MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„The Multiverse of Associations Between in vivo Brain Volume and Intelligence: Meta-Analytical Updates and Extensions"

verfasst von / submitted by

Daniel Gerdesmann, BSc BA

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2020 / Vienna 2020

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 840
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Psychologie UG2002
Betreut von / Supervisor:	Mag. Dr. Jakob Pietschnig

Danksagung

Das Ende meiner Studienzeit naht. Bevor ein neuer Lebensabschnitt beginnt, gebührt jenen Menschen, die mich in dieser Zeit begleitet, unterstützt und mir eine Menge denkwürdiger Momente beschert sowie jenen die zur Anfertigung dieser Arbeit beigetrageben haben mein herzlicher Dank.

Ich danke meinem Betreuer Dr. Jakob Pietschnig, dass er mir das Thema dieser Masterarbeit anvertraut hat. Ich danke ihm und Magdalena Siegel, MSc für interessante Seminare und die Hinführung zu Herausforderungen wie Meta-Analysen, Open Science und R. Ich möchte mich auch bei allen Forscher*innen bedanken, welche die Früchte ihrer Arbeit (paper, packages, tutorials …) anderen frei zur Verfügung stellen. Ohne dieses Engagement hätte ich diese Masterarbeit nicht verfassen können.

All den liebenswürdigen Menschen, denen ich in meiner Studienzeit begegnet bin, sage ich danke, zum Beispiel Robert, Vladi, Karol, Crisi und den Psychoforum Leuten. Dank euch kann ich auf viele schöne Momente zurückblicken.

Valéria gilt doppelter Dank. Für ihre Liebe und Unterstützung, aber auch für ihr beinhartes Lektorat. Dank ihr erhielten der vorliegende Text und die zahlreichen e-Mails, welche ich an Forscher*innen versandt habe so etwas wie Lesbarkeit.

Ich widme diese Arbeit meiner Muddi, Sabina Gerdesmann. Vom Arbeiterkind zur *self-made woman*, zwei Kinder durchs Studium gebracht und dabei herzensgut geblieben – kann man mal machen. Respekt.

Content

Introduction

Is brain size associated with intelligence? Is it true that big-brained people are smarter on average? The search for answers to this question has had an eventful history. From early pioneering work with all sorts of obstacles, through dark chapters of colonialism, racism and even genocide, to successes in the development of measuring instruments, heated debates about the value of this question, curiosities such as the "heavy-weight champion" among celebrities´ brains (2kg, Ivan Turgenev) to modern imaging technology and genome-wide association studies, following this history is like looking through a burning glass at the study of differential psychology. Even though that we have nowadays a good picture of the association between brain volume and intelligence after almost 200 years of research, not all questions have been answered yet. This master´s thesis is another attempt to provide some of them.

The first systematic examinations took place in the 1830s in Western Europe and the Russian Empire (Tiedemann, 1836; Morton, 1849; Vein & Maat-Schieman, 2008). Since then, this question occupied some influential minds (e.g. Broca, 1861; Darwin, 1871). These early efforts have been difficult and ineffective. Neither brain volume, nor intelligence were directly quantifiable. A popular stream of research was the evaluation of skull and brain characteristics after death, which led to inferences about the behavioral qualities of their deceased owners. Posthumous brain examinations of famous intellectuals considered to be highly intelligent in their lifetime have been especially popular (Vein & Maat-Schieman, 2008). One of the first large-scale studies linking brain size and intelligence (the used proxies have been head height and academic achievement) was conducted by Galton (1889). He concluded that despite some measurement problems, there seems to be some evidence for an association. In 1905 Alfred Binet and Theodore Simon published their famous IQ test (Binet & Simon, 1916), the first measurement breakthrough for researchers interested in brain volume and intelligence. Although there had been other intelligence tests before, their test has been the first one to be demonstrably valid (Boake, 2002). Nevertheless, the first reviewers of evidence on the topic of brain size and intelligence have concluded that on the one hand measurement problems (e.g. reliability) were a handicap to the correct assessment of the association between brain size and intelligence, and that on the other hand there was a rather insignificant, if any, correlation between them (Whipple, 1914; Paterson, 1930; see Jensen & Sinha, 1993).

In the wake of the horrors of the Second World War research activities on the topic were relatively sporadic. The racial connotations (see section “Ethnicity”) and the lack of any substantial advances in the measure of in vivo brain volume were discouraging aspects. In the early 1970s the interest raised again (e.g. van Valen, 1974), although met with criticism and doubts about the value of any such pursue (e.g. Gould & Gold, 1996). With the advent of highly precise and harmless in vivo imaging technology in the 1980s came the second measurement breakthrough. Researchers were able to measure in vivo brain volume directly (see Rushton & Ankney, 1996). Since then, numerous narrative reviews have been published, unanimously concluding that there indeed is a relevant positive association between brain volume and intellectual intelligence (Jensen & Sinha, 1993; Vernon et al., 2000; Gignac et al., 2003; Miller & Penke, 2007; Ruston & Ankney, 1996, 2000, 2009). Still, effect size estimates as well as discussion of influential aspects (moderators) have yielded different results.

In 2005 the first meta-analysis concerning brain volume and IQ was published (McDaniel, 2005). A meta-analysis applies statistical methods aiming at obtaining a refined weighted average across studies addressing the same question, while offering the possibility to explore potential moderators and effect size inflation due to various sources of bias. Succeeding the two above mentioned measurement breakthroughs, analytical benefits of the meta-analytic approach were the last missing pieces of the puzzle. McDaniel´s meta-analysis consisted of 37 correlations based on 1530 participants. In 2015 Pietschnig et al. conducted a meta-analysis that expanded the body of data (148 correlations based on 8036 participants) and the scope of analysis (e.g. including clinical samples as well as verbal and performance subdomains of intelligence) substantially. Subsequently, an additional meta-analysis applying different methods to a subset of the Pietschnig et al. (2015) data was conducted (Gignac & Bates, 2017). But still, the estimates differed substantially in size. The result of McDaniel´s (2005) analysis was an overall effect of $r = .33$. This dropped in the Pietschnig et al. (2015) analysis to $r = .24$ and then increased again in the Gignac and Bates (2017) analysis to $r = .39$. It should not go unnoticed that the latter two meta-analyses were based on the very same data set. The examination of potential influences of other variables (moderators) has yielded different results as well. Whereas McDaniel (2005) identified sex and age to be significant moderators, Pietschnig et al. (2015) concluded the opposite. Some moderators have been examined by only one research team and, considering the differences above, need conceptual replication. Ethnicity as a possible moderator has been discussed thoroughly in the literature (e.g. Rushton & Ankney, 2007), but has never been included in a meta-analysis due to lack of data (McDaniel, 2005). Additionally, there was no agreement on the extent of dissemination bias in the used literature.

To sum up, the general association between in vivo brain volume and intelligence has been firmly established, a lot of questions remain unanswered though. How strong is the association? Do age, sex or other variables moderate the effect? How are (sub-)domains of intelligence linked to brain volume? In order to look into these questions, this thesis contains a (further) meta-analysis based on an update of the Pietschnig et al. (2015) data pool and alternative analysis procedures. In the following sections, before the aims and hypotheses will be devised, the operationalization and measurement of brain volume and intelligence will be discussed in order to provide context for their bivariate association. These sections focus on basic concepts and measurement related considerations which have been usually skipped in research papers for the sake of brevity.

Brain Volume

The following section presents a short overview of the imaging technologies used in primary studies which have been included in the three above-mentioned meta-analyses. Those technologies are complex and develop fast. This thesis refers interested readers to the excellent review of imaging technology in Bigler (2017) as a starting point. Erin Bigler is one of the pioneers in the field of brain behavior association research and co-authored the first studies associating in vivo brain volume and intelligence via in vivo imaging (Yeo et al., 1987; Willerman et al., 1991) along with five subsequent studies on this topic.

Before the emergence of in vivo imaging techniques in the 1980s, brain volume was assessed with several surrogate measures both before and after death of subjects. Posthumous methods included filling skulls with pellets or water to calculate cranial capacity (Sahin, 2012). As direct in vivo assessment was not possible, researchers used external head measures like head circumference (e.g. Murdoch & Sullivan, 1923). All these surrogate measures provide a reasonable estimate of brain volume but are not as precise and reliable as needed in order to examine the association of brain volume and intelligence thoroughly. For example, head circumference correlates highly with brain volume in children, but only moderately in adults (Bartholomeusz et al., 2002).

The most precise estimates of in vivo brain volume come from studies that used either *Computed Tomography* (CT) or *Magnetic Resonance Imaging* (MRI). CT was developed in the 1970s. A CT machine sends x rays in direction of the brain, and detectors behind it measure the attenuation by brain tissue and structures (Coffrey, 2000). Computers relate this information to the density of tissue, fluids and bones, and construct a series of images. In the 1980s MRI technology emerged. MRI is a vastly complex technology. This basic description illustrates it: “MRI […] measures radio-frequency signals emitted from hydrogen atoms after the application of electromagnetic (radio-frequency) waves, localizing the signal using spatially varying magnetic gradients. Contrast from each voxel (a three-dimensional pixel) depends on the density of protons within the voxel and properties of the local tissue microenvironment that are either directly related to the magnetic properties of hydrogen or that can be detected through manipulation of magnetic fields.” (Lerch et al., 2017, 314). In a comparison of both methods, MRI scores with higher spatial resolution and much greater flexibility. The brain can be examined with MRI in a variety of ways from every angle. This technology provides the most accurate results. Nevertheless, the use of CT has some practical advantages in certain situations. CT scans are less expensive, faster, and available in almost every hospital. The speed advantage can be useful in populations where head movements are not easily controlled (e.g. infants). There are also more personnel that can perform CT scans. However, these advantages are limited to clinical samples only, as subjects are exposed to irradiation. Exposing healthy participants to irradiation for purely scientific purposes is ethically unacceptable. Moreover, the weight of these benefits is steadily decreasing due to technological advancements in MRI technology (Zijl & Knutsson, 2019). The outcomes of both approaches relating to the brain size are usually measured in mm^3 or ml (mass, at 1ml ~ 1g).

The quality of images from both methods depends on several hardware and analysis factors. The former includes the capability of the used scanner, scanner calibration and head movement (Coffey, 2000). Other confounds may be as complex as magnetic field inhomogeneities (Lerch et al., 2017). The latter factors include the effectiveness of the used computer program to process data, the criteria to define structures of interest, and the skills and unbiasedness of the image quality rater (Coffey, 2000). Additionally, difficulties in automatic separation of dura mater and cortical volume can arise as well as problems with spatial normalization due to the unique sulcal structure of every individual brain (Lerch et al., 2017). Researchers´ chosen settings can also influence the results. Slice thickness settings for example can influence estimates when using CT (Sahin, 2012).

There are clearly many factors which may influence the assessment of brain volume (Shinohara et al., 2017). Nonetheless the results are still exceptionally precise and reliable (McGuire et al., 2017; Maclaren et al., 2014; Madan & Kensinger, 2017; Reid et al., 2017). In addition to technical factors, there are some practical concerns which are easier to specify and can have a substantial impact. The operationalization of brain volume in studies is heterogenous. *Total brain volume* (TBV), synonyms are *whole brain volume* (WBV) and *total* *tissue volume*, is variably operationalized as the sum of grey matter (GM) and white matter (WM) or GM, WM and cerebrospinal fluid (CSF). Occasionally the cerebellum is excluded. This fact alone is noteworthy as the cerebellum has a considerable size (Sereno et al., 2020) and is linked to intelligence (Hogan et al., 2011). Another way to operationalize brain volume is *intracranial volume* (ICV) or its synonym *total intracranial volume* (TIV). ICV measures usually contain all brain structures within the skull, including the meninges, ventricles and brain stem. The maximal axial limit is the *foramen magnum*. There are several ways to derive an estimate of brain volume (see Lyden, 2015, 20). For example, some studies use *intracranial area* (ICA) to obtain a quick estimate of ICV. Both correlate highly at about r = .88 (Ferguson et al., 2005).

The comparison of studies using different operationalizations may pose a threat to internal validity. Fortunately, the use of heterogenous operationalizations diminishes since dedicated software like *Freesurfer* (Fischl, 2012) makes standardized, automatic, and reliable segmentation of brain structures readily available while also reducing dependence on personnel skills. Naturally, automatization software also allows a lot of different settings influencing the outcome (Haller et al., 2016), but establishes standard settings enhancing comparability of the average neuroscientific study. To further enhance comparability and resolve issues of small sample sizes, a current trend are efforts to accumulate large public data sets and implement big data and machine learning techniques (van Zijl & Knutsson, 2019). A considerable number of studies included in the previous meta-analyses are from the manual and semi-automatic era (Bigler, 2017).

In conclusion, in vivo imaging techniques are relatively precise and reliable methods to assess brain volume. Especially MRI keeps improving and provides vast possibilities for researchers. Large consortia are working hard on resolving issues regarding low sample sizes and measurement heterogeneity (see Bigler, 2017). Measurement issues regarding brain volume are noteworthy but minimal compared to the average psychological instrument. This may mean that operationalization differences pose the biggest, albeit modest, threat to internal validity.

Intelligence

Not only the history of research on brain volume, but some intelligence research as well got caught up in ugly phantasies about human hierarchies based on race, sex and social status. This turbulent past was often a focus of criticism of intelligence concepts (e.g. Gould & Gold, 1996). There were some authors who argued that intelligence is just another concept which justifies and solidifies social differences. Although it was important to demonstrate the harm some research had done in the past, some critical remarks were one-sided or false (see Fletcher & Hattie, 2011). Nowadays, basic concepts of intelligence are of little dispute. Intelligence test scores predict major life outcomes showing consistent results across lifetime (Goriounova & Mansvelder, 2019).

Many definitions of intelligence exist. Gilles Gignac describes shortly and accessibly intellectual intelligence "as an entity's maximal capacity to achieve a novel goal successfully using perceptual-cognitive abilities." (Gignac, 2018, 440). By adding an operational definition of psychometric intelligence, he bridges the often-criticized gap of theory and application. Gignac defines psychometric intelligence "as an entity's maximal capacity to complete a novel, standardized task with veridical scoring using perceptual-cognitive abilities." (Gignac, 2018, 440). This interpretation gives a good account on what was measured in the primary studies included in the three previous meta-analyses. Worthy of note is that only intellectual intelligence is considered in this thesis. Other constructs like emotional or social abilities are beyond the scope, although interesting brain-behavior studies exist (e.g. Tan et al., 2014). Neither executive functioning is considered here. Intelligence and executive functioning share some conceptual similarities and partly the same underlying biological foundation (Duggan et al., 2014), but they remain distinct concepts.

Intellectual intelligence is usually conceptualized hierarchically. The most prominent concept is the *Cattell-Horn-Carroll* (CHC) model of intelligence (see Schneider & McGrew, 2012), yet there are other models (e.g. Johnson & Bouchard, 2005). The CHC model has three levels. The first level consists of several related narrow abilities combined into broader ability types or intelligence domains. These domains all contribute to *g*, residing at the top of the hierarchy. The general factor of intelligence – *g* is a phenomenon important for subsequent analyses. It describes the observation that persons doing well in one test tend to do well in others, and an underlying factor of general intelligence is thus displayed. The extraction of *g* is accomplished by applying factor analysis to a variety of tests and is stable across different analysis methods (Jensen & Weng, 1994). The existence of *g* is relatively undisputed and has been observed in various samples around the world (Warne & Burningham, 2019).

One of the most prominent and persistent intelligence tests are the Wechsler intelligence scales (Wechsler, 1939). The majority of studies included in the previous three meta-analyses have applied these scales. The structure of the Wechsler scales is very similar to the CHC model of intelligence. The individual subtests of the whole scale represent the narrow abilities. These are for example vocabulary, arithmetic, and symbol search tests.

The second level is represented by four cognitive domains or indices comprising those individual tasks tapping into a specific group of abilities. These four indices are the *Verbal Comprehension Index* (VCI), the *Working Memory Index* (WMI), the *Perceptual Organization Index* (POI), and the *Processing Speed Index (PSI)*.

On a third level, the VCI and WMI combine to an index of verbal IQ, and the POI and PSI to an index of performance IQ. The fourth and last level at the top of the hierarchy is called full-scale IQ and refers to the concept of *g* (for a comprehensive look at the Wechsler scales, see Deary, 2020). The subsequent meta-analyses in this thesis are based on this categorization structure. Study outcomes are classified as either reflecting the association of brain volume and full-scale, verbal, or performance IQ.

Research Questions and Hypotheses

Goals

As mentioned in the introduction, we already know a lot about the association between brain volume and intelligence. At the same time there are both some contradictory findings and open questions. This thesis aims to solidify the understanding of this association, strengthen the confidence in examined effects, work out inconsistencies, and add analysis ideas. The leading research questions are formulated as follows:

How does the association of in vivo brain volume and intellectual intelligence quantify for each population? How trustworthy is the accumulated data? How do researchers´ specification choices affect summary effects? Does interpreting the lower or upper bound of overall effects lead to different conclusions? What do results mean for the neuroscientific research of human intellectual intelligence?

In order to outline the path to answers to these questions, a list of objectives of this master´s thesis follows. These goals were derived from the questions which remained unanswered after reading several publications on the topic (especially the previous meta-analyses). The first goal (1) is to once again estimate the strength of the association between brain volume and intelligence based on updated data and a variation of analysis procedures. The three previous meta-analyses came to different results. In order to discuss the reasons for these differences, (2) I will investigate how differences in data construction (which data was analyzed) and analysis procedures (how were they analyzed) have affected the outcome. Not only the general strength of the association between brain volume and intelligence was assessed differently, but also the potential influence of relevant variables. One question, for example, was whether the correlation changes with the age of the subjects. Some of these variables were either assessed unanimously or only examined by one research group. The results concerning the influence of these variables are (3) to be replicated conceptually. Differences in conclusions about potentially influential variables are (4) to be worked off. One variable was discussed extensively in the relevant literature, but never considered in a meta-analysis, because not enough data were available. This variable is (5) to be included. Another point of contention in previous meta-analyses was the extent to which various forms of bias (mainly publication bias) could influence the interpretation of the results. Therefore (6) it should be evaluated how bias threats interpretation on the basis of the updated data. This will include a replication of a *decline effect* (Schooler, 2011; Pietschnig et al., 2019) observed by Pietschnig et al. (2015). In a final step (7), it will be discussed to what extent these goals have been achieved and what do the results mean for the knowledge about the association between brain volume and intelligence. The next section discusses which variables actually matter in the pursuit of the goals outlined above.

Hypotheses

The goals stated above translate into hypotheses specified in this section. They are necessary to make the questions asked scientifically auditable. This section is structured according to the above-mentioned state of agreement or disagreement on different aspects. The section "Replication" deals with questions or variables that have been unanimously assessed or have only been examined once. The section "Inconsistencies" shows controversial topics. The section "Ethnicity" presents the variable of the same name which could not be included in any previous meta-analysis due to lack of data. The last section is dedicated to the question why previous meta-analyses came to different results about the general strength of the association between brain volume and intelligence, and how the influence of data and analysis procedures can be investigated. All hypotheses formulated here were preregistered. It means that they have been published online before the data collection and analysis were carried out. This preregistration includes a presentation of the topic, the hypotheses, and the exact analysis procedures that were used, and is available on https://osf.io/r6gnk. Further documentation as well as the complete data set on which all analyses are based will be uploaded to this webpage at the latest one year after publication of this master´s thesis. A preregistration corresponds to the *Open Science* criteria and should relief some concern about psychological research reporting habits, which have been increasingly discussed in recent years (see Ioannidis et al., 2014).

Replication

Positive Association Between In vivo Brain Volume and Intelligence. Most authors of reviews on the association between brain volume and intellectual intelligence came to the same conclusion: the association is significant and positive (van Valen, 1974; Vernon et al., 2000; Gignac et al., 2003; Miller & Penke, 2007; Ruston and Ankney, 1996, 2000, 2009; McDaniel, 2005; Pietschnig et al., 2015; Gignac & Bates, 2017). The following hypotheses are the foundation of all subsequent analyses. The hypotheses were preregistered as follows:

H1.1: *There is a positive association between in vivo brain volume and full-scale IQ.*

H1.2: *There is a positive association between in vivo brain volume and verbal IQ.*

H1.3: *There is a positive association between in vivo brain volume and performance IQ.*

One of the main research goals of this thesis is to explore how the previous meta-analysts have arrived at different effect sizes for the association between brain volume and full-scale IQ. The means to achieve this goal are devised in section "Exploration: Researchers´ Degrees of Freedom".

Differences by Domain. Pietschnig et al. (2015) expected the correlations between in vivo brain volume and full-scale IQ to be stronger than between brain volume and verbal or performance IQ. They explained their expectation with *g* theory (Jensen, 1998). As *g* should consist of all relevant domains of intelligence, full-scale intelligence tests map this broadness better, and correlate stronger with *g* than subdomain tests. Pietschnig et al. (2015) observed lower summary effects by domain based on healthy samples (full-scale: $r = .26$; verbal: $r = .18$; performance: $r = .22$). Especially the association between brain volume and verbal IQ was lower. It was not examined whether these results were statistically significant. Most studies provided correlations for more than one domain based on the same participants. To include all these results in one standard analysis would violate the assumption of independent effects. However, there are methods providing a solution for these cases. I used one of them (see section "Robust Variance Estimation Meta-Regression"). The other previous meta-analysts, McDaniel (2005), Gignac and Bates (2017), concentrated on full-scale intelligence. Therefore, a replication of these associative differences by intelligence domain as well as a robustness test is needed. These replicative hypotheses were preregistered as follows:

H2.1: *The association between in vivo brain volume and full-scale IQ is significantly larger than between in vivo brain volume and verbal IQ.*

H2.2: *The association between in vivo brain volume and full-scale IQ is significantly larger than between in vivo brain volume and performance IQ.*

Health-Status. Whether participants are considered healthy or not is an assessment comparative to the research question. In context of brain volume and intelligence, participants are classified healthy if they can be considered representative for the general healthy population concerning brain structure and intelligence. They have no clinical condition which is likely to affect assessments or cognitive processing (e.g. schizophrenia). For example, a participant with mildly high blood pressure belongs to a healthy sample, because her condition will not have an extraordinary effect on the assessments and is highly prevalent in the general population. Sometimes classification in healthy or clinical samples is not straightforward. Being born preterm is a regular phenomenon, and, although it might influence both variables (Arhan et al., 2017; Boberg & Wallström, 2015), it is not considered a clinical condition in this thesis. Being born extremely preterm is less common, does have a stronger impact on both variables (Bjuland et al., 2014; McCoy et al., 2014; Grunewaldt et al., 2014) and is thus considered a clinical condition. The classification at which gestational age births are viewed as preterm and extremely preterm is a bit arbitrary. However, in most cases classification is straightforward (and done by primary researchers labelling their samples as healthy or not). A substantial number of healthy samples in the Pietschnig et al. (2015) data were control groups.

The comparison of results from healthy and clinical samples showed one of the most notable effects in previous meta-analyses. Pietschnig et al. (2015) were the first researchers to include clinical samples in their analyses. They have discovered a statistically significant difference of $r = .06$ based on full-scale intelligence data (healthy $r = .26$; patients $r = .20$). In a narrative review Rushton and Ankney (1996) have reported a larger difference ($r = .40$ for healthy samples, and $r = .20$ for clinical samples). The “headline” correlation of brain volume and intelligence from Pietschnig et al. (2015) was based on both populations. The result was $r = .24$. Gignac and Bates (2017) criticized this combined analysis of healthy and clinical samples mainly for two reasons: (1) intelligence testing of participants affected by various conditions might not be accurate to their true potential and (2) the various clinical conditions might affect associations between brain volume and intelligence in different ways. Both objections are legitimate, but one may argue that Pietschnig et al. (2015) merely selected a different approach of reporting. Their headline correlation based on both sample types (healthy and clinical) comprised all obtainable effect sizes representing the general population without any restrictions (other than data availability). They proceeded to report the subgroup differences between healthy and clinical samples. Whereas this reporting order is perfectly sensible, it may have not been tactically wise, because readers tend to skim reports and miss important subgroups differences (Borenstein, 2019, 209). Although concerns about internal validity in the clinical groups raised by Gignac and Bates (2017) are important to consider, a gross overall comparison between subgroups is better than none. Since Pietschnig et al. (2015) have been the only meta-analysts who had taken clinical samples into account, subgroup differences in the association between brain volume and intelligence have to be replicated. The hypothesis regarding health-status was preregistered as:

H3: *The association between in vivo brain volume and intelligence is significantly larger in non-clinical samples than in samples of patients.*

Correlation of Applied IQ Measurement With *g*. Pietschnig et al. (2015), and Gignac and Bates (2017) observed that studies in which full-scale intelligence has been measured with a wider range of different ability tests yielded higher correlations for brain volume and IQ. Both research groups reasoned that the better reflection of all intelligence abilities in humans is the most likely explanation. In the section “Intelligence” above *g* was characterized as a factor of general intelligence. The phenomenon of higher effect sizes based on extensive intelligence measurement can thus be seen as founded in higher correlations of those measurements with *g*. For instance, the complete Wechsler Adult Intelligence Scale IV correlates extremely high with *g* (~ $r = .95$, Wechsler, 2008). The vocabulary subtest of the same scale correlates about $r = .7$ with *g* (Hunt, 2010). The theoretical basis for hypotheses 2.1, 2.2 and the correlation of the applied IQ measuring with *g* is the same in principle.

The methods that led to the detection of the moderating effect of the correlation with *g* differed between research groups. Pietschnig et al. (2015) ran a regression analysis distinguishing between Wechsler type tests and other tests. Gignac and Bates (2017) used a refined approach specifically designed to detect this effect. They constructed a rating system which should, in absent of empirical information in the literature, approximate the correlation with *g*. The rating categories were number of intelligence dimension assessed, number of tests and testing time. Each category could be rated from 1 (“poor”) to 4 (“excellent”).

In order to replicate this effect, I used the same rating system with one exception; testing time was not taken into consideration anymore. Although some tests are of a considerable length, they test only one dimension (e.g. sustained attention tests). There are some other factors (e. g. adaptive testing) which may influence the testing time, however, are not linked to *g*. Gignac explains the importance of testing time in more detail, e.g. refuting the argument with the sustained attention test by pointing out that this is a core feature of *g*, and therefore longer intelligence testing will result in a better reflection of *g* (Gignac, 2018). Both approaches are reasonable, and, in the end, a little variation is a useful sensitivity check. Results should not differ substantially, because the rating system was slightly changed. The following hypothesis was preregistered:

H4: *Higher correlations of applied intelligence measurements with g, in absence of information about the correlation with g assessed with the number of tests and the number of tested dimensions, are associated with larger positive associations between in vivo brain volume and general intelligence.*

Decline Effect. The decline effect refers to a decrease of effect sizes over time as evidence accumulates starting with the first study addressing a specific research question (Schooler, 2011). There are various alleged causes of declining effect sizes (Protzko & Schooler, 2017). Sometimes an observed effect does decrease genuinely over time. In other cases, decreases may be rooted in strategic research behavior or publication bias and have false positives as a result (e.g. the “Mozart effect”, see Pietschnig et al., 2010). In their meta-analysis about brain volume and intelligence, Pietschnig et al. (2015) observed an *inflated decline effect* for healthy samples based on full-scale IQ data, meaning that there was a true effect, but the observed effect size was smaller than previously reported. This finding is in line with a recent investigation of widespread declining effects in intelligence research (Pietschnig et al., 2019). Nujiten et al. (2019) found no compelling evidence for disproportionally numerous decline effects in that area.

The following preregistered hypothesis addresses the question, if a decline effect persists in the updated data:

H5*: The magnitude of effect size estimates from included studies diminishes systematically over time (from earlier to recent studies).*

Inconsistencies

Age. Ageing seems to impact brain volume and intelligence in the same way. In absolute terms, brain volume and general intelligence increase to early adulthood, and then decline steadily. However, the developmental trajectories differ slightly. Over the lifespan, *total brain volume* increases until the age of 13, followed by a slight decrease until the age of 18 (Hedman et al., 2012). From then, brain volume increases (or at least does not change) until 35, after which a relatively steady decline takes place, accelerating from age 65. Brain tissue volume decreases, ventricle volumes and sulcal cerebrospinal fluid increase. General intelligence decreases from the twenties to older age, but patterns differ by domain (Deary et al., 2014). Whereas the intelligence domain *processing speed* declines consistently over time, *reasoning* decreases not linearly and *vocabulary knowledge* increases. A steeper decline in fluid than crystalline intelligence is usually observable.

Because of these slightly different developmental trajectories of brain volume and intelligence, their associations could be affected by age. Pietschnig et al. (2015) observed no effect for the association between brain volume and either full-scale IQ, neither verbal nor performance IQ. McDaniel (2005) observed an effect of age, if sex was also considered. He reported higher correlations for female adults than for male children. Both used categorial variables (children vs. adults) for their moderator tests. Pietschnig et al. (2015) coded also the mean age of samples but decided against using it, because that would have led to data loss. It is possible that these variable operationalizations were too insensitive to detect effects. Another potential obstacle of those moderator tests may have been that the operationalization of brain volume has not been taken into account (TBV or ICV). Linking ICV and intelligence may not be adequate to detect an impact of age, especially in older age, because ICV measurements do not reflect decreases in brain volume as effective as measurements of TBV (Caspi et al., 2020). In sum, evidence for a potential effect of age is conflicting. Hence the meta-analysis based on the largest data set (Pietschnig et al., 2015) did not detect any effects of age I devised the following preregistered hypotheses as null effects:

H6.1*: Participants´ age has no significant effect on the association between in vivo brain volume and full-scale IQ.*

H6.2: *Participants´ age has no significant effect on the association between in vivo brain volume and verbal IQ.*

H6.3: *Participants´ age has no significant effect on the association between in vivo brain volume and performance IQ.*

Sex. On average, males have larger brain volumes than females (Ruigrok et al., 2014). These differences persist after accounting for body height (Ankney, 1992). Evidence for divergent averages of scores in full-scale intelligence tests between sexes is conflicting, and ranges from non-existent (Deary et al., 2003; 2007) to a small advantage for males (Nyborg, 2005; Daseking et al., 2017). In any case, sex differences in brain volumes are much more pronounced than alleged differences in general intelligence. The association of brain volume and intelligence could thus be varied for sexes. The evidence for sex as a moderator is mixed, too. Whereas McDaniel (2005) reported higher correlations for females (*r* = .40) than for males (*r* = .34), the larger meta-analysis from Pietschnig et al. (2015) showed no effect of sex.

Another matter is the association of brain volume and domain intelligence. Some studies reported subtle differences in confined intelligence subdomains (e.g. Strand et al., 2006). If the average IQ scores vary by domain, the association between brain volume and domain intelligence could be dissimilar, too. Burgaleta et al. (2012) for example reported that sex differences in brain volume were not related to general intelligence, but to visuo-spatial skills. The only meta-analysis considering domains unveiled no evidence in this direction (Pietschnig et al., 2015). To formalize the test of these questions, the following preregistered hypotheses have been devised:

H7.1: *Participants´ sex has no significant effect on the association between in vivo brain volume and full-scale IQ.*

H7.2: *Participants´ sex has no significant effect on the association between in vivo brain volume and verbal IQ.*

H7.3: *Participants´ sex has no significant effect on the association between in vivo brain volume and performance IQ.*

Lynn (1994, 2017) proposed that differences in IQ only begin at age 16 and develop to a sex gap favoring males of approximately 4 IQ points in adulthood. He pointed out that this influence of age could explain the conflicting evidence regarding sex differences in IQ. The results of a recently published large-scale study (*n* > 10000) supported Lynn’s theory (Arribas-Aguila et al., 2019). McDaniel (2005) observed indeed more pronounced sex differences in the association between brain volume and intelligence in children than in adults. In order to reflect these possibilities an age - sex interaction effect will be considered in subsequent analyses (see section “Moderators”). I considered this possibility only after the preregistration, since I have not been aware of Lynn´s theory before.

Ethnicity

Ethnicity is a social category used to describe a group of people who identify with each other on the basis of shared nationality, language and culture (Betancourt & López, 1993, 631). Ethnicity is sometime used interchangeable with the term race, which is a concept trying to describe groups based on phenotypic similarities (Braje & Hall, 2015). From early on, research that links brain volume and intelligence has been accompanied by attempts to establish ethnicity or race as important factors (Rushton & Ankney, 1996). Efforts in measuring brain size and intelligence were embedded in their times. Especially colonialist motives, and social interpretations of Darwin´s evolution theory influenced thinking about this topic in 19th and first half of the 20th century. Unfortunately, this led in some cases to efforts aiming to demonstrate race hierarchies among humans, and some skull "data" have been gathered under horrific circumstances. For example, some German researchers profited from the colonial oppression in Namibia (which ultimately led to a genocide) by receiving skulls from deceased Nama and Herero (Heller & Pesmen, 2020; Adhikari, 2008). However, most of this dark research history was concerned with hard differences between population groups, either in cranial capacity or intelligence (or alleged proxies), assuming a stable link between brain volume and intelligence across ethnicities. In this thesis only relative differences in the magnitude of the association of brain volume and intelligence are of interest. This means not asking if certain populations differ in their average brain volume and IQ, but if the link between brain volume and intelligence (i.e. higher brain volume relates to higher intelligence) is of the same magnitude in all humans, regardless of ethnicity or race. Ironically, this much less controversial question has gained little attention. In fact, so little that McDaniel (2005) could not investigate it due to lack of data.

