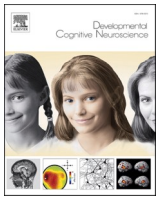
Developmental Cognitive Neuroscience 47 (2021) 100904

Contents lists available at ScienceDirect

Developmental Cognitive Neuroscience

journal homepage: www.elsevier.com/locate/dcn



Bayesian evidence synthesis in case of multi-cohort datasets: An illustration by multi-informant differences in self-control

Sofieke T. Kevenaar [a,b,*], Maria A.J. Zondervan-Zwijnenburg [c], Elisabet Blok [d], Heiko Schmengler [c,e], M. (Ties) Fakkel [c], Eveline L. de Zeeuw [a,b], Elsje van Bergen [a,b], N. Charlotte Onland-Moret [f], Margot Peeters [c], Manon H.J. Hillegers [d], Dorret I. Boomsma [a], Albertine J. Oldehinkel [e]

[a] Netherlands Twin Register, Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands
[b] Research Institute LEARN!, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands
[c] Utrecht University, Utrecht, the Netherlands
[d] Erasmus Universiteit, Rotterdam, the Netherlands
[e] University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[f] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands

ARTICLE INFO

Keywords:
Multiple cohorts
Multiple informants
Self-control
Bayesian evidence synthesis
Multiple imputation by chained equations (MICE)

ABSTRACT

The trend toward large-scale collaborative studies gives rise to the challenge of combining data from different sources efficiently. Here, we demonstrate how Bayesian evidence synthesis can be used to quantify and compare support for competing hypotheses and to aggregate this support over studies. We applied this method to study the ordering of multi-informant scores on the ASEBA Self Control Scale (ASCS), employing a multi-cohort design with data from four Dutch cohorts. Self-control reports were collected from mothers, fathers, teachers and children themselves. The available set of reporters differed between cohorts, so in each cohort varying components of the overarching hypotheses were evaluated. We found consistent support for the partial hypothesis that parents reported more self-control problems than teachers. Furthermore, the aggregated results indicate most support for the combined hypothesis that children report most problem behaviors, followed by their mothers and fathers, and that teachers report the fewest problems. However, there was considerable inconsistency across cohorts regarding the rank order of children's reports. This article illustrates Bayesian evidence synthesis as a method when some of the cohorts only have data to evaluate a partial hypothesis. With Bayesian evidence synthesis, these cohorts can still contribute to the aggregated results.

1. Introduction

There is a growing awareness of the limited reliability of single-study findings, in Developmental Cognitive Neuroscience and other fields of empirical research (Open Science Collaboration, 2015). This awareness has contributed to the call for replication and the need to synthesize findings across studies. Consortia, such as the Consortium on Individual Development (CID), have been established to combine research efforts of different groups to study a particular subject. This raises the challenge to do so in a way that includes and does justice to each study's unique qualities, and still allows conclusions based on accumulated evidence.

A common way to synthesize research findings is meta-analysis, where the results of several previously conducted studies concerning a particular research question, topic, or theory are combined (Rosenthal and DiMatteo, 2002). Meta-analysis has notable advantages, such as the possibility to base the analysis on summary statistics, but has also limitations. Three limitations are (1) that meta-analysis does not allow additional inference on the level of the individual studies, (2) that meta-analysis is prone to the effects of searching strategies and publication bias, (3) and that meta-analysis can only include studies employing comparable models and parameters.

In this article, we apply the alternative strategy of Bayesian evidence synthesis to reach robust conclusions by combining results derived from different sources. Here, the different data sources are four Dutch population cohort studies. Bayesian evidence synthesis can be used to combine results by aggregating their evidence for competing hypotheses

* Corresponding author at: Department of Biological Psychology, Vrije Universiteit, Van der Boechortstraat 7-9, 1081 BT, Amsterdam, the Netherlands.
E-mail address: s.t.kevenaar@vu.nl (S.T. Kevenaar).

https://doi.org/10.1016/j.dcn.2020.100904
Received 28 January 2020; Received in revised form 16 December 2020; Accepted 17 December 2020
Available online 26 December 2020

(Kuiper, Buskens, Raub & Hoijtink, 2012; Zondervan-Zwijnenburg et al., 2019). In this manner, studies covering various contexts and measurement instruments can be combined (Zondervan-Zwijnenburg et al., 2019, 2020). This approach is also suitable to combine the results of structural equation modelling (Zondervan-Zwijnenburg et al., 2019, 2020). The main assumptions of Bayesian evidence synthesis are that all sources of information provided by individual studies contribute to the overarching research question, and that all samples are representative of the population of interest (Veldkamp et al., 2020).

In the current study, we demonstrate that Bayesian research synthesis can be used even if not all parameters relevant to the hypotheses are estimated in all cohorts. More specifically, our overarching hypothesis concerns the ordering of mean raters obtained from four raters of child self-control: teachers, fathers, mothers and children. However, some cohorts only have data of three or fewer raters, and provide partial information concerning the ordering of the mean ratings. So while the comprehensive hypotheses may concern the ordering of several means, the information provided by some cohort may be limited to a subset of the means. For example, consider the assessment of differences among multiple neuropsychological tasks that are assumed to assess the same process, brain areas that are activated by a task, or, as in our case, informants that rate a specific trait or state. In these cases, the Bayesian synthesis approach offers the advantage that it enables statements about the support for specific hypotheses concerning the ordering of parameters, and the possibility to aggregate results, given incomplete information (results) in one or more of the studies. To our best knowledge, this application of Bayesian evidence synthesis is new.

We demonstrate the opportunities and challenges of Bayesian evidence synthesis for a comparison of multiple groups using multi-informant scores of self-control. Self-control is a key topic within the Dutch Consortium on Individual Development (CID). Self-control is the ability to enforce appropriate subdominant responses and inhibit inappropriate dominant impulses (Friedman and Miyake, 2004; Nigg, 2017). Self-control is viewed as an effortful, top-down process in behavioral control. It has been related to, inter alia, the dorsal anterior cingulate cortex, dorsolateral prefrontal cortex, and cortical structures (Bridgett et al., 2015). We assessed self-control in 8- to 12-year-old children using the self-control scale (ASCS) in the Achenbach System of Empirically Based Assessment (ASEBA), which was filled in by four different informants: mothers, fathers, teachers and the children themselves. The ASCS was constructed by Willems et al. (2018) based on items of the ASEBA checklists, which are available in parent-, teacher- and self-report versions (Achenbach et al., 2017; Willems et al., 2018). It is well-established that in completeing questionnaires like the ASEBA scales, different raters have different perspectives, and consequently provide different information (see for example Van der Ende et al., 2012). Here, we make use of Bayesian evidence synthesis to assess hypotheses regarding differences between the raters with respect to the ASCS. We assessed the support for competing hypotheses regarding the ordering of the informants in four CID cohorts: the Netherlands Twin Register (NTR), Tracking Adolescents' Individual Lives Survey (TRAILS), Generation R (GenR), and YOUth, in primary school-aged children aged 8–12 years. The competing informative hypotheses and the literature supporting these hypotheses are discussed in Section 2.3.

