down approach. Accordingly, we advise using the top-down approach scales in both ADHD and anxiety/depression for data harmonization in other studies.

To prevent psychometric problems, some CBCL items were excluded from the combined scale in the top-down approach because the semantic similarity analysis showed that they had similar wording as SDQ items. Accordingly, the researchers who want to put CBCL items on the harmonized scale will not have item parameters for excluded items and that can lead to information loss, and the reliability of the scale can be jeopardized. We plotted test information curves based on all CBCL items and based only on CBCL items from the combined scale (Supplementary Figs. 9 and 10) to investigate the impact of exclusion of these items on test information (reliability). In those figures, we can see three lines: black (all CBCL items), blue (CBCL items after exclusion of semantically similar or very similar items using item parameters from a model consisting only of these CBCL items), and red (CBCL items after exclusion of semantically same or very similar items using item parameters from combined scale). Furthermore, from this figure, we can observe two different types of effects. The first type of effect is a consequence of the exclusion of CBCL items which are semantically same or very similar to some SDQ items and it can be observed as a difference between the black and blue lines, while the second type of effect is a consequence of a change of the remaining CBCL item parameters after addition of SDQ items in the combined scale and it can be observed as a difference between the blue and red lines.

In the case of anxiety/depression, we can clearly see the information loss after excluding 3 CBCL items, but because of the high number of remaining items, the reliability is still acceptable (Supplementary Fig. 9), and the correlation between harmonized scores based on item parameters from these two versions of scales (CBCL all items and CBCL items from combined scale) is still very high: 0.98. Besides that, the results showed that the addition of SDQ items in the combined scale did not lead to large changes in CBCL item parameters (Supplementary Fig. 9).

Regarding ADHD, the exclusion of 1 CBCL item which is semantically very similar to one SDQ item leads to loss of information. The reliability is still acceptable, but because

of the small number of items, even the exclusion of one item affects it (Supplementary Fig. 10). The Pearson coefficient of correlation between harmonized scores based on item parameters from these two scale versions (CBCL all items and CBCL items from combined scale) is still very high: 0.99.

In summary, the reliability of the scale is still acceptable in the case of both anxiety/depression and ADHD, but in both cases, there is a loss of information after the exclusion of similar items.

The solution to the problem with the loss of information lies in using item parameters of excluded items to prevent information loss and to keep reliability intact. In that case, researchers will need item parameters of CBCL items that are excluded from the combined scale due to semantic similarity. For that reason, we have tried to predict them based on a linear transformation. To check whether such a linear transformation would be reasonable, we created scatterplots to examine the changes in item parameters on the scale with only CBCL items after combining them with SDQ items. On the scatter plot (Supplementary Figs. 11 and 12) we see two lines: one regression line, describing the linear relationship between item parameters before and after the combination with SDQ, and a 45-degree line (intercept 0 and slope 1) that represents the perfect world where the item parameters on one scale are the same as on the combined scale. If there are no changes in item parameters of CBCL items after combining them with SDQ items, researchers can simply use existing item parameters from the initial scale (scale consisting of *all* CBCL items) for CBCL items which are excluded from the combined scale. In contrast, if the item parameters from the initial and combined scale cannot be described with a 45-degree line, we can use the linear equation from regression analysis to 'predict' item parameter values of CBCL items that are excluded on the combined scale.

We observed changes in item parameters of CBCL items after combining them with SDQ items in the case of both anxiety/depression and ADHD (i.e., the dots are not on the 45-degree line; Supplementary Figs. 11 and 12). For both phenotypes, the discrimination parameters are changed which leads to changes in threshold parameters also. These changes are not large, but they are statistically significant in most of the cases (i.e., the 45-degree line is not in the 95% confidence region for the regression line), but they still seem linear. Accordingly, we cannot advise using item parameters from the initial scale (scale consisting of all CBCL items) for CBCL items which are excluded from the combined scale, because there are changes in item parameters of CBCL items after combining them with SDQ items. Instead of that, we can use linear regression in which CBCL item parameters from the initial scale are predictors, while CBCL item parameters from the combined scale are criterion variables. We calculated regression coefficients and based on these regression coefficients and item parameters from the initial scale (Supplementary Tables 1 and 3) we predicted item parameters of CBCL items excluded from the combined scale (Tables 5 and 6). This way, researchers will have item parameters for all CBCL items: they can use all relevant CBCL items, and they will not have a problem with information loss due to exclusion of some items as they would in the previous option.