Research using categorization terms like ethnicity and race is challenging. Both terms are imprecise and arbitrarily defined and contain a high risk of being misunderstood (Heinz et al., 2014). Especially race is defined arbitrarily. It is not clear which phenotypic differences are relevant to separate human races or how many there are. If these terms are used in a context with a noticeable biological framing like brain volume and intelligence, biological inferences can be made without these concepts supporting any theoretically sound basis for such inferences. A related problem is the confusion with social variables like socioeconomic status (SES). Jensen & Sinha (1993) showed that people of color in the United States tend to have an increased prevalence of premature births. Premature birth is linked to lower brain volume (e.g. Bjuland et al., 2014), so the association between brain volume and intelligence under consideration of ethnicity could be confounded by SES. The effect of SES on brain volume and intelligence is an interesting research area of itself, but when considering ethnicity, we usually think of natural (i.e. genetic) variation. Reasons for why the association between brain volume and intelligence could vary naturally are sparse. One may argue that the only testable contender when using a bivariate correlational design is climate. The *Bergmann´s rule* (Bergmann, 1848) states that the body volume of animals tends to increase in colder climate zones. This leads to a lower body surface area to volume ratio, which helps to save energy for keeping warm. The rule does apply for most mammals and birds, and has been observed in humans, too. If body volume increases in colder climates, brain volume could be enlarged, too, without being related to higher intelligence. That would lead to a variation in effect strength between brain volume and intelligence. Aspects of the *Cold Winter Theory* do argue in this direction, although more based on the notion that living in colder climates favored the development of higher intelligence (Lynn, 1991). In sum, reading the existing literature on ethnicity in the context of brain volume and intelligence produces two paradoxes: (1) Ethnicity is mentioned extensively in context with brain volume and intelligence, but research on the actual association between them considering ethnicity is sparse. Researchers mostly focused on hard differences in one variable while using the other for leverage. (2) Researchers try to capture biological, climatic, and genetic influences with social categories.

Nevertheless, the topic of brain volume and intelligence is strongly connected with these types of questions, so I did not want to disregard them altogether. Keeping these considerations in mind, two goals were pursued: (1) review if data availability has improved and (2) explore if moderator analyses based on ethnicity and race conceptualizations can supply further insight into the association of brain volume and intelligence. The following preregistered hypothesis was devised:

H8: *Participant´s ethnicity has a significant effect on the association between in vivo brain volume and full-scale intelligence.*

Exploration: Researchers´ Degrees of Freedom

Besides the potential effects of certain variables on the association between brain volume and intelligence, the methods to look into this topic may also affect outcomes. There are a lot of ways to study a phenomenon. In reference to brain volume and intelligence, we have seen above that researchers can use different intelligent scales or composite their own. They can operationalize and estimate brain volume in different ways, and they can compute their association with different types of correlations and regressions, controlling for potential confounds or not. They may also assemble the sample they like to test and choose what to report and what not. Each of those options and their combinations may affect the outcome. These choices have been termed *researchers´ degrees of freedom* (Simmons et al., 2011). The potential implications for the replicability of observed effects gained attention in the last decade (Wicherts et al., 2016). In the worst-case scenario opportune use of the possibilities to assemble data and analyze it can lead to increased rates of false positives (Ioannidis et al., 2005). Furthermore, effect sizes can be inflated (Simonsohn et al., 2014). As primary studies, meta-analyses are not protected against this kind of bias. There are vast possibilities to decide where to look for data, which to include and how to analyze it (Voracek et al., 2019). In the introduction I mentioned that all previous meta-analysts found different summary effects. Especially the difference between the Pietschnig et al. (2015), and the Gignac and Bates (2017) analyses is striking, since the same pool of data was used. The results differed between $r = .24$ and $r = .39$, depending on what parts of the data were analyzed, and how. Gignac and Bates (2017) decided to include a subsample of the data accumulated by Pietschnig et al. (2015) comprising healthy adult participants only. The upside is, we already know that it is no worst-case scenario, where the very existence of an effect is in question. Nevertheless, the differences, based on the same data set, are unsatisfactory. Where do they stem from? How can we decide what estimate is more accurate? The obvious way is to look for inconsistencies, errors, and implausibility in these analyses, evaluate their quality and decide which to rely on. There are two pitfalls with this approach. (1) Besides obvious mistakes, there are many reasonable ways to do the work. Comparing them can lead to long, tiresome and unfruitful discussions about which choices are most appropriate (Voracek et al., 2019). (2) There are sometimes no criteria on which basis to decide, because approaches are just equally appropriate or there is not enough supporting evidence for one or the other. This is the case for the previous three meta-analyses on the association between brain volume and intelligence. All are of high quality and justified their choices reasonably. For example, McDaniel (2005), and Gignac and Bates (2017) used psychometric meta-analytic methods (Hunter & Schmidt, 2015), whereas Pietschnig et al. (2015) used methods in tradition of Hedges and Olkin (1985). In the briefest possible way to compare both approaches, the former is primarily concerned with the underestimation of effects due to measurement errors, the latter provides a vast array of methods to detect bias and safeguard conclusions with sensitivity analyses. There are endless arguments for and against certain specifications, but in the end, both methods are accepted, validated approaches, their application was accurate, and a discussion will not yield a satisfactory result.

Fortunately, Voracek et al. (2019) developed a method to resolve this problem. Their approach is a meta-analytic adaption of solutions for primary studies (Simonsohn et al., 2015; Steegen et al., 2016). The idea is to incorporate every reasonable choice made by meta-analysts and show what consequence each decision has on the overall effect. With this information at hand, readers can see the range of reasonable estimates, as well as decide which parameters are relevant for themselves and ultimately, which point or range estimates to prefer. The approach has not seen much use as it was developed only recently. The implementation is one of the main goals of this thesis. Due to the lack of experience using it and the lack of interpretation guidelines, results are considered exploratory (as preregistered). Specifics are denoted in section “Exploring the Multiverse”.

Methods

The following sections describe how the meta-analyses in this thesis were performed. The first three sections "Inclusion and Exclusion Criteria", "Information Sources", and "Data Collection" explain which study results (data) were searched for in order to update the data pool, where they were searched for, and how this information was processed into a data set. The following section "Summary Measures" describes which metrics were used as effect measures and which transformations were applied. The sections on "Methods of Synthesis" deal with the meta-analytical calculation methods. A number of different approaches have been chosen, each with its advantages and disadvantages. The main goal is to obtain a refined weighted average across study results and evaluate differences between them. In the following, the section "Moderators" deals with the statistical means to test the influence of the above-mentioned variables (e.g. age). The penultimate section "Dissemination Bias" is dedicated to methods for the detection of effect sizes inflations due to various sorts of bias. The last section "Exploring the Multiverse" deals with the above-mentioned *researchers' degrees of freedom* and their impact on the estimation of the association between brain volume and intelligence.

Inclusion and Exclusion Criteria

For the insurance of a reasonable scope of the meta-analyses the potentially eligible studies needed to meet eight inclusion criteria first. Since most of those criteria were consistent with the previously conducted meta-analyses a direct comparison of results is warranted. There were no deviations from the preregistered inclusion and exclusion criteria. In order to be included each respective study was required to (1) assess the association between in vivo brain volume and intelligence. However, it was not relevant whether this has been the primary goal of the study. (2) In vivo brain volume has had to be measured by either MRI or CT. Studies had to provide measurements of the whole brain volume (TBV or ICV). If both have been reported in the study, TBV took precedence over ICV. Partial measurements (of brain areas) were excluded. (3) Intelligence have had to be measured directly via standardized tests. Standardized meant that the test has been administered in a standardized way, has been objectively scored and offered norms allowing comparisons to the norm sample. Only measurements of intellectual abilities were included, whereas constructs like social or emotional intelligence were excluded. (4) Effect sizes had to be based on individual participant data. Associations based on group means (e.g. from high and low IQ groups) were not eligible. (5) There were no population restrictions. Both clinical and non-clinical participants across all ages and both sexes were included. There were (6) no restrictions on location, type of report or language either. Studies with abstracts or full texts in languages other than English and German were translated with the free version of the online translator *DeepL* (https://www.deepl.com/translator). This translator was used approximately 15 times while screening Chinese studies, and dissertations in various languages. In the end all eligible studies included in subsequent meta-analyses have been written in English or German. 7) If more than one publication of the same study was found the alternative with the best proximity to the goals of this thesis and most extensive data display was chosen. In case no informed choice was possible, the earlier publication was included. The same procedure was applied to studies from different authors, who had analyzed the same data. For instance, if a study displayed effect sizes based on sex separated samples, it took preference over a study displaying an effect size based on a mixed sample, because this was advantageous for moderator analyses regarding sex. 8) In order to avoid a dependent data structure study effect sizes were coded separately by intelligence domain (full scale IQ, verbal IQ and performance IQ). If more than one effect size fitting the same domain were available, the one based on the most comprehensive domain assessment was chosen. All things being equal, verbal comprehension took priority over working memory and perceptual organization over processing speed. An integration of available dependent (domain specific) effect sizes in one data sheet was also achieved via *robust variance estimation* (RVE) meta-regression. More information about the RVE approach is displayed in the section of the same name.

If a potentially eligible study displayed no effect size or enough information to compute one the authors were contacted. When necessary information was not obtainable the study was not considered further. Study authors reporting only that the association was non-significant, were contacted, too. If no additional information could be obtained, the effect size was set to zero (Pigott, 2009, 408-409). A codebook delineating all coded variables and abbreviations is available at https://osf.io/75b8s/.

Information Sources

The basis for the data update was the data set from Pietschnig et al. (2015). These data are openly accessible (supplemental material to their publication). In order to find any additional eligible studies online databases *PubMed*, *ISI Web of Science*, *Scopus* and *Google Scholar* were searched using the following search string:

(brain size AND intelligen*) OR (brain volume AND intelligen*) OR (brain size AND IQ) OR (brain volume AND IQ).

Some databases (especially for grey literature) did not support a search with all terms simultaneously via parentheses. In those cases, the search terms were entered individually. Sometimes the search string contained additional specifications (e.g. to exclude non-relevant collections from the search). Comprehensive information including the full search strings, dates, number of hits and extracted search files are available on https://osf.io/hfkmj/. Results from *PubMed*, *ISI Web of Science* and *Scopus* were exported and integrated in an Excel file for screening. *Google Scholar* did not offer this feature. The first 250 hits of the *Google Scholar* search were screened in a browser. Step 2 involved a forward citation search of all three published meta-analysis on the subject (McDaniel, 2005; Pietschnig et al., 2015; Gignac & Bates, 2017) using the *Google Scholar* feature. *ISI Web of Science* and *Scopus* provided similar functions but yielded together far less hits than *Google Scholar*. In step 3 an extensive search for grey literature was conducted. It contained a wide range of sources, covering (1) grey literature data bases, search engines and repositories, (2) sources dedicated to theses and dissertations, (3) conference materials, (4) registries for active studies and (5) contact to experts. A list of all resources is provided in the same document containing the search information (see above). Lastly, reference lists of all as eligible identified studies were searched for additional studies.

Study selection involved a screening of both the study title and abstract. Full text from all studies passing the screening were obtained and checked for eligibility.

Data Collection

The above-mentioned openly accessible data sheets from Pietschnig et al. (2015) contain the following variables: study (first author), sample sex, sample type (healthy or clinical), test measure, year, sample age, male ratio, number of participants, effect size estimate (r), number of corrections (to the correlation coefficient), reported vs. personal communication, children/adolescents vs. adults and study goal. I adopted these variables and their categories. In order to achieve the stated goals of this thesis I added the following variables: study ID, effect size ID, review coverage, the standard deviation of the sample age mean, ethnicity (coded categorially and as ratios), page of effect size in study, brain volume measurement tool (MRI or CT), type of whole brain volume measurement (TBV or ICV), IQ domain, sample standard deviation of the intelligence test and the corresponding population standard deviation, u-ratios, ratings concerning the number of tests used, number of dimensions assessed, alleged correlation with *g*, a combined rating of the correlation with *g*, and lastly the reliability of the intelligence test application as examined by the study authors. A coding book containing data sheets and a coding manual explaining all variables and their categories is available at https://osf.io/75b8s/.

The whole process of study selection and data collection was conducted by me alone. I had minimal prior experience in conducting a systematic literature search for a meta-analysis and coding data, gained in a seminar during my master. Reliabilities of the study selection or data collection processes were not assessed. However, the coding of eligible studies was repeated one time to minimize potential coding errors.

Summary Measures

Sometimes studies display their results in different metrics and indices. Before meta-analyzing study outcomes, they must be transformed to the same metric. All eligible studies considered in this thesis have used correlation or standardized regression coefficients (β) as their effect size metric. No conversions among effect size metrics were necessary. This was mostly because the inclusion and exclusion criteria specified effect sizes based on group means as not eligible. Effect sizes had to be based on individual participant data. Therefore, conversions from means and their standard deviations, t-tests or χ^2 values to correlation coefficients were not eligible for inclusion. In subsequent analyses, the effect sizes were transformed and/or corrected in some way to. These transformations and corrections were not applied due to different metrics but to benefit analyses mathematically and analytically. These procedures are described in the following sections.

Methods of Synthesis

The following sections describe the meta-analytical methods used to calculate a weighted average of the correlations of brain volume and intelligence reported in the literature. Another goal of the following first method was to investigate heterogeneity among these correlations, e.g. whether there were extreme correlations and whether these had a large influence on the weighted mean (summary effect). Different methods were applied, since each of them has certain advantages, which are even more useful in combination. The first two methods below are those used by previous meta-analysts. The main differences between the methods are their primary concerns (what they are particularly good at, what is the focus), the weighting of the individual study results, how heterogeneity in the data is estimated, and which transformations or corrections are usually applied. The main basis for all meta-analytical methods were study results based on healthy samples. Data from clinical samples were used in an appropriate place for comparison with results from healthy samples, or for comparison with results from previous meta-analyses. Analyses for full-scale IQ were repeated for verbal and performance IQ. All analyses were performed using the statistical programming language R. "Packages" refer to functions written by users that enable or simplify certain analyses.

The first method, a "Hedges and Olkin Meta-Analysis" has the advantage of providing a huge range of analysis tools, options and sensitivity analyses. It is a good start to get first results and to get to know the data thoroughly. The second method, "Psychometric Meta-Analysis", is primarily concerned with the question of how measurement errors in the included studies could influence the summary effect. For this purpose, the individual correlations are "corrected" before synthesizing effect sizes. The third method "Robust Variance Estimation Meta-Regression" makes it possible to process dependent data in an analysis. What this means and why it was useful is briefly explained at the beginning of that section. The fourth method "Bayesian Meta-Analysis" is based on a slightly different conception of statistical testing than the previous models and was performed as an explorative complementary analysis. All methods used have in common that they were based on a random-effects model. This approach assumes that the true effect differs between studies. In contrast, the fixed-effect model assumes that all studies measure the same true effect and differ only in their respective sampling error. The random effects model was chosen, because I assumed that the included studies were representative of all studies assessing the association between in vivo brain volume and intellectual intelligence, and the goal was to make inferences about that larger universe of studies. The true effect size was expected to vary across studies. These criteria ruled out the use of fixed-effect (fixed true effect) and fixed-effects (no generalization beyond included studies) models. Congenial to this choice, prior meta-analyses found considerable between-study variance.

Appendix A lists all primary used programs, packages, and R codes.

Hedges and Olkin Meta-Analysis

The first approach resembled the path taken by Pietschnig et al. (2015). A random effects meta-analysis in the tradition of Hedges and Olkin (1985) was conducted based on independent effect sizes.

Fisher's *r*-to-*z* transformed correlation coefficients (Fisher, 1921) were used for the computation of results. These were obtained with the "ZCOR" command of the *escalc* function within the *metafor* package (Viechtbauer, 2010). This is a standard procedure accounting for the skewed distribution of the Pearson correlation. The transformation leads to a stabilization of variance. For ease of interpretation results were transformed back to the *r* metric prior to reporting. Some researchers criticized this procedure to introduce a substantial upward bias (see Hunter & Schmidt, 2015). Therefore, a sensitivity analysis with the "UCOR" (correlation corrected for its slight negative bias, see Olkin & Pratt, 1958) command was conducted. To calculate sample variances, the *escalc* function within the *metafor* package was used. Effect sizes were weighted according to study precision, defined as the inverse standard error: 1 / (*n* – 3). Precision of the effect size estimates was displayed with 95 % confidence intervals (CI). The Knapp-Hartung adjustment (Knapp & Hartung, 2003; Sidik & Jonkman, 2002) was used, as generally advised when using random-effects models (Inhout et al., 2014; Jackson et al., 2017). With this adjustment, the CIs are computed based on a *t*-distribution, not a *z*-distribution. This leads to more appropriate and usually wider CIs. Because the t-distribution works with *k* – 1 degrees of freedom, the differences between CIs based on *z*- or *t*-distributions are expected to be minimal when the number of studies is large.

Since the choice of a random-effects model assumes that there is between-study heterogeneity, this must be taken into account in the calculation of the summary effect. For this purpose, a τ^2-estimator is used. As such, the *restricted maximum likelihood estimator* (REML) was utilized. The *Paule-Mandel estimator* (PM) was utilized for a sensitivity analysis. Those two estimators were chosen, because the data from Pietschnig et al. (2015) had properties concerning the number of studies and differences of study sizes, which can favor either estimator. The REML is regarded a solid choice in a wide range of contexts. It is advantageous when large studies are included (Viechtbauer et al., 2005) and when study sizes differ substantially (Langan et al., 2018). Both properties apply to the Pietschnig et al. (2015) data. The PM estimator can outperform the REML estimator, if the number of studies is large and heterogeneity substantial (Veroniki et al., 2016). This applies to the data, too. Therefore, both were used.

Heterogeneity was described by reporting *Cochran´s Q*, τ, τ^2, I^2 and the *prediction interval*. *Cochran´s Q* is a robustness test that tells us if heterogeneity is present or not. I^2 describes the proportion of variance in the observed effects that is due to variance in true effects (Borenstein et al., 2017). The *prediction interval* displays the absolute variation of true effects. A *normal quantile-quantile* (QQ) plot showed if the residual heterogeneity in true effects was normally distributed (Wang & Bushman, 1998). To enquire the contribution of each study to the overall heterogeneity, a *Baujat* plot (Baujat et al., 2002) was used. A *leave-one-out analysis* was conducted to evaluate influences of individual studies on the overall effect size. All possible subsets of studies were examined with a *GOSH* plot (Olkin et al., 2012) to explore potential subgroup effects. To assess potential distorting effects of outliers, nine outlier evaluation statistics described by Viechtbauer and Cheung (2010) were used via the "influence" command in *metafor*. If distorting effects of individual effect sizes occurred, the possible implications of the presence of outliers were discussed, but no numeric alterations applied. In the preregistration I have explained how I would handle missing data. No situation arose in which dedicated methods were used. The data were analyzed as they were. Missing data were considered to be *missing at random* (MAR).

Psychometric Meta-Analysis

The second approach resembled the paths taken by McDaniel (2005), and Gignac and Bates (2017). A random-effects psychometric meta-analysis (Hunter & Schmidt, 2015) was conducted. This type of meta-analysis originated in the area of personnel selection (Schmidt, 2015), a field with some instruments of low reliability and unbalanced samples. As Hunter and Schmidt (2015) have demonstrated several measurement errors, also called "artifacts", can have an impact on results. For example, if an instrument of low reliability was used in a study, results will be attenuated. The main difference to the above-mentioned approach is the desire to correct for these potential measurement errors. In my view, a good way to think about these corrections is a simulation, asking how results could have looked like if studies have had used perfect measurement instruments on perfectly balanced samples under the assumption of perfect construct validity.

There are 4*2 corrections possible (Hunter & Schmidt, 2015). In this thesis only a correction for range departure in intelligence measures was applied. As stated in the section "Brain Volume", MRI measurement properties are excellent, especially in reference to psychological standards. For the most prominent "artifacts" unreliability and range departure, there was not much to correct. Reliabilities are usually very high and near to 1 in studies using recent automatic extraction software. Range departure could not be corrected since there are no firmly established standard deviations of brain volume means for all populations (e.g. for every age group).

Reliability of intelligence measures is usually also quite high but does fluctuate more between different intelligence tests or versions. It could have been worth the effort to try to correct for attenuation due to unreliability, but almost no study had reported reliability coefficients from their measurements. Previous psychometric meta-analysts have not done this either (McDaniel, 2005; Gignac & Bates, 2017). McDaniel has not done it as he thought the usually high reliabilities leave not much to correct for. Gignac and Bates have not wanted to use reliability coefficients reported in test manuals, because “reliability is a property of test scores derived from a particular sample, rather than a property of a test” (Gignac & Bates, 2017, 27), and data loss would have been considerable. I would like to state further that the moderator analysis in reference to correlation of applied intelligence measurement with *g* does already give us some idea about how reliability might have impacted results. Brief tests rated “fair” have lower reliabilities than tests rated “excellent”. Their reliability, for example a full Wechsler scale, is near to 1. Nevertheless, there are several reasons why a correction could have been worthwhile. (1) Tests for adults usually have higher reliabilities than for children, even if this difference is very subtle. (2) Domain tests were not rated or corrected for, which makes their interpretation less informed than those of full-scale IQ tests. (3) The rating procedure does only show differences in correlation magnitude, but not to what extent differences were due to artifacts or the correlation with *g*. To conclude, there are some arguments to correct unreliability, however in light of the resulting data loss due to missing information the benefits seem doubtful.

A correction for direct range departure (concerning the intelligence measures) was applied to effect sizes and their standard errors. Some samples did deviate from the normative standard deviation of test scores (usually SD = 15). In case the range was restricted, effects may have been attenuated. If the range was enhanced effects may have been overestimated. This is reflected in u-ratios, which were computed by dividing the sample standard deviation by the population standard deviation. Ratios greater than one represent range enhancement, ratios lower than 1 range restriction. In order to obtain corrected correlations, the *Case II formula* (Thorndike, 1949) was used. There are some mathematically identical variations of this formula. I used the R package *psychmeta* (Dahlke & Wiernik, 2019) which applied the following formula for univariate direct range departure:

$$\rho_{TP_a} = \left[\frac{\rho_{XY_i}}{u_X \sqrt{\rho_{YY_i}} \sqrt{\left(\frac{1}{u_X^2} - 1\right) \frac{\rho_{XY_i}^2}{\rho_{YY_i}} + 1}} \right] / \sqrt{\rho_{XX_a}}$$

This version of the formula can correct for range departure in both variables simultaneously. Since range departure was only considered in one variable, the u-ratio of the other is set to 1 (no correction).

To estimate the range departure corrected correlation standard errors the formula from Kelley (1923) was used. The formula was:

$$\boldsymbol{SE_R = [(1 - r^2)/(n - 2)]^{1/2}\{R(1 - R^2)/[r(1 - r^2)]\},}$$

where *r* is the observed correlation, *R* the corrected correlation and *n* the sample size (Duan & Dunlap, 1997, 256). Information regarding range departure was available in Gignac and Bates (2017) for healthy adult samples from the Pietschnig et al. (2015) data. I adopted this information in cases where they had obtained it via personal communication.

Some meta-analytic computational features were different from the first approach. The correlation coefficients were corrected for their slight negative bias via the “UCOR” command in *metafor*. Studies (effect sizes) were weighted according to their number of participants. To estimate between-study heterogeneity, the *Hunter & Schmidt estimator* (HS) was used. The HS method to estimate τ^2 is the standard estimator to be used in psychometric meta-analyses. It has a downward bias (Viechtbauer, 2005; Hunter & Schmidt, 2015). While results of both approaches to assess heterogeneity were compared, the parameters derived from the REML and PM estimators were considered less biased in that regard.

Robust Variance Estimation Meta-Regression

In a third approach, studies were able to provide more than one effect size based on the same participants within one intelligence domain. Data dependencies were modelled using RVE meta-regression (Hedges et al., 2010). For example, when a study reported one correlation between in vivo brain volume and the score of the VCI of a Wechsler test, and another correlation based on the WMI score, both were integrated in the verbal intelligence data sheet. In the other meta-analytic approaches only one correlation has been coded to retain data independence. Another cause of dependence were multiple effect sizes from the same participants at different times (cohort waves). Both types primarily concerned the verbal and performance domains. The use of RVE meta-regression is considered best practice in meta-analyses dealing with dependent effect sizes (Pigott & Polanin, 2020).

One strength of the approach is that it is agnostic to the type of dependence in the data (Tanner-Smith et al., 2016). It can be used to model correlated effects and hierarchical dependence. In order to better illustrate what correlated effects are it should be recalled that persons who do well in one type of intelligence tests tend to do well in others. Thus, their scores in e.g. verbal comprehension and vocabulary tests will correlate. Intelligence is a good example for hierarchical dependence, too. We have conceptualized intelligence hierarchically with distinct levels of domains. If we would desire to analyze full-scale, verbal, and performance IQ together, full-scale IQ would be hierarchically higher, and some variance of lower level test scores would be part of the full-scale IQ scores. It is also useful to incorporate hidden data dependency. Researchers favor a certain style in conducting studies. Study outputs from the same researchers may thus be more similar to each other than expected when outputs from different researchers are compared.

For the estimation of summary effects between brain volume and intelligence I left the data sheets separated by domain. Combining data in one sheet was only necessary to evaluate if correlations differed statistically significant by domain. Although both types of data dependency were present, I decided to use the formula for correlated effects to determine study weights (Tanner-Smith et al., 2016). The choice between a hierarchical- and correlated effects model only affects efficiency, not inference. A correlated effects model was considered more efficient, because data dependence was primarily caused by correlations between participants´ domain intelligence scores from different (sub-)domains. The heterogeneity statistic τ^2 (REML) and the weighting statistic ω^2 are calculated via simplistic methods of moments estimators and are primarily needed for the estimation of inverse variance weights (Tanner-Smith et al., 2016). These statistics were not interpreted. Fisher's z-transformed correlation coefficients were used.

Other methods for modeling dependent effect sizes, full multivariate methods and multilevel meta-analyses, could not have been used. Full multivariate methods require extensive knowledge about the relation between variables, e.g. through a correlation matrix reported in a study. These were not provided in many studies. The use of those methods would have thus led to substantial data loss. RVE meta-regression leads to close approximations of full multivariate methods when the number of studies is large (Hedges, 2019). Multilevel meta-analyses, called three-level meta-analyses also, were not appropriate, since the same participants provided data for multiple effect sizes within the same intelligence domain (Tanner-Smith et al., 2016). This violated the model assumption of independent sampling errors within clusters (domains).

Bayesian Meta-Analysis

In order to complement the frequentist meta-analytic approaches displayed above a Bayesian meta-analyses was conducted via the *bayesmeta* package (Röver, 2020). The debate about the merits of Bayesian over frequentist inference received a new boost with the advent of the replication crisis in psychology (e.g. Wagenmakers et al., 2018). The most important difference in the context of this master thesis between these two approaches is that a Bayesian framework allows the inclusion of previous knowledge or assumptions. Since the meta-analysis of Pietschnig et al. (2015) at the latest, we know that brain volume and intelligence correlate. We also know that the effect size is approximately in the range of $r = .20$ to $r = .40$. We can incorporate this knowledge in a Bayesian meta-analysis and compare the results to other approaches. The goals were thus (1) to obtain an additional overall estimate incorporating prior information, (2) to define the probability that the estimate is below or above a certain value, and (3) to explore the possibilities of using Bayesian inference in addition to the used frequentist approaches. In contrast to the other approaches, the Bayesian meta-analysis was an exploratory endeavor.

When applying a Bayesian meta-analysis is it important to specify what kind of prior information or assumptions were used before fitting the model. Otherwise, we could play around with specifications and report the one that flatters our analytical skills the most. Prior specifications for the summary effect were $\mu \sim N(0.3, 1)$ with sensitivity analyses $\mu \sim N(0.2, 1)$ and $\mu \sim N(0.4, 1)$. The values of 0.2, 0.3 and 0.4 represent the range of summary effects observed in previous meta-analyses. There was no reason to doubt that the summary effects would be normally distributed. Specifications concerning heterogeneity were $\tau \sim HC(0, 0.2)$ with sensitivity analysis $\tau \sim HC(0, 0.5)$. HC is an abbreviation for the Half-Cauchy distribution, which was chosen due to its favorable properties examining heterogeneity in a meta-analysis (Harrer et al., 2019). The values 0.2 and 0.5 for τ represent moderate to high heterogeneity among effect sizes not explained by sampling error. These values were also derived from previous meta-analyses. Fisher´s *z*-transformed correlation coefficients were used. Shortest credible intervals served as indicators of precision. In the preregistration I announced that I would like to conduct the Bayesian meta-analysis on the basis of the newly accumulated studies only. Reconsidering, there was no reason to limit the analysis to new data. A strong point of Bayesian methods is the possibility to update analyses as soon as new data are available. Analyses were based on the entire data set separated by sample type and included all three intelligence domains.

Dissemination Bias

Dissemination bias refers to the problem that not all studies or study outcomes are published, equally accessible or visible. As with any missing data problem, if studies easily accessible and visible differ from unpublished or hidden ones systematically for other reasons than study quality, results of systematic reviews could be biased (Mueller et al., 2016). Dissemination bias can arise on the level of whole studies (publication bias), or individual outcomes (outcome bias) that were not reported. There are a lot of reasons, why results may not be equally accessible. Some results may have been not reported, because statistical significance has not reached a certain threshold, the magnitude or direction of an effect was not considered interesting, results differed from expectations from funding parties, study authors have not published in a recognized journal due to various reasons, or current preferences and trends drove reporting in a certain direction (see Vevea et al., 2019).

I tried to tackle this problem in various ways. (1) Dissemination bias was avoided through an extensive search for grey literature and contact to authors and experts. (2) In-depth use of statistical methods allowed the detection and the assessment of the potential impact of dissemination bias. All these procedures were executed based on the meta-analytic approach in tradition of Hedges and Olkin (1985). This approach offers by far the widest (readily available) arsenal of methods to examine dissemination bias. The use of several methods is necessary because of the different potential causes of dissemination bias, and considered to best practice (Carter et al., 2019). The focus of these analyses was on published results based on healthy samples. Fisher´s *r*-to-*z* transformed correlation coefficients were used, except for analyses based on *p*-values. These utilized raw correlation coefficients, since this does not assume researchers´ *p*-values were derived from analyses using *z*-transformed correlations. All analyses were repeated for full-scale, verbal, and performance IQ data.

In previous meta-analyses researchers have reached different conclusions about the extent and impact of dissemination bias in the data. Whereas McDaniel (2005) has merely expressed concerns, Pietschnig et al. (2015) have found substantial bias leading to an overestimation of summary effects. Gignac and Bates (2017) have found no substantial bias in a healthy adult subset of the data.

Dissemination bias analyses started with a power-enhanced funnel plot (sunset plot, Kossmeier et al., 2020b), created with the R package *metaviz* (Kossmeier et al., 2020a).

Additional to basic funnel plot features e.g. the distribution of effect sizes in regard to summary effect and *p*-values, the sunset plot displayed the median power of all studies, the true effect size necessary such that the median power of the studies would have been 33% or 66%, results of a *test of excess significance* (Ioannidis & Trikalinos, 2007), and the *R-Index for expected replicability* (Schimmack, 2016). It provided information on whether low-powered significant studies were overrepresented, if studies generally were too successful in finding statistically significant results compared to expectations based on power, and about the replicability of the results. The x-axis of the sunset plot was set to display the Pearson correlation scale. The y-axis was set to display the standard error of effect sizes. Statistical power was calculated in reference to the meta-analytic summary effect obtained through the Hedges and Olkin meta-analysis.

As the next step, two versions of the *Sterne and Egger regression* (Sterne & Egger, 2005) and the *trim-and-fill* method (Duval & Tweedie, 2000) were applied. The *Sterne and Egger regression* showed, if funnel plot asymmetry was present, meaning if more studies reported positive effects compared to negative effects than expected by statistical probability. A weighted regression with a multiplicative dispersion term and a random-effects meta-regression model were computed to check if conclusions differed depending on the regression approach. In both cases, the regression analysis was based on the standard error of effect sizes. A one-tailed α level of .10 was used. Applying the *trim-and-fill* method allowed to visualize "missing" studies due to asymmetry. The left side of the funnel plot, the area of negative correlations, was of interest because researchers have more incentive to report positive correlations between brain volume and intelligence. Negative correlations are counterintuitive and may not have been reported. Recomputing the summary effect including these studies gave an impression about how the summary effect may have been affected by this selection mechanism. Both procedures were displayed in another, contour-enhanced funnel-plot to avoid overcrowding information in the sunset plot. I used the *metafor* for the *Sterne and Egger regression*, and *trim-and-fill analyses*.