2. Methods

2.1. Participants

The participants came from four of the cohort studies that are part of the Consortium on Individual Development: The Netherlands Twin Register (NTR; Bartels et al., 2007; Ligthart et al., 2019), Generation R (GenR; Kooijman et al., 2016), Tracking Adolescents' Individual Lives Survey (TRAILS; Huisman et al., 2008; Oldehinkel et al., 2015), and YOUth (Onland-Moret et al., 2020). The NTR is a national register based in Amsterdam in which twins, other multiples and their families participate. It was established in 1987 and includes children and adults. Children are registered by their parents at birth or any time after birth. About every two years, parents, and, once the children are old enough, teachers and the children themselves, are invited to fill out questionnaires about the children's health and behavior (Bartels et al., 2007; Ligthart et al., 2019). The NTR sample used in the present study largely overlaps with the sample used by Willems et al. (2018) to develop the ASCS. GenR is a cohort study that follows individuals born in Rotterdam from fetal life to adulthood. Mothers with a delivery date between April 2002 and July 2006 were enrolled in the study. During the primary school years, questionnaires were administered twice (Kooijman et al., 2016). TRAILS concerns a population cohort, established in 2000/2001, which has followed children from the Northern parts of the Netherlands from the age of 11 onwards (Oldehinkel et al., 2015). Finally, YOUth is a prospective cohort study established in 2015. In the primary school years, questionnaires were administered at ages 6, 9 and 12 (Onland--Moret et al., 2020).

During development, children display different levels of behavioral problems (Verhulst and Van der Ende, 1995). The developmental trends may be informant-specific, that is, trends may be characterized by parameters, such as intercept and slope(s), that vary over informants (Van der Ende and Verhulst, 2005). We do not formally test the development of informant differences here, but explore the presence of such differences by defining two age groups: a younger group consisting of 8.5–10.5-year-olds and an older age group of 10.5–12.5-year-olds. Table 1 breaks down, by cohort and age group, the number of individuals, number of ASCS observations (total and per informant), mean age, and percentage of boys. As this table shows, in some cohorts, some raters are missing, i.e., there is systematic missingness in the ratings. Self-reports were especially scarce in the younger age group, because pre-adolescents often are not asked to report on their own behavior. Within each age group, the same participant was only included once. In all cohorts except the TRAILS cohort, the participants could be present in both the younger and the older age group (i.e., given longitudinal designs, children participated repeated at different ages). This does not pose a problem, because the data are analyzed and results are aggregated within age groups only. In case of multiple participants in the same nuclear family (e.g. siblings), we randomly selected one to be included in the analyses.

2.2. Measures

Self-control was measured using the ASEBA self-control scale (ASCS; Willems, 2018). The ASEBA system includes questionnaires for different informants: the Child Behavior Checklist 6–18 (CBCL) for parents, the Teacher's Report Form (TRF) for teachers, and the Youth Self-Report (YSR) for the children. In these questionnaires, problem behaviors are rated on a three-point scale with the response options *not true (0)*, *somewhat or sometimes true (1),* and *very true or often true (2)*. In all cohorts, the ASCS was administered as part of the entire ASEBA. The content of the eight items in the ASCS are displayed in Table 2. Four items come from the attention problem scale (item 4, 8, 41, and 78), three from the aggressive behavior scale (item 86, 87, and 95), and one from the rule breaking behavior scale (item 28). The sum scores of the ASCS range from 0 to 16. The psychometric properties of the scale are reported in Willems et al., 2018. The inter-rater reliability for each of the participating cohorts is displayed in Supplementary Table S1. Inter-rater reliability was highest between mother and father ratings, and lowest between self- and mother-ratings. Table 3 contains the ASCS means and standard deviations for each age group, informant and cohort.

2.3. Bayesian evidence synthesis

Bayesian evidence synthesis consists of four steps, which are explained in detail below. In the first step, informative hypotheses are formulated, based on available literature. The second step is to fit the

Table 1
Number of ASCS observations, means and standard deviations (SD) of age, and percentage boys per informant, cohort and age.

Age group	Cohort	Mother	Father	Teacher	Self	Mean (SD) age	% boys	Total observations	Number of individuals (N)*
Younger (8.5–10.5)	NTR	9904	6821	6971	–	9.79 (0.43)	49.7	23,696	12,514
	GenR	4516	3269	713	–	9.50 (0.27)	51.6	8498	4972
	TRAILS	232	–	–	252	10.32 (0.13)	49.0	484	259
	YOUth	504	–	201	–	9.47 (0.58)	42.9	705	513
Older (10.5–12.5)	NTR	6403	4633	5355	562	12.08 (0.23)	50.4	16,953	9095
	GenR	102	90	–	–	11.11 (0.53)	54.0	192	154
	TRAILS	1713	–	–	1935	11.24 (0.52)	49.0	3648	1953
	YOUth	139	–	73	–	10.82 (0.20)	47.1	212	140

* Note that there are missing data because but not all participants have data from all available informants. See Table 5 for the sample sizes used in the analyses.

Table 2
Items of the ASEBA self-control scale (ASCS).

Item number	Item
4	Fails to finish things he/she starts
8	Can't concentrate, can't pay attention for long
28	Breaks rules at home, school or elsewhere
41	Impulsive or acts without thinking
78	Inattentive or easily distracted
86	Stubborn, sullen or irritable
87	Sudden changes in mood or feelings
95	Temper tantrums or hot temper

Table 3
Means (SD) of the ASEBA self-control scale (ASCS) per informant, cohort, and age group.