Increasing Statistical Power of the Results Using Harmonized Scores in Behaviour Genetic Studies

The results showed that the statistical power to detect a SNP at the genome-wide significance level explains 1% of the true phenotypic variance with an allele frequency of 0.5 when using sum scores was 9.54% in the case of ADHD (N = 1330; scenario I). The use of harmonized scores instead of sum scores changed statistical power to 23.64% (N = 1330; scenario II). In the case of both harmonized and sum scores, statistically significant effect sizes were positive, that is, they were in the same direction.

Discussion

This study aimed to construct common metrics for anxiety, depression, and ADHD as measured by the Child Behaviour Checklist (CBCL) and Strengths and Difficulties Questionnaire (SDQ). We used a top-down (theoretical) and bottom-up (data-driven) approach. In the top-down approach, we used existing scales related to anxiety/depression and ADHD, while in the bottom-up approach we conducted a factor analysis to identify anxiety/depression and ADHD scales and to examine whether the theoretical (top-down) approach is supported by the data-driven (bottom-up) approach.

When it comes to the theoretical approach, existing anxiety/depression scales consisting only of the CBCL, and only of the SDQ items showed good measurement properties in the Raine data set. Also, a combined anxiety/depression scale consisting of both CBCL and SDQ items is a good scale. In the case of ADHD, separate CBCL and SDQ scales, as well as combined scale consisting of both CBCL and SDQ items, are good quality scales.

Regarding the bottom-up (data-driven) approach, the psychometric analysis showed that all the items in the anxiety/depression scale consisting only of the CBCL, and only of the SDQ items show a good fit. Both of the scales, analysed separately, are good. The psychometric analysis showed that the items from the scale consisting of both the CBCL and the SDQ items have good item fit. We concluded that CBCL and SDQ items form a good-quality scale that operationalizes anxiety/depression. When it comes to ADHD, the scale consisting only of the CBCL items show a good fit as well as the scale consisting only of SDQ items. The results showed

that the combined ADHD scale consisting of both the CBCL and SDQ items is good, too.

A comparison between the top-down and bottom-up scales showed a large overlap. The top-down approach where we started from existing subscales is therefore supported by a purely data-driven approach. Accordingly, we can advise using the top-down approach scales in both anxiety/depression and ADHD for data harmonization in other studies (Tables 3 and 4). Based on these item parameters, EAP estimates can be calculated for each participant which can function as harmonized scores.

One problem we stumbled on is that some items are very similar across questionnaires. For psychometric reasons, such items had to be deleted from one of the scales (CBCL). This, however, leads to loss of information: content-wise as well as in the number of items, affecting both validity and reliability. We showed that an approach where latent trait levels are estimated using parameters only of CBCL items from the combined scale (without the use of CBCL items that are excluded due to psychometric reasons) leads to loss of information. Accordingly, we devised an approach using linear regression to overcome this problem. We obtained item parameters of excluded items using a linear transformation so that other researchers also have item parameters for these items (Tables 3 and 4). This restores both reliability and validity.

The results also showed that the use of harmonized scores solves one of the biggest problems in the meta-analysis of behavioural measures. Namely, one of the main difficulties in the meta-analytic approach is the use of different measurement instruments across studies for assessing a particular phenotype (van den Berg et al., 2014). Consequently, effect sizes cannot be meaningfully compared across studies (van den Berg & de Moor, 2020). Instead, researchers need to use standardized regression coefficients or *p*-values, but that leads to loss of information about the absolut size of the effect, in the case of standardized regression coefficients, and both the absolut size and direction of the effect, in the case of *p*-values. The use of harmonized scores solves this problem because it allows using of unstandardized regression coefficients as a measure of the effect size. They provide us with information on both the direction and size of the effect, that is, they allow meaningful comparison of the absolute effect sizes across studies. In addition, the previous studies showed that the use of harmonized scores can increase the size of the sample and, accordingly, statistical power of the results (e.g. van den Berg et al., 2014). In this study, we showed that the use of harmonized scores leads to the higher statistical power of the results than the use of sum scores, even if the size of the sample is the same in both cases.