In order to investigate the possibility of *p-hacking*, *p-curve* (Simonsohn et al., 2014), *p-uniform* (van Assen et al., 2015) and *p-uniform** (van Aert & van Assen, 2018) were applied. The *p-curve* analysis was performed using the website www.p-curve.com. For *p-uniform* and *p-uniform** I used the package *puniform* (van Aert, 2020). The term *p-hacking* describes a form of strategic research behavior. Results which are statistically significant might be easier to publish, gain more attention and affirm researchers´ interests or theories. In order to obtain significant results researchers could intentionally or unknowingly focus on analyses producing them while discarding statistically insignificant results. When effect sizes included in a meta-analysis have been influenced by this type of behavior, overall results may have been biased. The method of *p-curve* assumes, that if the summary effect estimate indeed reflects a true effect, the distribution of *p*-values must be right-skewed (more very small *p*-values). If the distribution is left-skewed, meaning there are more barely significant *p*-values, some researchers could have looked for significant effect sizes to report. *P-uniform* looks for similar *p*-value patterns with different algorithms. The drawback of both methods are their limitations in working with heterogeneous data (van Aert et al., 2016; McShane et al., 2016). *P-uniform** is designed to make it more robust when model assumptions, such as no heterogeneity in the data, are not met. Comparison with selection model approaches showed enhanced performance in case of heterogeneity (van Aert & van Assen, 2018). *P-uniform** is a relatively new method and needs further evaluation. Therefore, methods were used that have proven their effectiveness in case of heterogeneity. Two further selection model approaches were applied. The first approach was based on *p*-values (Vevea & Hedges, 1995), whereas the second one on the standard error of the effect sizes (Copas & Shi, 2001). With the use of selection model approaches researchers can specify different scenarios in which effect sizes may have been suppressed, and compute adjusted estimates in reference to these specifications (Hedges & Vevea, 2005). I set the α cut-offs for the *p*-value based analytic model from Vevea and Hedges (1995) to .010, .025, .050, .100, .250, .500, and .750 representing highly significant, the positive tail in a two-tailed test, barely significant, potential trends to significance and statistically insignificant thresholds. I used the *weightr* package (Coburn & Vevea, 2019) for this latter approach, and *metasens* (Schwarzer et al., 2020) to conduct the Copas and Shi analysis.

Two cumulative meta-analyses conducted, too. In one case, study results were sorted by publication year, in the other case by sample size. Cumulative meta-analyses help to examine the stability of the meta-analytic results in relation to one variable. The former was conducted to identify a possible time trend. Pietschnig et al. (2015) observed declining sizes from early studies to recent ones. The latter was conducted to determine if study precision (size) is related to the magnitude of the effect size estimate, a potential sign of publication bias.

Moderator analyses tested possible effects of dissemination bias related variables. A meta-regression with the type of report (reported in a journal or grey literature or through personal communication) as the predictor was conducted to examine effects of publication choices on effect size estimates. Lastly, a meta-regression with publication year as the predictor served as an addition to the cumulative meta-analyses ordered according to publication year.

Moderators

The correlation of brain volume and intelligence may have been influenced by several factors. As explained in the chapter "Hypotheses", the correlation may have changed with the age, sex or health-status of the subjects. A variable that influences a correlation is called a moderator variable. In the above-mentioned section can be seen which potential moderators were investigated. This section explains the methods used to do this. Categorical variables (variables with two or more levels) were investigated mainly with subgroup analyses, continuous variables (e.g. mean age) with meta-regressions. I used the *metafor* package for all following analyses.

Subgroup Analysis

Potential categorial moderators were assessed with a series of mixed effects subgroup analyses. These are called mixed effects analyses, because the within subgroup estimates were based on random-effects and the between group analyses were based on fixed-effect analyses. This is the crucial difference between subgroup analyses and meta-regressions (see below). In the former, the possibility is conceded that the heterogeneity patterns between the groups analyzed differ. For the latter, it is assumed that they are identical. The following variables were incorporated in subgroup analyses according to hypotheses: sample type (healthy or clinical), age (children or adults), sex (females or males), ethnicity (White, African, Hispanic/Latin, Asian). Other subgroup analyses were conducted for sensitivity or bias checks: type of report (published in a peer-reviewed journal or grey literature or personal communication), type of brain volume measurement (TBV or ICV). Analyses were mostly based on healthy samples. Data from clinical samples was incorporated to test hypothesis 3.

Meta-Regression

Univariate. Continuous moderators were assessed with weighted linear meta-regressions. Tested moderators were the proportion of males in samples, the publication year of studies, and the correlation of applied intelligence measurement with *g*. Defined as the percentage of male participants in a sample, the male ratio analysis complemented the subgroup analysis of sex categorially coded. The year of publication was tested as a predictor in order to identify a decline effect found in a previous meta-analysis (Pietschnig et al., 2015).

It complemented the cumulative meta-analysis ordered by publication year (see section "Dissemination Bias"). To test hypothesis 2 an RVE meta-regression approach was used.

Multiple. First, a potential interaction effect of age and sex was tested. Second, a theory-guided hierarchical weighted multiple mixed effects meta-regression included a combined assessment of potential moderators with reference to overall model fit. This way it was possible to examine potential moderators in one model, allowing comparison of effects. Multicollinearity was checked with an intercorrelation matrix. Noticeable correlations were assessed regarding meaning, possible influence and solutions. The presence of multicollinearity was formally defined as *variance inflation factors* (VIF) being above 4. Predictors with largest VIFs were dropped until all VIFs were below 4. The following predictors were included (in that order): correlation of applied intelligence measurement with *g*, study year, type of report (block 1); male ratio, ethnicity, mean age (block 2); study goal, number of included covariates in study (block 3). A permutation test, conducted with the "permutest" function within the *metafor* package, assessed the robustness of the final model. The whole procedure was repeated based on a subset excluding studies which only reported ICV. Results were compared, especially the effect on the predictor age. An age sex interaction effect was tested as well.

Exploring the Multiverse

This section describes methods used to explore the impact of specification choices made by previous meta-analysts including my own. Reasons why this was necessary are listed in section "Exploration: Researchers´ Degrees of Freedom". In the following three methods are discussed: *combinatorial meta-analysis*, *multiverse analysis* and *specification curve analysis*. The context in which they have been developed is briefly described, as how they can be applied to a meta-analytic context. I followed the guide by Voracek et al. (2019) and used their openly accessible R code in an adapted form. This code is available at https://osf.io/nkv46/.

Combinatorial Meta-Analysis

Combinatorial meta-analysis (Olkin et al., 2012) is a brute-force method aiming at analyzing all possible subsets of available data. The goal is to provide a sensitivity analysis asking if summary effects are substantially inflated by varying combinations of data subsets. Combinatorial meta-analyses can be regarded as sweeping variants of leave-one-out analyses. Leave-one-out analyses help to evaluate if a particular study has an inflating effect.

Combinatorial meta-analyses evaluate if there are study subsets yielding substantially different results. Two peculiarities of this approach must be noted. First, this method does compute all possible subsets, meaning that there will be (lots of) implausible combinations, too. Secondly, it can be computed for only one meta-analytical method at a time. This means that the method focuses on data, not analysis procedures.

The total amount of possible subsets is defined as 2^*k*-1. In the case of our full-scale IQ data based on healthy samples, there were 2^{122}-1 possible subsets of data. An inconceivably high number. Therefore, a random sample of 100,000 subsets was analyzed. I used the code by Voracek et al. (2019) which is designed to undersample intermediate subset sizes and oversample extreme subset sizes (containing either very few or most of the available studies). Under- and oversampling lead to a more rigid stress test. This is the only computational difference to the GOSH plots computed with the *metafor* package (see section "Hedges and Olkin Meta-Analysis"). The latter just fits 100,000 models based on random subsets.

Multiverse & Specification-Curve Analyses

What if we do not want to incorporate every possible subset, but reasonable ones only? If we believe, for example, that age plays a role in the relationship between brain volume and intelligence, then we will hardly construct the subset we want to analyze randomly, but in accordance to our intentions, believes or interests. Gignac and Bates (2017) have focused their analysis on adult samples. However, McDaniel (2005), and Pietschnig et al. (2015) have taken children into account as well. In 2016, Steegen et al. have introduced a well named method to look into the influence of specific data choices in primary studies. A *multiverse analysis* includes not only data construction options, but also data cleaning processes, such as handling outliers and missing data. Results from all possible combinations of these reasonable specifications are computed. The output is usually a histogram showing the distribution of *p*-values according to specifications. If the distribution of *p*-values is right-skewed, meaning most specification combinations yield highly significant results, we can assume with great confidence that a found effect is not the result of inflation based on data decisions.

Not only that data choices are important, but there are many ways how to analyze it, too. *Specification-curve* (Simonsohn et al., 2015) is a method originally focusing on analysis specifications. The intention is the same as in multiverse analysis with a focus on analytical decisions, for example which statistical methods and corrections are chosen. Graphical displays show how effects are influenced by each analysis combination. *Specification-curve* also adds inferential statistical procedures to further examine the results. This is not typically done in multiverse analyses, where inference relies on visual interpretation of histograms. The next section explains how these methods developed for primary studies can be applied to meta-analysis.

Specifications in Meta-Analysis

Data and analysis decisions also influence meta-analyses. A good example comes from previous meta-analysts working on brain volume and IQ. Gignac and Bates (2017) have reanalyzed the data which Pietschnig et al. (2015) had accumulated, and by choosing different data construction, data cleaning and analysis procedures, they have yielded a difference in results which has not been trivial in size (r = .39; r = .24). If such differences can occur when using the same data set, meta-analysts who collect data individually may obtain even more pronounced differences in results.

Voracek et al. (2019) have transferred the *multiverse*, and *specification curve analysis* to the meta-analytic context. The first step is to identify reasonable data and analysis specifications which might influence results. Following their title “Which data to meta-analyze, and how?”, Voracek et al. (2019) have termed data specifications “which” factors and analysis specifications “how” factors. The first step of their approach is identical for *multiverse* and *specification curve analysis*, as all (data and analysis) specifications can be included in both approaches. The integration of specifications in the R-code provided by Voracek et al. (2019) yields an Excel sheet with all possible combinations of specifications and their results. These data are used to create graphical outputs. Differences between multiverse and specification curve analyses are the different graphical outputs to be used, and that *specification curve* has inferential statistics aiding interpretation.

When performing a specification analysis, it must be decided which specifications will be considered. These decisions are degrees of freedom and must be substantiated. I concentrated on decisions of previous meta-analysts, including my own. I think these specifications are a good representation of reasonable choices to conduct a meta-analysis on the topic, leaving not much else to do. In order to be transparent, I also report which specifications were not included and explain why.

The following data or “which” factors were considered: (1) age group (adults only, children/adolescents, or both combined), (2) sample type (healthy samples only, clinical samples, or both combined) and (3) rating groups according to the correlation of the applied

IQ test with *g* (“abbreviated”, “full”, or all IQ tests). All incorporated data specifications made up for $3^3 = 27$ combinations. Verbal and performance IQ analyses did not include the rating factor. Therefore $3^2 = 9$ factors applied to verbal and performance IQ analyses. A coding approach by Gignac and Bates (2017) was not considered. They have chosen to code effect sizes preferably based on samples not separated in males and females. McDaniel (2005), and Pietschnig et al. (2015) have preferably coded effect sizes based on sex-separated samples. This coding choice was also applied in this thesis. Gignac and Bates (2017) explained their choice by saying that they were not interested in differences between the sexes. There are no apparent reasons why this could have an effect on results. Additionally, the number of affected correlations would have been small (less than 20%). Another factor which was not included was the handling of outliers by Gignac and Bates (2017). They have chosen to Winsorize sample sizes of studies identified as outliers. There was no outlier with substantial influence on summary effect in this thesis according to results from leave-one-out analyses. Furthermore, McDaniel (2005) has not included effect sizes based on estimates from brief intelligence tests, like the National Adult Reading Test (NART; Nelson, 1982). Potential effects are already considered in the correlation with *g* moderator analysis, so including this specification was unnecessary. McDaniel (2005) has also imputed missing standard deviations of intelligence scores with the mean range restriction in the data. This was infeasible to do with the updated data, because of the vastly different sample sizes easily inflating results. Nowadays, there are refined methods for imputing that kind of information, based on artifact distributions. Only about 50% of all studies contained information on range departure. Especially the lack of this information in some large-scale studies weighted heavily (e.g. Takeuchi et al., 2018; Cox et al., 2019; Mathias et al., 2020). Imputations based on a rather poorly informed distribution with very large differences in sample sizes did not seem appropriate.

The following meta-analytic analysis, or “how” factors were included: (1) The effect size to be used in analysis (*r*-to-*z* transformed coefficients, correlations corrected for their slight negative bias, correlations corrected for range departure, or raw correlation coefficients) and (2) the meta-analytic method of synthesis (the Hedges and Olkin method with the REML estimator and inverse variance weights, the Hunter and Schmidt method with the HS estimator and sample sizes as weights, or an unweighted approach resembling the numerous narrative reviews). The same “how” factors were applied to verbal and performance analyses. Not included were the RVE meta-regression and Bayesian approaches used in this thesis. Results of both approaches were highly similar to those obtained by the Hedges-Olkin approach. Analysis factors made up for 3*4 = 12 combinations. Together, data and analysis specification made up for 27*12 = 324 ways to construct and analyze the data. For verbal and performance IQ there were 9*12 = 108 combinations. Analyses were restricted to unique combinations with at least two studies.

Results

Study Selection

Efforts to update the data yielded 48 eligible citations, comprising 116 independent effect sizes (68 effect sizes based on full-scale IQ, 26 on verbal IQ and 22 on performance IQ). The total number of individual participants from newly accumulated studies was 21071 for full-scale IQ, 2545 for verbal IQ and 2265 for performance IQ. The number of total individual participants tripled due to the inclusion of studies based on very large data sets compared to the Pietschnig et al. (2015) data. Although my search was limited to studies published 2012 or later, I found some earlier eligible studies. The number and sample sizes were low, indicating that Pietschnig et al. (2015) covered the previous time frame comprehensively. My search included results until May 2020.

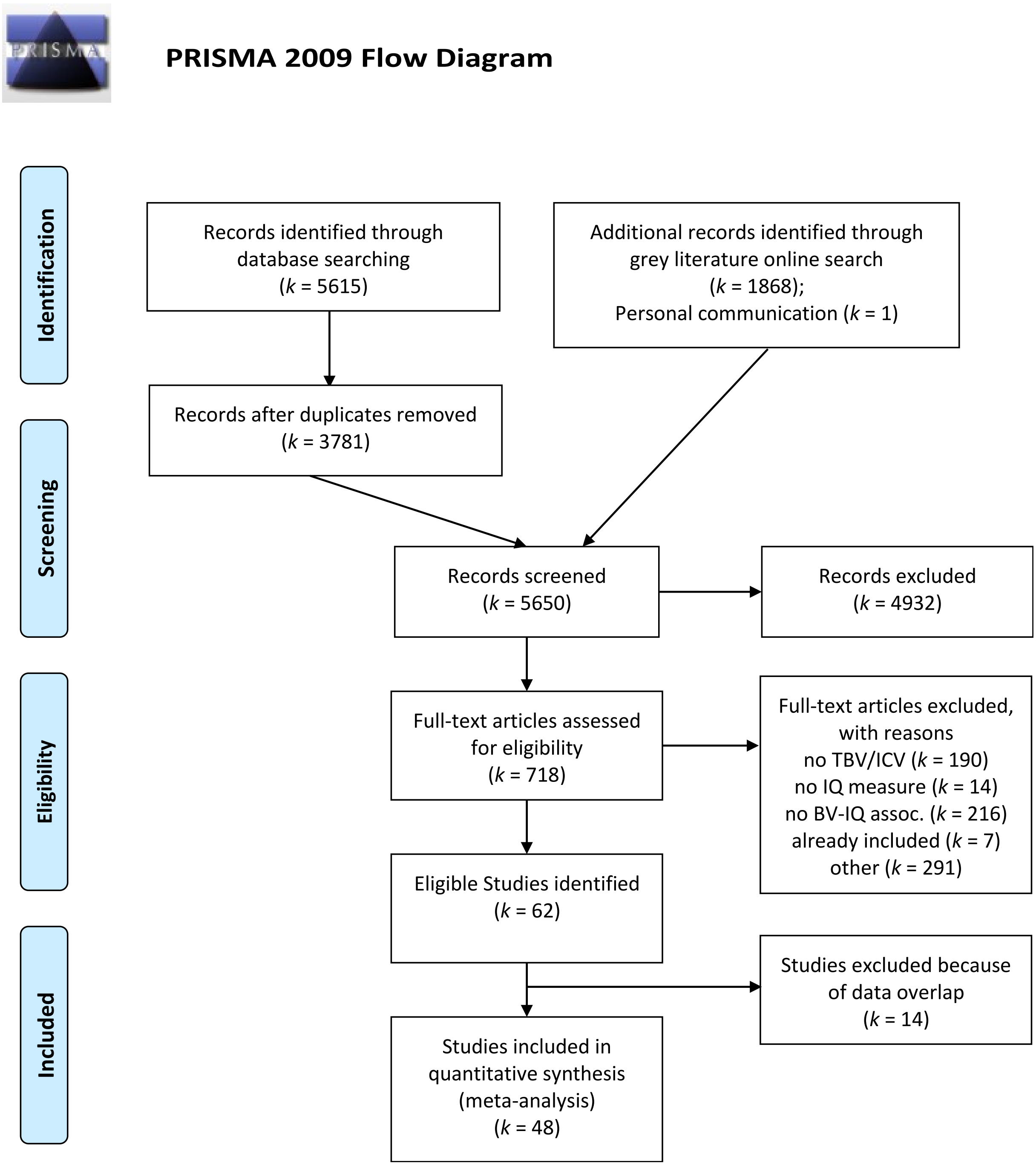
Data for the RVE meta-regressions (modeling data dependence) comprised more effect sizes than stated above. Characteristics of these additional effect sizes are discussed below in section “Robust Variance Estimation Meta-Regression”.

The study selection process is illustrated in Figure 1. Considering the total number of screened citations and the ratio of this number and the overall yield, the process was characterized by high recall and low precision. This was true for standard data base searches, as well as for grey literature.

I applied some minor changes to the Pietschnig et al. (2015) data set. Two effect sizes concerning verbal IQ associated with Egan et al. (1994) were deleted. The same correlations were already included with the Egan et al. (1995) results. Furthermore, one effect size (also verbal IQ) associated with Raz et al. (1995) was deleted. I could not find the effect size in the paper and suspected it to be a coding error ($r = .9$). Five effect sizes from Witelson et al. (2006) were also excluded as brain volume was evaluated postmortem. One effect size associated with Shapleske (2002) was deleted. It was based on four participants. Variances cannot be computed for sample sizes below five.

As part of the strategy to search for grey literature, I contacted experts, authors of studies with good proximity to the here posed questions but lacking necessary information, and authors of eligible studies missing information for further analyses (e.g. studies reporting an effect as not significant). In my e-mails I asked for that specific information and/or general advice on further relevant studies, while attaching a reference list with already included studies and a data sharing agreement.

Figure 1


PRISMA
PRISMA 2009 Flow Diagram
Identification
Screening
Eligibility
Included
Records identified through database searching (*k* = 5615)
Additional records identified through grey literature online search (*k* = 1868); Personal communication (*k* = 1)
Records after duplicates removed (*k* = 3781)
Records screened (*k* = 5650)
Records excluded (*k* = 4932)
Full-text articles assessed for eligibility (*k* = 718)
Full-text articles excluded, with reasons
no TBV/ICV (*k* = 190)
no IQ measure (*k* = 14)
no BV-IQ assoc. (*k* = 216)
already included (*k* = 7)
other (*k* = 291)
Eligible Studies identified (*k* = 62)
Studies excluded because of data overlap (*k* = 14)
Studies included in quantitative synthesis (meta-analysis) (*k* = 48)

Note. Template retrieved from www.prisma-statement.org; see Moher et al. (2009).

The data sharing agreement was intended as a convenience option for busy researchers, who were willing to provide data so I could extract information on my own. 26% of my e-mails were answered. I am most grateful to all responders in a busy time. I would like to thank Wai Kwong Tang from Hong Kong University to make the effort to look for results associated with Lin (2016). Unfortunately, the obtained correlations could not be used, because they were based on GM-ICV ratios. I also would like to thank Birgitte Fagerlund from Copenhagen University for six eligible correlations associated with Jensen et al. (2019) provided via e-mail (October 10, 2020).

Study Characteristics

A list of all studies considered in this thesis is provided in Table 1. It contains 434 effect sizes in total. These include dependent effect sizes used in the RVE approach. The variables displayed are a selection intended to provide an overview. Data with all variables are available at https://osf.io/y6msp/

Results of Individual Studies

Figure 2 shows a rainforest plot (Schild & Voracek, 2015) based on full-scale IQ data from healthy samples. Studies are ordered by publication year from early to recent studies. The rainforest plot shows a consistent pattern of positive associations between in vivo brain volume and full-scale IQ. From approximately 2010 the “raindrops” gets thicker in color and diminish in width indicating growing sample sizes. Results from the most recent studies come from large samples. Figure 3 shows results for full-scale IQ data from clinical samples. The pattern of effect sizes is not as consistent as for healthy samples. The majority of studies yielded positive associations between in vivo brain volume and full-scale IQ, too, but with more variation in effect sizes. There is also an increased number of negative correlations. Rainforest plots for verbal and performance IQ show comparable patterns and are available in Appendix B.

Table 1

Details of Included Studies

Study	Year	Review	Sample type	Mean age	Male ratio	Reporting	IQ domain	Measure	Type of test	*n*	*r*
Yeo et al.	1987	2	patients	38.4	34.00%	reported	FSIQ	CT	WAIS	41	.07
Yeo et al.	1987	2	patients	38.4	34.00%	reported	performance	CT	WAIS	41	.06
Yeo et al.	1987	2	patients	38.4	34.00%	reported	verbal	CT	WAIS	41	.12
Willerman et al.	1991	1	healthy	18.9	0.00%	reported	FSIQ	MRI	WAIS-R	20	.33
Willerman et al.	1991	1	healthy	18.9	100.00%	reported	FSIQ	MRI	WAIS-R	20	.51
Andreasen et al.	1993	2	healthy	38	0.00%	reported	performance	MRI	WAIS-R	30	.30
Andreasen et al.	1993	2	healthy	38	100.00%	reported	performance	MRI	WAIS-R	37	.43
Andreasen et al.	1993	2	healthy	38	0.00%	reported	verbal	MRI	WAIS-R	30	.43
Andreasen et al.	1993	2	healthy	38	100.00%	reported	verbal	MRI	WAIS-R	37	.33
Andreasen et al.	1993	1	healthy	38	0.00%	reported	FSIQ	MRI	WAIS-R	30	.44
Andreasen et al.	1993	1	healthy	38	100.00%	reported	FSIQ	MRI	WAIS-R	37	.40
Raz et al.	1993	2	healthy	43.8	59.00%	reported	verbal	MRI	V3	29	.10
Raz et al.	1993	1	healthy	43.8	59.00%	reported	FSIQ	MRI	CFIT	29	.43
Castellanos et al.	1994	2	healthy	12.1	100.00%	reported	verbal	MRI	WISC-R	46	.33
Castellanos et al.	1994	1	healthy	12.1	100.00%	reported	FSIQ	MRI	WISC-R	46	.33
Harvey et al.	1994	2	healthy	31.6	55.00%	reported	verbal	MRI	NART	34	.69
Harvey et al.	1994	2	patients	35.6	38.00%	reported	verbal	MRI	NART	26	.38
Harvey et al.	1994	2	patients	31.1	77.00%	reported	verbal	MRI	NART	48	.24
Jones et al.	1994	2	healthy	31.7	64.00%	reported	verbal	CT	NART or […]	67	.30
Wickett et al.	1994	1	healthy	25	0.00%	reported	FSIQ	MRI	MAB FS	40	.40
Wickett et al.	1994	2	healthy	25	0.00%	reported	performance	MRI	MAB	40	.28
Wickett et al.	1994	2	healthy	25	0.00%	reported	verbal	MRI	MAB	40	.44
Bigler	1995	2	patients	29.4	71.00%	reported	FSIQ	MRI	WAIS-R	72	-.03
Egan et al.	1995	1	healthy	22.5	100.00%	reported	FSIQ	MRI	WAIS-R	40	.31
Egan et al.	1995	2	healthy	22.5	100.00%	reported	performance	MRI	WAIS-R	40	.22
Egan et al.	1995	2	healthy	22.5	100.00%	reported	verbal	MRI	WAIS-R	40	.21
Haier et al.	1995	2	patients	26.39	54.00%	reported	FSIQ	MRI	WAIS-R	28	.65
Kareken et al.	1995	3	healthy	27.66	63.00%	reported	performance	MRI	WAIS-R	68	.26
Kareken et al.	1995	3	patients	29.75	63.00%	reported	performance	MRI	WAIS-R	68	.18
Kareken et al.	1995	3	healthy	27.66	63.00%	reported	verbal	MRI	COWA […]	68	.24
Kareken et al.	1995	3	patients	29.75	63.00%	reported	verbal	MRI	COWA […]	68	.36
Kareken et al.	1995	1	healthy	27.66	63.00%	PC	FSIQ	MRI	WAIS-R	68	.30
Raz et al.	1995	2	patients	35.2	77.00%	reported	FSIQ	MRI	WPPSI-R + BCS	11	-.24
Reiss et al.	1995	2	healthy	11.28	42.00%	PC	FSIQ	MRI	WISC-R or […]	87	.00

Table 1
Details of Included Studies

Study	Year	Review	Sample type	Mean age	Male ratio	Reporting	IQ domain	Measure	Type of test	*n*	*r*
Reiss et al.	1995	2	patients	10.8	35.00%	reported	FSIQ	MRI	WISC-R or […]	51	.25
Reiss et al.	1996	1	healthy	10.6	0.00%	PC	FSIQ	MRI	unknown	57	.37
Reiss et al.	1996	2	healthy	10.1	100.00%	PC	FSIQ	MRI	unknown	12	.52
Blatter et al.	1997	2	patients	NA	NA	reported	performance	MRI	WAIS-R	21	.47
Blatter et al.	1997	2	patients	NA	NA	reported	verbal	MRI	WAIS-R	22	.57
Mori et al.	1997	2	patients	70.2	38.00%	reported	performance	MRI	WAIS-R	60	.37
Mori et al.	1997	2	patients	70.2	38.00%	reported	verbal	MRI	WAIS-R	60	.37
Mori et al.	1997	2	patients	70.2	38.00%	reported	FSIQ	MRI	WAIS-R	60	.40
Paradiso et al.	1997	2	healthy	24.8	53.00%	reported	performance	MRI	WAIS-R	62	.32
Paradiso et al.	1997	2	healthy	24.8	53.00%	reported	verbal	MRI	WAIS-R	62	.27
Paradiso et al.	1997	3	healthy	24.8	53.00%	reported	verbal	MRI	WAIS-R	62	.11
Paradiso et al.	1997	2	healthy	24.8	53.00%	reported	FSIQ	MRI	WAIS-R	62	.38
Flashman et al.	1998	2	healthy	27	53.00%	reported	performance	MRI	WAIS-R	90	.26
Flashman et al.	1998	2	healthy	27	53.00%	reported	verbal	MRI	WAIS-R	90	.16
Flashman et al.	1998	1	healthy	27	53.00%	reported	FSIQ	MRI	WAIS-R	90	.25
Gur et al.	1999	3	healthy	25	0.00%	reported	performance	MRI	WAIS-R […]	40	.57
Gur et al.	1999	3	healthy	27	100.00%	reported	performance	MRI	WAIS-R […]	40	.35
Gur et al.	1999	2	healthy	25	0.00%	reported	verbal	MRI	WAIS-R […]	40	.40
Gur et al.	1999	2	healthy	27	100.00%	PC	verbal	MRI	WAIS-R […]	40	.00
Gur et al.	1999	1	healthy	25	0.00%	reported	FSIQ	MRI	WAIS-R	40	.40
Gur et al.	1999	1	healthy	27	100.00%	reported	FSIQ	MRI	WAIS-R […]	40	.39
Leonard et al.	1999	2	healthy	42	100.00%	PC	performance	MRI	WAIS-R	33	.00
Leonard et al.	1999	2	patients	43	100.00%	PC	performance	MRI	WAIS-R	37	.00
Leonard et al.	1999	2	healthy	42	100.00%	PC	verbal	MRI	WAIS-R	33	.00
Leonard et al.	1999	2	patients	43	100.00%	PC	verbal	MRI	WAIS-R	37	.00
Tan et al.	1999	1	healthy	22	0.00%	reported	FSIQ	MRI	CFIT	54	.62
Tan et al.	1999	1	healthy	22	100.00%	reported	FSIQ	MRI	CFIT	49	.28
Warwick et al.	1999	2	healthy	21.5	0.00%	PC	verbal	MRI	Quick IQ Test	13	.00
Warwick et al.	1999	2	healthy	21.5	100.00%	PC	verbal	MRI	Quick IQ Test	25	.00
Warwick et al.	1999	2	patients	21.6	0.00%	PC	verbal	MRI	Quick IQ Test	11	.00
Warwick et al.	1999	2	patients	21.8	100.00%	PC	verbal	MRI	Quick IQ Test	10	.00
Warwick et al.	1999	2	patients	21.8	100.00%	PC	verbal	MRI	Quick IQ Test	10	.00
Warwick et al.	1999	2	patients	21.63	100.00%	reported	verbal	MRI	Quick IQ Test	45	.31
Warwick et al.	1999	2	patients	21.55	0.00%	reported	verbal	MRI	Quick IQ Test	24	.53

Table 1

Details of Included Studies

Study	Year	Review	Sample type	Mean age	Male ratio	Reporting	IQ domain	Measure	Type of test	*n*	*r*
Garde et al.	2000	1	healthy	80.7	0.00%	PC	FSIQ	MRI	WAIS	22	.22
Garde et al.	2000	1	healthy	80.7	100.00%	PC	FSIQ	MRI	WAIS	46	.07
Isaacs et al.	2000	2	healthy	7.75	73.00%	PC	FSIQ	MRI	WISC-III	11	-.03
Isaacs et al.	2000	2	healthy	7.75	73.00%	PC	performance	MRI	WISC-III	11	-.18
Isaacs et al.	2000	2	healthy	7.75	73.00%	PC	verbal	MRI	WISC-III	11	-.04
Isaacs et al.	2000	2	healthy	7.75	38.00%	PC	FSIQ	MRI	WISC-III	8	.55
Isaacs et al.	2000	2	healthy	7.75	38.00%	PC	performance	MRI	WISC-III	8	.35
Isaacs et al.	2000	2	healthy	7.75	38.00%	PC	verbal	MRI	WISC-III	8	.57
Kumra et al.	2000	2	patients	12.3	81.00%	PC	FSIQ	MRI	WISC-III or […]	27	.00
Kumra et al.	2000	2	patients	14.4	57.00%	PC	FSIQ	MRI	WISC-III or […]	44	.00
Lawson et al.	2000	2	patients	NA	NA	reported	FSIQ	MRI	WISC-III or […]	47	.43
Pennington et al.	2000	1	healthy	19.06	44.00%	reported	FSIQ	MRI	WISC-R or […]	36	.31
Pennington et al.	2000	2	healthy	16.97	58.00%	reported	FSIQ	MRI	WISC-R or […]	96	.42
Schoenemann et al.	2000	2	healthy	23.2	0.00%	reported	verbal	MRI	MAB	36	.12
Schoenemann et al.	2000	1	healthy	23.2	0.00%	PC	FSIQ	MRI	RSPM	72	.21
Wickett et al.	2000	1	healthy	24.97	100.00%	reported	FSIQ	MRI	MAB	68	.35
Wickett et al.	2000	2	healthy	24.97	100.00%	reported	performance	MRI	MAB	68	.31
Wickett et al.	2000	2	healthy	24.97	100.00%	reported	verbal	MRI	MAB	68	.33
Castellanos et al.	2001	1	patients	9.7	0.00%	reported	FSIQ	MRI	WISC-R or […]	40	.36
Coffey et al.	2001	2	healthy	74.85	38.00%	reported	performance	MRI	WAIS-R […]	318	.06
Coffey et al.	2001	2	healthy	74.85	38.00%	reported	verbal	MRI	Verbal fluency	319	-.06
Aylward et al.	2002	2	healthy	18.9	92.00%	reported	performance	MRI	unknown	83	.09
Aylward et al.	2002	2	patients	18.8	87.00%	reported	performance	MRI	unknown	67	.10
Aylward et al.	2002	2	healthy	18.9	92.00%	reported	verbal	MRI	unknown	83	-.01
Aylward et al.	2002	2	patients	18.8	87.00%	reported	verbal	MRI	unknown	67	.08
Aylward et al.	2002	1	healthy	NA	100.00%	PC	FSIQ	MRI	unknown	46	-.13
Aylward et al.	2002	1	healthy	NA	NA	PC	FSIQ	MRI	unknown	30	.08
Aylward et al.	2002	1	patients	18.8	87.00%	reported	FSIQ	MRI	unknown	67	.10
MacLullich et al.	2002	2	healthy	67.8	100.00%	reported	verbal	MRI	NART	97	.30
MacLullich et al.	2002	1	healthy	67.8	100.00%	reported	FSIQ	MRI	RSPM	95	.39
Nosarti et al.	2002	1	healthy	14.9	65.00%	PC	FSIQ	MRI	unknown	42	.37
Shapleske et al.	2002	1	healthy	33.3	100.00%	PC	FSIQ	MRI	unknown	23	.13
Collinson et al.	2003	2	healthy	16.4	60.00%	PC	FSIQ	MRI	WISC-R or […]	22	-.13
Collinson et al.	2003	2	patients	16.8	67.00%	PC	FSIQ	MRI	WISC-R or […]	32	-.27