Age	Cohort	Rater / informant			
		Mother Mean (SD)	Father Mean (SD)	Teacher Mean (SD)	Self Mean (SD)
Younger (8.5–10.5)	NTR	3.36 (3.17)	2.88 (2.97)	2.26 (2.93)	–
	GenR	2.89 (2.87)	2.94 (2.79)	3.11 (3.66)	–
	TRAILS	4.62 (3.25)	–	–	3.81 (2.85)
	YOUth	4.08 (3.25)	–	2.08 (2.52)	–
Older (10.5–12.5)	NTR	3.01 (3.00)	2.66 (2.86)	2.02 (2.75)	4.21 (3.06)
	GenR	3.09 (3.05)	3.64 (2.86)	–	–
	TRAILS	4.65 (3.33)	–	–	3.95 (2.65)
	YOUth	3.92 (3.38)	–	2.41 (3.16)	–

model of interest in all datasets separately. In the third step, Bayesian informative hypothesis testing is employed. The fourth and final step involves the actual Bayesian evidence synthesis, in which the support for each hypothesis is aggregated across all cohorts.

2.3.1. Formulation of competing informative hypotheses

Bayesian evidence synthesis starts with a specification of a set of informative hypotheses about the model parameters (Hoijtink, 2012). When formulating informative hypotheses, the inclusion of all plausible hypotheses supported by literature, expert knowledge, or other sources is recommended. Whereas the classical frequentist null hypothesis testing tests if one or more model parameters deviate significantly from a given value (usually zero), informative hypotheses may also stipulate an ordering of parameters or range constraints.

We formulated competing informative hypotheses based on literature on informant differences in the measurement of self-control. Informants see the children in different contexts (e.g., at school or at home) and may have different relationships with the child. These differences may give rise to differences in perspective on the child's behaviour, and to differences in reference (i.e., a teacher may rate a child relative to other children in the class, whereas a father may rate a child relative to its siblings). Thus, informants have different perspectives on the child's behavior, and may display varying levels of agreement concerning the child's behavior. Several studies have focused on informant differences in problem behaviors, with diverging results. For self-control assessed with the ASCS, Willems et al. (2018) reported the highest average scores for self-reports, followed by, respectively, mother-, father-, and teacher-reports in data from 7- to 16-year-olds in the Netherlands Twin Register (i.e., $\mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$). Note that their data partly overlap with the NTR data used in the present study. Comparable results were found for the ASEBA total problems scale (Grigorenko et al., 2010; Rescorla et al., 2013; Van der Ende and Verhulst, 2005) attention problems (Bartels et al., 2018), and rule-breaking behaviors, (Bartels et al., 2018; Noordhof et al., 2008). With regard to self- and mother-ratings of aggressive problems, Noordhof et al. (2008) reported the opposite pattern (i.e., $\mu_{self} < \mu_{mother}$). Noordhof's sample overlapped with the TRAILS data used in the present study. An alternative hypothesis is that the means of all raters are equal (i.e., $\mu_{self} = \mu_{mother} = \mu_{father} = \mu_{teacher}$). This cannot be ruled out as in most studies the mean differences between the raters were not tested. Thus, based on literature discussed above we formulated the following competing hypotheses, which were evaluated across cohorts:

H1. $\mu_{self} = \mu_{mother} = \mu_{father} = \mu_{teacher}$;

H2. $\mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$;

H3. $\mu_{self} < \mu_{mother} < \mu_{father} < \mu_{teacher}$;

Hc. complement of H1 – H3; any ordering not specified by the three hypotheses above. This hypothesis is included to test if there is any support for possible configurations of differences in means not included in the set H1 to H3.

2.3.2. Model fitting in each cohort separately

The second step is to fit the model of interest in all datasets separately. That is, we fitted a within-subjects linear model, in which we estimated the mean ASCS sum scores of the informants seperately in each cohort and age group.

2.3.3. Bayesian informative hypothesis testing

After specification of the competing informative hypotheses and fitting of the model, the relative support for each of the hypotheses is evaluated for each cohort separately, by means of Bayesian informative hypothesis testing (Hoijtink, 2012). Contrary to the frequentist approach - where only support against the null hypothesis is obtained - the Bayesian approach quantifies support for each of the competing hypotheses, including the null-hypothesis, in terms of posterior model probabilities.

We note that the available data in each cohort determines which components of the hypotheses can be tested. Table 4 contains an overview of which components of each hypothesis are tested in each cohort and age group. For example, the support for H1 in NTR younger age group represents the support for $\mu_{mother} = \mu_{father} = \mu_{teacher}$ only, i.e., does

Table 4
Partial hypotheses tested by each cohort, each age group and missing data approach.

Complete case analyses					
Age	Cohort	H1: $\mu_{self} = \mu_{mother} = \mu_{father} = \mu_{teacher}$	H2: $\mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$	H3: $\mu_{self} < \mu_{mother} < \mu_{father} < \mu_{teacher}$	Hc
Younger (8.5–10.5)	NTR	$\mu_{mother} = \mu_{father} = \mu_{teacher}$	$\mu_{mother} > \mu_{father} > \mu_{teacher}$	$\mu_{mother} < \mu_{father} < \mu_{teacher}$	Yes
	GenR	$\mu_{mother} = \mu_{father} = \mu_{teacher}$	$\mu_{mother} > \mu_{father} > \mu_{teacher}$	$\mu_{mother} < \mu_{father} < \mu_{teacher}$	Yes
	TRAILS	$\mu_{self} = \mu_{mother}$	$\mu_{self} > \mu_{mother}$	$\mu_{self} < \mu_{mother}$	No
	YOUth	$\mu_{mother} = \mu_{teacher}$	$\mu_{mother} > \mu_{teacher}$	$\mu_{mother} < \mu_{teacher}$	No
Older (10.5–12.5)	NTR	$\mu_{self} = \mu_{mother} = \mu_{father} = \mu_{teacher}$	$\mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$	$\mu_{self} < \mu_{mother} < \mu_{father} < \mu_{teacher}$	Yes
	GenR	$\mu_{mother} = \mu_{father}$	$\mu_{mother} > \mu_{father}$	$\mu_{mother} < \mu_{father}$	No
	TRAILS	$\mu_{self} = \mu_{mother}$	$\mu_{self} > \mu_{mother}$	$\mu_{self} < \mu_{mother}$	No
	YOUth	$\mu_{mother} = \mu_{teacher}$	$\mu_{mother} > \mu_{teacher}$	$\mu_{mother} < \mu_{teacher}$	No
Analyses based on imputed data					
All (8.5–12.5)	NTR	$\mu_{mother} = \mu_{father} = \mu_{teacher}$	$\mu_{mother} > \mu_{father} > \mu_{teacher}$	$\mu_{mother} < \mu_{father} < \mu_{teacher}$	Yes
	GenR	$\mu_{mother} = \mu_{father}$	$\mu_{mother} > \mu_{father}$	$\mu_{mother} < \mu_{father}$	No
	TRAILS	$\mu_{self} = \mu_{mother}$	$\mu_{self} > \mu_{mother}$	$\mu_{self} < \mu_{mother}$	No

not include childrens' self-reports. Hc, the fail-safe hypothesis capturing orderings not specified by the other hypotheses, can only be tested in cohorts with three or four informants (i.e. GenR in the young age group and NTR in both age groups in the complete case analyses and only in NTR in the analyses based on imputed data), because in cohorts with fewer informants all combinations were covered by the specified hypotheses.