These findings can help future researchers to harmonize data from different samples and/or different questionnaires. The findings can be useful in various ways, both in CAPICE and other research consortia. It can be helpful for researchers to combine data from different groups of respondents with different questionnaires to obtain the larger sample sizes, to be able to compare research results across subpopulations or to increase generalizability, the validity or statistical power of research results (Fortier et al., 2010, 2011; Hamilton et al., 2011; Smith-Warner et al., 2006; Thompson, 2009; van den Berg et al., 2014). For example, researchers can use CBCL item parameters from a combined scale to estimate latent trait levels among participants who filled in only CBCL items, and they can use SDQ item parameters from a combined scale to estimate latent trait levels for participants who filled in only SDQ items. Furthermore, based on those estimations they can compare results between persons who filled in only CBCL or only SDQ items. In the situation where they have participants` answers on both questionnaires, they can use the results of this study to estimate participants` latent trait levels based on both questionnaires together, thereby increasing the measurement reliability. These findings can also be used in longitudinal studies where CBCL data are gathered at age *x* and SDQ data are gathered at age *y*.

In the supplementary material, we describe how to use these results in practice. The procedure is very simple and a detailed example R script is provided that shows how item responses can be used to obtain harmonized scores for anxiety/depression and ADHD on the respective common metrics using the Computerized Adaptive Testing with Multidimensional Item Response Theory (mirtCAT) package (Chalmers, 2016) of R. In Supplementary Table 16, we presented harmonized scores for various example data vectors. For instance, response pattern A includes only CBCL items with a total sum score of 6. Response pattern B includes only SDQ items with the same sum score of 6. However, based on our combined scale, we see that the scores on the common metric are different, with also different standard errors (reliability).

One important limitation of the current study is that the quality of the harmonization depends on the extent that there is measurement invariance across different populations: other cohorts from other countries and using different languages, and different ages. Future research should compare item parameters from different cohorts, to determine to what extent the results from this harmonization effort extend beyond Australian 10-year-olds.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10862-021-09925-9.

Acknowledgements We are grateful to the Raine Study participants and their families and we thank the Raine Study team for cohort coordination and data collection. The core management of the Raine Study is funded by The University of Western Australia, Curtin University, Telethon Kids Institute, Women and Infants Research Foundation, Edith Cowan University, Murdoch University, The University of Notre Dame Australia, and the Raine Medical Research Foundation. AJOW is supported by an Investigator Grant from the National Health and Medical Research Council (APP1173896).

Funding This work has been supported by the CAPICE project, funded by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Sklodowska—Curie grant agreement no. 721567.

Declarations

Ethical Approval The broader Raine Study has ethics approval from The University of Western Australia Human Research Ethics Committee.

Informed Consent Informed consent was provided by all participants in the Raine study. Participant assent and parental consent was provided for minors. The population data used in this study was provided as de-identified data.

Conflict of Interest Miljan Jović, Kratika Agarwal, Andrew Whitehouse, and Stéphanie M. van den Berg declare no conflicts of interest.

Experiment Participants All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4-18 and 1991 profile*. University of Vermont.