Table 1
Details of Included Studies

Study	Year	Review	Sample type	Mean age	Male ratio	Reporting	IQ domain	Measure	Type of test	*n*	*r*
Collinson et al.	2003	2	healthy	16.4	60.00%	PC	performance	MRI	WISC-R or […]	22	-.17
Collinson et al.	2003	2	patients	16.8	67.00%	PC	performance	MRI	WISC-R or […]	32	-.19
Collinson et al.	2003	2	healthy	16.4	60.00%	PC	verbal	MRI	WISC-R or […]	22	-.09
Collinson et al.	2003	2	patients	16.8	67.00%	PC	verbal	MRI	WISC-R or […]	32	-.28
Giedd	2003	1	healthy	NA	0.00%	PC	FSIQ	NA	unknown	8	.46
Giedd	2003	1	healthy	NA	100.00%	PC	FSIQ	NA	unknown	7	.17
Giedd	2003	1	healthy	NA	0.00%	PC	FSIQ	NA	unknown	7	-.67
Giedd	2003	1	healthy	NA	100.00%	PC	FSIQ	NA	unknown	7	.67
Giedd	2003	1	healthy	NA	0.00%	PC	FSIQ	NA	unknown	39	.34
Giedd	2003	1	healthy	NA	100.00%	PC	FSIQ	NA	unknown	63	.27
Kesler et al.	2003	2	patients	26.16	52.00%	reported	verbal	MRI	WAIS-R	25	.57
Kesler et al.	2003	2	patients	26.16	52.00%	reported	FSIQ	MRI	WAIS-R	25	.47
Yurgelun-Todd et al.	2003	3	healthy	14.6	0.00%	reported	performance	MRI	WAIS-III	24	.07
Yurgelun-Todd et al.	2003	3	healthy	14.5	100.00%	reported	performance	MRI	WAIS-III	13	.48
Yurgelun-Todd et al.	2003	2	healthy	14.6	0.00%	reported	verbal	MRI	Shipley	24	.17
Yurgelun-Todd et al.	2003	2	healthy	14.5	100.00%	reported	verbal	MRI	Shipley	13	.19
Yurgelun-Todd et al.	2003	3	healthy	14.6	0.00%	reported	verbal	MRI	WAIS-III	24	.19
Yurgelun-Todd et al.	2003	3	healthy	14.5	100.00%	reported	verbal	MRI	WAIS-III	13	.55
Yurgelun-Todd et al.	2003	2	healthy	14.6	0.00%	reported	FSIQ	MRI	Shipley	24	.20
Yurgelun-Todd et al.	2003	2	healthy	14.5	100.00%	reported	FSIQ	MRI	Shipley	13	.26
Frangou et al.	2004	1	healthy	15.05	50.00%	reported	FSIQ	MRI	WISC-III or […]	40	.41
Isaacs et al.	2004	2	healthy	15.9	0.00%	PC	FSIQ	MRI	Wechsler	38	.24
Isaacs et al.	2004	2	healthy	15.9	100.00%	PC	FSIQ	MRI	Wechsler	38	.27
Isaacs et al.	2004	2	healthy	14.86	50.00%	PC	FSIQ	MRI	Wechsler	16	.49
Isaacs et al.	2004	2	healthy	15.9	0.00%	PC	performance	MRI	Wechsler	38	.21
Isaacs et al.	2004	2	healthy	15.9	100.00%	PC	performance	MRI	Wechsler	38	.15
Isaacs et al.	2004	2	healthy	15.6	0.00%	PC	verbal	MRI	Wechsler	38	.20
Ivanovic et al.	2004	2	healthy	18	0.00%	reported	performance	MRI	WAIS-R	49	.38
Ivanovic et al.	2004	2	healthy	18	100.00%	reported	performance	MRI	WAIS-R	47	.52
Ivanovic et al.	2004	2	healthy	18	0.00%	reported	verbal	MRI	WAIS-R	49	.33
Ivanovic et al.	2004	2	healthy	18	100.00%	reported	verbal	MRI	WAIS-R	47	.55
Ivanovic et al.	2004	1	healthy	18	0.00%	reported	FSIQ	MRI	WAIS-R	49	.37
Ivanovic et al.	2004	1	healthy	18	100.00%	reported	FSIQ	MRI	WAIS-R	47	.55
Rojas et al.	2004	2	healthy	43.62	47.00%	PC	FSIQ	MRI	WAIS-R or […]	17	.31

Table 1
Details of Included Studies

Study	Year	Review	Sample type	Mean age	Male ratio	Reporting	IQ domain	Measure	Type of test	*n*	*r*
Rojas et al.	2004	2	patients	30.3	87.00%	PC	FSIQ	MRI	WAIS-R or […]	15	.07
Rojas et al.	2004	2	healthy	43.62	47.00%	PC	performance	MRI	WAIS-R or […]	17	.27
Rojas et al.	2004	2	patients	30.3	87.00%	PC	performance	MRI	WAIS-R or […]	15	.15
Rojas et al.	2004	2	healthy	43.62	47.00%	PC	verbal	MRI	WAIS-R or […]	17	.19
Rojas et al.	2004	2	patients	30.3	87.00%	PC	verbal	MRI	WAIS-R or […]	15	.30
Toulopoulou et al.	2004	2	patients	42.23	50.00%	reported	verbal	MRI	WAIS-R	201	.28
Toulopoulou et al.	2004	2	patients	42.23	50.00%	reported	FSIQ	MRI	WAIS-R	201	.28
Waiter et al.	2004	2	healthy	15.5	100.00%	PC	performance	MRI	WISC-III-R	16	.23
Waiter et al.	2004	2	patients	15.4	100.00%	PC	performance	MRI	WISC-III-R	16	.10
Waiter et al.	2004	2	healthy	15.5	100.00%	PC	verbal	MRI	WISC-III-R	16	.20
Waiter et al.	2004	2	patients	15.4	100.00%	PC	verbal	MRI	WISC-III-R	16	-.17
Waiter et al.	2004	2	healthy	15.5	100.00%	PC	FSIQ	MRI	WISC-III-R	16	.13
Waiter et al.	2004	2	patients	15.4	100.00%	PC	FSIQ	MRI	WISC-III-R	16	-.06
Antonova et al.	2005	2	healthy	33.72	58.00%	PC	verbal	MRI	WAIS-III	43	.24
Antonova et al.	2005	2	patients	40.49	60.00%	PC	verbal	MRI	WAIS-III	44	.16
Lodygensky et al.	2005	2	healthy	8.42	57.00%	PC	FSIQ	MRI	WISC-R	21	.46
Lodygensky et al.	2005	2	patients	8.58	53.00%	PC	FSIQ	MRI	WISC-R	60	.35
Thoma et al.	2005	2	healthy	23.5	100.00%	reported	FSIQ	MRI	RPM + […]	19	.27
Debbané et al.	2006	2	healthy	15.1	43.00%	PC	FSIQ	MRI	WISC-III or […]	41	.16
Debbané et al.	2006	2	patients	16.7	37.00%	PC	FSIQ	MRI	WISC-III or[…]	43	.16
Rojas et al.	2006	2	healthy	21.41	100.00%	PC	FSIQ	MRI	WAIS-III or […]	23	.46
Rojas et al.	2006	2	patients	20.79	100.00%	PC	FSIQ	MRI	WAIS-III or […]	24	.30
Rojas et al.	2006	2	healthy	21.41	100.00%	PC	performance	MRI	WAIS-III or […]	23	.09
Rojas et al.	2006	2	patients	20.79	100.00%	PC	performance	MRI	WAIS-III or […]	24	.31
Rojas et al.	2006	2	healthy	21.41	100.00%	PC	verbal	MRI	WAIS-III or […]	23	.55
Rojas et al.	2006	2	patients	20.79	100.00%	PC	verbal	MRI	WAIS-III or […]	24	.28
Staff et al.	2006	1	healthy	79.5	61.00%	PC	FSIQ	MRI	RSPM	102	-.10
Staff et al.	2006	2	healthy	79.5	61.00%	PC	verbal	MRI	NART	102	-.14
Voelbel et al.	2006	2	healthy	10.77	100.00%	PC	performance	MRI	WISC-III	13	.06
Voelbel et al.	2006	2	patients	10.16	100.00%	PC	performance	MRI	WISC-III	38	-.02
Voelbel et al.	2006	2	patients	10.16	100.00%	PC	verbal	MRI	WISC-III	38	.08
Voelbel et al.	2006	2	patients	10.08	100.00%	PC	performance	MRI	WISC-III	12	-.48
Voelbel et al.	2006	2	healthy	10.77	100.00%	PC	verbal	MRI	WISC-III	13	-.15
Voelbel et al.	2006	2	patients	10.08	100.00%	PC	verbal	MRI	WISC-III	12	.23

Table 1
Details of Included Studies

Study	Year	Review	Sample type	Mean age	Male ratio	Reporting	IQ domain	Measure	Type of test	*n*	*r*
Voelbel et al.	2006	2	healthy	10.77	100.00%	PC	FSIQ	MRI	WISC-III	13	-.11
Voelbel et al.	2006	2	patients	10.16	100.00%	PC	FSIQ	MRI	WISC-III	38	.02
Voelbel et al.	2006	2	patients	10.08	100.00%	PC	FSIQ	MRI	WISC-III	12	-.14
Wozniak et al.	2006	2	healthy	12.4	46.20%	PC	FSIQ	MRI	WISC-III or […]	13	.59
Wozniak et al.	2006	2	patients	12.3	50.00%	PC	FSIQ	MRI	WISC-III or […]	14	.41
Chiang et al.	2007	2	healthy	NA	NA	reported	performance	MRI	WAIS	16	.41
Chiang et al.	2007	2	patients	29.2	45.00%	reported	performance	MRI	WAIS	39	.10
Chiang et al.	2007	2	healthy	NA	NA	reported	verbal	MRI	WAIS	16	-.44
Chiang et al.	2007	2	patients	29.2	45.00%	reported	verbal	MRI	WAIS	39	-.02
DeBoer et al.	2007	2	healthy	10.5	NA	PC	performance	MRI	WISC-III or […]	20	-.22
DeBoer et al.	2007	2	patients	10.75	NA	PC	performance	MRI	WISC-III or […]	21	.38
DeBoer et al.	2007	2	healthy	10.5	NA	PC	verbal	MRI	WISC-III or […]	20	-.20
DeBoer et al.	2007	2	patients	10.75	NA	PC	verbal	MRI	WISC-III or […]	21	.30
DeBoer et al.	2007	2	healthy	10.5	NA	PC	FSIQ	MRI	WISC-III or […]	20	-.55
DeBoer et al.	2007	2	patients	10.75	NA	PC	FSIQ	MRI	WISC-III or […]	21	.25
Doernte	2007	3	healthy	58.5	0.00%	grey	verbal	MRI	HAWIE-R	18	-.23
Doernte	2007	3	healthy	58.5	100.00%	grey	verbal	MRI	HAWIE-R	17	.18
Doernte	2007	3	patients	59.1	0.00%	grey	verbal	MRI	HAWIE-R	12	-.02
Doernte	2007	3	patients	59.1	100.00%	grey	verbal	MRI	HAWIE-R	23	-.01
Fine et al.	2007	2	healthy	40.1	45.00%	PC	FSIQ	MRI	WASI	44	-.11
Fine et al.	2007	2	healthy	10.47	63.00%	PC	FSIQ	MRI	WASI	24	.23
Luders et al.	2007	2	healthy	28.48	45.00%	reported	FSIQ	MRI	WAIS-R	62	.28
Nakamura et al.	2007	2	healthy	40.8	90.00%	PC	performance	MRI	WAIS-III	43	.29
Nakamura et al.	2007	2	patients	40.6	90.00%	PC	performance	MRI	WAIS-III	44	.34
Nakamura et al.	2007	2	healthy	40.8	90.00%	PC	verbal	MRI	WAIS-III	44	.40
Nakamura et al.	2007	2	patients	40.6	90.00%	PC	verbal	MRI	WAIS-III	44	.26
Nakamura et al.	2007	2	healthy	40.8	90.00%	PC	FSIQ	MRI	WAIS-III	44	.38
Nakamura et al.	2007	2	patients	40.6	90.00%	PC	FSIQ	MRI	WAIS-III	43	.32
Narr et al.	2007	3	healthy	28.24	46.20%	reported	FSIQ	MRI	WAIS	63	.36
Schottenbauer et al.	2007	2	healthy	34.32	0.00%	PC	performance	MRI	WAIS-R	22	.30
Schottenbauer et al.	2007	2	healthy	37.77	100.00%	PC	performance	MRI	WAIS-R	35	.17
Schottenbauer et al.	2007	2	patients	40.9	0.00%	PC	performance	MRI	WAIS-R	68	.29
Schottenbauer et al.	2007	2	patients	39.65	100.00%	PC	performance	MRI	WAIS-R	203	.17
Schottenbauer et al.	2007	2	healthy	34.32	0.00%	PC	verbal	MRI	WAIS-R	22	.54

Table 1
Details of Included Studies

Study	Year	Review	Sample type	Mean age	Male ratio	Reporting	IQ domain	Measure	Type of test	*n*	*r*
Schottenbauer et al.	2007	2	healthy	37.77	100.00%	PC	verbal	MRI	WAIS-R	35	.38
Schottenbauer et al.	2007	2	patients	40.9	0.00%	PC	verbal	MRI	WAIS-R	68	.43
Schottenbauer et al.	2007	2	patients	39.66	100.00%	PC	verbal	MRI	WAIS-R	202	.28
Schottenbauer et al.	2007	2	healthy	34.32	0.00%	PC	FSIQ	MRI	WAIS-R	22	.60
Schottenbauer et al.	2007	2	healthy	37.77	100.00%	PC	FSIQ	MRI	WAIS-R	35	.33
Schottenbauer et al.	2007	2	patients	40.96	0.00%	PC	FSIQ	MRI	WAIS-R	69	.34
Schottenbauer et al.	2007	2	patients	39.64	100.00%	PC	FSIQ	MRI	WAIS-R	205	.28
Schumann et al.	2007	2	healthy	13.1	100.00%	reported	performance	MRI	WASI	22	.25
Schumann et al.	2007	2	healthy	13.1	100.00%	reported	verbal	MRI	WASI	22	.38
Schumann et al.	2007	2	healthy	13.1	100.00%	reported	FSIQ	MRI	WASI	22	.41
Amat et al.	2008	2	healthy	31.5	56.00%	PC	performance	MRI	WAIS-R	27	.18
Amat et al.	2008	2	healthy	31.5	56.00%	PC	verbal	MRI	WAIS-R	27	-.29
Amat et al.	2008	2	healthy	31.5	56.00%	PC	FSIQ	MRI	WAIS-R	27	-.11
Choi et al.	2008	3	healthy	21.6	54.30%	reported	FSIQ	MRI	WAIS-R	164	.35
Ebner et al.	2008	2	healthy	32.45	51.00%	PC	verbal	MRI	MWT-B	37	-.13
Ebner et al.	2008	2	patients	34.52	68.00%	PC	verbal	MRI	MWT-B	44	.15
Raz et al.	2008	2	healthy	51.11	43.00%	PC	FSIQ	MRI	CFIT	55	.18
Raz et al.	2008	2	patients	59.75	25.00%	PC	FSIQ	MRI	CFIT	32	-.02
Raz et al.	2008	2	healthy	51.11	43.00%	PC	verbal	MRI	V2 & V3	55	.13
Raz et al.	2008	2	patients	59.75	25.00%	PC	verbal	MRI	V2 & V3	31	.15
Castro-Fornieles et al.	2009	2	healthy	14.6	11.00%	PC	performance	MRI	WISC-R	9	.55
Castro-Fornieles et al.	2009	2	patients	14.5	8.00%	PC	performance	MRI	WISC-R	12	.38
Castro-Fornieles et al.	2009	2	healthy	14.6	11.00%	PC	verbal	MRI	WISC-R	9	.43
Castro-Fornieles et al.	2009	2	patients	14.5	8.00%	PC	verbal	MRI	WISC-R	12	.11
Miller et al.	2009	2	healthy	12.08	NA	reported	verbal	MRI	WJIII	11	-.65
Miller et al.	2009	2	patients	9.25	NA	reported	verbal	MRI	WJIII	5	.84
Miller et al.	2009	2	patients	16.53	NA	reported	verbal	MRI	WJIII	6	.76
Miller et al.	2009	2	healthy	9.25	33.00%	reported	FSIQ	MRI	WJIII	12	.23
Miller et al.	2009	2	healthy	12.08	NA	reported	FSIQ	MRI	WJIII	11	-.11
Miller et al.	2009	2	patients	16.53	63.00%	reported	FSIQ	MRI	WJIII	16	-.30
Qiu et al.	2009	2	healthy	10.5	53.00%	PC	performance	MRI	WISC-III or […]	66	.12
Qiu et al.	2009	2	patients	10.4	57.00%	PC	performance	MRI	WISC-III or […]	47	.20
Qiu et al.	2009	2	healthy	10.5	53.00%	PC	verbal	MRI	WISC-III or […]	66	.35
Qiu et al.	2009	2	patients	10.4	57.00%	PC	verbal	MRI	WISC-III or […]	47	.21

Table 1
Details of Included Studies

Study	Year	Review	Sample type	Mean age	Male ratio	Reporting	IQ domain	Measure	Type of test	*n*	*r*
Qiu et al.	2009	2	healthy	10.5	53.00%	PC	FSIQ	MRI	WISC-III or […]	66	.26
Qiu et al.	2009	2	patients	10.4	57.00%	PC	FSIQ	MRI	WISC-III or […]	47	.26
Shenkin et al.	2009	2	healthy	78.4	29.00%	reported	verbal	MRI	CWA	107	.13
Shenkin et al.	2009	2	healthy	78.4	29.00%	reported	FSIQ	MRI	MHT + […]	99	.21
Van Leeuwen et al.	2009	2	healthy	9.1	50.00%	reported	performance	MRI	WISC-III	209	.28
Van Leeuwen et al.	2009	3	healthy	9.1	50.00%	reported	performance	MRI	WISC-III	209	.12
Van Leeuwen et al.	2009	2	healthy	9.1	50.00%	reported	verbal	MRI	WISC-III	209	.33
Van Leeuwen et al.	2009	2	healthy	9.1	50.00%	reported	FSIQ	MRI	RSPM	209	.20
Weniger et al.	2009	2	healthy	33	0.00%	PC	performance	MRI	HAWIE-R	25	.24
Weniger et al.	2009	2	patients	32	0.00%	PC	performance	MRI	HAWIE-R	10	.23
Weniger et al.	2009	2	healthy	33	0.00%	PC	verbal	MRI	HAWIE-R	25	.00
Weniger et al.	2009	2	patients	32	0.00%	PC	verbal	MRI	HAWIE-R	13	.35
Weniger et al.	2009	2	patients	32	0.00%	PC	performance	MRI	HAWIE-R	13	.16
Weniger et al.	2009	2	patients	32	0.00%	PC	verbal	MRI	HAWIE-R	10	-.17
Weniger et al.	2009	2	patients	32	0.00%	PC	FSIQ	MRI	HAWIE-R	10	.02
Weniger et al.	2009	2	healthy	33	0.00%	PC	FSIQ	MRI	HAWIE-R	25	.15
Weniger et al.	2009	2	patients	32	0.00%	PC	FSIQ	MRI	HAWIE-R	13	.27
Zeegers et al.	2009	2	patients	3.72	91.00%	reported	FSIQ	MRI	unknown	21	.06
Zeegers et al.	2009	2	patients	3.44	92.00%	reported	FSIQ	MRI	unknown	10	.73
Betjemann et al.	2010	2	healthy	11.4	52.00%	reported	performance	MRI	WISC-R	142	.42
Betjemann et al.	2010	2	healthy	11.4	52.00%	reported	verbal	MRI	WISC-R	142	.14
Hermann	2010	2	healthy	33.34	42.00%	PC	performance	MRI	Wechsler	67	.33
Hermann	2010	2	patients	36.09	35.00%	PC	performance	MRI	Wechsler	77	.09
Hermann	2010	2	healthy	33.34	42.00%	PC	verbal	MRI	Wechsler	67	.23
Hermann	2010	2	patients	36.09	35.00%	PC	verbal	MRI	Wechsler	77	.28
Hermann	2010	2	healthy	33.34	42.00%	PC	FSIQ	MRI	Wechsler	67	.31
Hermann	2010	2	patients	36.09	35.00%	PC	FSIQ	MRI	Wechsler	77	.21
Hogan et al.	2010	2	healthy	68.69	53.00%	PC	FSIQ	MRI	RSPM	234	.11
Hogan et al.	2010	2	healthy	68.69	53.00%	PC	verbal	MRI	NART	235	.00
Isaacs et al.	2010	2	healthy	15.75	0.00%	PC	performance	MRI	WISC-III or […]	24	.00
Isaacs et al.	2010	2	healthy	15.75	100.00%	reported	performance	MRI	WISC-III or […]	26	.19
Isaacs et al.	2010	2	healthy	15.75	0.00%	PC	verbal	MRI	WISC-III or […]	24	.00
Isaacs et al.	2010	2	healthy	15.75	100.00%	reported	verbal	MRI	WISC-III or […]	26	.48
Isaacs et al.	2010	2	healthy	15.75	0.00%	PC	FSIQ	MRI	WISC-III or […]	24	.00

Table 1
Details of Included Studies

Study	Year	Review	Sample type	Mean age	Male ratio	Reporting	IQ domain	Measure	Type of test	*n*	*r*
Isaacs et al.	2010	2	healthy	15.75	100.00%	reported	FSIQ	MRI	WISC-III or […]	26	.36
Lange et al.	2010	2	healthy	10.88	0.00%	reported	FSIQ	MRI	WASI	166	.22
Lange et al.	2010	2	healthy	10.95	100.00%	reported	FSIQ	MRI	WASI	143	.23
Wallace et al.	2010	2	healthy	11.8	48.00%	reported	performance	MRI	WASI	649	.14
Wallace et al.	2010	2	healthy	11.8	48.00%	reported	verbal	MRI	WASI	649	.13
Wallace et al.	2010	2	healthy	11.8	48.00%	reported	FSIQ	MRI	WASI	649	.14
Ashtari et al.	2011	2	healthy	18.5	100.00%	reported	FSIQ	MRI	WRAT-III	14	.57
Ashtari et al.	2011	2	patients	19.3	100.00%	reported	FSIQ	MRI	WRAT-III	14	.29
Chen et al.	2011	3	healthy	22.56	44.00%	reported	FSIQ	MRI	WASI	27	.02
Chen et al.	2011	3	patients	23	27.00%	reported	FSIQ	MRI	WASI	37	.41
Chen et al.	2011	3	patients	23.07	47.00%	reported	FSIQ	MRI	WASI	30	.68
Kievit et al.	2011	2	healthy	21.1	36.00%	PC	FSIQ	MRI	WAIS-III	80	.29
Kievit et al.	2011	2	healthy	21.1	36.00%	PC	verbal	MRI	WAIS-III	80	.23
Tate et al.	2011	2	patients	81.7	43.00%	PC	FSIQ	MRI	Shipley	194	.00
Aydin et al.	2012	2	healthy	15.1	100.00%	reported	FSIQ	MRI	WISC-R	30	.40
Aydin et al.	2012	2	healthy	15.1	100.00%	reported	performance	MRI	WISC-R	30	.34
Aydin et al.	2012	2	healthy	15.1	100.00%	reported	verbal	MRI	WISC-R	30	.26
Burgaleta et al.	2012	2	healthy	19.88	44.00%	reported	FSIQ	MRI	APM, […]	100	.17
Bigler et al.	2013	3	patients	10.66	58.00%	reported	performance	MRI	WISC-IV: PSI	47	.00
Bigler et al.	2013	3	patients	10.67	68.00%	reported	performance	MRI	WISC-IV: PSI	32	.00
Bigler et al.	2013	3	patients	10.14	58.00%	reported	performance	MRI	WISC-IV: PSI	27	.00
Royle et al.	2013	2	healthy	72.47	100.00%	reported	FSIQ	MRI	WAIS-III	293	.26
Royle et al.	2013	2	healthy	72.6	0.00%	reported	FSIQ	MRI	WAIS-III	327	.27
Royle et al.	2013	3	healthy	72.47	100.00%	reported	performance	MRI	WAIS-III	293	.25
Royle et al.	2013	3	healthy	72.6	0.00%	reported	performance	MRI	WAIS-III	327	.25
Royle et al.	2013	3	healthy	72.47	100.00%	reported	performance	MRI	WAIS-III	293	.14
Royle et al.	2013	3	healthy	72.6	0.00%	reported	performance	MRI	WAIS-III	327	.18
Royle et al.	2013	3	healthy	72.47	100.00%	reported	performance	MRI	WAIS-III	293	.22
Royle et al.	2013	3	healthy	72.6	0.00%	reported	performance	MRI	WAIS-III	327	.33
Royle et al.	2013	3	healthy	72.47	100.00%	reported	performance	MRI	WAIS-III	293	.17
Royle et al.	2013	3	healthy	72.6	0.00%	reported	performance	MRI	WAIS-III	327	.34
Royle et al.	2013	3	healthy	72.47	100.00%	reported	verbal	MRI	WAIS-III	293	.10
Royle et al.	2013	3	healthy	72.6	0.00%	reported	verbal	MRI	WAIS-III	327	.22
Royle et al.	2013	3	healthy	72.47	100.00%	reported	verbal	MRI	WAIS-III	293	.11

Table 1

Details of Included Studies

Study	Year	Review	Sample type	Mean age	Male ratio	Reporting	IQ domain	Measure	Type of test	*n*	*r*
Royle et al.	2013	3	healthy	72.6	0.00%	reported	verbal	MRI	WAIS-III	327	.23
Zelko et al.	2013	3	healthy	14.9	53.00%	reported	performance	MRI	WAIS or […]	36	.30
Zelko et al.	2013	3	healthy	14.9	53.00%	reported	performance	MRI	WAIS or […]	36	-.12
Zelko et al.	2013	3	patients	14.6	49.00%	reported	performance	MRI	WAIS or […]	108	.21
Zelko et al.	2013	3	patients	14.6	49.00%	reported	performance	MRI	WAIS or […]	108	.09
Zelko et al.	2013	3	healthy	14.9	53.00%	reported	verbal	MRI	WAIS or […]	36	.04
Zelko et al.	2013	3	healthy	14.9	53.00%	reported	verbal	MRI	WAIS or […]	36	.33
Zelko et al.	2013	3	patients	14.6	49.00%	reported	verbal	MRI	WAIS or […]	108	.23
Zelko et al.	2013	3	patients	14.6	49.00%	reported	verbal	MRI	WAIS or […]	108	.26
Zelko et al.	2013	3	healthy	14.9	53.00%	reported	FSIQ	MRI	WAIS or […]	36	.25
Zelko et al.	2013	3	patients	14.6	49.00%	reported	FSIQ	MRI	WAIS or […]	108	.23
Bjuland et al.	2014	3	patients	20.1	41.00%	reported	performance	MRI	WAIS-II	43	.48
Bjuland et al.	2014	3	patients	20.1	41.00%	reported	performance	MRI	WAIS-II	43	.48
Bjuland et al.	2014	3	patients	20.1	41.00%	reported	verbal	MRI	WAIS-III	43	.44
Bjuland et al.	2014	3	patients	20.1	41.00%	reported	verbal	MRI	WAIS-III	43	.54
Bjuland et al.	2014	3	healthy	20.3	42.00%	reported	FSIQ	MRI	WAIS-III	60	.36
Bjuland et al.	2014	3	patients	20.1	41.00%	reported	FSIQ	MRI	WAIS-III	43	.56
Grunewaldt et al.	2014	3	patients	10.17	34.80%	reported	verbal	MRI	WISC-III	21	.00
Grunewaldt et al.	2014	3	patients	10.17	34.80%	reported	FSIQ	MRI	WISC-III	21	.00
Jenkins et al.	2014	3	healthy	11.7	42.00%	reported	FSIQ	MRI	WASI or […]	102	.19
MacDonald et al.	2014	3	healthy	11.6	100.00%	reported	performance	MRI	WASI	142	.29
MacDonald et al.	2014	3	healthy	11.3	100.00%	reported	performance	MRI	WASI	161	.19
MacDonald et al.	2014	3	healthy	11.6	100.00%	reported	verbal	MRI	WASI	142	.13
MacDonald et al.	2014	3	healthy	11.3	100.00%	reported	verbal	MRI	WASI	161	.18
MacDonald et al.	2014	3	healthy	11.6	100.00%	reported	FSIQ	MRI	WASI	142	.23
MacDonald et al.	2014	3	healthy	11.3	0.00%	reported	FSIQ	MRI	WASI	161	.22
McCoy et al.	2014	3	patients	13	100.00%	reported	FSIQ	MRI	WISC-IV	10	.59
McCoy et al.	2014	3	patients	13	0.00%	reported	FSIQ	MRI	WISC-IV	16	.62
Zhu et al.	2014	3	healthy	20.41	41.00%	reported	FSIQ	MRI	WAIS-R	316	.10
Boberg et al.	2015	3	healthy	8.00	55.00%	grey	FSIQ	MRI	WISC-IV	10.00	.69
Boberg et al.	2015	3	healthy	8.30	50.00%	grey	FSIQ	MRI	WISC-IV	21.00	.00
Grazioplene et al.	2015	3	healthy	26.26	51.00%	reported	performance	MRI	WAIS-IV	285	.30
Grazioplene et al.	2015	3	healthy	21.73	54.00%	reported	performance	MRI	WASI	125	.04
Grazioplene et al.	2015	3	healthy	22.94	100.00%	reported	performance	MRI	WAIS-III	107	.04