The R package bain (version 0.2.2) was used to compute Bayes Factors to assess the support of two competing hypotheses (Gu et al., 2019). For example, a Bayes Factor of $BF_{12} = 10$ means that the support in the data for hypothesis 1 is 10 times greater than the support for hypothesis 2 (Lavine and Schervish, 1999). A priori, all hypotheses were considered equally likely in our study, so were assigned the same prior model probability. Given equal priors, Bayes Factors can be easily translated to posterior model probabilities (PMPs), which express the relative support for each of the tested hypotheses (Kuiper et al., 2012). The closer to zero the PMP of a specific hypothesis is, the less likely it is that the hypothesis is true. The PMPs add up to 1.0 over all hypotheses (Lavine & Chervish, 1999). PMPs were calculated for each cohort individually, so the PMPs express support for the partial hypothesis in each cohort. For example, in the younger age group the PMP of Hypothesis 1 reflects support for $\mu_{mother} = \mu_{father} = \mu_{teacher}$ in NTR, $\mu_{mother} = \mu_{father} = \mu_{teacher}$ in GenR, $\mu_{self} = \mu_{mother}$ in TRAILS, and $\mu_{mother} = \mu_{teacher}$ in YOUth. The hypothesis that received most support was considered to describe the data the best in that cohort and age group. If the PMPs of two hypotheses differed less than 0.1, we judged the hypotheses to be equally likely.

2.3.4. Bayesian evidence synthesis

In the final step, the cohort-specific PMPs are aggregated across cohorts to obtain the posterior model probabilities that represent the relative probability of a hypothesis being supported by all cohorts simultaneously (Kuiper et al., 2012; Zondervan-Zwijnenburg et al., 2019). Hence, the approach adopted makes it possible to compare the quantified support for each hypothesis both within studies, and accumulated over studies. By combining the cohort-specific PMPs that each represent relative support for different components of a specific hypothesis, the aggregated PMP covers the full hypothesis, because every informant is available in the combined partial hypotheses at least once, there is enough overlap in informants across cohorts, and the cohorts are representative of the same population. For example, in the younger age group the synthesized support for Hypothesis 1 ($\mu_{self} = \mu_{mother} = \mu_{father} = \mu_{teacher}$) represents support for $\mu_{mother} = \mu_{father} = \mu_{teacher}$ in NTR and GenR and for $\mu_{self} = \mu_{mother}$ in TRAILS and for $\mu_{mother} = \mu_{teacher}$ in YOUth. While this is justified statistically, it is important to realize that the overall support represents a combination of different components tested in different cohorts, and that some components (e.g. the comparison between mother- and teacher-reports) are tested in more cohorts than other components. We used equal prior model probabilities for all hypotheses as a starting point for the first cohort. For the subsequent cohorts, the PMP of the previous cohort was used as a prior model probability, until all cohorts were added. The order of updating is irrelevant for the final results. The details of this procedure can be found in Kuiper et al. (2012).

Because larger sample sizes lead to more precision, Bayes Factors based on larger samples show clearer evidence for or against the hypotheses of interest. This is reflected in greater differences in the PMPs of hypotheses in cohorts with larger sample sizes. This stronger evidence will have a larger impact on the final PMP. The impact of a cohort on the result is thus determined by the strength of the BF, which can be affected by sample size.

In addition to sample size, PMPs of a given hypothesis close to zero also affect the aggregated results over all cohorts. A hypothesis with a near-zero PMP (i.e., close to zero support) in one or more of the cohorts is likely near zero support in the results aggregated results, even if this hypothesis is well supported by other cohorts (i.e., PMP appreciably greater than zero). This is because the support is used as a multiplier in the updating process. In theory, this is a desirable quality of the method because the goal is to reach robust, broadly supported conclusions. However, the updated results over cohorts may provide a picture that appears to be at variance with the results of the individual cohorts.

2.4. Missing data

In the current study, we had to deal with missing data within and across cohorts and with missing data on the item level and on the sum score level. There are several ways to deal with missing data. Here we provide an account of what we considered to be the best strategy to handle the missing data in the present study.

On the item level, we allowed for missingness in three or fewer items. That is, within each cohort, we computed sum scores of the ASCS only if three or fewer items were missing. We used person-mean imputation in calculating the sum scores of a particular person at a particular age per rater (as suggested in Willems et al., 2018).

To handle the missing data at the sum score level, we used two missing data handling methods, complete case analysis and multiple imputation, and analyzed the data given both methods. Both methods have their own advantages and disadvantages. We used both methods to establish that our conclusions did not depend on the method used. It is important to distinguish between sum scores that are not available at all in a certain cohort (for example, self-reports in YOUth), and actual missing data on sum scores that were available in that cohort (for example, a participant for whom mother-report was missing in YOUth). We call the former *systematic* missingness and the latter *incidental* missingness. Here, we applied two methods to handle *incidental* missingness Systematic missingness does not call for imputation. *Given systematic* missingness (e.g., self-reports in YOUth), we tested the partial hypotheses based on the available data.

In the complete case analysis, also known as listwise deletion, a participant with any missing data was excluded. Depending on the cohort and the age group, this resulted in a reduction of the sample sizes ranging from 12 % to 95 % and may result in in bias (depending on the exact cause of the *incidental* missingness). On the other hand, this complete case approach enabled us to test our hypotheses in the younger and older age groups separately, thus providing an indication of stability of the results over the two age groups. Furthermore, there was no loss of informants in the complete case analysis, as only participants that had data of all available informants for that cohort and age group were included in this method.