Achenbach, T. M., Howell, C. T., Quay, H. C., Conners, C. K., & Bates, J. E. (1991). National survey of problems and competencies among four- to sixteen-year-olds: Parents' reports for normative and clinical samples. *Monographs of the Society for Research in Child Development, 1*, i–130. https://doi.org/10.2307/1166156

Achenbach, T. M., & Ruffle, T. M. (2000). The Child Behavior Checklist and related forms for assessing behavioral/emotional problems and competencies. *Pediatrics in Review, 21*(8), 265–271. https://doi.org/10.1542/pir.21-8-265

Allen, K., & Prior, M. (1995). Assessment of the validity of easy and difficult temperament through observed mother-child behaviours. *International Journal of Behavioral Development, 18*(4), 609–630. https://doi.org/10.1177/016502549501800403

Andrich D., & Marais, I. (2019). A course in Rasch measurement theory. *Measuring in the Educational, Social and Health Sciences*. Springer, Singapore.

Anxiety and Depression Association of America: ADAA. (2020). Anxiety and Depression in children. Available from https://adaa.org/living-with-anxiety/children/anxiety-and-depression. Accessed 6 May 2020.

Begum, F., Ghosh, D., Tseng, G. C., & Feingold, E. (2012). Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Research, 40*(9), 3777–3784. https://doi.org/10.1093/nar/gkr1255

Caspi, A., Henry, B., McGee, R. O., Moffitt, T. E., & Silva, P. A. (1995). Temperamental origins of child and adolescent behavior problems: From age three to age fifteen. *Child Development, 66*(1), 55–68. https://doi.org/10.1111/j.1467-8624.1995.tb00855.x

Cattell, R. B., & Vogelmann, S. (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research, 12*(3), 289–325. https://doi.org/10.1207/s15327906mbr1203_2

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications. *Journal of Statistical Software, 71*(5), 1–39. https://doi.org/10.18637/jss.v071.i05

Chivers, P., Hands, B., Parker, H., Bulsara, M., Beilin, L. J., Kendall, G. E., & Oddy, W. H. (2010). Body mass index, adiposity rebound and early feeding in a longitudinal cohort (Raine Study). *International Journal of Obesity, 34*(7), 1169–1176. https://doi.org/10.1038/ijo.2010.61

Danielson, M. L., Bitsko, R. H., Ghandour, R. M., Holbrook, J. R., Kogan, M. D., & Blumberg, S. J. (2016). Prevalence of parent-reported ADHD diagnosis and associated treatment among US children and adolescents. *Journal of Clinical Child & Adolescent Psychology, 47*(2), 199–212. https://doi.org/10.1080/15374416.2017.1417860

De Moor, M. H., Van Den Berg, S. M., Verweij, K. J., Krueger, R. F., Luciano, M., Vasquez, A. A., … & Boomsma, D. I. (2015). Meta-analysis of genome-wide association studies for neuroticism, and the polygenic association with major depressive disorder. *JAMA Psychiatry, 72*(7), 642–650. https://doi.org/10.1001/jamapsychiatry.2015.0554

Dulcan, M. (1997). Practice parameters for the assessment and treatment of children, adolescents, and adults with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 36*(10), 85S-121S.

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Lawrence Eribaum Associates. Inc., Publishers.

Faraone, S. V., Rostain, A. L., Blader, J., Busch, B., Childress, A. C., Connor, D. F., & Newcorn, J. H. (2019). Practitioner Review: Emotional dysregulation in attention-deficit/hyperactivity disorder–implications for clinical recognition and intervention. *Journal of Child Psychology and Psychiatry, 60*(2), 133–150. https://doi.org/10.1111/jcpp.12899

Fortier, I., Burton, P. R., Robson, P. J., Ferretti, V., Little, J., L'heureux, F., … & Hudson, T. J. (2010). Quality, quantity and harmony: The DataSHaPER approach to integrating data across bioclinical studies. *International Journal of Epidemiology, 39*(5), 1383–1393. https://doi.org/10.1093/ije/dyq139

Fortier, I., Doiron, D., Little, J., Ferretti, V., L'Heureux, F., Stolk, R. P., … & Burton, P. R. (2011). Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. *International Journal of Epidemiology, 40*(5), 1314–1328. https://doi.org/10.1093/ije/dyr106

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry, 38*(5), 581–586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry, 40*(11), 1337–1345. https://doi.org/10.1097/00004583-200111000-00015