Table 1
Details of Included Studies

Study	Year	Review	Sample type	Mean age	Male ratio	Reporting	IQ domain	Measure	Type of test	*n*	*r*
Grazioplene et al.	2015	3	healthy	26.26	51.00%	reported	verbal	MRI	WAIS-IV	285	.18
Grazioplene et al.	2015	3	healthy	21.73	54.00%	reported	verbal	MRI	WASI	125	.04
Grazioplene et al.	2015	3	healthy	22.94	100.00%	reported	verbal	MRI	WAIS-III	107	.10
Grazioplene et al.	2015	3	healthy	26.26	51.00%	reported	FSIQ	MRI	WAIS-IV	285	.28
Grazioplene et al.	2015	3	healthy	21.73	54.00%	reported	FSIQ	MRI	WASI	125	.04
Grazioplene et al.	2015	3	healthy	22.94	100.00%	reported	FSIQ	MRI	WAIS-IV	107	.08
Lefebvre et al.	2015	3	healthy	17	83.00%	reported	performance	MRI	unknown	284	.18
Lefebvre et al.	2015	3	patients	16.6	88.00%	reported	performance	MRI	unknown	254	.17
Lefebvre et al.	2015	3	healthy	17	83.00%	reported	verbal	MRI	unknown	354	.22
Lefebvre et al.	2015	3	patients	16.6	88.00%	reported	verbal	MRI	unknown	318	.08
Lefebvre et al.	2015	3	healthy	17	83.00%	reported	FSIQ	MRI	unknown	354	.23
Lefebvre et al.	2015	3	patients	16.6	88.00%	reported	FSIQ	MRI	unknown	318	.04
Paul et al.	2015	3	healthy	24.57	0.00%	reported	verbal	MRI	Span, […]	90	.25
Paul et al.	2015	3	healthy	24.07	100.00%	reported	verbal	MRI	Span, […]	121	.18
Paul et al.	2015	3	healthy	24.57	0.00%	reported	FSIQ	MRI	BOMAT, […]	90	.14
Paul et al.	2015	3	healthy	24.07	100.00%	reported	FSIQ	MRI	BOMAT, […]	121	.13
Walters et al.	2015	3	patients	17.32	100.00%	reported	FSIQ	MRI	WAIS or […]	178	.19
Ballester-Plane et al.	2016	3	patients	25.1	67.00%	reported	performance	MRI	WASI	30	.72
Ballester-Plane et al.	2016	3	patients	25.1	67.00%	reported	verbal	MRI	PPVT-III	30	.71
Ballester-Plane et al.	2016	3	patients	25.1	67.00%	reported	FSIQ	MRI	RCPM	30	.73
Bathelt et al.	2016	3	healthy	9.93	54.00%	reported	FSIQ	MRI	WASI-II	63	.07
Bathelt et al.	2016	3	patients	9.35	64.70%	reported	FSIQ	MRI	WASI-II	139	.02
Bohlken et al.	2016	3	healthy	32.7	42.00%	reported	performance	MRI	WAIS-III	164	.31
Bohlken et al.	2016	3	healthy	32.7	42.00%	reported	performance	MRI	WAIS-III	164	.12
Bohlken et al.	2016	3	healthy	32.7	42.00%	reported	verbal	MRI	WAIS-III	164	.18
Bohlken et al.	2016	3	healthy	32.7	42.00%	reported	verbal	MRI	WAIS-III	164	.26
Bohlken et al.	2016	3	healthy	32.7	42.00%	reported	verbal	MRI	WAIS-III	164	.00
Bohlken et al.	2016	3	healthy	32.7	42.00%	reported	FSIQ	MRI	WAIS-III	164	.26
Ferreira et al.	2016	3	healthy	45.1	49.00%	reported	performance	MRI	WAIS-III	73	.33
Ferreira et al.	2016	3	healthy	45.1	49.00%	reported	verbal	MRI	WAIS-III	73	.36
Ferreira et al.	2016	3	healthy	45.1	49.00%	reported	verbal	MRI	WAIS-III	73	.50
Gregory et al.	2016	3	healthy	14.7	42.90%	reported	FSIQ	MRI	Matrix, […]	662	.24
Monson et al.	2016	3	patients	7.5	50.00%	reported	performance	MRI	WASI	134	.31
Monson et al.	2016	3	patients	7.5	50.00%	reported	verbal	MRI	WASI	134	.11

Table 1
Details of Included Studies

Study	Year	Review	Sample type	Mean age	Male ratio	Reporting	IQ domain	Measure	Type of test	*n*	*r*
Monson et al.	2016	3	patients	7.5	50.00%	reported	FSIQ	MRI	WASI	134	.26
Nikolaidis et al.	2016	3	healthy	21.15	34.00%	reported	verbal	MRI	Memory, […]	71	.13
Nikolaidis et al.	2016	3	healthy	21.15	34.00%	reported	FSIQ	MRI	RAPM, […]	71	.44
Treit et al.	2016	3	healthy	11.9	48.00%	reported	FSIQ	MRI	WRIT or […]	66	.09
Treit et al.	2016	3	patients	12.5	53.00%	reported	FSIQ	MRI	WRIT or […]	50	.21
Amaral et al.	2017	3	healthy	3	100.00%	reported	FSIQ	MRI	MSEL	49	.35
Amaral et al.	2017	3	patients	3.075	100.00%	reported	FSIQ	MRI	MSEL	19	-.18
Amaral et al.	2017	3	patients	3.133	100.00%	reported	FSIQ	MRI	MSEL	110	.01
Arhan et al.	2017	3	healthy	9.2	46.00%	reported	performance	MRI	WISC-R	46	.77
Arhan et al.	2017	3	healthy	9.2	46.00%	reported	performance	MRI	WISC-R	46	.24
Arhan et al.	2017	3	healthy	9.2	46.00%	reported	performance	MRI	WISC-R	46	.04
Arhan et al.	2017	3	healthy	9.2	46.00%	reported	performance	MRI	WISC-R	46	.04
Arhan et al.	2017	3	healthy	9.2	46.00%	reported	performance	MRI	WISC-R	46	.27
Arhan et al.	2017	3	healthy	9.2	46.00%	reported	verbal	MRI	WISC-R	46	.71
Arhan et al.	2017	3	healthy	9.2	46.00%	reported	verbal	MRI	WISC-R	46	.54
Arhan et al.	2017	3	healthy	9.2	46.00%	reported	verbal	MRI	WISC-R	46	.38
Arhan et al.	2017	3	healthy	9.2	46.00%	reported	verbal	MRI	WISC-R	46	.45
Arhan et al.	2017	3	healthy	9.2	46.00%	reported	verbal	MRI	WISC-R	46	.24
Arhan et al.	2017	3	healthy	9.2	46.00%	reported	verbal	MRI	WISC-R	46	.31
Arhan et al.	2017	3	healthy	9.2	46.00%	reported	FSIQ	MRI	WISC-R	46	.51
Martinez et al.	2017	3	healthy	19.6	0.00%	reported	performance	MRI	DAT-SR, […]	40	.39
Martinez et al.	2017	3	healthy	20.2	100.00%	reported	performance	MRI	DAT-SR, […]	40	.24
Martinez et al.	2017	3	healthy	19.6	0.00%	reported	verbal	MRI	DAT-VR, […]	40	.28
Martinez et al.	2017	3	healthy	20.2	100.00%	reported	verbal	MRI	DAT-VR, […]	40	-.04
Ritchie et al.	2017	3	healthy	92.1	45.00%	reported	FSIQ	MRI	WAIS-III	34	.23
Ritchie et al.	2017	3	healthy	92.1	45.00%	reported	performance	MRI	WAIS-III	34	.19
van der Linden et al.	2017	3	healthy	28.82	0.00%	reported	FSIQ	MRI	Matrices, […]	503	.26
van der Linden et al.	2017	3	healthy	28.82	100.00%	reported	FSIQ	MRI	Matrices, […]	393	.25
van der Vlugt et al.	2017	3	patients	7	83.00%	reported	FSIQ	CT	MSEL or […]	70	.00
Vreeker et al.	2017	3	healthy	44.6	49.00%	reported	FSIQ	MRI	WAIS-III	160	.28
Annink et al.	2018	3	patients	9.79	48.00%	reported	FSIQ	MRI	WISC-III	52	.43
Jensen et al.	2018	3	healthy	24.91	59.00%	PC	FSIQ	MRI	WAIS-III	56	.30
Jensen et al.	2018	3	patients	24.69	57.40%	PC	FSIQ	MRI	WAIS-III	54	.14
Jensen et al.	2018	3	healthy	24.91	59.00%	PC	performance	MRI	WAIS-III	56	.19

Table 1

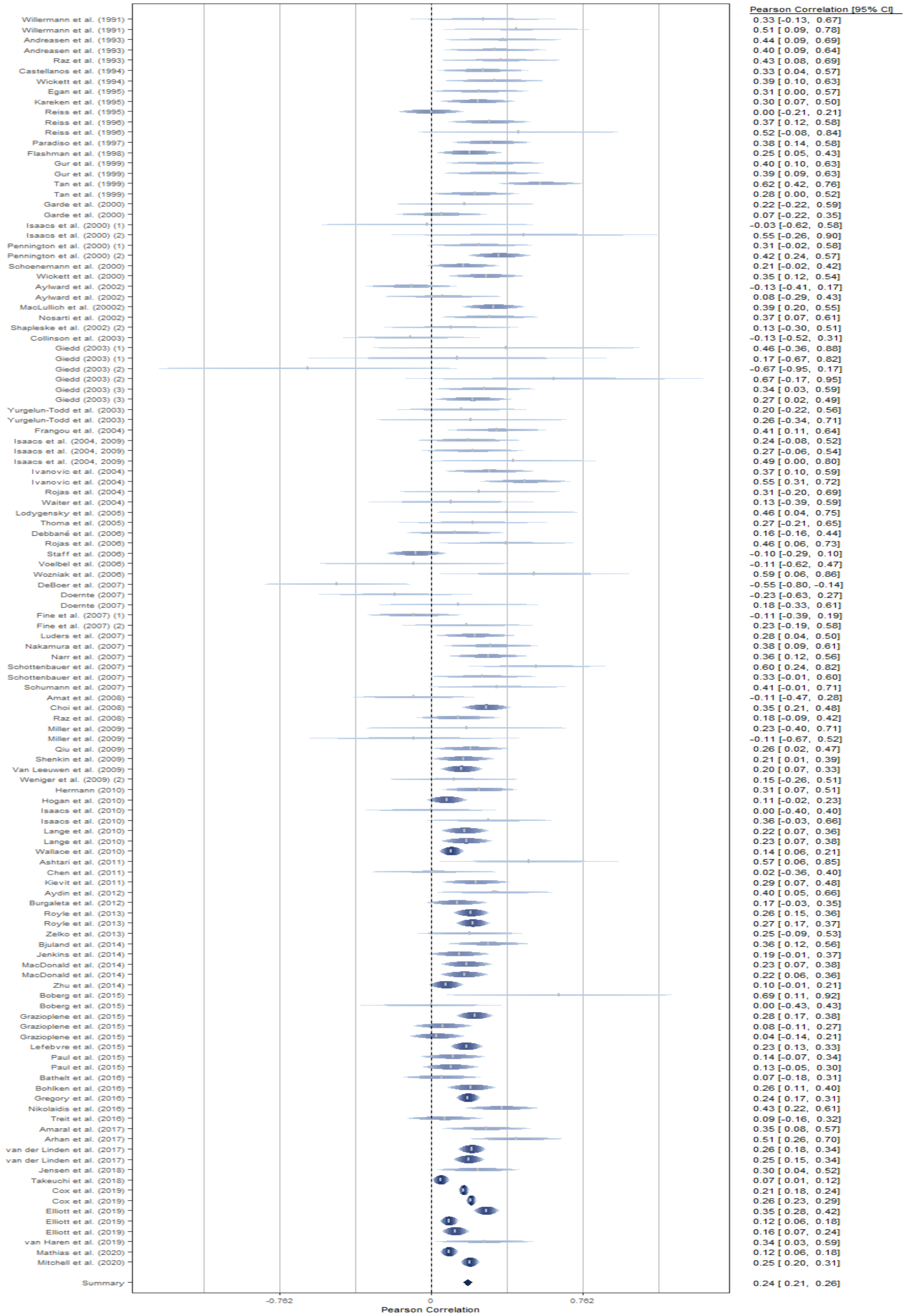
Details of Included Studies

Study	Year	Review	Sample type	Mean age	Male ratio	Reporting	IQ domain	Measure	Type of test	*n*	*r*
Jensen et al.	2018	3	patients	24.69	57.40%	PC	performance	MRI	WAIS-III	54	.20
Jensen et al.	2018	3	healthy	24.91	59.00%	PC	verbal	MRI	WAIS-III	56	.30
Jensen et al.	2018	3	patients	24.69	57.40%	PC	verbal	MRI	WAIS-III	54	.09
Lammers et al.	2018	3	patients	72	61.00%	reported	FSIQ	MRI	Memory, […]	282	.27
Mankovsky et al.	2018	3	patients	62.3	34.00%	reported	performance	MRI	processing speed	93	.08
Mankovsky et al.	2018	3	patients	62.3	34.00%	reported	verbal	MRI	RAVL […]	93	.02
Nygaard et al.	2018	3	patients	18.96	60.00%	reported	FSIQ	MRI	WASI	82	.30
Sreedharan et al.	2018	3	patients	10.8	66.00%	reported	FSIQ	MRI	WISC	30	.00
Takeuchi et al.	2018	3	healthy	20.8	58.00%	reported	FSIQ	MRI	Tanaka B	1319	.07
Tozer et al.	2018	3	patients	70.01	65.00%	reported	performance	MRI	BIRT, […]	121	.28
Tozer et al.	2018	3	patients	70.01	65.00%	reported	FSIQ	MRI	span, […]	121	.23
Ahn et al.	2019	3	patients	32.97	42.00%	reported	FSIQ	MRI	K-WAIS-R	38	.00
Cox et al.	2019	3	healthy	63.13	100.00%	reported	FSIQ	MRI	Matrix, […]	3900	.21
Cox et al.	2019	3	healthy	63.13	0.00%	reported	FSIQ	MRI	Matrix, […]	4192	.26
de Zwarte et al.	2019	3	patients	27.49	60.00%	reported	FSIQ	MRI	WAIS-III	516	.29
de Zwarte et al.	2019	3	patients	52.85	32.00%	reported	FSIQ	MRI	GIT	85	.06
Elliott et al.	2019	3	healthy	45	48.00%	reported	FSIQ	MRI	WAIS-IV	596	.35
Elliott et al.	2019	3	healthy	22.23	47.00%	reported	FSIQ	MRI	Shipley	1163	.12
Elliott et al.	2019	3	healthy	20.26	47.00%	reported	FSIQ	MRI	WASI	515	.16
Hiraiwa et al.	2019	3	patients	9.43	52.00%	reported	FSIQ	MRI	WISC-IV	27	.34
van Haren et al.	2019	3	healthy	12.74	53.00%	reported	FSIQ	MRI	WISC-III or […]	40	.34
van Haren et al.	2019	3	patients	13.77	30.00%	reported	FSIQ	MRI	WISC-III or […]	40	.53
van Haren et al.	2019	3	patients	14.52	56.00%	reported	FSIQ	MRI	WISC-III or […]	66	.39
Mathias et al.	2020	3	healthy	39.6	43.00%	reported	FSIQ	MRI	Verbal Learning	1216	.12
Mitchell et al.	2020	3	healthy	22.3	38.00%	reported	FSIQ	MRI	MAB + […]	1097	.25

Note. NA = info not available; Review: 1 = included in McDaniel (2005), 2 = additional studies gathered by Pietschnig et al. (2015), 3 = newly accumulated studies; Reporting: reported = published in a journal article, grey = published as thesis/dissertation, PC = result obtained via personal communication; FSIQ = full-scale IQ; Measure: technology used to measure in vivo brain volume, either CT or MRI. Type of test: IQ test, sometimes only example of used tests displayed (indicated by brackets), full information explaining all abbreviations are available in codebook and data files at https://osf.io/y6msp/. Published study outcomes with *r* = exactly 0 represent correlations set to zero, because no eligible numerical value was available.

Figure 2

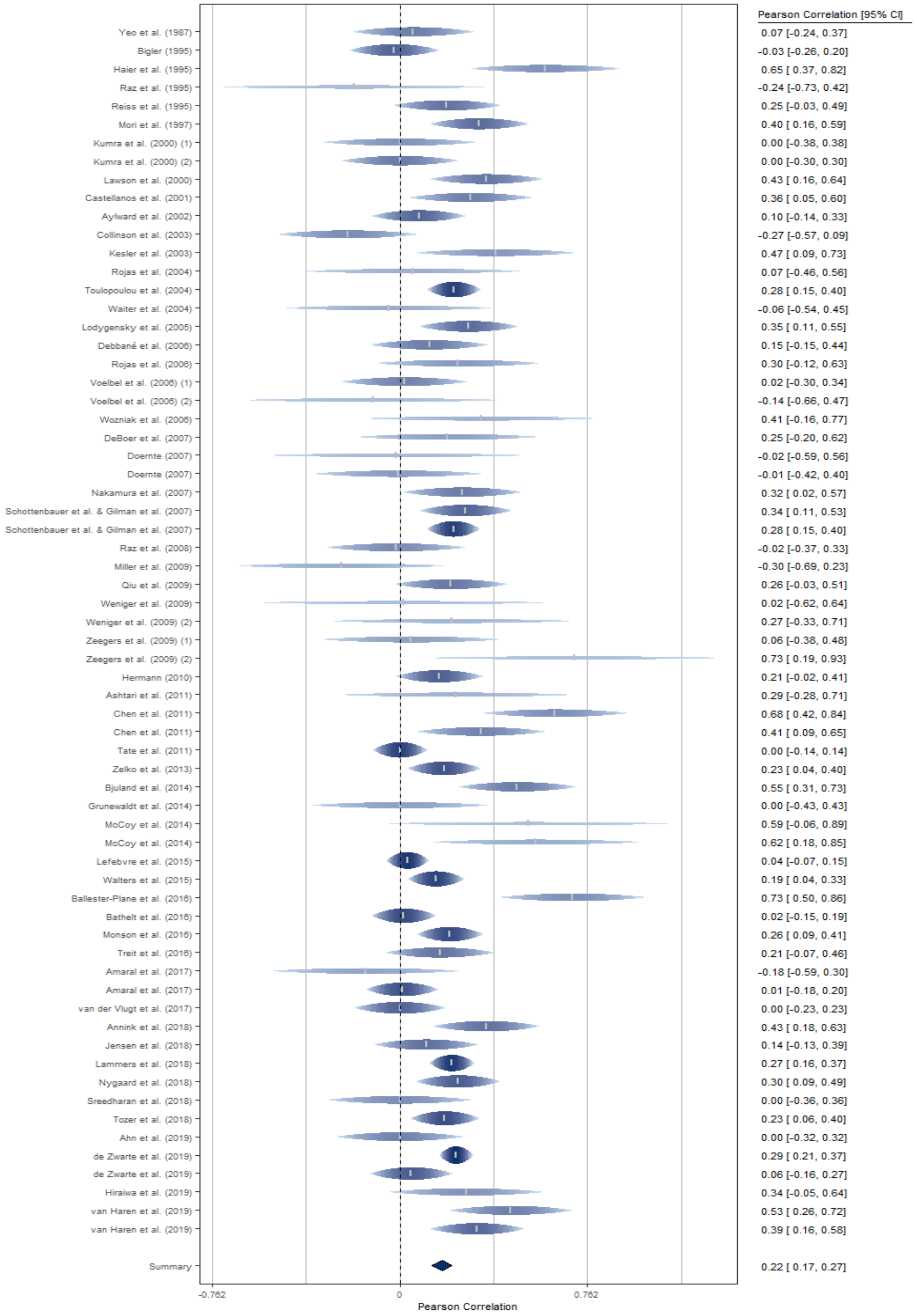
Rainforest Plot for Associations of In Vivo Brain Volume and Full-scale IQ Based on Healthy Samples



Note. Summary effect is based on a random effects model and represented by the diamond; symbol size and coloring of raindrops are varied according to relative study weight within analysis.

Figure 3

Rainforest Plot for Associations of In Vivo Brain Volume and Full-scale IQ Based on Clinical Samples



Note. Summary effect is based on a random effects model and represented by the diamond; symbol size and coloring are varied according to relative study weight within analysis.

Results of Synthesis

The meta-analytical results are reported below. Some non-crucial results and graphs are not presented to avoid cluttering. All exact results are noted by comment in the respective R script. These can be found at https://osf.io/y6msp/. All plots are also available there. Appendix C lists all results obtained by the different approaches for a better overview.

Hedges and Olkin Meta-Analysis

Summary Effects. Synthesis of 122 correlation coefficients based on healthy samples using the REML estimator yielded a highly significant overall effect of $r = .24$ ($p < .0001$, 95% CI [.21, .26]) for full-scale IQ. The effect for verbal IQ was $r = .19$ ($k = 73$, $p < .0001$, 95% CI [.14, .23]) and for performance IQ $r = .22$ ($k = 49$, $p < .0001$, 95% CI [.18, .26]). Results for clinical samples were $r = .22$ ($k = 66$, $p < .0001$, 95% CI [.16, .27]) for full-scale IQ, $r = .21$ ($k = 44$, p < .0001, 95% CI [.15, .27]) for verbal IQ and $r = .19$ ($k = 32$, $p < .0001$, 95% CI [.13, .26]) for performance IQ. In order to allow comparison with results from Pietschnig et al. (2015), summary effects based on mixed (healthy and patient) samples were computed, too. The results were $r = .23$ ($k = 188$, $p < .0001$, 95% CI [.21, .26]) for full-scale IQ, $r = .19$ ($k = 118$, $p < .0001$, 95% CI [.16, .23]) for verbal IQ and $r = .21$ ($k = 81$, $p < .0001$, 95% CI [.18, .24]) for performance IQ. Results did not differ beyond the third decimal in all analyses using the PM estimator. The meta-analytic results from correlations corrected for their slight negative bias (“UCOR”) instead of Fisher´s *r*-to-*z* transformed correlations (“ZCOR”) were marginally higher (deviation $\leq .03$).

Heterogeneity. Based on full-scale IQ data from healthy samples, the *Cochran´s Q* test for heterogeneity was highly significant ($Q(121) = 257.08$, $p < .0001$). Total heterogeneity estimates were $\tau = .086$ and $\tau^2 = .007$. The percentage of total variation across effect sizes due to the variation of true effects was moderate according to standard guidelines (Higgins & Thompson, 2002; $I^2 = 56.35\%$, 95% CI [42.76%, 78.53%]). The true effect size varied substantially across observed effect sizes (95% PI [.07, .39]). A QQ plot (Figure 4) showed no obvious pattern of non-normally distributed residual heterogeneity of true effect sizes. Several studies were identified as contributing exceptionally to heterogeneity with the use of a *Baujat* plot (Figure 4). Most had negatable effect on the overall effect estimate due to small sample size (e.g. an effect size associated with de Boer (2007), $r = -0.55$, $n = 20$). An effect size from a study by Tan et al. (1999) was the only one deviating strongly and having a (small) effect on the overall effect estimate (effect size number 023, $r = .64$, $n = 54$). This effect size was also identified as an outlier by using the “influence” command in the *metafor* package (Figure 4). The function identified two effect sizes associated with Cox et al. (2019) as influential, too, because of their large sample sizes (n = 3900 and n = 4192). A *leave-one-out* analysis indicated negatable impact of individual studies on the overall estimate. All estimates were within the range r = .23 to r = .24. Sampling of all possible subsets of studies yielded an overall estimate range of GOSH [.17, .33]. The corresponding plot (Figure 4) shows a unimodal pattern centered around the summary effect, which indicates that there are no distorting effects of deviant subsets.

The assessment of heterogeneity in the verbal IQ data based on healthy samples produced similar results. Total heterogeneity estimates were τ = .126 and τ^2 = .016. The percentage of total variation across effect sizes due to the variation of true effects was moderate (I^2 = 52.04%, 95% CI [38.25%, 75.92%]). An effect size associated with Harvey et al. (1994, n = 34, r = .69) contributed exceptionally to heterogeneity, however, did have a negatable impact on the overall effect size. The overall effect range was GOSH [.06., .31].

Heterogeneity analyses for performance IQ data based on healthy samples produced lower results for heterogeneity descriptors. The results were τ = .076 and τ^2 = .006 and I^2 = 29.33%, 95% CI [0%, 45.35%]. The PM estimator yielded an even lower result (I^2 = 13.85%). Two effect sizes with an exceptional contribution to heterogeneity were identified. Coffey et al. (2000) offered one of them (n = 318, r = .06), Betjemann (2010) the other (n = 142, r = .42). Both effect sizes had negatable impact on the overall result. The overall effect range was GOSH [.08, .37]. Plots assessing heterogeneity in verbal and performance IQ data based on healthy samples can be found in Appendix D.

Figure 4

Collection of Plots Assessing Heterogeneity in Full-Scale IQ Data Based on Healthy Samples

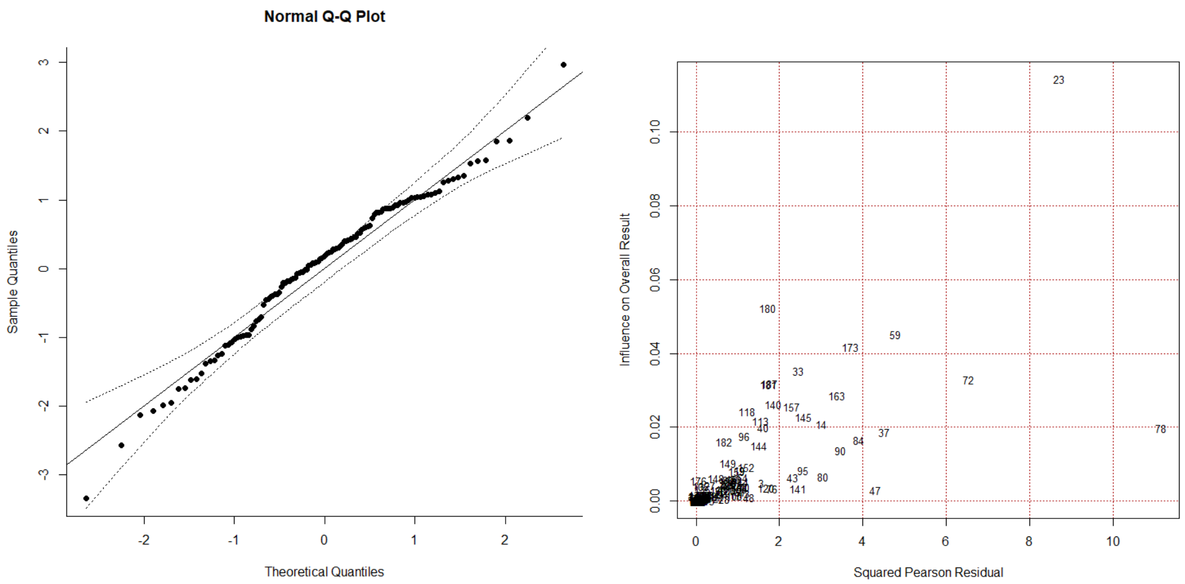
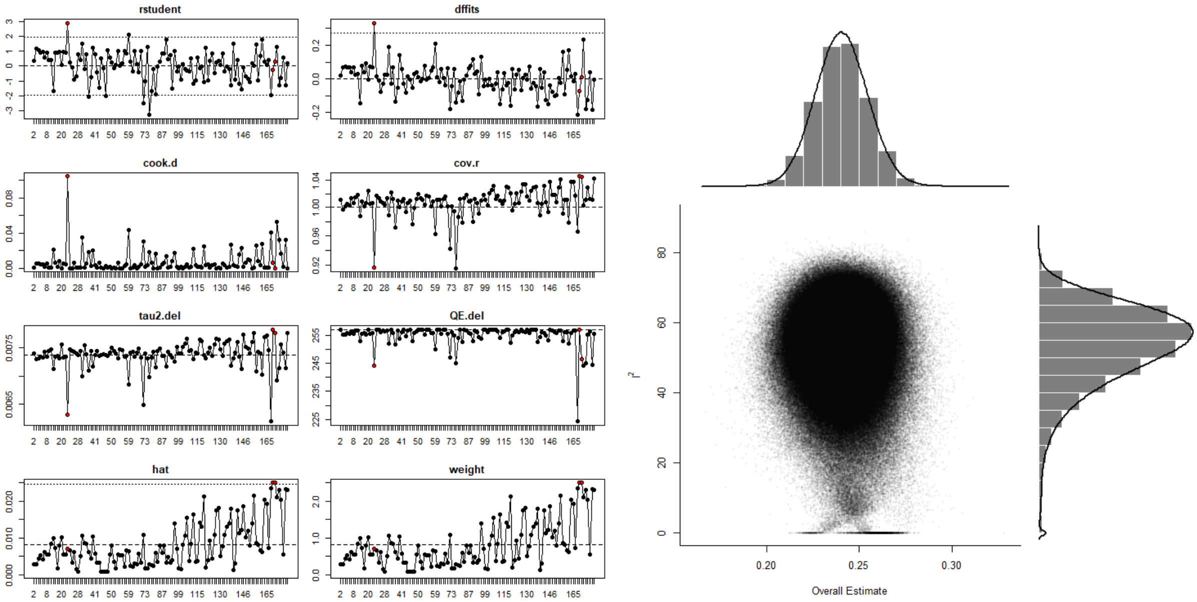


Figure 4 (continued)



Note. Plots from left to right*:* *normal QQ* plot, *Baujat* plot, influence diagnostics, GOSH plot. All plots were created with the *metafor* package.

Psychometric Meta-Analysis

The “bare-bones” meta-analysis (without corrections) of full-scale IQ data based on healthy samples yielded an overall estimate of $r = .22$ ($k = 122$, $p < .0001$, 95% CI [.15, .28]). Heterogeneity estimates were very similar to the results of the Hedges-Olkin meta-analysis, with $\tau^2 = .0084$ and $I^2 = 62.97\%$. After correcting each effect size individually for range departure and computing the corresponding standard errors, another meta-analysis was carried out. The result was $r = .28$ ($k = 64$, $p < .0001$, 95% CI [.23, .33]). Since sample standard deviations for IQ scores were not obtainable for half of all effect sizes, data loss was considerable. Bare-bones meta-analytic results for verbal IQ were $r = .16$ ($k = 73$, $p < .0001$, 95% CI [.09, .23]) and $r = .20$ ($k = 49$, $p < .0001$, 95% CI [.16, .25]) for performance IQ. After applying range departure corrections, results increased to $r = .24$ ($k = 31$, $p < .0001$, 95% CI [.17, .31]) for verbal IQ and $r = .28$ ($k = 28$, $p < .0001$, 95% CI [.22, .33]) for performance IQ.

The bare-bones meta-analysis of full-scale IQ data based on clinical samples brought similar results compared the Hedges-Olkin approach. The results were $r = .21$ ($k = 66$, $p <$ .0001, 95% CI [.13, .28]) for full-scale IQ, $r = .16$ ($k = 73$, $p < .0001$, 95% CI [.09, .23]) for verbal IQ and $r = .20$ ($k = 49$, $p < .0001$, 95% CI [.16, .25]) for performance IQ. However, the meta-analysis of range departure corrected correlations produced lower estimates than the bare-bones meta-analyses. Results were $r = .20$ ($k = 32$, $p < .0001$, 95% CI [.11, .29]) for full-scale IQ, $r = .16$ ($k = 16$, $p = .007$, 95% CI [.05, .27]) for verbal IQ and $r = .20$ ($k = 15$, $p = .006$, 95% CI [.07, .33]) for performance IQ. In contrast to studies based on healthy samples, the average clinical sample standard deviation of IQ scores was enhanced in range leading to corrected coefficients lower than uncorrected ones.

Robust Variance Estimation Meta-Regression

Data for the RVE models were similar for full-scale IQ effect sizes compared to the other approaches not modeling dependency. The following two effect sizes were added. (1) A study from Ritchie et al. (2017) used data from a later wave (approximately 20 years) from the same sample as in Staff et al. (2006). (2) Vreeker et al. (2017) used data from the Dutch Bipolar Cohort Study, which other researchers have done, too (e.g. Bohlken et al., 2016; de Zwarte et al., 2019). The amount of data overlap was unclear, so the study of Vreeker et al. (2017) was included in the RVE data sheet only. Differences in terms of added effect sizes were larger in the verbal RVE sheet (19 added) and performance RVE sheet (23 added). The main reason for dependency among effect sizes were added tests from the same sample in a different intelligence dimension (e.g. an effect size based on the Wechsler WMI was added next to an effect size based on the VCI).

Fitting an RVE meta-analytic model based on 92 studies comprising 124 full-scale IQ effect sizes produced an overall estimate of $r = .24$ ($p < .01$, 95% CI [.21, .26]). For verbal IQ the result was $r = .19$ ($p < .01$, 95% CI [.14, .23]) based on 63 studies comprising 92 effect sizes. The overall estimate for performance IQ was $r = .22$ ($p < .01$, 95% CI [.18, .27]) based on 46 studies comprising 72 effect sizes.

Results for clinical samples were $r = .22$ ($k = 56$ (66), $p < .01$, 95% CI [.17, .28]) for full-scale IQ, $r = .21$ ($k = 36$ (48), $p < .01$, 95% CI [.15, .28]) for verbal IQ, and $r = .21$ ($k = 28$ (35), $p < .01$, 95% CI [.14, .27]) for performance IQ.