The second method was multiple imputation. In case of a percentage of missing data greater than 50 %, the ratings of the informant were discarded from further analyses (see the sample sizes per informant relative to the total number of individuals in Table 1). We adopted this strategy, because we believe that imputation quality cannot be guaranteed when more than half of the data is missing. Consequently, the multiple imputation approach included substantially more participants, but fewer informants than the complete case analysis approach (see Table 4). In the YOUth cohort, following this procedure, the remaining data was limited to only one informant, so that the informative hypotheses could not be evaluated in this cohort. In sum, multiple imputation maximized the sample size and reduced the number of partial hypotheses that could be tested. We pooled the data of the two age groups in carrying out multiple imputation to optimize the total number of participants. If we would have decided to impute and analyze the data for the age groups separately, some of the cohorts would have again included a very small number of participants. In case a participant had participated repeatedly, we randomly selected one assessment. Multiple imputation was performed using the R-package mice (multiple imputation by chained equations, version 3.7.0; Van Buuren and Groothuis-Oudshoorn, 2011) in R (version 3.6.1; R core team, 2019). Sum scores were imputed for each cohort separately by means of predictive mean matching (Van Buuren, 2018). The predicted value of the target variable was calculated by the specified imputation model. For each missing value, the method identifies a set of donors from the complete cases, who have predicted values closest to the predicted value for the missing value. One of these donors is randomly selected, and the observed value of the donor is used to replace the missing value (van Buuren, 2018). Imputations were based on the gender of the child and the other informants' ASCS scores. An initial predictor matrix for imputation was created based on minimum correlations of 0.20 between all combinations of variables. For each imputation, 15 iterations were performed and missing data points were imputed 50 times (Azur et al., 2011). The within-subject linear regressions were performed on each imputed dataset, and the results pooled by the R-package semTools (version 0.5.2; Jorgensen et al., 2019). The final sample sizes given the two methods, the complete case analyses and the analyses based on imputed data, are given in Table 5.

3. Results

The means and sample sizes for the complete case analyses and for the analyses based on imputed data can be found in Table 5.

The top part of Table 6 shows the posterior model probabilities (PMPs) of each hypothesis, within each cohort and age group given the first missing data approach, i.e., the complete case analysis. Note that in all the analyses, each cohort tests a component of the hypotheses of

Table 6
Posterior model probabilities (PMPs) of the hypotheses concerning the rank ordering of mean ASCS scores from different informants for the complete case analyses (age groups 8.5–10.5 and 10.5–12.5 years) and for the analyses based on imputed data (ages 8.5–12.5 years).

Complete case analyses				
Age 8.5–10.5	Informants	H1	H2	H3
NTR	m, f, t	< 0.001	1.000	< 0.001
GenR	m, f, t	< 0.001	1.000	< 0.001
TRAILS	s, m	0.089	< 0.001	0.910
YOUth	m, t	< 0.001	1.000	< 0.001
Aggregated		< 0.001	1.000	< 0.001
Age 10.5–12.5	Informants	H1	H2	H3
NTR	s, m, f, t	< 0.001	1.000	< 0.001
GenR	m, f	0.736	0.086	0.178
TRAILS	s, m	< 0.001	< 0.001	1.000
YOUth	m, t	< 0.001	1.000	< 0.001
Aggregated		< 0.001	1.000	< 0.001
Analyses based on imputed data				
Age 8.5–12.5	Informants	H1	H2	H3
NTR	m, f, t	< 0.001	1.000	< 0.001
GenR	m, f	0.033	< 0.001	0.967
TRAILS	s, m	0.078	< 0.001	0.922
Aggregated		< 0.001	1.000	< 0.001

Note: H1: $\mu_{self} = \mu_{mother} = \mu_{father} = \mu_{teacher}$; H2: $\mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$; H3: $\mu_{self} < \mu_{mother} < \mu_{father} < \mu_{teacher}$. The aggregated support reflects the support for the combined partial hypotheses. To obtain the aggregated PMPs, we used the unrounded PMPs.

Table 5
Means (with 95 % confidence intervals (CI)) and sample size for the complete case analyses (age groups 8.5–10.5 and 10.5–12.5 years) and for the analyses based on imputed data (ages 8.5–12.5 years).

Complete case analyses						
Age	Cohort	Mother Mean (95 % CI)	Father Mean (95 % CI)	Teacher Mean (95 % CI)	Self Mean (95 % CI)	Sample size
8.5–10.5	NTR	3.21 (3.11–3.32)	2.85 (2.74–2.95)	2.09 (1.99–2.18)	–	3229
	GenR	2.77 (2.41–3.13)	2.96 (2.61–3.32)	1.89 (1.53–2.25)	–	230
	TRAILS	4.63 (4.20–5.06)	–	–	3.90 (3.52–4.28)	225
	YOUth	3.86 (3.40–4.31)	–	2.16 (1.80–2.52)	–	192
10.5–12.5	NTR	3.27 (2.83–3.70)	2.91 (2.50–3.31)	1.86 (1.45–2.27)	3.96 (3.52–4.40)	186
	GenR	3.31 (2.04–4.24)	3.33 (2.28–4.38)	–	–	38
	TRAILS	4.64 (4.49–4.80)	–	–	3.96 (3.83–4.08)	1695
	YOUth	4.28 (3.50–5.05)	–	2.43 (1.70–3.16)	–	72
Analyses based on imputed data						
Age	Cohort	Mother Mean (95 % CI)	Father Mean (95 % CI)	Teacher Mean (95 % CI)	Self Mean (95 % CI)	Sample size
8.5–12.5	NTR	3.29 (3.21–3.37)	3.04 (2.96–3.11)	2.17 (2.10–2.25)	–	15,884
	GenR	2.89 (2.80–2.99)	3.00 (2.90–3.10)	–	–	4778
	TRAILS	4.27 (4.14–4.40)	–	–	4.02 (3.91–4.13)	2205
	YOUth	–	–	–	–	–

interest, i.e., partial hypotheses. First, we evaluated support for the hypotheses H1, H2 and H3. At age 8.5–10.5, the support for the components of hypothesis 2 was the greatest in NTR ($\mu_{mother} > \mu_{father} > \mu_{teacher}$), GenR ($\mu_{mother} > \mu_{father} > \mu_{teacher}$) and YOUth, ($\mu_{mother} > \mu_{teacher}$). In TRAILS, partial hypothesis 3 ($\mu_{self} < \mu_{mother}$) received most support. The aggregated support was greatest for hypothesis 2 ($\mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$). At age 10.5–12.5, the aggregated support was again strongest for hypothesis 2, but there was more variation in support across cohorts. In NTR, which included all four informants at this age, the support for hypothesis 2 was greatest. In GenR, hypothesis 1 ($\mu_{mother} = \mu_{father}$) received most support and in TRAILS, hypothesis 3 ($\mu_{self} < \mu_{mother}$) received most support.