Guenther, F., Dudschig, C., & Kaup, B. (2015). LSAfun: An R package for computations based on Latent Semantic Analysis. *Behavior Research Methods, 47*(4), 930–944. https://doi.org/10.3758/s13428-014-0529-0

Hamilton, C. M., Strader, L. C., Pratt, J. G., Maiese, D., Hendershot, T., Kwok, R. K., Hammond, J. A., Huggins, W., Jackman, D., Pan, H., & Nettles, D. S. (2011). The PhenX Toolkit: Get the most from your measures. *American Journal of Epidemiology, 174*(3), 253–260. https://doi.org/10.1093/aje/kwr193

Haro, J. M., Ayuso-Mateos, J. L., Bitter, I., Demotes-Mainard, J., Leboyer, M., Lewis, S. W., … & Walker-Tilley, T. (2014). ROAMER: Roadmap for mental health research in Europe. *International Journal of Methods in Psychiatric Research, 23*(S1), 1–4. https://doi.org/10.1002/mpr.1406

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486–504.

Hettema, J. M., Prescott, C. A., Myers, J. M., Neale, M. C., & Kendler, K. S. (2005). The structure of genetic and environmental risk factors for anxiety disorders in men and women. *Archives of General Psychiatry, 62*(2), 182–189. https://doi.org/10.1001/archpsyc.62.2.182

Howard, A. L., Robinson, M., Smith, G. J., Ambrosini, G. L., Piek, J. P., & Oddy, W. H. (2011). ADHD is associated with a "Western" dietary pattern in adolescents. *Journal of Attention Disorders, 15*(5), 403–411. https://doi.org/10.1177/1087054710365990

Hussong, A. M., Curran, P. J., & Bauer, D. J. (2013). Integrative data analysis in clinical psychology research. *Annual Review of Clinical Psychology, 9*, 61–89. https://doi.org/10.1146/annurev-clinpsy-050212-185522

Karam, R. G., Breda, V., Picon, F. A., Rovaris, D. L., Victor, M. M., Salgado, C. A. I., ... & Caye, A. (2015). Persistence and remission of ADHD during adulthood: A 7-year clinical follow-up study. *Psychological Medicine, 45*(10), 2045–2056. https://doi.org/10.1017/S0033291714003183

Kendler, K. S., Myers, J. M., Maes, H. H., & Keyes, C. L. (2011). The relationship between the genetic and environmental influences on common internalizing psychiatric disorders and mental well-being. *Behavior Genetics, 41*(5), 641–650. https://doi.org/10.1007/s10519-011-9466-1

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking* (2nd ed.). Springer.

Löwe, B., Spitzer, R. L., Williams, J. B., Mussell, M., Schellberg, D., & Kroenke, K. (2008). Depression, anxiety and somatization in primary care: Syndrome overlap and functional impairment. *General Hospital Psychiatry, 30*(3), 191–199.

Luningham, J. M., McArtor, D. B., Hendriks, A. M., van Beijsterveldt, C. E. M., Lichtenstein, P., Lundström, S., … & Lubke, G. H. (2019). Data Integration Methods for Phenotype Harmonization in Multi-Cohort Genome-Wide Association Studies With Behavioral Outcomes. *Frontiers in Genetics, 10*, 1227. https://doi.org/10.3389/fgene.2019.01227

McKnight, C. M., Newnham, J. P., Stanley, F. J., Mountain, J. A., Landau, L. I., Beilin, L. J., … & Mackey, D. A. (2012). Birth of a cohort—the first 20 years of the Raine study. *Medical Journal of Australia, 197*(11), 608.