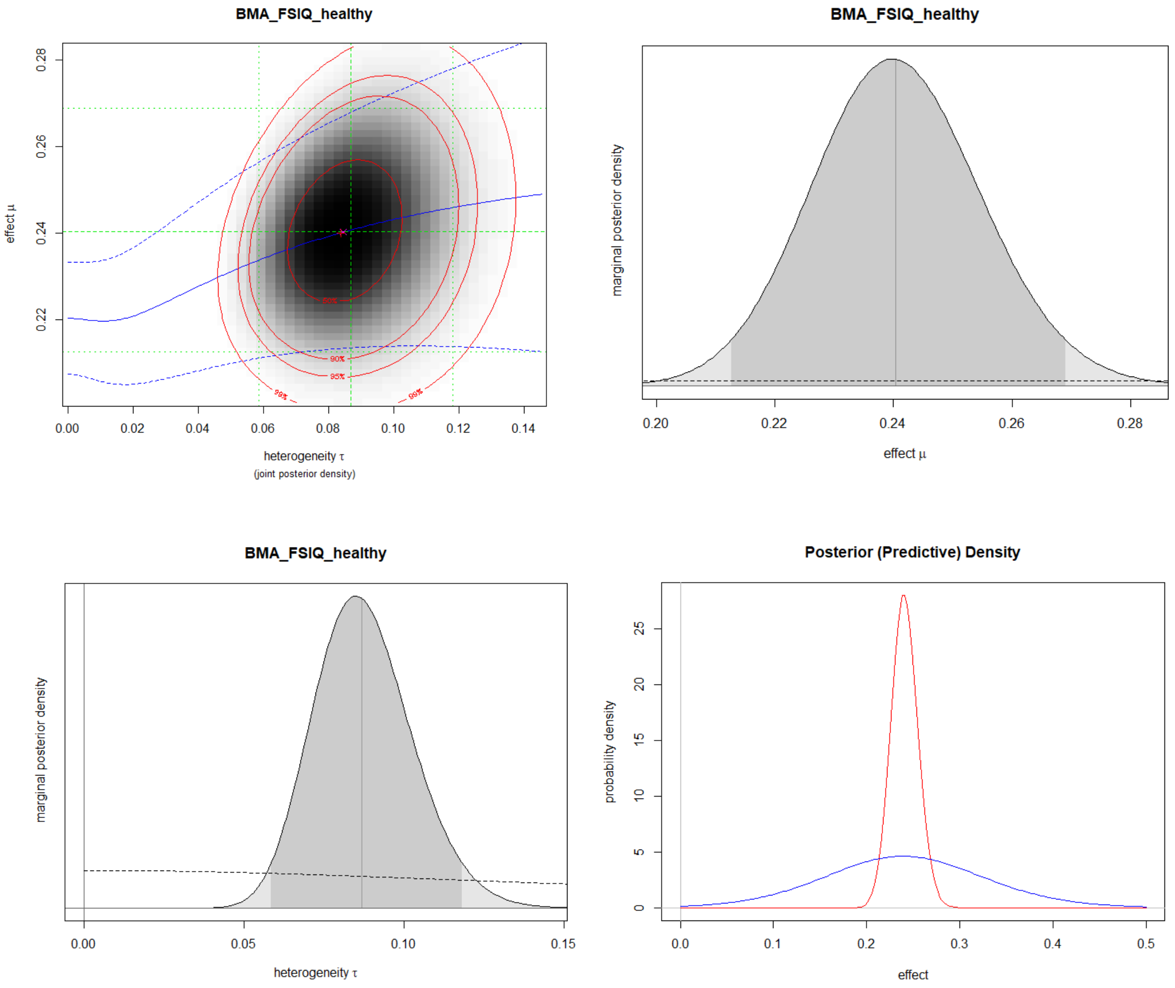
Changing the value for ρ or removing the small sample bias correction did not affect the outcome beyond the third decimal in all analyses.

Bayesian Meta-Analysis

The results of the Bayesian meta-analysis were also similar to those of the other meta-analyses. The result of the full-scale IQ data ($k = 122$) based on healthy samples was $r = .24$. The shortest credible interval was 95% CI [.21, .26]. Changing informative prior specifications (μ and τ values) did not affect results. Figure 5 shows four plots in which it is easy to see that the probability distributions for both the mean effect and the heterogeneity are normally distributed within a limited range of values. The fourth plot shows a comparison between the posterior and predictive (i.e. “future” results) probability density. The predictive probability density was consistent with the prediction interval displayed in section “Hedges and Olkin Meta-Analysis”. Results for clinical samples, verbal IQ, and performance IQ were nearly identical to the Hedges and Olkin approach (results annotated in R script “BayesianMetaAnalysis” available at https://osf.io/e24zq/). Overall, the inclusion of preliminary information had no influence on the results, probably due to the relatively large amount of data.

Figure 5

Collection of Descriptive Plots for the Distributions of the Effects and Heterogeneity Based on Full-Scale IQ Data Comprising Healthy Samples



Note. Plots from left to right: plot displaying the density distribution in reference to summary effects and heterogeneity; plot of summary effect posterior density, plot of heterogeneity posterior density, plot showing posterior (red line) and predictive (blue line) probability densities.

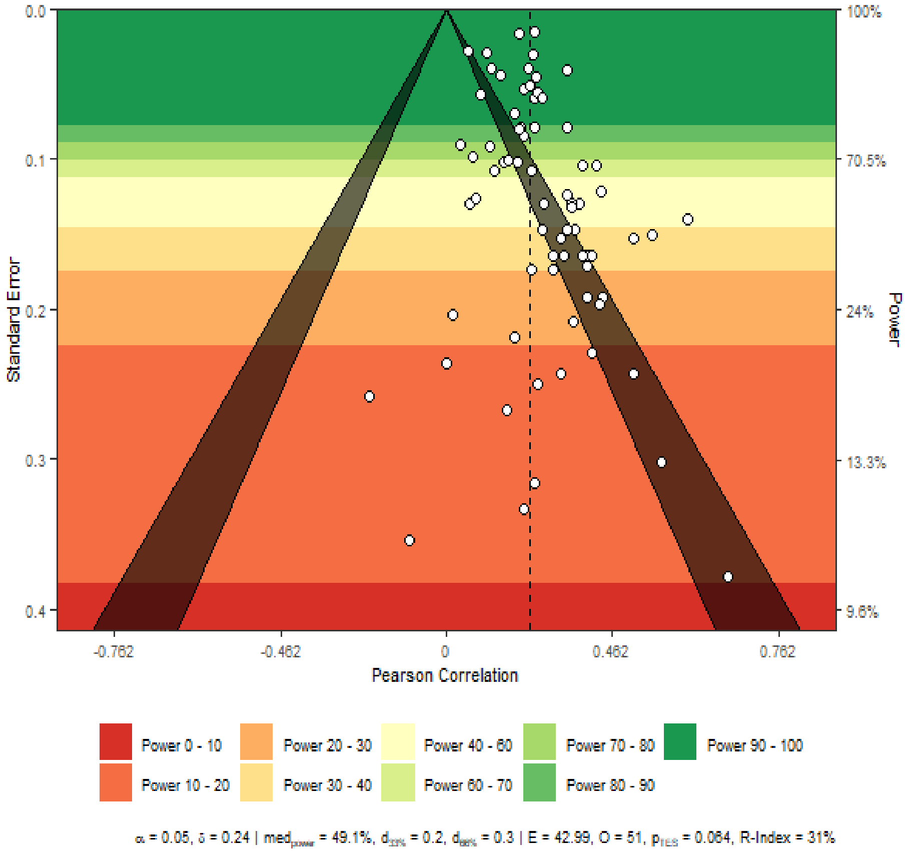
Dissemination Bias

First, a power-enhanced funnel plot (Kossmeier et al., 2020b) was used to obtain an overview of how well published full-scale IQ studies based on healthy samples were powered. Figure 6 shows that most studies had lower power than desirable. The median power was 49.1%. The test for excess significance suggested that studies were more successful in finding significant results compared the expected number of significant results based on their power (8 more significant results than expected, *p* = .064). The Replicability-Index was 31%, indicating low chance of replication for the average individual effect size.

Next a contour-enhanced funnel plot (Figure 7) was created to assess funnel plot asymmetry and the impact of potentially missing studies due to publication bias. The Egger´s regression line is askew to right side of the funnel plot, indicating asymmetry. Tests of robustness of this asymmetry were statistically significant (*p* = .008 for the weighted regression with a multiplicative dispersion term and *p* = .006 for the mixed-effects meta-regression model). The trim-and-fill analysis suggested 16 potentially missing effect sizes. Recalculating the overall estimate including these supposedly missing effect sizes revealed a negatable impact (*r* = .22, 95% CI [.19, .25]).

Figure 6

Sunset Plot of Published Full-Scale IQ Studies Based on Healthy Samples

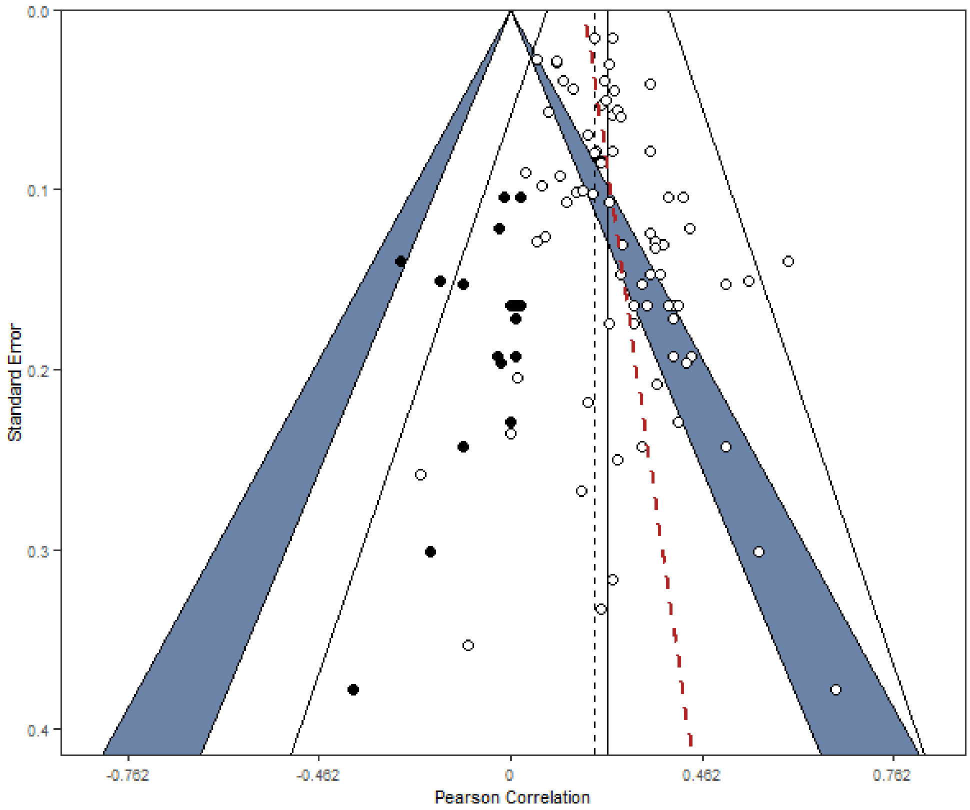


The *p-curve* analysis did not indicate the presence of *p-hacking*. Figure 8 shows a right-skewed distribution of *p*-values, meaning substantially more highly than barely significant results. Binominal and continuous tests confirmed this impression (binominal test: $p < .0001$; full *p*-curve: $z = 16.16$, $p < .0001$; half *p*-curve: $z = 15.07$, $p < .0001$). Results from *p-uniform* did not indicate the presence of *p-hacking*, too. The adjusted estimate was $r = .25$ The test for publication bias was statistically insignificant ($p = .958$). The adjusted estimate from the *p-uniform** analysis was $r = .22$. The H0 of no publication bias was not rejected ($p = .154$). Applying a selection model based on *p*-values (Vevea & Hedges, 1995) yielded an adjusted estimate of $r = .21$. The weight function might not have been well informed, since few studies results are not highly significant.

A selection model based on the standard error of effect sizes (Copas & Shi, 2001) suggested the presence of selection bias inflating the summary effect. The model´s adjusted estimate was $r = .21$ assuming 32 missing studies. The hypothesis that no selection remained unexplained did not reach significance ($p = .135$).

Figure 7

Contour-Enhanced Funnel Plot of Full-Scale IQ Studies Based on Healthy Samples



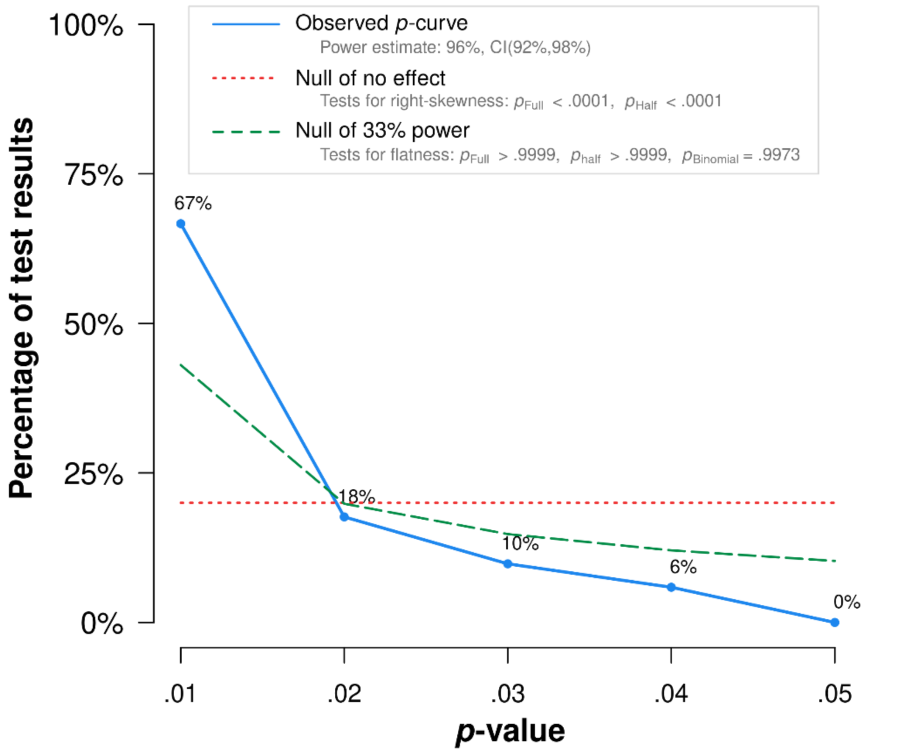
Note. The red dashed line represents the Egger´s regression, the dashed black line the estimate from the trim-and-fill analysis, the black continuous line the meta-analytic summary effect. The black dots on the left side represent potentially missing studies due to publication bias as computed by the trim-and-fill analysis.

Standard errors are closely related to sample size. A cumulative meta-analysis ordered by sample size showed that studies with comparatively medium number of participants reported higher effect sizes than studies with small or large samples (plot available at https://osf.io/47nwj/).

Lastly, a potential moderating effect regarding the publication year of each study was examined. Fitting a meta-regression with year as the predictor produced a significant result (p = .008, R^2 = 13.34%). The slope was -.005 indicating a slight decrease of effect sizes per year (Figure 9). Results were more pronounced when only considering published results. A cumulative meta-analysis ordered from early to recent publication confirmed this finding (Figure 10). The summary effect from around 2009 to 2015 was significantly reduced by the studies during this period.

Figure 8

p-Curve Analysis of Published Full-Scale IQ Studies Based on Healthy Samples



Note: The observed p-curve includes 51 statistically significant (p < .05) results, of which 44 are p < .025. There were 23 additional results entered but excluded from p-curve because they were p > .05.

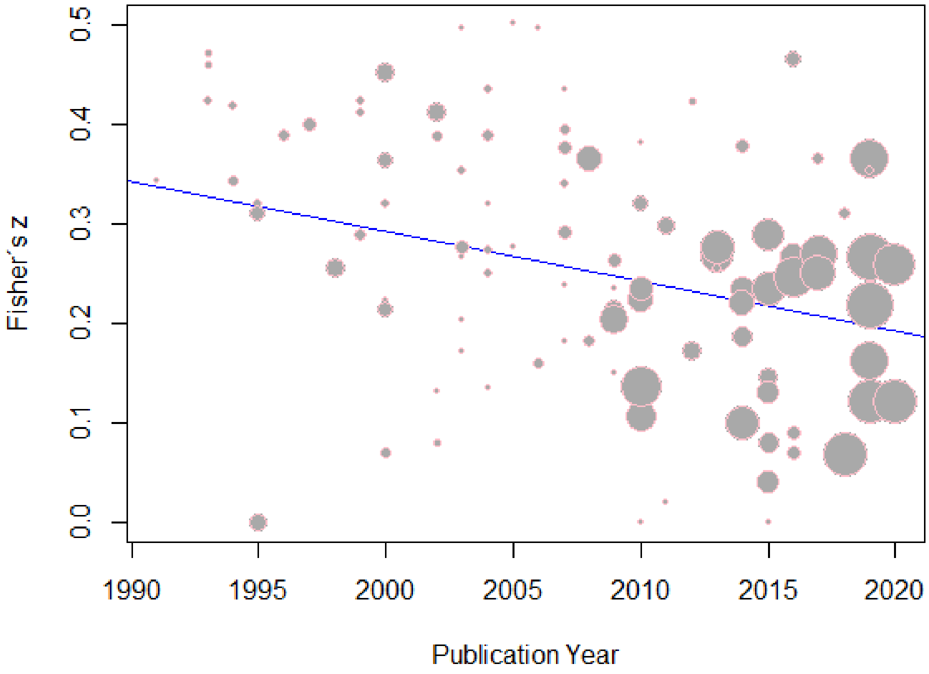
Note. The blue line shows the distribution of observed p-values. The red dashed line represents the expected distribution under the null hypothesis of no effect. The green dashed line shows a scenario of a true effect and underpowered studies (33% power).

In the verbal IQ data, most analyses did not indicate the presence of dissemination bias. Only a *trim-and-fill* analysis suggested six missing studies. The adjusted estimate of r = .17 (95% CI [.11, .23]). was only slightly below the meta-analytic summary effect though. Publication year as a predictor did not reach statistical significance (slope = -.005, p = .169, R^2 = 4.12%).

Results for performance IQ were comparable. The *trim-and-fill* analysis and the selection model based on standard errors of effect sizes suggested nine missing studies. The adjusted estimates were r = .19 (95% CI [.14, .24]) and r = .20 respectively. An influence of publication year was not observed (slope = -.001, p = .761, R^2 = 0 %).

Figure 9

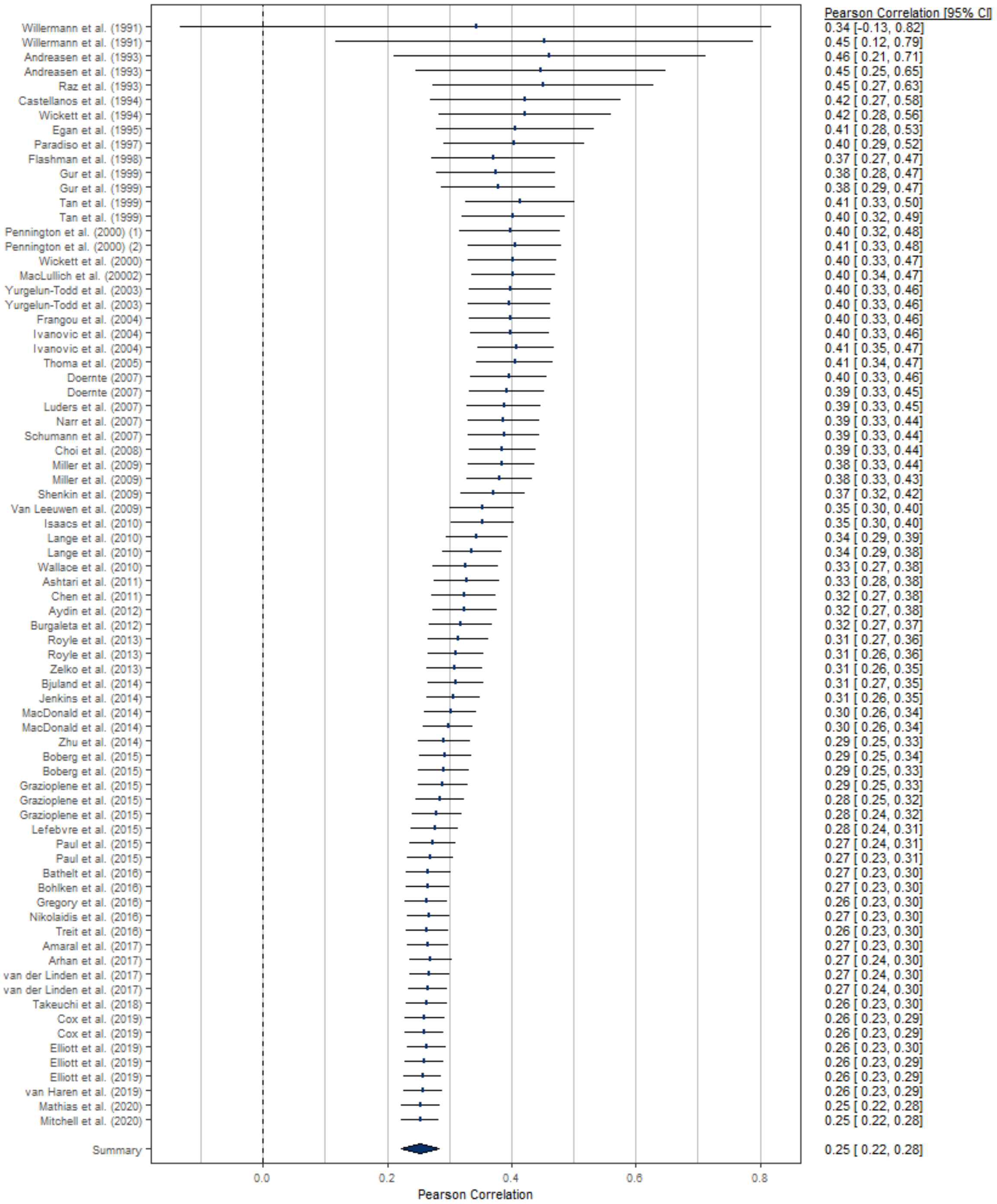
Bubble Plot of Meta-Regression with Predictor Publication Year Based on Full-Scale IQ Data from Healthy Samples



Note. Bubbles represent individual study outcomes and are varied in size according to relative weight in the meta-regression. The blue line shows the slope.

Figure 10

Cumulative Forest Plot of Studies Ordered by Year of Publication Based on Full-Scale IQ Data from Healthy Samples



Note. The plot shows a sequence of random-effects meta-analyses starting with the first published study by Willerman et al. (1991) and adding the other studies one at a time. Each correlation (on the right side) corresponds to the summary effect of the study pool up to a given study. Correlations obtained via personal communication are not included.

Moderators

Subgroup Analysis

Table 2 summarizes results from subgroup analyses of categorial moderators. The only statistically significant subgroup differences in the full-scale IQ data concerned the rating groups of the correlation with *g*. The summary effect for studies based on IQ tests rated "fair" was $r = .23$ (CI [.09, .37]), for ratings "elevated" $r = .20$, (CI [.17, .23]), and for ratings "high" $r = .31$, (CI [.27, .34]). A further noticeable difference was observed between reported correlations ($r = .26$, CI [.23, .29]) and those obtained via personal communication ($r = .20$, CI [.13, .26]). It should not go unnoticed that the number of total participants per reporting group was unevenly distributed (reported: $n = 21455$; PC: $n = 1838$). Results from subgroup analyses regarding ethnicity were not interpretable due to lack of data. Most study authors reported no information on ethnicity. Those who did tested predominantly white samples, or samples with whites as the majority. Other ethnic categories comprised few samples (less than five per category).

In the verbal IQ data, comparing summary effects of studies using either TBV or ICV as their brain volume operationalization showed a trend to statistically significant differences. Correspondingly, comparison of samples of children/adolescents and adults reached significance when only studies using TBV were considered ($k = 46$, $Q = 4.17$, $p = .041$). The comparison of studies using TBV or ICV may have been affect by an uneven distribution of the number of studies and number of total participants. The summary effect of reported correlations was higher than the summary effect of correlations obtained via personal communication in the verbal IQ data as well.

The difference between reporting groups was also observed in the performance IQ data. No other noticeable differences emerged. Descriptive information and summary effects for all subgroups are provided in Appendix E.

Table 2

Results of Subgroup Comparisons Based on Healthy Samples

	Full-scale IQ		
	k	*Q*	*p*
Healthy vs. clinical samples	188	0.29	.591
Reported r vs. in grey literature vs. via personal communication*	122	2.90	.240
Children vs. adults*	122	0.05	.829
TBV vs. ICV	91	0.37	.546
Females vs. males	60	0.22	.643
Fair vs. elevated vs. high correlation with g	166	20.5	< .001

	Verbal IQ		
	k	*Q*	*p*
Healthy vs. clinical samples	118	0.59	.442
Reported r vs. in grey literature vs. via personal communication*	73	3.29	.193
Children vs. adults*	73	1.10	.295
TBV vs. ICV	60	3.07	.080
Females vs. males	37	0.01	.920
Fair vs. elevated vs. high correlation with g	-	-	-

	Performance IQ		
	k	*Q*	*p*
Healthy vs. clinical samples	81	0.46	.497
Reported r vs. in grey literature vs. via personal communication*	49	3.41	.065
Children vs. adults*	49	0.27	.607
TBV vs. ICV	38	1.90	.168
Females vs. males	25	0.05	.802
Fair vs. elevated vs. high correlation with g	-	-	-

Note. Subgroup comparison of healthy vs. clinical samples utilized the whole data, all other results are based on healthy samples. Only few studies were categorized as grey literature. Repeating analyses with only reported *r* and personal communication did not change results in the verbal IQ data meaningfully but produced a higher *Q* value and a trend to significance (*p* = .107) in the full-scale IQ data. There were no studies categorized as grey literature in the performance IQ data. Recalculating the children vs. adults comparison based on studies using TBV operationalization did not change results in full-scale or performance IQ, but produced a significant result in the verbal IQ data (*p* = .041).

Meta-Regression

Univariate. Fitting a meta-regression based on healthy samples with male ratio as a predictor showed no statistically significant effects in all three intelligence domains (FSIQ: slope < -.001, p = .998; VIQ: slope .004, p = .953; PIQ: slope -.047, p = .463). Considering the mean age of a sample as a predictor produced two statistically insignificant results for full-scale and performance IQ (FISQ: slope -.001, p = .388; PIQ: slope -.0018, p = .177). Mean age had an effect on the association between brain volume and verbal IQ (slope -.003, p = .020). Ratings of the correlation between applied intelligence measurement and *g* had a notable influence on effect sizes (F(2,104) = 7.57, p = .001). Effect sizes differed particularly between rating group 4 (excellent correlation with g; r = .31) and ratings groups 2 and 3 (fair: r = .22; elevated r = .20). Heterogeneity patterns between the rating group differed, so a mixed-effects subgroup analysis was conducted as a sensitivity analysis. The results are displayed in section “Subgroup Analysis”.

Utilizing an RVE meta-regression approach showed that neither differences of summary effects for full-scale and performance IQ (p = .739), nor full-scale and verbal IQ (p = .128) were statistically significant. Although the comparison was based on numerous studies (104 studies comprising 286 correlations), one may interpret the latter result as a trend to statistical significance, considering the relatively weak power of RVE meta-regressions (Tanner-Smith et al., 2016).

Multiple. First, a potential interaction effect of age and sex was examined. Fitting a meta-regression with a mean age * male ratio term showed no evidence for such an effect based on full-scale IQ data (k = 112, F(1, 110) = 0.53, p = .467, R^2 = 0%). A subgroup comparison of girls, boys, men, and women confirmed these results (Q(3) = 3.49, p = .322). Results for verbal and performance IQ were comparable, although some variance was “explained” (R^2 = 9.11%, p = .162, and R^2 = 12.83%, p = .140 respectively).

A hierarchical multiple meta-regression for full-scale IQ data based on healthy samples was performed. In a first step, the year of publication, the correlation with *g* and the type of report were included as predictors in the model. The rating group "4" (excellent correlation with *g*) and the year of publication were significant predictors (R^2 = 53.84%). In a second step, the average age and sex ratio were added to the model. Rating group 4 and the year of publication remained the only significant predictors (R^2 = 55.27%). In a third step, the number of corrections of effect strengths and the objective of the study (by-product or not) were added (R^2 = 46.09%). Rating group 4, study goal "other" and number of corrections “3” were significant predictors. If we take economy and explained variance into account, model 1 had the best model fit. VIFs were checked at each step. In all three models, the two levels of type of report appeared with high VIFs (> 17). VIFs of all other predictors were unremarkable. Since the type of report proved to be an important predictor in the univariate analysis, I decided not to remove it permanently. Instead, each sub-step of the analysis was performed without this variable and results were compared. Small differences were visible without changing the inference or conclusion for individual variables as for the model fit.

The procedure was repeated for verbal and performance IQ, except that the correlations with *g* was not included as a predictor. For verbal IQ model 2 had the best model fit ($R^2 = 38.04\%$). Publication year and sample mean age were significant predictors. The only model explaining any variance ($R^2 = 4.51\%$) for performance IQ was model 2. Mean age showed a trend to significance.

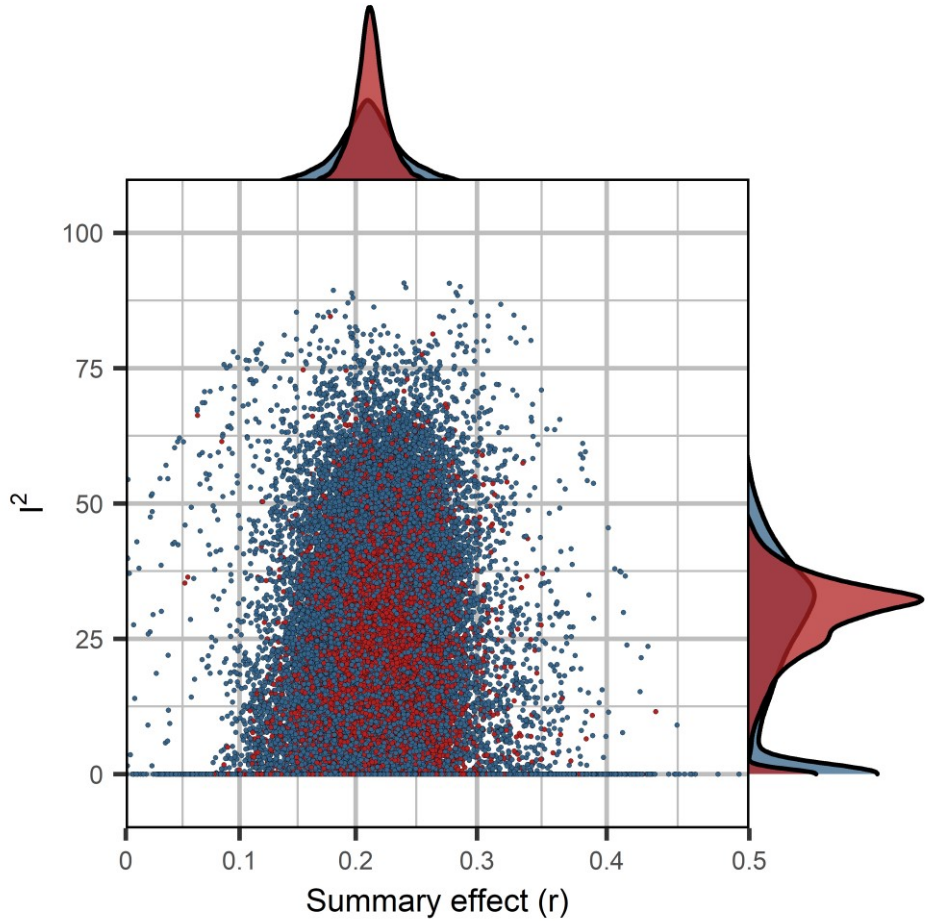
Permutation tests (1000 iterations) confirmed the robustness of the results in all analyses.

Specification Analyses

Figure 11 shows the results of the combinatorial meta-analysis for performance IQ studies based on healthy samples. There is little difference in numerical output discernable compared to the GOSH plot generated with *metafor* (see Appendix D). The oversampling of subsets with particularly many or fewer studies did not influence subset patterns. A feature of the GOSH plot according to Voracek et al. (2019) is the ability to mark subsets that contain a particular study result (this is applicable to the *metafor* GOSH output as well). These specific subsets can then be compared with all the others. This is useful, for example, if a first study on an issue has produced particularly extreme results that could not be confirmed in the following (i.e. "winner's curse"). Although this was not the case in the first MRI study on the association between brain volume and performance IQ, subsets with the study by Andreasen et al. (1993) were highlighted for illustration purposes. The comparison between subsets with and without this study shows little influence on the summary effect or proportion of variation due to the variation of true effects (I^2). The overall GOSH plot shows a unimodal pattern centered around the results of the Hedges-Olkin meta-analysis ($r = .22$, $I^2 = 29.33\%$). There are no branching patterns to indicate different results when only certain subsets are considered. The subsets that are close to the x-axis are based on very few study results and do not give cause for concern. Overall, the meta-analytical results were stable under all possible combinations of studies. The same applied to the full-scale and verbal IQ data.