Subsequently, to evaluate any patterns not captured by our informative hypotheses, we evaluated support for any hypothesis other than our hypotheses H1 to H3, we evaluated the support for hypothesis Hc in the cohorts and age groups with at least three informants, i.e., GenR at age 8.5–10.5 and NTR at both age groups in the complete case analyses and only in NTR in the analyses based on multiple imputation. In these cohorts, there was little support for the Hc hypothesis (PMP of Hc $\leq$ 0.001), but for age 8.5–10.5, Hc received most support, with a PMP of 0.738 (Table 7). A post hoc inspection of the mean values in Table 5 suggests that the Hc hypothesis represents the hypothesis $\mu_{mother} = \mu_{father} > \mu_{teacher}$ here.

The bottom part of Table 6 shows the posterior model probabilities for each hypothesis based on the imputed datasets. The general pattern is similar to that of the complete case analyses. Overall, hypothesis 2 again received most support. In NTR, hypothesis 2 ($\mu_{mother} > \mu_{father} > \mu_{teacher}$) received most support. In GenR, hypothesis 3 ($\mu_{mother} < \mu_{father}$) was judged to be the best hypothesis and as was the case for TRAILS ($\mu_{self} < \mu_{mother}$).

Summarizing, we found the strongest evidence for the hypothesis that children themselves report most self-control problems, followed by mothers, fathers and teachers (i.e., H2 $\mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$). However, we found some inconsistent results across cohorts. The most consistent difference between informants was that parents reported less self-control problems than teachers did. Although this hypothesis (i.e. $\mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$) received the strongest overall support, it was not the preferred ordering when considering each study separately. Again, it is important to realize that the synthesized result demonstrates which hypothesis is best supported by all cohorts simultaneously, and that this can be different from the hypothesis that is most often preferred within cohorts.

4. Discussion

The trend towards large-scale collaborative studies involving consortia, such as CID, gives rise to the challenge of combining data from different sources efficiently in a manner that facilitates comprehensive hypothesis testing. Here, we presented Bayesian evidence synthesis as a method to combine data from different sources and to quantify support for competing informative hypotheses, both within and across cohorts. We illustrated the use of Bayesian evidence synthesis in the situation that different components of the hypotheses were tested in different cohorts.

Overall, our results show most support for the hypothesis that children on average report most problem behaviors, followed by their mothers and fathers, and that on average, teachers report the fewest problems (H2: $\mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$). The most consistent evidence was found for the conclusion that parents report more self-control problems than teachers. The aggregated findings should be interpreted in relation to the findings within each cohort. Observing different findings across cohorts may call for (post hoc) inspection of the exact differences between the cohorts that gave rise to the inconsistent results. In Bayesian evidence synthesis, we assume that the samples are representative of the same target population, in our case, the population of 8- to 12-year-old Dutch children. In our illustration, the cohorts are all assumed to be selected from the general Dutch population, but differ, for example, in the regions of the Netherlands covered and the periods of data collection. Furthermore, one of the cohorts included twins. It is important to take into account differences between the samples and how these might relate to the concept under investigation when interpreting differences in results. Differences in cohort samples should be evaluated in the light of their relevance with regards to the phenomenon of interest, so the implications of sample differences vary from study to study.

Results from the analyses on the complete cases and on the imputed data favored the same hypothesis. The approaches we used to handle missing data have advantages and disadvantages, but the aggregated results supported the same ordering pattern of means. This indicates that the conclusions about the ordering of the means do not depend on the missing data approach.

The ordering of the sum scores of the different informants was the same in 8.5–10.5-year-olds and 10.5–12.5-year-olds, indicating a constant rank ordering in the two age groups. On the cohort level, the only difference in best supported hypothesis between the younger and older age group concerned GenR. This difference likely is due to the fact that teacher data was available only in the 8.5–10.5 group. A post hoc inspection of the mean differences suggests that H2 (partial hypothesis $\mu_{mother} > \mu_{father} > \mu_{teacher}$) in GenR was likely to be preferred in the younger age group, in view of the big difference in teacher ratings and parent ratings. In the 10.5–12.5 group, only ratings of mothers and fathers were available, and these differed much less than the differences between parents and teachers. Post hoc inspection of the means suggest that the differences in means between parents are much smaller, hence, H1 (partial hypothesis $\mu_{mother} = \mu_{father}$) receives most support here. Hence, which components of the hypotheses are tested in a specific sample can have an impact on which hypothesis received the most support.

A novel aspect of Bayesian evidence synthesis is that it can accommodate partial hypotheses given the available data in the cohorts. We illustrated that this method can be used if the information in cohorts is limited to partial hypotheses, while the synthesized information for all cohorts did address the (complete) hypotheses of interest. In previous studies that used Bayesian research synthesis to combine results over cohorts, all aspects of the hypotheses were tested in all cohorts, even though the measurement instrument might differ (Veldkamp et al., 2020; Zondervan-Zwijnenburg et al., 2019, 2020). Statistically, Bayesian research synthesis is suitable to assess and combine the support for partial hypothesis. As mentioned above, it is important to interpret the support for each hypothesis in a particular cohort as the support for the particular component of the hypothesis that was actually tested in that cohort. In the present application, combining the support for partial hypotheses with Bayesian evidence synthesis was feasible because there

Table 7
Posterior model probabilities for the hypotheses concerning the rank ordering of the mean ASCS scores from different raters, including the catch-all hypothesis (Hc).

Complete case analyses				
Age 8.5–10.5	H1	H2	H3	Hc
NTR	< 0.001	1.000	< 0.001	< 0.001
GenR	< 0.001	0.262	< 0.001	0.738
Aggregated	< 0.001	1.000	< 0.001	< 0.001
Age 10.5–12.5	H1	H2	H3	Hc
NTR	< 0.001	1.000	< 0.001	< 0.001
Analyses based on imputed data				
Age 8.5–12.5	H1	H2	H3	Hc
NTR	< 0.001	1.000	< 0.001	< 0.001

was sufficient overlap between the partial hypotheses that were tested in each cohort. While the different cohorts each addressed only a part of the hypothesized orderings, together the data contained information with regard to all comparisons between informants. Put simply, the present overlap between the partial hypotheses was sufficient to arrive at a comprehensive interpretation of the aggregated PMPs.

Bayesian evidence synthesis has several advantages. One advantage is that this approach, in contrast to meta-analysis, is not influenced by publication bias as it is not dependent on published results (Sutton et al., 2000). If the hypotheses cover all orderings, all hypotheses are considered equally likely a priori, and no datasets are excluded based on published findings, Bayesian evidence synthesis is not affected by publication bias. Furthermore, Bayesian evidence synthesis does not require previous investigations to form hypotheses, as it is equally suitable to address new research questions. Here, we included data of all Dutch cohorts that track children's self-control with the ASCS. As we included a complement hypothesis (Hc), assigned equal prior model probability to all hypotheses and, to our best knowledge, included all ASCS data collected in the Netherlands, publication bias plays no role in the current study. A disadvantage of Bayesian evidence synthesis is that, contrary to classical meta-analysis, it requires access to the raw data. However, we note that the analysis of individual participant data is more reliable than aggregate data in meta-analysis (Riley et al., 2010).