Muris, P., Meesters, C., & van den Berg, F. (2003). The strengths and difficulties questionnaire (SDQ). *European Child & Adolescent Psychiatry, 12*(1), 1–8. https://doi.org/10.1007/s00787-003-0298-2

Nadder, T. S., Silberg, J. L., Eaves, L. J., Maes, H. H., & Meyer, J. M. (1998). Genetic effects on ADHD symptomatology in 7-to 13-year-old twins: Results from a telephone survey. *Behavior Genetics, 28*(2), 83–99. https://doi.org/10.1023/A:1021686906396

Newnham, J. P., Evans, S. F., Michael, C. A., Stanley, F. J., & Landau, L. I. (1993). Effects of frequent ultrasound during pregnancy: A randomised controlled trial. *Lancet (london, England), 342*(8876), 887–891. https://doi.org/10.1016/0140-6736(93)91944-h

Ortuno-Sierra, J., Chocarro, E., Fonseca-Pedrero, E., i Riba, S. S., & Muñiz J. (2015). The assessment of emotional and behavioural problems: Internal structure of the Strengths and Difficulties Questionnaire. *International Journal of Clinical and Health Psychology, 15*(3), 265–273. https://doi.org/10.1016/j.ijchp.2015.05.005

Park, J. Y., Cornillie, F., van der Maas, H. L., & Van Den Noortgate, W. (2019). A multidimensional IRT approach for dynamically monitoring ability growth in computerized practice environments. *Frontiers in Psychology, 10*, 620. https://doi.org/10.3389/fpsyg.2019.00620

Polanczyk, G., De Lima, M. S., Horta, B. L., Biederman, J., & Rohde, L. A. (2007). The worldwide prevalence of ADHD: a systematic review and metaregression analysis. *American Journal of Psychiatry, 164*(6), 942–948. Available from https://doi.org/10.1176/ajp.2007.164.6.942. Accessed 29 Aug 2020.

Rajula, H. S. R., Manchia, M., Agarwal, K., Akingbuwa, W. A., Allegrini, A., Diemer, E., … & Middeldorp, C. M. (2021). Overview of CAPICE – Childhood and Adolescence Psychopathology: Unravelling the Complex Etiology by a large Interdisciplinary Collaboration in Europe - an EU Marie Skłodowska-Curie International Training Network. *European Child and Adolescent Psychiatry*. https://doi.org/10.1007/s00787-020-01713-2

Schmitz, M., Polanczyk, G., & Rohde, L. A. (2007). ADHD: Remission in adolescence and predictors of persistence into adulthood. *Jornal Brasileiro De Psiquiatria, 56*(1), 25–29.

Silberg, J., Pickles, A., Rutter, M., Hewitt, J., Simonoff, E., Maes, H., … & Eaves, L. (1999). The influence of genetic factors and life stress on depression among adolescent girls. *Archives of General Psychiatry, 56*(3), 225–232. https://doi.org/10.1001/archpsyc.56.3.225

Silberg, J., Rutter, M., Neale, M., & Eaves, L. (2001). Genetic moderation of environmental risk for depression and anxiety in adolescent girls. *The British Journal of Psychiatry, 179*(2), 116–121. https://doi.org/10.1192/bjp.179.2.116

Silove, D., Manicavasagar, V., O'connell, D., & Morris-Yates, A. (1995). Genetic factors in early separation anxiety: Implications for the genesis of adult anxiety disorders. *Acta Psychiatrica Scandinavica, 92*(1), 17–24. https://doi.org/10.1111/j.1600-0447.1995.tb09537.x

Simon, V., Czobor, P., Bálint, S., Mészáros, A., & Bitter, I. (2009). Prevalence and correlates of adult attention-deficit hyperactivity disorder: Meta-analysis. *The British Journal of Psychiatry, 194*(3), 204–211. https://doi.org/10.1192/bjp.bp.107.048827

Smith-Warner, S. A., Spiegelman, D., Ritz, J., Albanes, D., Beeson, W. L., Bernstein, L., … & Hunter, D. J. (2006). Methods for pooling results of epidemiologic studies: The Pooling Project of Prospective Studies of Diet and Cancer. *American Journal of Epidemiology, 163*(11), 1053–1064. https://doi.org/10.1093/aje/kwj127

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. *Journal of Graduate Medical Education, 4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Sullivan, P. F., Daly, M. J., & O'Donovan, M. (2012). Genetic architecture of psychiatric disorders: The emerging picture and its implications. *Nature Reviews Genetics, 13*(8), 537–551. https://doi.org/10.1038/nrg3240