Figure 11

GOSH Plot of the Combinatorial Meta-Analysis for Full-Scale IQ Data Based on Healthy Samples

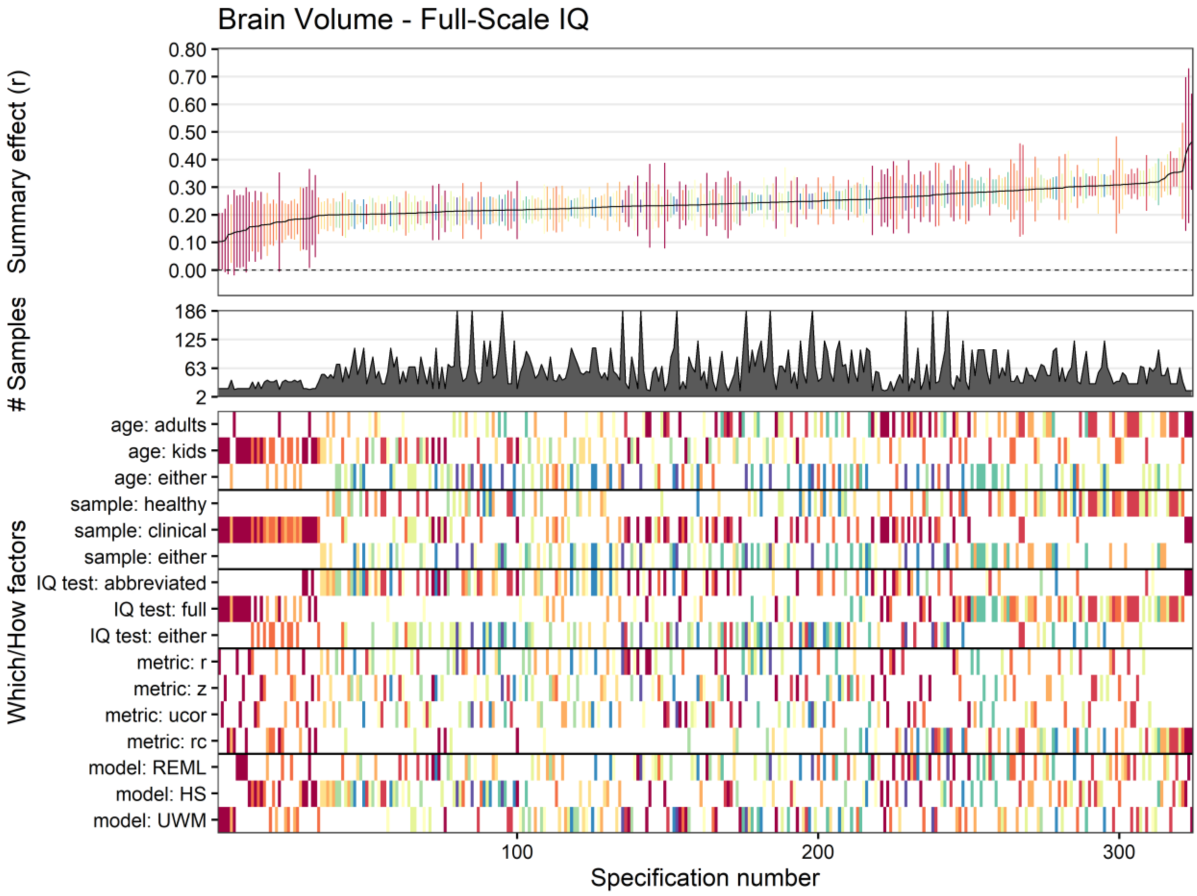


Note. The plot shows random-effects meta-analytic summary effects on the x-axis and the relative between-study variance statistic I^2 on the y-axis for 100000 random study subsets. Subsets containing the first MRI investigation of the association between in vivo brain volume and intelligence (Andreasen et al., 1993) is highlighted red. Distributional densities are shown on the top (summary effects) and on the right side (I^2 values) of the plot.

Next, the influence of all reasonable meta-analytical data and analysis specifications on the summary effect was examined. Figure 12 shows a combination of several descriptive meta-analytic specification plots. These should be read vertically. The upper panel of the plot shows the effect sizes resulting from the corresponding specifications together with the respective 95% confidence intervals. The panel in the middle shows the number of included study results. The lower panel shows the same information considering the individual specifications. The color spectrum ranges from red to green to violet. The former stands for relatively few included study results, the latter for many.

Figure 12

Descriptive Meta-Analytical Specification Plot for Full-Scale IQ Data



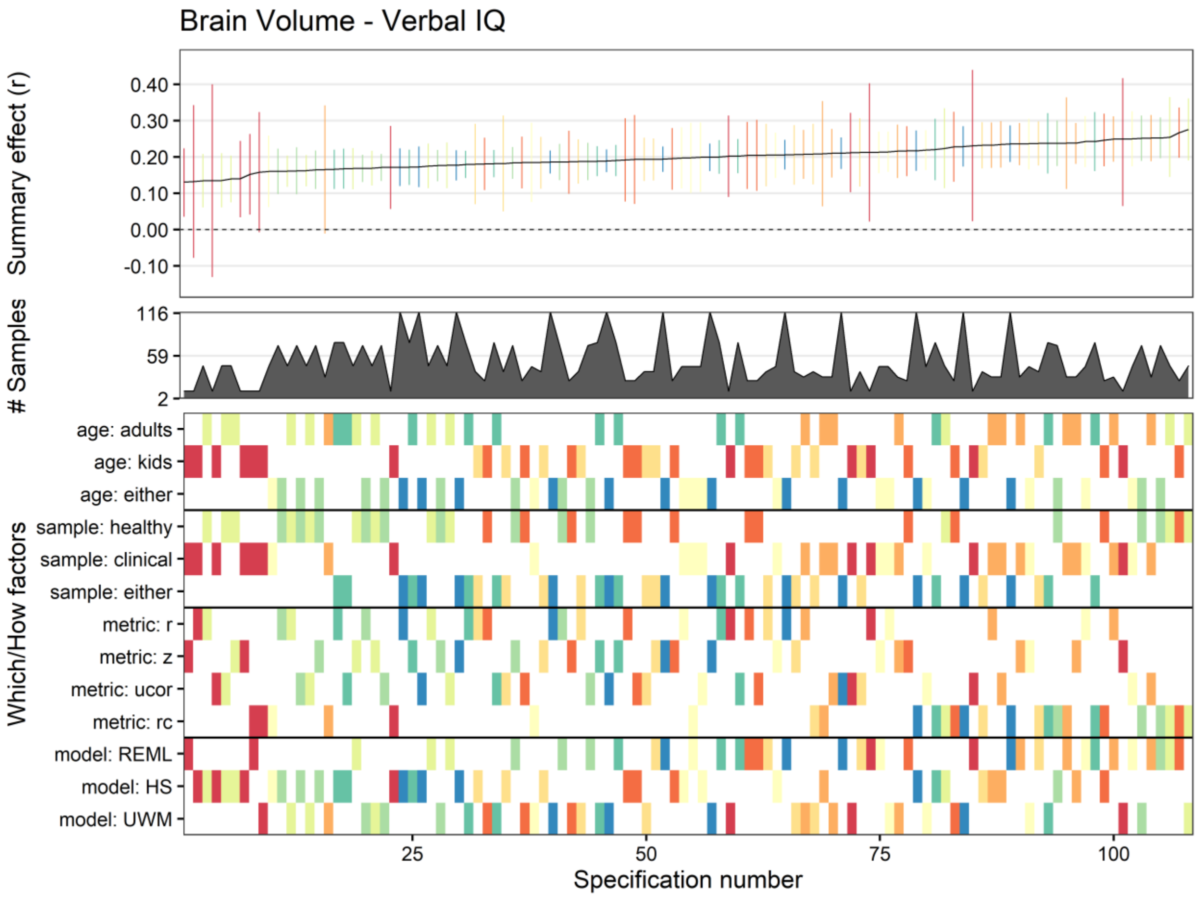
Note. Vertical columns in the lower half of the plot represent which and how factors that constitute a given specification. Coloring displays the number of studies a specification is based on (red colors indicate less, blue colors a larger number of studies). The panel in the middle likewise shows on how many studies a specification is based on. The top panel displays the corresponding effect sizes along with their 95% confidence intervals.

How can we interpret the plot? The range of effect sizes in the upper panel is the most important information. Depending on the combination of specifications a summary effect of $r = .10$ to $r = .37$ was observed. However, the outermost results are based on very few study results leading to wide confidence intervals. If only results based on a reasonable number of studies are considered, the range shrinks to about $r = .20$ to $r = .35$. This is approximately the range of results achieved by Pietschnig et al. (2015) and McDaniel (2005). The result of Gignac and Bates (2017; $r = .39$) can no longer be achieved on the basis of the updated data. In order to find out which specifications have led to higher effect sizes the lower panel is useful. We see that clinical samples as well as abbreviated IQ tests lead to lower effect sizes. In these specifications only white space is visible vertically below the higher effect sizes in the upper plot (no specifications). In contrast, meta-analyses that take only extensive IQ tests ("full IQ test") and range departure corrected correlations ("rc") into account often lead to higher effect sizes. For all other specifications no special patterns are visible. This visual interpretation fits well with the results from previous sections, which showed the influence of the same variables. Note that there is no benefit in taking an average of the effect size range or the center of their distribution as all specifications are equally reasonable. Doing so would invite fruitless discussions about the ideal route to a result which we sought to avoid.

There was a total of 108 possible combinations of specifications for the association between brain volume and verbal IQ (Figure 13). No rating categories are available for correlation with *g* in the verbal IQ data.

Figure 13

Descriptive Meta-Analytical Specification Plot for Verbal IQ Data



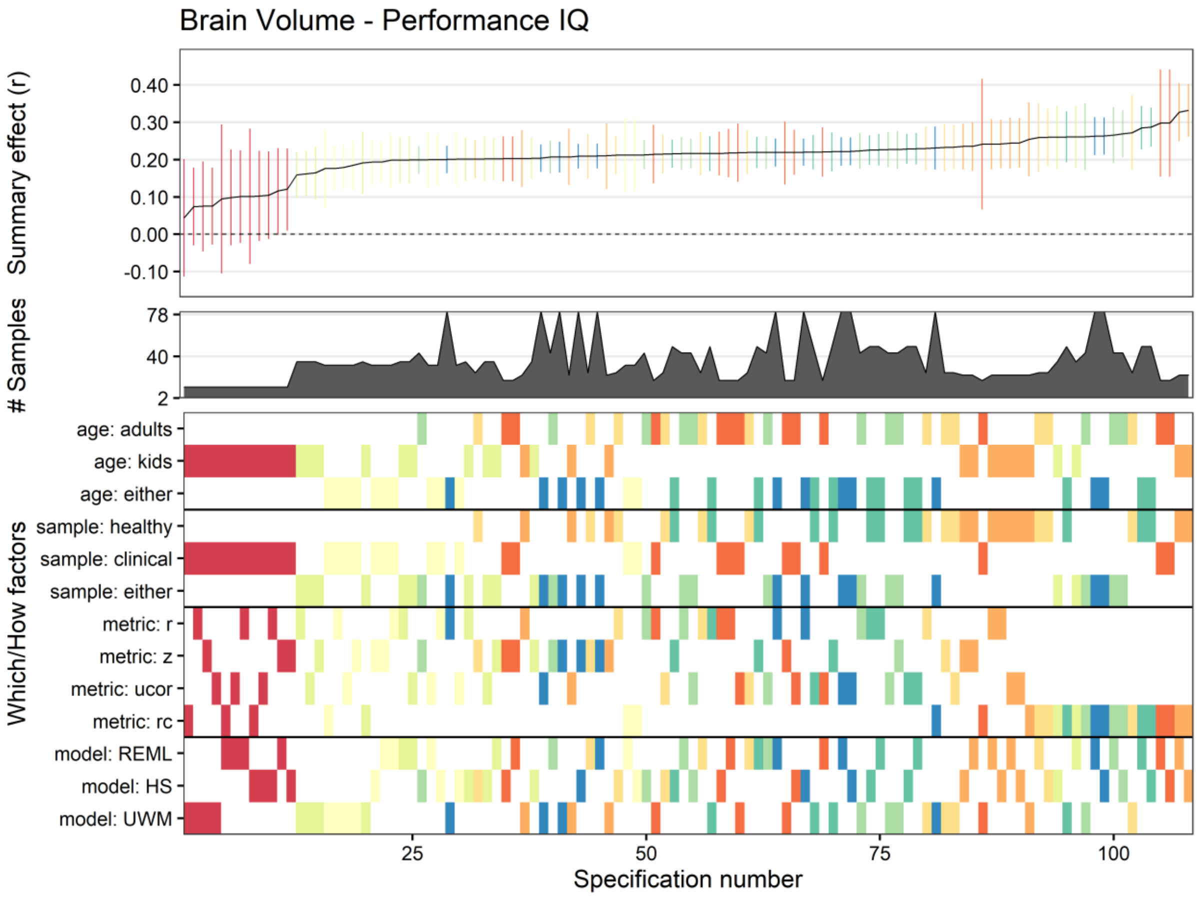
Note. Vertical columns in the lower half of the plot represent which and how factors that constitute a given specification. Coloring displays the number of studies a specification is based on (red colors indicate less, blue colors a larger number of studies). The panel in the middle likewise shows on how many studies a specification is based on. The top panel displays the corresponding effect sizes along with their 95% confidence intervals.

These specifications were thus dropped. If we ignore the inaccurate estimates on the two outsides, the range of the resulting effect sizes was approximately between *r* = .16 and *r* = .26. It is interesting to note that the effect size of the Hedges and Olkin Meta-Analysis (*r* = .18) is in the lower part of the range. If we had only reported this result, this would have been limited information. At the variable level the range departure correction of correlation coefficients was associated with higher effect sizes. In comparison with the results for the full-scale IQ data, the trend for lower effect sizes associated with clinical samples turned towards higher effects sizes. The age categories as well as other metrics did not suggest any specific patterns for verbal IQ either.

The range for performance IQ was approximately *r* = .17 to .29 (Figure 14). 108 combinations of specifications were possible. Besides the slightly increased range, the specification patterns were the same as for verbal IQ.

Figure 14

Descriptive Meta-Analytical Specification Plot for Performance IQ Data



Note. Vertical columns in the lower half of the plot represent which and how factors that constitute a given specification. Coloring displays the number of studies a specification is based on (red colors indicate less, blue colors a larger number of studies). The panel in the middle likewise shows on how many studies a specification is based on. The top panel displays the corresponding effect sizes along with their 95% confidence intervals.

Discussion

Discussion of Meta-Analytic Results

Overall, the results of this thesis suggest that the association between in vivo brain volume and intelligence is stable regarding intelligence domains, populations, and the type of data construction or analysis. There is some variation in effect sizes due to meta-analytic specifications. The range of results for full-scale IQ data from reasonable specifications is r = .20 to r = .35. Variation in summary effect sizes come primarily from the application of range departure corrections and the consideration of the type of test (shortened intelligence measurement or complete battery). The meta-analytical method which is used to determine the overall effect is of little influence. These outcomes show that the meta-analytical estimate of McDaniel (2005; r = .33) are at the upper end of the possible effect sizes able to be observed. The estimate from Pietschnig et al. (2015; r = .24) is approximately in the middle of the distribution. The result of Gignac and Bates (2017; r = .39) is outside the range and can probably be regarded as an overestimation of the correlation. Comparing the range of effect sizes with typical results from differential psychology research reveals a medium to strong correlation (Gignac & Szodorai, 2016). From a traditional perspective, the correlation is of small to medium strength (Cohen, 1988). Whichever view one prefers, brain volume is one of the strongest predictors of intelligence in the context of brain-behavior research (Richie et al., 2015).

But how trustworthy are these results in terms of dissemination bias? In consensus with Pietschnig et al. (2015), extensive analyses found signs of dissemination bias in the full-scale IQ data based on published results from healthy samples. This bias is likely not caused by *p-hacking*. A focus on p-values was infeasible for primary researchers in the light of low to medium average power and the bivariate correlational study design. The detected bias seems more related to sample size, a known problem in the field of neuroscientific research (Button et al., 2013). Although the median power (49.1%) of published studies is an encouraging improvement compared to power estimates in neuroscientific research, it is nevertheless insufficient from a statistical point of view, and lead to several statistically insignificant results. The problem is that some of those results were not reported. The funnel plot asymmetry as well as the “missing” studies indicators from the *trim-and-fill*, and Copas and Shi (2001) analyses support this interpretation. The comparison of the published and unpublished correlations show that the latter are lower on average. Not reporting them may have thus been one cause of the effect size inflation. This reporting behavior is understandable from the perspective of primary study authors. Especially when the primary goal of a study is not to give an estimate of the association of brain volume and IQ, not reporting null or negative effects can happen easily (as well as reporting larger effects although not originally intended).

Several protective factors prevented a more substantial effect size inflation. Two of them are comprehensive literature screening and a relatively high number of included studies (Mathur & VanderWeele, 2019). McDaniel (2005) as well as Pietschnig et al. (2015) have carried out a careful literature screening in combination with (very successful) efforts to obtain unpublished results. Nevertheless, dissemination bias analyses still gave rise to concern. The inclusion of several large-scale studies with the data update in this thesis represents a further protective factor. Efforts of national and international consortia generated sample sizes that would have been impossible to accumulate by individual research teams. Although the majority of the applied methods, especially those that deal well with heterogeneity, still show an upward effect size inflation of about $r = .02$ to $r = .04$ based on published full-scale IQ data, this inflation is reduced when considering unpublished results as well. The remaining extent of bias is less than the variation due to estimate imprecision or the influence of individual studies. In general, the meta-analytic investigation of brain volume and intelligence is an encouraging example of how international cooperation combined with *Open Science* practices, careful literature searches, and the improvement of meta-analytic methods can together effectively reduce threats due to various sources of bias.

Attempts to conceptually replicate moderator effects that have been observed unanimously or only once in the past show mixed results.

First, differences in effect sizes between full-scale, verbal and performance IQ are not statistically significant. Nevertheless, a tendency for smaller effect sizes for brain volume and verbal IQ may be observed. Pietschnig et al. (2015) have argued that the smaller correlations could be due to the lower saturation in *g*. However, it is difficult to evaluate this assumption based on these results. The performance IQ tests contained in the data would have to show higher average *g* loadings than verbal IQ tests in order to explain the difference in the effect sizes between brain volume and verbal or performance IQ. This is conceivable, but not certain. An evaluation is difficult because many different tests have been used. To distinguish between perceptual organization tests and processing speed tests in the performance IQ data in the style of the Wechsler scales would be an indirect approach. The former usually have a higher saturation in *g* (e.g. van der Linden, 2017). Correlations between brain volume and perceptual organization tests must thus be stronger than those between brain volume and processing speed. To test this assumption, I conducted a supplementary RVE meta-regression analysis. Correlations based on tests categorized as measuring perceptual organization abilities are higher indeed than those based on processing speed tests (results are displayed in Appendix F). Correlations based on verbal comprehension or working memory tests do not differ from each other systematically. These results indicate that the saturation in *g* might play a role in domain differences. However, this approach is rather indirect. It is of relatively weak power and cannot rule out alternative explanations like systematic variations in subtest reliability. The results presented here do not allow a final assessment. Generally, it can be said that the differences in effect sizes between brain volume and full-scale, and verbal IQ are rather small, and that there is no noticeable difference in full-scale and performance IQ correlations.

Second, a difference in effect size between brain volume and intelligence based on healthy or clinical samples is no longer discernible when utilizing uncorrected correlations. This can be explained by the wide range of effect sizes within the clinical population. The correlation disappears completely in some diagnostic groups within the autism spectrum (Amaral et al., 2017) or even becomes negative in the case of megalocephaly (Petersson et al., 1999), whereas it is strong for patients with cerebral-palsy (Ballester-Plane et al., 2016). Some rather strong correlations based on clinical samples were added in the course of the data update in this thesis (e.g. Bjuland et al., 2014; van Haren et al., 2019; Annink et al., 2018), leading to a little higher summary effect compared to Pietschnig et al. (2015). However, using range departure corrected correlations leads to more pronounced differences in summary effects. The average standard deviation of IQ scores from healthy samples was restricted in range but was enhanced in clinical samples. Hypothesis 3 is therefore not to be rejected. The association between brain volume and IQ is usually weaker under clinical conditions. The value of this information is limited since the variation between conditions is high. For a precise effect size determination, separate analyses by diagnostic groups must be performed.

However, some general patterns of effect size differences per condition are identifiable. Brain volume and intelligence are stronger correlated in conditions where maturation is hindered due to confined space (volume). Examples are microencephaly, children extremely born preterm, patients suffering from cerebral palsy, and developmental delay due to alcohol abuse during pregnancy. Every gain in brain volume is a reduction of this confinement and benefits maturation. Brain volume is therefore used as a developmental marker in those contexts (e.g. Katušić et al., 2020). The correlation between brain volume and intelligence disappears when clinical conditions lead to an enlargement of brain volume. In this case gains in brain volume do not enhance efficiency of the brain, and compartments of cerebrospinal fluids are disproportionally enlarged compared to grey and white matter tissue (e.g. de Zwarte et al., 2019). For conditions where brain volume is not obviously affected matters are more complex. The results of this thesis demonstrate that on average the correlation of brain volume and intelligence is weakened. Reasons for weakened correlations depend on the context of measurement (e.g. state of condition, medication) and on the way a clinical condition does affect cognition in general. A comparison of offspring from schizophrenic and bipolar patients shows that brain volume is more affected by schizophrenia than bipolar disorder (van Haren et al., 2019). The correlation between brain volume and IQ does not differ much between bipolar patients and controls (Vreeker et al., 2017). The comparison of healthy and clinical samples therefore provides interesting starting points for why brain volume and intelligence are correlated in the first place.

Third, the use of extensive intelligence batteries, usually a full Wechsler scale, is associated with higher effect sizes. However, a gradual sequence as in Gignac and Bates (2017) was not observed. Studies utilizing intelligence tests that correlate quite well with *g* (rating "3", 2-8 subtests assessing 2-3 intelligence dimensions) reported on average lower effect sizes than studies with rating "2" (1-2 subtests assessing 1-2 dimensions). Also, the differences between the rating groups are not as large as in Gignac and Bates (2017). The difference between rating groups 2 and 4 was remarkably high ($r = .18$) in their meta-analysis. The difference is $r = .10$ in this master's thesis.

Gignac and Bates (2017) have described the influence of test extensiveness as "measurement quality". From my point of view, there are some arguments against this label. (1) Classical measurement quality criteria, for example the experience and competence of test givers, the appropriateness of the testing environment, or the condition of participants on the day of testing have not been considered. The number of subtests and domains involved have been assessed. The authors acknowledge that classic measurement quality criteria as mentioned above may act as a confound, because comprehensive IQ tests might have been administered more likely by trained personnel than abbreviated IQ tests or subtests (Gignac & Bates, 2017, 28). (2) The label measurement quality implies that primary researchers have made either a better or worse job measuring intelligence. While this is true if one only considers how broad the assessment was, measurement quality is an infelicity chosen expression from the perspective of primary researchers in this context. Research goals, economic constraints and ethical considerations (like burden of total testing time) are different across studies. A full-scale intelligence battery may not be cost effective (when the main goal is not an estimate of the association between brain volume and intelligence) or ethically desirable. (3) The label does not enhance concept precision. Other terms like “measurement modality” might be conceivable, but there is no benefit in using those. To call a spade a spade, I suggest naming the variable operationalization criteria (number of subtests and domains). Calling this influence correlation with *g* as I did is partly warranted as there is likely a moderate influence of *g* loadings on the association between brain volume and intelligence (Woodley of Menie et al., 2016), but the exact composition of this moderating influence cannot be clearly determined. It remains unresolved what part *g* theory and other factors such as test quality criteria (e.g. reliability) or the type of variable operationalization (coding) have.

Fourth, a decline effect was observed in the updated full-scale IQ data, 29 years after the first MRI study by Willerman et al. (1991). Especially study results between 2009 and 2016 based on larger samples have amended the summary effect downwards. This *inflated decline effect* is most likely rooted in the average underpowered design of earlier studies and missing zero or negative correlations due to selective reporting. Both factors are typically associated with decline effects (Protzko & Schooler, 2017). Other explanations such as changing measurement or analysis procedures seem unlikely as there were no dramatic changes in any of them in the context of in vivo brain volume, intelligence and bivariate correlational study designs. Interestingly, the effect of publication year is reduced in the verbal IQ and vanishes in the performance IQ data. This agrees well with the fact that dissemination bias analyses for verbal and performance IQ data generally showed a reduced impact compared to full-scale IQ data. The concentration of effect size inflation on headline effects suggests that strategic research and reporting behavior may have contributed to an overestimation of effects in earlier studies.

Fifth, analyses in this thesis suggest that the correlation between brain volume and full-scale IQ based on healthy samples is not influenced by age. No matter how the variable age was operationalized, whether categorically with two levels, or continuously as mean age, analyses showed no systematic influence. The same applies to performance IQ. However, for verbal IQ data an influence is observable. The younger the mean age of the sample, the higher the correlation between brain volume and verbal intelligence. This result can be interpreted in such a way that with increasing age educational experiences (building up knowledge) gain in importance and the relevance of brain volume thus decreases a little. A more detailed analysis of the influence of certain verbal IQ subtests would provide further insight. General intelligence, processing speed and visuospatial ability for example decrease, however memory does not decrease as much (Hoogendam et al., 2014). Even within different memory functions diverging patterns of aging can be detected. In old age some components of working memory are better preserved than others. Episodic and prospective memory, and the ability to divide attention decreases with age, while implicit and semantic memory and sustained attention remain stable (Oschwald et al., 2019). Examining if brain volume operationalization (TBV or ICV) influenced results lead to a more pronounced result for verbal IQ but otherwise was not inconspicuous.

The association between brain volume and IQ could very well vary with age. For example, the indirect factors (compensatory scaffolding) of the *revised scaffolding theory of aging and cognition* (STAC-r; Reuter-Lorenz & Park, 2014) suggest a variation in the association between brain volume and intelligence with age, since brain tissue loss is not necessarily accompanied by worse cognitive performance. Longitudinal studies have shown that a reduction in brain volume does need not be associated with a reduction in cognitive performance (Jäncke et al., 2020), at least in a limited period of time. There are some other factors such as lifestyle and health which might be important for further consideration. For example, sporting activities protect against accelerated loss of brain volume (Pruimboom et al., 2015). The decline in fluid intelligence could even be entirely due to deteriorating health (Bergman & Almkvist, 2013). Reasons for why no influence of age was detected may also lie aside from theoretical considerations in coding decisions. It might have been that the categorial coding of children/adolescents vs. adults was too insensitive to detect age effects. Mean age is somewhat more informative, but standard deviations of age means were sometimes large. In order to test this possibility, I conducted a supplementary mixed effects meta-regression based on a refined categorization of age groups (results can be found in Appendix G). There were indeed some differences between younger age groups and older adults (> 35 years), especially between adolescents and elderly, but group sizes were too small to have great confidence in these results (results were statistically insignificant). They rather suggest that moderator tests of age might be (in a small degree) sensitive to variable operationalization choices. The focus on studies using TBV was not possible due to lack of data. Although a greater level of detail might answer some remaining questions, age seems to have limited influence on the association between brain volume and (performance) IQ.

Sixth, there is no difference in the association between brain volume and intelligence domains for females and males. Neither the use of categories, nor using male ratio of samples as a predictor revealed noticeable effects. As there are no other coding options, and data availability is satisfactory, we can be fairly certain that one´s sex alone does not influence the association between brain volume and intelligence. An analysis based on correlation coefficients corrected for range departure did not change the results. This finding is in line with previous results (Burgaleta et al., 2012; Escorial et al., 2015; Pietschnig et al., 2015). In contrast, van der Linden et al. (2017) have observed small differences in general intelligence between females and males in a large sample (*n* = 896), which were mediated by brain size. Although there may be small variations within individual samples, the majority of the results suggests that there is no difference in brain volume and intelligence association for females and males. Some researchers have suggested that there might be an interaction effect of sex and age (Lynn, 1994; McDaniel, 2005). The analyses in this thesis provide no support for the existence of such an effect.

Brain volume and intelligence correlate. The brain volume of males is larger on average. Females and males show no (or very small) differences in general intelligence. How is this possible? Some study results indicate anatomical and functional differences that compensate for a difference in brain volume (see van der Linden et al., 2017). Recent research has expanded insights into these differences by using transcriptomic analysis methods (Liu et al., 2020). The exact mechanisms of this compensation are not yet understood.

Seventh, a reopening of the topic of ethnicity/race which has accompanied research into the association between brain volume and intelligence for so long failed due to a lack of data as well as conceptual problems. On the one hand, the overwhelming majority of study participants of in vivo brain volume studies is white. There are only eight correlations based on other ethnic categories available in the entire full-scale IQ dataset. On the other hand, using race or ethnicity might not be adequate to obtain meaningful results in this context. Most theories on differences in brain volume and intelligence include the factor climate zone. So even if we had enough data for group comparisons based on race, this comparison would be ineffective because each of the categories includes several climate zones. With the help of the concept of ethnicity it would at least be possible to define population groups more precise and to make targeted comparisons. For example, Kura et al. (2014) compared a certain ethnic minority, called Ainu, with the rest of the Japanese population in terms of brain volume. The problem with this approach is that it is very data intensive. A large number of data sets from all over the world in which in vivo brain volume and intelligence were measured would be needed. This is far from reality, especially in structurally weak regions. The data set collected for this thesis contains almost exclusively studies that have differentiated participants into White, Black, Hispanic or Asian following the North American habit. There are hardly any classifications according to ethnic criteria. Almost no study from Europe covered the race or ethnicity of their participants. These problems are compounded by the difficulty of controlling potential confusions, such as SES (Jensen & Sinha, 1993). Not to be forgotten, studies based on social classification variables in the context of intelligence are sensitive, and results need to be robust. This may mean that large-scale genetic association studies are a better starting point for meaningful results. These avoid both the data problem and conceptual difficulties. Genetic clusters or geographic variations of allele frequencies offer a much finer construction of variables (Heinz et al., 2014).

Generally, the association of brain volume and intelligence is expected to apply to people from all continents. In addition to the studies used in this thesis, further studies using surrogate measures to estimate brain volume support this assumption (Bakhiet et al., 2016; Hein et al., 2014; Ivanovic et al., 2014). Whether the association of brain volume and intelligence varies between ethnic groups is not yet foreseeable.

One of the main objectives of this thesis was the application of methods to examine the influence of meta-analytical specifications on the result. Since this thesis used those methods as one of the first, some remarks regarding their application may be useful. (1) Many other methods were used; however, none was able to model data and analytical specifications simultaneously. The execution of different meta-analytical approaches side by side is not an alternative, because this way one data set only can be used at a time, or the presentation of results quickly becomes cluttered. Interesting alternatives are currently being developed (e.g. *Bayesian model averaging* for meta-analysis, Heck et al., 2019), however they are not yet as flexible in the inclusion of data and analysis procedures. Bayesian methods may have advantages when the data pool is small. (2) The approach of Voracek et al. (2019) is very flexible. There is no limit to the number or type of specifications that can be modeled. The only practical limit is computational feasibility. (3) The application serves several purposes. On the one hand the possible result space of the summary effect is determined, on the other hand influential specifications can be identified more easily. This implies a time saving in the research process by anticipating repetitions due to other specifications and avoids tedious discussions about the best way forward. In addition, it offers the possibility to use several tools at the same time in a clear and concise way, and to perform the ultimate robustness check of your results. (4) The code from Voracek et al. (2019) is relatively easy to use. In principle you only have to enter relevant specifications and adjust some parameters. (5) The graphical outputs of the *specification curve* analyses are in my opinion more flexible and detailed. The *p*-value histogram from the *multiverse analysis* is suitable to inspect the evidence against the H0. Depending on the size of the data pool the inferential statistical validation by bootstrapping in the *specification curve* plots is more robust. If effect sizes are to be evaluated the descriptive output of the *specification curve* analysis is most suitable. It provides the opportunity to interpret the influence of individual specifications. An advantage of *p*-value histograms might be that they are easier to read. (6) According to Voracek et al. (2019) the increased use of *multiverse* and *specification curve* analyses could reveal whether data or analysis specifications have more influence on the result in a meta-analytical context. In this thesis it was one specification each that influenced the summary effect in somewhat equal parts. The "which" factor is the selection of a set of study results based on the scope of the intelligence testing. The "how" factor is whether correlation coefficients were corrected for range departure. In sum, *multiverse* and *specification curve* analyses by Voracek et al. (2019) enrich the meta-analytical arsenal of methods. The prospective power of specification analyses can be used to clarify or even anticipate inconsistencies and to report the entire range of results of all reasonable meta-analytical specifications.

Limitations

By applying a variety of diverging methods based on a relatively large data set, the results presented in this thesis stand, from an analytical point of view, on solid ground. However, there are some limitations worth considering. First, neither the literature search nor the coding was reviewed by a second person. The reliability of these processes was also not determined intrapersonally. Neither complies with current recommendations (e.g. Higgins et al., 2009). However, the entire coding process was repeated, and inconsistencies were corrected. Since any analysis can only be as good as the data set on which it is based, an external review of the data update is desirable.