A major advantage of Bayesian evidence synthesis is that it provides the degree of support for a set of competing hypotheses both at the within-study level and across studies. This highlights inconsistencies between cohorts and allows one to address the robustness of the overall findings (see also, Zondervan-Zwijnenburg et al., 2020). Moreover, the Bayesian approach answers the focal question of which hypothesis is most plausible given the data. Furthermore, new data can be added to the analyses, because the evaluation of the hypotheses depends on posterior model probabilities, and are not affected by order of data entering. So, the results can be updated if additional data become available, facilitating the growth of knowledge by the accumulation of evidence.

A point of attention is that we only specified and tested hypotheses that were supported by literature. In theory, it is possible to specify additional (novel) hypotheses. For example, our results in some cohorts suggest that there might be no meaningful differences in self-control problem scores of mothers and fathers. In future research, we recommend including, for example, $\mu_{self} > \mu_{mother} = \mu_{father} > \mu_{teacher}$, where the ordering between the parents is not of interest.

The differences that we found between informants implies that different informants provide different information concerning self-control. One may wish to calculate self-control scores based on the ratings of all informants (e.g., an average), but, given the differences between raters, this involves a loss of information. We note that in general one should consider the issue of measurement invariance in the comparison and interpretation of (differences in) test scores. In the present case, the interpretation of the differences between the informants in terms of differences with respect self-control on the conceptual level is based on the tacit, but testable assumption that the self-control test scores are measurement invariant with respect to informant. New datasets, preferably covering parts of the hypotheses that were underrepresented thus far, can easily be added to increase the reliability of the support and accumulate the evidence. Altogether, we feel that Bayesian evidence synthesis is a promising approach to get the most information out of the data available.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We warmly thank all participating families and teachers in the Dutch cohorts which supplied data. We also warmly thank Herbert Hoijtink for his valuable contributions to this paper and Conor Dolan for his valuable comments. All authors and cohorts are part of the Consortium on Individual Development (CID). CID is funded through the Gravitation Program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024-001-003). NTR is funded by 'Decoding the gene-environment interplay of reading ability' (NWO: 451-15-017); 'Netherlands Twin Registry Repository: researching the interplay between genome and environment' (NWO: 480-15-001/674); 'Twin-family study of individual differences in school achievement' (NWO: 056-32-010); 'Longitudinal data collection from teachers of Dutch twins and their siblings' (NWO: 481-08-011) and BBMRI-NL. Participating centers of TRacking Adolescents' Individual Lives Survey (TRAILS) include various departments of the University Medical Center and University of Groningen, the University of Utrecht, the Radboud Medical Center Nijmegen, and the Parnassia Bavo group, all in the Netherlands. TRAILS has been financially supported by various grants from the Netherlands Organization for Scientific Research (NWO), ZonMW, GB-MaGW, the Dutch Ministry of Justice, the European Science Foundation, BBMRI-NL, and the participating universities. The general design of Generation R Study is made possible by financial support from the Erasmus Medical Center, Rotterdam, ZonMw, the Netherlands Organization for Scientific Research (NWO), and the Ministry of Health, Welfare and Sport, and is conducted by the Erasmus Medical Center in close collaboration with the Faculty of Social Sciences of the Erasmus University Rotterdam, and the Stichting Trombosedienst & Artsenlaboratorium Rijnmond (STAR-MDC), Rotterdam.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.dcn.2020.100904.

References

Achenbach, T.M., Ivanova, M.Y., Rescorla, L.A., 2017. Empirically based assessment and taxonomy of psychopathology for ages 1½–90+ years: developmental, multi-informant, and multicultural findings. Compr. Psychiatry 79, 4–18. https://doi.org/10.1016/j.comppsych.2017.03.006.

Azur, M.J., Stuart, E.A., Frangakis, C., Leaf, P.J., 2011. Multiple imputation by chained equations: what is it and how does it work? Int. J. Methods Psychiatr. Res. 20 (1), 40–49. https://doi.org/10.1002/mpr.329.

Bartels, M., Beijsterveldt, C., Derks, E., Stroet, T., Polderman, T., Hudziak, J., Boomsma, D., 2007. Young netherlands twin register (Y-NTR): a longitudinal multiple informant study of problem behavior. Twin Res. Hum. Genet. 10 (1), 3–11. https://doi.org/10.1375/twin.10.1.3.

Bartels, M., Hendriks, A., Mauri, M., Krapohl, E., Whipp, A., Bolhuis, K., Roetman, P., 2018. Childhood aggression and the co-occurrence of behavioural and emotional problems: results across ages 3–16 years from multiple raters in six cohorts in the EU-ACTION project. Eur. Child Adolesc. Psychiatry 27 (9), 1105–1121. https://doi.org/10.1007/s00787-018-1169-1.

Bridgett, D.J., Burt, N.M., Edwards, E.S., Deater-Deckard, K., 2015. Intergenerational transmission of self-regulation: a multidisciplinary review and integrative conceptual framework. Psychol. Bull. 141 (3), 602. https://doi.org/10.1037/a0038662.

Friedman, N.P., Miyake, A., 2004. The relations among inhibition and interference control functions: a latent-variable analysis. J. Exp. Psychol. Gen. 133 (1), 101. https://doi.org/10.1037/0096-3445.133.1.101.

Grigorenko, E.L., Geiser, C., Slobodskaya, H.R., Francis, D.J., 2010. Cross-informant symptoms from CBCL, TRF, and YSR: trait and method variance in a normative sample of Russian youths. Psychol. Assess. 22 (4), 893. https://doi.org/10.1037/a0020703.

Gu, X., Hoijtink, H.J.A., Mulder, J., Van Lissa, C.J., 2019. Bain: Bayes Factors for Informative Hypotheses. R Package Version 0.2.1. https://CRAN.R-project.org/package=bain.

Hoijtink, H., 2012. Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists. CRC Press. https://doi.org/10.1201/b11158.

Jorgensen, T.D., Pornprasertmanit, S., Schoemann, A.M., Rosseel, Y., Miller, P., Quick, C., Selig, J., 2019. Package 'semTools'.