Tenenbaum, R. B., Musser, E. D., Morris, S., Ward, A. R., Raiker, J. S., Coles, E. K., & Pelham, W. E. (2019). Response inhibition, response execution, and emotion regulation among children with attention-deficit/hyperactivity disorder. *Journal of Abnormal Child Psychology, 47*(4), 589–603. https://doi.org/10.1007/s10802-018-0466-y

Thompson, A. (2009). Thinking big: Large-scale collaborative research in observational epidemiology. *European Journal of Epidemiology, 24*(12), 727–731. https://doi.org/10.1007/s10654-009-9412-1

Uto, M., & Ueno, M. (2018). Empirical comparison of item response theory models with rater's parameters. *Heliyon, 4*(5), e00622. https://doi.org/10.1016/j.heliyon.2018.e00622

van den Berg, S. M., & de Moor, M. H. M. (2020). Molecular genetic research on personality. In: Saudino, K., & Ganiban, J. M. (Eds.) Behavior Genetics of Temperament and Personality. New York: Springer-Verlag.

van den Berg, S. M., De Moor, M. H., McGue, M., Pettersson, E., Terracciano, A., Verweij, K. J., … & Boomsma, D. I. (2014). Harmonization of Neuroticism and Extraversion phenotypes across inventories and cohorts in the Genetics of Personality Consortium: An application of Item Response Theory. *Behavior Genetics, 44*(4), 295–313. https://doi.org/10.1007/s10519-014-9654-x

van den Berg, S. M., de Moor, M. H., Verweij, K. J., Krueger, R. F., Luciano, M., Vasquez, A. A, … & Milaneschi, Y. (2016). Meta-analysis of genome-wide association studies for extraversion: Findings from the genetics of personality consortium. *Behavior Genetics, 46*(2), 170–182. https://doi.org/10.1007/s10519-015-9735-5

van den Berg, S. M., Glas, C. A., & Boomsma, D. I. (2007). Variance decomposition using an IRT measurement model. *Behavior Genetics, 37*(4), 604–616. https://doi.org/10.1007/s10519-007-9156-1

van den Berg, S. M., Willemsen, G., de Geus, E. J., & Boomsma, D. I. (2006). Genetic etiology of stability of attention problems in young adulthood. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 141*(1), 55–60. https://doi.org/10.1002/ajmg.b.30251

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Vos, T., Allen, C., Arora, M., Barber, R. M., Bhutta, Z. A., Brown, A., … & Murray, C. J. L. (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: A systematic analysis for the Global Burden of Disease Study 2015. *The Lancet, 388*(10053), 1545–1602. https://doi.org/10.1016/S0140-6736(16)31678-6

Wetzel, E., & Carstensen, C. H. (2014). Reversed thresholds in partial credit models: A reason for collapsing categories? *Assessment, 21*(6), 765–774. https://doi.org/10.1177/1073191114530775

World Health Organisation: WHO. (1992). *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines.* Available from http://apps.who.int/iris/handle/10665/37958. Accessed 6 May 2020.

World Health Organization: WHO. (2017). *Depression and other common mental disorders: global health estimates* (No. WHO/MSD/MER/2017.2). Available from https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf?sequence=1. Accessed 6 May 2020.

World Health Organization: WHO. (2020). Depression. Retrieved from https://www.hhs.gov/answers/public-health-and-safety/what-is-the-difference-between-isolation-and-quarantine/index.html. Accessed 6 May 2020. Retrieved 6 May 2020.

Zeggini, E., & Ioannidis, J. P. (2009). Meta-analysis in genome-wide association studies. *Pharmacogenomics 10*(2), 191–201. https://doi.org/10.2217/14622416.10.2.191

Zhang, Y. (2016). On The Use of P-Values in Genome Wide Disease Association Mapping. *Journal of Biometrics & Biostatistics, 7*(3). https://doi.org/10.4172/2155-6180.1000297