Kooijman, M.N., Kruithof, C.J., van Duijn, C.M., Duijts, L., Franco, O.H., van IJzendoorn, M.H., Moll, H.A., 2016. The Generation R Study: design and cohort update 2017. Eur. J. Epidemiol. 31 (12), 1243–1264. https://doi.org/10.1007/s10654-016-0224-9.

Kuiper, R., Buskens, V., Raub, W., Hoijtink, H., 2012. Combining statistical evidence from several studies: a method using Bayesian updating and an example from research on trust problems in social and economic exchange. Sociol. Methods Res. 42, 60–81. https://doi.org/10.1177/0049124112464867.

Lavine, M., Schervish, M.J., 1999. Bayes factors: what they are and what they are not. Am. Stat. 53 (2), 119–122. https://doi.org/10.1080/00031305.1999.10474443.

Ligthart, L., van Beijsterveldt, C.E.M., Kevenaar, S.T., de Zeeuw, E., van Bergen, E., Bruins, S., Boomsma, D.I., 2019. The Netherlands Twin Register: longitudinal research based on twin and twin-family designs. Twin Res. Hum. Genet. https://doi.org/10.1017/thg.2019.93.

Nigg, J.T., 2017. Annual Research Review: on the relations among self-regulation, self-control, executive functioning, effortful control, cognitive control, impulsivity, risk-taking, and inhibition for developmental psychopathology. J. Child Psychol. Psychiatry 58 (4), 361–383. https://doi.org/10.1111/jcpp.12675.

Noordhof, A., Oldehinkel, A.J., Verhulst, F.C., Ormel, J., 2008. Optimal use of multi-informant data on co-occurrence of internalizing and externalizing problems: the TRAILS study. Int. J. Methods Psychiatr. Res. 17 (3), 174–183. https://doi.org/10.1002/mpr.258.

Oldehinkel, A.J., Rosmalen, J.G.M., Buitelaar, J.K., Hoek, H.W., Ormel, J., Raven, D., Hartman, C.A., 2015. Cohort profile update. The TRacking Adolescents' Individual Lives Survey (TRAILS). Int. J. Epidemiol. 44 (1) https://doi.org/10.1093/ije/dyu225, 76-76n.

Onland-Moret, N.C., Buizer-Voskamp, J.E., Albers, M.E., Brouwer, R.M., Buimer, E.E., Hessels, R.S., Kemner, C., 2020. The YOUth study: rationale, Design, and study procedures. Dev. Cogn. Neurosci. 46, 100868 https://doi.org/10.1016/j.dcn.2020.100868.

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349 (6251). https://doi.org/10.1126/science.aac4716.

R Core Team, 2019. R: a Language and Environment for Statistical Computing. URL. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org.

Rescorla, L.A., Ginzburg, S., Achenbach, T.M., Ivanova, M.Y., Almqvist, F., Begovac, I., Döpfner, M., 2013. Cross-informant agreement between parent-reported and adolescent self-reported problems in 25 societies. J. Clin. Child Adolesc. Psychol. 42 (2), 262–273. https://doi.org/10.1080/15374416.2012.717870.

Riley, R.D., Lambert, P.C., Abo-Zaid, G., 2010. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 340, c221. https://doi.org/10.1136/bmj.c221.

Rosenthal, R., DiMatteo, M.R., 2002. Meta-analysis. Stevens' Handbook of Experimental Psychology. https://doi.org/10.1002/0471214426.pas0410.

Sutton, A.J., Duval, S.J., Tweedie, R.L., Abrams, K.R., Jones, D.R., 2000. Empiricalassessment of effect of publication bias on meta-analyses. BMJ 320 (7249), 1574–1577. https://doi.org/10.1136/bmj.320.7249.1574.

Van Buuren, S.V., 2018. Flexible Imputation of Missing Data, 2nd ed. CRC Press. https://doi.org/10.1201/9780429492259.

Van Buuren, S., Groothuis-Oudshoorn, K., 2011. Mice: multivariate imputation by chained equations in R. J. Stat. Softw. 45 (3), 1–67. https://doi.org/10.18637/jss.v045.i03.

Van der Ende, J., Verhulst, F.C., 2005. Informant, gender and age differences in ratings of adolescent problem behaviour. Eur. Child Adolesc. Psychiatry 14 (3), 117–126. https://doi.org/10.1007/s00787-005-0438-y.

Van der Ende, J., Verhulst, F.C., Tiemeier, H., 2012. Agreement of informants on emotional and behavioral problems from childhood to adulthood. Psychol. Assess. 24 (2), 293. https://doi.org/10.1037/a0025500.

Veldkamp, S.A.M., Zondervan-Zwijnenburg, M.A.J., van Bergen, E., Barzeva, S.A., Tamayo Martinez, N., Becht, A.I., Van Beijsterveldt, C.E.M., Meeus, W., Branje, S., Hillegers, M.H.J., Oldehinkel, A.J., Hoijtink, H.J.A., Boomsma, D.I., Hartman, C., 2020. Effect of Parental Age on Their Children's Neurodevelopment. https://doi.org/10.1080/15374416.2020.1756298.

Verhulst, F.C., Van der Ende, J., 1995. The eight-year stability of problem behavior in an epidemiologic sample. Pediatr. Res. 38 (4), 612. https://doi.org/10.1203/00006450-199510000-00023.

Willems, Y.E., Dolan, C.V., van Beijsterveldt, C.E., de Zeeuw, E.L., Boomsma, D.I., Bartels, M., Finkenauer, C., 2018. Genetic and environmental influences on self-control: assessing self-control with the ASEBA self-control scale. Behav. Genet. 48 (2), 135–146. https://doi.org/10.1007/s10519-018-9887-1.

Zondervan-Zwijnenburg, M.A.J., Veldkamp, S.A.M., Neumann, A., Barzeva, S.A., Nelemans, S.A., Van Beijsterveldt, C.E.M., Branje, S., Meeus, W.H.J., Hillegers, M.H.J., Tiemeier, H., Hoijtink, H.J.A., Oldehinkel, A.J., Boomsma, D.I., 2019. The impact of parental age on child behavior problems: updating evidence from multiple cohorts. Child Dev. 91 (3), 964–982. https://doi.org/10.1111/cdev.13267.

Zondervan-Zwijnenburg, M.A.J., Richards, J.S., Kevenaar, S.T., Becht, A.I., Hoijtink, H.J.A., Oldehinkel, A.J., Boomsma, D.I., 2020. Robust longitudinal multi-cohort results: the development of self-control during adolescence. Dev. Cogn. Neurosci., 100817 https://doi.org/10.1016/j.dcn.2020.100817.