There are also some open questions from an analytical point of view. For example, it was not possible to clearly determine how the saturation in *g* influences differences between intelligence domains or the extent of intelligence testing. To exclude alternative explanations, a correction for unreliability and construct validity in the manner of psychometric meta-analyses would be helpful. Although reliability measures have hardly been reported in included studies, the use of reliability measures from the respective manuals could at least exclude systematic variations due to reliability differences between IQ tests.

Gignac and Bates (2017) point out in the discussion of their meta-analysis that the measurement of brain volume could be evaluated in a similar way as the measurement of intelligence. In addition to range departure corrections, the capability of the MRI device used, corrections for measurement artifacts (e.g. head movement) and the data extraction method could be considered. An influence of the measurement modality on the outcome could thus be determined. Even though the determination of brain volume by MRI is precise, there have been improvements of this technology over the time span of the included studies. However, it is unclear how a rating procedure could be constructed to capture differences in the measurement modality and fit the reporting habits in the literature. This aspect may resolve itself over time. More and more large data sets with harmonized measurement procedures are published. Due to their great weight in meta-analyses it is questionable whether the effort described above is worth it. It is also unclear to what extent current trends in the implementation of machine learning will determine measurement or analysis modalities in the future.

Why Is Brain Volume Associated with Intelligence?

The results of this master thesis show that the association of brain volume and intelligence is surprisingly stable. Neither age nor sex influence the association, and it remains noticeable even under many clinical circumstances. But why do brain volume and intelligence correlate? The most common theory of their association is quite simple. Larger brains contain more (cortical) neurons and therefore have more computational power to solve complex tasks (see van der Linden et al., 2017). Some examples were already mentioned rendering this explanation problematic, if it stands alone. An IQ difference between females and males is absent, or at least much less obvious than the difference in average brain volumes (Ruigrok et al., 2014). In some forms of autism larger brain volumes are also not associated with increases in intelligence (Amaral et al., 2017). Lower intelligence due to extremely large brains is a particularly striking illustration of the problem (Petersson et al., 1999). Considering other structural brain properties like grey matter cortical thickness and surface area, or white matter integrity (see Mathiesen, 2015) does not solve this problem either, since in a comparison brain volume explained by far the most variance (Ritchie et al., 2015). The number of neurons may have played a role from an evolutionary perspective (Pietschnig et al., 2015), but these examples show that other factors must be combined with the number of neurons. A useful refinement is the conceptualization of brain volume as a function of cortical neuron numbers and degree of myelination (Roth & Dicke, 2005). Although this concept does not fully resolve the outlined problems, it opens possibilities to integrate functional properties like network flexibility and dynamics (Barbey, 2017). Together with theories trying to identify those structures and networks considered to be particularly important for intelligence performance (for a review see Jausovec, 2019), the role of brain volume may be clearer to determine. But the quest for a full picture does not end there. Further factors could be neurogenesis (Hill et al., 2019), neuro-hormonal regulation (Saniotis, 2020), and cell properties (see Goriounova & Mansvelder, 2019). Understanding a variety of these factors and their complex interactions is likely to be necessary in order to understand how exactly the size of the brain relates to intelligence. A promising array of research are *genome-wide association studies* (GWAS). GWAS examine genetic variations in a genome by associating a phenotype (e.g. brain volume) to alleles within genomic loci. Jansen et al. (2019) identified 67 shared genes as well as five genomic loci that may drive the genetic correlation between brain volume and intelligence. These genes are involved mainly in regulating cell growth. Further research in that area might provide true insight. Whether deciphering this fascinating complexity will ultimately result in brain size being seen as a "poor proxy" (Woodley of Menie et al., 2016, 218) for all things we did not understand or whether brain size provides direct functional benefits is not yet clear. Recent research favors the view of a causal influence of brain volume on IQ (Lee et al., 2019), and found a functional equivalent of their association in the sense that fMRI signals of intelligent people wander in a larger space (Dizaji et al., 2019).

Conclusion

The association between in vivo brain volume and intelligence in humans is robust across age, sex, intelligence domain, various clinical conditions, and meta-analytical data and analyses specifications. The careful search for grey literature and the success of consortia in collecting large samples has efficiently reduced the impact of dissemination bias. This gives reason for optimism in the study of brain-behavior associations. However, an *inflated decline effect* is observed over the entire time span of the included publications in this thesis.

The use of *specification curve* in a meta-analytical context according to Voracek et al. (2019) enabled the modeling of all data and analysis specifications used by meta-analysts on the topic. The summary effect ranges from approximately $r = .20$ to $r = .35$. The exclusive consideration of study results based on extensive IQ tests, and the use of correlation coefficients corrected for range departure yield higher effect sizes. Data subsets by age and sex, the meta-analytic model, and analyses using various *r* metrics have little impact.

It could not have been clarified to what extent the saturation in *g* leads to slight differences in effect sizes between domains, and whether other factors such as fluctuating reliability between used tests play a role.

Although there are still some open questions, the correlative study design is reaching its limits when investigating the association between in vivo brain volume and intelligence.

Progress in genetic research methods, and improved data availability and precision of brain-behavior studies offer the prospect of elucidating causal relations between brain volume and intelligence as well as the complex interplay of factors at different levels of neuroscientific research of individual differences in which brain volume and intelligence are embedded.

References

*References marked with an asterisk indicate studies included in the meta-analysis.

Adhikari, M. (2008). 'Streams of blood and streams of money': new perspectives on the annihilation of the Herero and Nama peoples of Namibia, 1904-1908. *Kronos, 34*(1), 303-320.

*Ahn, J. I., Yu, S. T., Sung, G., Choi, T. K., Lee, K. S., Bang, M., & Lee, S. H. (2019). Intra-individual variability in neurocognitive function in schizophrenia: relationships with the corpus callosum. *Psychiatry Research: Neuroimaging, 283*, 1-6. https://doi.org/10.1016/j.pscychresns.2018.11.005

*Amaral, D. G., Li, D., Libero, L., Solomon, M., Van de Water, J., Mastergeorge, A., Naigles, L., Rogers, S., & Wu Nordahl, C. (2017). In pursuit of neurophenotypes: The consequences of having autism and a big brain. *Autism Research, 10*(5), 711-722. https://doi.org/10.1002/aur.1755

*Amat, J. A., Bansal, R., Whiteman, R., Haggerty, R., Royal, J., & Peterson, B. S. (2008). Correlates of intellectual ability with morphology of the hippocampus and amygdala in healthy adults. *Brain and Cognition, 66*(2), 105-114. https://doi.org/10.1016/j.bandc.2007.05.009

*Andreasen, N. C., Flaum, M., Swayze II, V., O'Leary, D. S., Alliger, R., Cohen, G., Ehrhardt, J., & Yuh, W. T. C. (1993). Intelligence and brain structure in normal individuals. *American Journal of Psychiatry*, 150, 130–1134.

Ankney, C. D. (1992). Sex differences in relative brain size: The mismeasure of woman, too? *Intelligence, 16*, 329–336. https://doi.org/10.1016/0160-2896(92)90013-H

*Annink, K. V., de Vries, L. S., Groenendaal, F., van den Heuvel, M. P., van Haren, N. E., Swaab, H., van Handel, M., Jongmans, M. J., Benders, M. J., & van der Aa, N. E. (2019). The long-term effect of perinatal asphyxia on hippocampal volumes. *Pediatric Research, 85*(1), 43-49. https://doi.org/10.1038/s41390-018-0115-8

*Antonova, E., Kumari, V., Morris, R., Halari, R., Anilkumar, A., Mehrotra, R., & Sharma, T. (2005). The relationship of structural alterations to cognitive deficits in schizophrenia: a voxel-based morphometry study. *Biological Psychiatry, 58*(6), 457-467. https://doi.org/10.1016/j.biopsych.2005.04.036

*Arhan, E., Gücüyener, K., Soysal, Ş., Şalvarlı, Ş., Gürses, M. A., Serdaroğlu, A., Demir, E., Ergenekon, E., Türkyılmaz, C., Önal, E., Koç, E., & Atalay, Y. (2017). Regional brain volume reduction and cognitive outcomes in preterm children at low risk at 9 years of age. *Child's Nervous System, 33*(8), https://doi.org/10.1007/s00381-017-3421-2

Arribas-Aguila, D., Abad, F. J., & Colom, R. (2019). Testing the developmental theory of sex differences in intelligence using latent modeling: Evidence from the TEA Ability Battery (BAT-7). *Personality and Individual Differences, 138*, 212-218. https://doi.org/10.1016/j.paid.2018.09.043

*Ashtari, M., Avants, B., Cyckowski, L., Cervellione, K. L., Roofeh, D., Cook, P., Gee, J., Sevy, S., & Kumra, S. (2011). Medial temporal structures and memory functions in adolescents with heavy cannabis use. *Journal of Psychiatric Research, 45*(8), 1055-1066. https://doi.org/10.1016/j.jpsychires.2011.01.004

*Aydin, K., Uysal, S., Yakut, A., Emiroglu, B., & Yılmaz, F. (2012). N-acetylaspartate concentration in corpus callosum is positively correlated with intelligence in adolescents. *Neuroimage, 59*(2), 1058-1064. https://doi.org/10.1016/j.neuroimage.2011.08.114

*Aylward, E. H., Minshew, N. J., Field, K., Sparks, B. F., & Singh, N. (2002). Effects of age on brain volume and head circumference in autism. *Neurology, 59*, 175–183. https://doi.org/10.1212/WNL.59.2.175

Bakhiet, S. F. A., Essa, Y. A. S., Dwieb, A. M. M., Elsayed, A. M. A., Sulman, A. S. M., Cheng, H., & Lynn, R. (2017). Correlations between intelligence, head circumference and height: Evidence from two samples in Saudi Arabia. *Journal of Biosocial Science*, *49*(2), 276-280. https://doi.org/10.1017/S0021932016000249

*Ballester-Plané, J., Laporta-Hoyos, O., Macaya, A., Póo, P., Meléndez-Plumed, M., Vázquez, É., Delgado, I., Zubiaurre-Elorza, L., Narberhaus, A., Toro-Tamargo, E., Russi, M. E., Tenoria, V., Segarra, D., & Pueyo, R. (2016). Measuring intellectual ability in cerebral palsy: The comparison of three tests and their neuroimaging correlates. *Research in Developmental Disabilities, 56*, 83-98. https://doi.org/10.1016/j.ridd.2016.04.009

Barbey, A. K. (2018). Network neuroscience theory of human intelligence. *Trends in Cognitive Cciences*, *22*(1), 8-20. https://doi.org/10.1016/j.tics.2017.10.001

Bartholomeusz, H. H., Courchesne, E., & Karns, C. M. (2002). Relationship between head circumference and brain volume in healthy normal toddlers, children, and adults. *Neuropediatrics, 33*(5), 239-241. https://doi.org/10.1055/s-2002-36735

*Bathelt, J., Scerif, G., Nobre, A. C., & Astle, D. E. (2019). Whole-brain white matter organization, intelligence, and educational attainment. *Trends in Neuroscience and Education, 15*, 38-47. https://doi.org/10.1016/j.tine.2019.02.004

Baujat, B., Mahé, C., Pignon, J. P., & Hill, C. (2002). A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials. *Statistics in Medicine, 21*(18), 2641-2652. https://doi.org/10.1002/sim.1221

Bergman, I., & Almkvist, O. (2013). The effect of age on fluid intelligence is fully mediated by physical health. *Archives of Gerontology and Geriatrics*, *57*(1), 100-109. https://doi.org/10.1016/j.archger.2013.02.010

Bergmann, C. (1848). *Über die Verhältnisse der Wärmeökonomie der Thiere zu ihrer Größe*. Vandenhoeck & Ruprecht.

Betancourt, H., & López, S. R. (1993). The study of culture, ethnicity, and race in American psychology. *American Psychologist, 48*(6), 629–637. https://doi.org/10.1037/0003-066X.48.6.629

*Betjemann, R. S., Johnson, E. P., Barnard, H., Boada, R., Filley, C. M., Filipek, P. A., Willcutt, E. G., DeFries, J. C., & Pennington, B. F. (2010). Genetic covariation between brain volumes and IQ, reading performance, and processing speed. *Behavior Genetics, 40*(2), 135-145. https://doi.org/10.1007/s10519-009-9328-2

*Bigler, E. D. (1995). Brain morphology and intelligence. *Developmental Neuropsychology, 11*(4), 377-403. https://doi.org/10.1080/87565649509540628

Bigler, E. D. (2017). Structural neuroimaging in neuropsychology: History and contemporary applications. *Neuropsychology, 31*(8), 934–953. https://doi.org/10.1037/neu0000418

*Bigler, E. D., Abildskov, T. J., Petrie, J., Farrer, T. J., Dennis, M., Simic, N., Taylor, H. G., Rubin, K. H., Vannatta, K., Gerhardt, C. A., Stancin, T., & Owen Yeates, K. (2013). Heterogeneity of brain lesions in pediatric traumatic brain injury. *Neuropsychology, 27*(4), 438–451. https://doi.org/10.1037/a0032837

Binet, A., & Simon, T. (1916). New methods for the diagnosis of the intellectual level of subnormals. (L'Année Psych., 1905, pp. 191-244). In A. Binet, T. Simon & E. S. Kite (Trans.), *The development of intelligence in children (The Binet-Simon Scale)* (p. 37–90). Williams & Wilkins Co. https://doi.org/10.1037/11069-002

*Bjuland, K. J., Rimol, L. M., Løhaugen, G. C., & Skranes, J. (2014). Brain volumes and cognitive function in very-low-birth-weight (VLBW) young adults. *European Journal of Paediatric Neurology, 18*(5), 578-590. http://dx.doi.org/10.1016/j.ejpn.2014.04.004

*Blatter, D. D., Bigler, E. D., Gale, S. D., Johnson, S. C., Anderson, C. V., Burnett, B. M., Ryser, D., Macnamara, S. E., & Bailey, B. J. (1997). MR-based brain and cerebrospinal fluid measurement after traumatic brain injury: correlation with neuropsychological outcome. *American Journal of Neuroradiology, 18*(1), 1-10.

Boake, C. (2002). From the Binet–Simon to the Wechsler–Bellevue: Tracing the history of intelligence testing. *Journal of Clinical and Experimental Neuropsychology, 24*(3), 383-405. https://doi.org/10.1076/jcen.24.3.383.981

*Boberg, R., & Wallström, S. (2015). A study of twins born preterm: Functional lateralization, cognition, and brain volumes in twin and single-born children at early school ages [Master´s thesis, Umea University]. Digitala Vetenskapliga Arkivet. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A816049&dswid=8975

*Bohlken, M. M., Brouwer, R. M., Mandl, R. C., Hedman, A. M., van den Heuvel, M. P., van Haren, N. E., Kahn, R. S., & Pol, H. E. H. (2016). Topology of genetic associations between regional gray matter volume and intellectual ability: Evidence for a high capacity network. *Neuroimage, 124*, 1044-1053. https://doi.org/10.1016/j.neuroimage.2015.09.046

Borenstein, M. (2019). *Common mistakes in meta-analysis and how to avoid them*. Biostat, Incorporated.

Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8*(1), 5-18. https://doi.org/10.1002/jrsm.1230

Braje, S. E., & Nagayama Hall, G. C. (2015). Cross-cultural issues in assessment. *The Encyclopedia of Clinical Psychology*, 1-9. https://doi.org/10.1002/9781118625392.wbecp435

Broca, P. (1861). Remarks on the seat of the faculty of articulated language, following an observation of aphemia (loss of speech). *Bulletin de la Société Anatomique, 6*, 330-57.

*Burgaleta, M., Head, K., Álvarez-Linera, J., Martínez, K., Escorial, S., Haier, R., & Colom, R. (2012). Sex differences in brain volume are related to specific skills, not to general intelligence. *Intelligence, 40*(1), 60-68. https://doi.org/10.1016/j.intell.2011.10.006

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365-376. https://doi.org/10.1038/nrn3475

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science, 2*(2), 115-144. https://doi.org/10.1177%2F2515245919847196

Caspi, Y., Brouwer, R. M., Schnack, H. G., van de Nieuwenhuijzen, M. E., Cahn, W., Kahn, R. S., Niessen, W. J., van der Lugt, A., & Hulstoff Pol, H. (2020). Changes in the intracranial volume from early adulthood to the sixth decade of life: A longitudinal study. *NeuroImage, 116842*. https://doi.org/10.1016/j.neuroimage.2020.116842

*Castellanos, F. X., Giedd, J. N., Berquin, P. C., Walter, J. M., Sharp, W., Tran, T., Vaituzis, A. C., Blumenthal, J. D., Nelson, J., Bastain, T. M., Zijdenbos, A., Evans, A. C., & Rapoport, J. L. (2001). Quantitative brain magnetic resonance imaging in girls with attention-deficit/hyperactivity disorder. *Archives of General Psychiatry, 58*(3), 289-295. https://doi.org/10.1001/archpsyc.58.3.289

*Castellanos, F. X., Giedd, J. N., Eckburg, P., Marsh, W. L., Vaittuzis, A. C., Kaysen, D., Hamburger, S. D., & Rapoport, J. L. (1994). Quantitative morphology of the caudate nucleus in attention deficit hyperactivity disorder. *American Journal of Psychiatry*, *151*(12), 1791–1796. https://doi.org/10.1176/ajp.151.12.1791

*Castro-Fornieles, J., Bargalló, N., Lázaro, L., Andrés, S., Falcon, C., Plana, M. T., & Junqué, C. (2009). A cross-sectional and follow-up voxel-based morphometric MRI study in adolescent anorexia nervosa. *Journal of Psychiatric Research, 43*(3), 331-340. https://doi.org/10.1016/j.jpsychires.2008.03.013

*Chen, X., Coles, C. D., Lynch, M. E., & Hu, X. (2012). Understanding specific effects of prenatal alcohol exposure on brain structure in young adults. *Human Brain Mapping, 33*(7), 1663-1676. https://doi.org/10.1002/hbm.21313

*Chiang, M. C., Reiss, A. L., Lee, A. D., Bellugi, U., Galaburda, A. M., Korenberg, J. R., Mills, D. L., Toga, A. W., & Thompson, P. M. (2007). 3D pattern of brain abnormalities in Williams syndrome visualized using tensor-based morphometry. *Neuroimage, 36*(4), 1096-1109. https://doi.org/10.1016/j.neuroimage.2007.04.024

*Choi, Y. Y., Shamosh, N. A., Cho, S. H., DeYoung, C. G., Lee, M. J., Lee, J. M., Lee, J-M., Kim, S., Cho, Z-H., Kim, K., Gray, J. R., & Lee, K. H. (2008). Multiple bases of human intelligence revealed by cortical thickness and neural activation. *Journal of Neuroscience, 28*(41), 10323-10329. https://doi.org/10.1523/JNEUROSCI.3259-08.2008

Coburn, K. M., & Vevea, J. L. (2019). weightr: Estimating weight-function models for publication bias. R package version 2.0.2. https://CRAN.R-project.org/package=weightr

Coffey, C. E. (2000). Anatomic imaging of the aging human brain: Computed Tomography and Magnetic Resonance Imaging. In: C. E. Coffey & J. L. Cummings (Eds.), *Textbook of Geriatric Neuropsychiatry* (2nd ed., pp. 181-238). The American Psychiatric Press.

*Coffey, C. E., Ratcliff, G., Saxton, J. A., Bryan, R. N., Fried, L. P., & Lucke, J. F. (2001). Cognitive correlates of human brain aging: a quantitative magnetic resonance imaging investigation. *The Journal of Neuropsychiatry and Clinical Neurosciences, 13*(4), 471-485. https://doi.org/10.1176/jnp.13.4.471

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

*Collinson, S. L., Mackay, C. E., James, A. C., Quested, D. J., Phillips, T., Roberts, N., & Crow, T. J. (2003). Brain volume, asymmetry and intellectual impairment in relation to sex in early-onset schizophrenia. *The British Journal of Psychiatry, 183*(2), 114-120. https://doi.org/10.1192/bjp.183.2.114

Copas, J. B., & Shi, J. Q. (2001). A sensitivity analysis for publication bias in systematic reviews. *Statistical Methods in Medical Research, 10*(4), 251-65. https://doi.org/10.1177%2F096228020101000402

*Cox, S. R., Ritchie, S. J., Fawns-Ritchie, C., Tucker-Drob, E. M., & Deary, I. J. (2019). Structural brain imaging correlates of general intelligence in UK Biobank. *Intelligence, 76*, 101376. https://doi.org/10.1016/j.intell.2019.101376

Dahlke, J., Wiernik, B. (2019). psychmeta: an R package for psychometric meta-analysis. *Applied Psychological Measurement, 43*(5), 415–416. https://doi.org/10.1177/0146621618795933

Darwin, C. (1871). The descent of man, and selection in relation to sex. Murray.

Daseking, M., Petermann, F., & Waldmann, H. C. (2017). Sex differences in cognitive abilities: Analyses for the German WAIS-IV. *Personality and Individual Differences, 114*, 145-150. https://doi.org/10.1016/j.paid.2017.04.003

Deary, I. J. (2014). The stability of intelligence from childhood to old age. *Current Directions in Psychological Science, 23*(4), 239-245. https://doi.org/10.1177%2F0963721414536905

Deary, I. J. (2020). *Intelligence: A very short introduction.* Oxford University Press.

Deary, I. J., Irwing, P., Der, G., & Bates, T. C. (2007). Brother–sister differences in the g factor in intelligence: Analysis of full, opposite-sex siblings from the NLSY1979. *Intelligence, 35*(5), 451-456. https://doi.org/10.1016/j.intell.2006.09.003

Deary, I. J., Thorpe, G., Wilson, V., Starr, J. M., & Whalley, L. J. (2003). Population sex differences in IQ at age 11: The Scottish mental survey 1932. *Intelligence, 31*(6), 533-542. https://doi.org/10.1016/S0160-2896(03)00053-9

*Debbané, M., Schaer, M., Farhoumand, R., Glaser, B., & Eliez, S. (2006). Hippocampal volume reduction in 22q11. 2 deletion syndrome. *Neuropsychologia, 44*(12), 2360-2365. https://doi.org/10.1016/j.neuropsychologia.2006.05.006

*DeBoer, T., Wu, Z., Lee, A., & Simon, T. J. (2007). Hippocampal volume reduction in children with chromosome 22q11.2 deletion syndrome is associated with cognitive impairment. *Behavioral and Brain Functions, 3*(1), 54. https://doi.org/10.1186/1744-9081-3-54

Dizaji, A. S., Khodaei, M. R., & Soltanian-Zadeh, H. (2019). Resting-state fMRI signals of intelligent people wander in a larger space. *BioRxiv*, 529362. https://doi.org/10.1101/529362

*Dörnte, J. (2007). Intrakranielle Volumenänderungen im Magnetresonanztomogramm (MRT) und neuropsychologische Veränderungen bei Patienten mit Mild Cognitive Impairment (MCI) [Doctoral dissertation, University of Göttingen]. Deutsche Nationalbibliothek. https://d-nb.info/989315606/34

Duan, B., & Dunlap, W. P. (1997). The accuracy of different methods for estimating the standard error of correlations corrected for range restriction. *Educational and Psychological Measurement, 57*(2), 254-265. https://doi.org/10.1177%2F0013164497057002005

Duggan E. C., Garcia-Barrera M. A. (2015). Executive Functioning and Intelligence. In: Goldstein S., Princiotta D., Naglieri J. (eds), *Handbook of Intelligence*. Springer, New York, NY. https://doi.org/10.1007/978-1-4939-1562-0_27

Duval, S. J., & Tweedie, R. L. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*(449), 89–98. https://doi.org/10.1080/01621459.2000.10473905

*Ebner, F., Tepest, R., Dani, I., Pfeiffer, U., Schulze, T. G., Rietschel, M., Maier, W., Träber, F., Block, W., Schild, H. H., Wagner, M., Steinmetz, H., Gaebel, W., Horner, W. G., Schneider-Axmann, T., & Falkai, P. (2008). The hippocampus in families with schizophrenia in relation to obstetric complications. *Schizophrenia Research, 104*(1-3), 71-78. https://doi.org/10.1016/j.schres.2008.06.007

Egan, V., Chiswick, A., Santosh, C., Naidu, K., Rimmington, J. E., & Best, J. J. (1994). Size isn't everything: A study of brain volume, intelligence and auditory evoked potentials. *Personality and Individual Differences, 17*(3), 357-367. https://doi.org/10.1016/0191-8869(94)90283-6

*Egan, V., Wickett, J. C., & Vernon, P. A. (1995). Brain size and intelligence: Erratum, addendum, and correction. *Personality and Individual Differences, 19*, 113–115. https://doi.org/10.1016/0191-8869(95)00043-6

*Elliott, M. L., Belsky, D. W., Anderson, K., Corcoran, D. L., Ge, T., Knodt, A., Prinz, J. A., Sugden, K., Williams, B., Ireland, D., Poulton, R., Caspi, A., Holmes, A., Moffitt, T., & Harriri, A. R. (2019). A polygenic score for higher educational attainment is associated with larger brains. *Cerebral Cortex, 29*(8), 3496-3504. https://doi.org/10.1093/cercor/bhy219

Enders, C. K. (2010). *Applied missing data analysis*. Guildford Press.

Escorial, S., Román, F. J., Martínez, K., Burgaleta, M., Karama, S., & Colom, R. (2015). Sex differences in neocortical structure and cognitive performance: a surface-based morphometry study. *Neuroimage, 104*, 355-365. https://doi.org/10.1016/j.neuroimage.2014.09.035

Ferguson, K. J., Wardlaw, J. M., Edmond, C. L., Deary, I. J., & MacLullich, A. M. (2005). Intracranial area: a validated method for estimating intracranial volume. *Journal of Neuroimaging, 15*(1), 76-78. https://doi.org/10.1111/j.1552-6569.2005.tb00289.x

*Ferreira, D., Bartrés-Faz, D., Nygren, L., Rundkvist, L. J., Molina, Y., Machado, A., Jungué, C., Barroso, J., & Westman, E. (2016). Different reserve proxies confer overlapping and unique endurance to cortical thinning in healthy middle-aged adults. *Behavioural Brain Research, 311*, 375-383. https://doi.org/10.1016/j.bbr.2016.05.061

*Fine, J. G., Semrud-Clikeman, M., Keith, T. Z., Stapleton, L. M., & Hynd, G. W. (2007). Reading and the corpus callosum: An MRI family study of Vol. and area. *Neuropsychology, 21*(2), 235–241. https://doi.org/10.1037/0894-4105.21.2.235

Fischl, B. (2012). FreeSurfer. *Neuroimage, 62*(2), 774-781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Fisher, R. A. (1921). On the" probable error" of a coefficient of correlation deduced from a small sample. *Metron, 1*(1), 1-32.

Fisher, Z., Tipton, E. & Zhipeng, H. (2017). Robumeta: robust variance meta-regression (version 2.0). https://CRAN.R-project.org/package=robumeta.

*Flashman, L. A., Andreasen, N. C., Flaum, M., & Swayze, V. W. (1998). Intelligence and regional brain volumes in normal controls. *Intelligence, 25*(3), 149–160. https://doi.org/10.1016/S0160-2896(97)90039-8

Fletcher, R. B., & Hattie, J. (2011). *Intelligence and intelligence testing.* Taylor & Francis.

*Frangou, S., Chitins, X., & Williams, S. C. R. (2004). Mapping IQ and gray matter density in healthy young people. *Neuroimage, 23*(3), 800–805. https://doi.org/10.1016/j.neuroimage.2004.05.027

Galton, F. (1889). Head growth in students at the University of Cambridge. *Nature*, *40*(1031), 318-318. https://doi.org/10.1038/040318a0

*Garde, E., Mortensen, E. L., Krabbe, K., Rostrup, E., & Larsson, H. B. W. (2000). Relation between age-related decline in intelligence and cerebral white-matter hyperintensities in healthy octogenarians: A longitudinal study. *Lancet, 356*(9230), 628–634. https://doi.org/10.1016/S0140-6736(00)02604-0

*Giedd, J. N. (2003). Personal communication to M.A. McDaniel, obtained through the meta-analysis of McDaniel (2005).

Gignac, G. E. (2018). Conceptualizing and measuring intelligence. In: V. Zeigler-Hill & T. K. Shackelford, *the SAGE handbook of personality and individual differences* (pp. 439-464). SAGE.

Gignac, G. E., & Bates, T. C. (2017). Brain volume and intelligence: the moderating role of intelligence measurement quality. *Intelligence, 64*, 18-29. https://doi.org/10.1016/j.intell.2017.06.004

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74-78. https://doi.org/10.1016/j.paid.2016.06.069

Gignac, G., Vernon, P. A., & Wickett, J. C. (2003). Factors influencing the relationship between brain size and intelligence. In: *The scientific study of general intelligence* (pp. 93-106). Pergamon.

Goriounova, N. A., & Mansvelder, H. D. (2019). Genes, cells and brain areas of intelligence. *Frontiers in Human Neuroscience, 13(44)*, 1-14. https://doi.org/10.3389/fnhum.2019.00044

Gould, S. J., & Gold, S. J. (1996). *The mismeasure of man.* WW Norton & company.

*Grazioplene, R. G., G. Ryman, S., Gray, J. R., Rustichini, A., Jung, R. E., & DeYoung, C. G. (2015). Subcortical intelligence: caudate volume predicts IQ in healthy adults. *Human Brain Mapping, 36*(4), 1407-1416. https://doi.org/10.1002/hbm.22710

*Gregory, M. D., Kippenhan, J. S., Dickinson, D., Carrasco, J., Mattay, V. S., Weinberger, D. R., & Berman, K. F. (2016). Regional variations in brain gyrification are associated with general cognitive ability in humans. *Current Biology, 26*(10), 1301-1305. https://doi.org/10.1016/j.cub.2016.03.021

*Grunewaldt, K. H., Fjørtoft, T., Bjuland, K. J., Brubakk, A. M., Eikenes, L., Håberg, A. K., Løhaugen, G. C. C., & Skranes, J. (2014). Follow-up at age 10 years in ELBW children - functional outcome, brain morphology and results from motor assessments in infancy. *Early Human Development, 90*(10), 571-578. http://dx.doi.org/10.1016/j.earlhumdev.2014.07.005

*Gur, R. C., Turetsky, B. I., Matsui, M., Lan, M., Bilker, W., Hughett, P., & Gur, R. E. (1999). Sex differences in brain gray and white matter in healthy young adults: Correlations with cognitive performance. *Journal of the Neuroscience, 19*(10), 4065–4072. https://doi.org/10.1523/JNEUROSCI.19-10-04065.1999

*Haier, R. J., Chueh, D., Touchette, P., Lott, I., Buchsbaum, M. S., MacMillan, D., Sandman, C., LaCasse, L., & Sosa, E. (1995). Brain size and cerebral glucose metabolic rate in nonspecific mental retardation and Down syndrome. *Intelligence, 20*(2), 191-210. https://doi.org/10.1016/0160-2896(95)90032-2

Haller, S., Falkovskiy, P., Meuli, R., Thiran, J. P., Krueger, G., Lovblad, K. O., Kober, T., Roche, A., & Marechal, B. (2016). Basic MR sequence parameters systematically bias automated brain volume estimation. *Neuroradiology, 58*(11), https://doi.org/1153-1160. 10.1007/s00234-016-1737-3

Hankin, R. K. S. (2006). Special functions in R: introducing the gsl package. *R News 6*(4).

Harrer, M., Cuijpers, P., Furukawa, T.A, & Ebert, D. D. (2019). Doing Meta-Analysis in R: A Hands-on Guide. https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/

*Harvey, I., Persaud, R., Ron, M. A., Baker, G., & Murray, R. M. (1994). Volumetric MRI measurements in bipolars compared with schizophrenics and healthy controls. *Psychological Medicine, 24*(3), 689-699. https://doi.org/10.1017/S0033291700027847

Heck, W. D, Gronau, F. Q, Wagenmakers &, E.-J. (2019). metaBMA: Bayesian Model Averaging for random and fixed effects meta-analysis. https://CRAN.R-project.org/package=metaBMA

Hedges, L. V. (2019). Stochastically dependent effect sizes. In: H. Cooper, L. V. Hedges, J. C. Valentine (eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 281-297). Russel Sage Foundation.

Hedges, L. V., Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*, 39–65. https://doi.org/10.1002/jrsm.5

Hedges, L. V., & Vevea, J. L. (2005). Selection method approaches. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (eds.) Publication bias in meta-analysis: prevention, assessment, and adjustments (pp. 145–174). Wiley.

Hedman, A. M., van Haren, N. E., Schnack, H. G., Kahn, R. S., & Hulshoff Pol, H. E. (2012). Human brain changes across the life span: a review of 56 longitudinal magnetic resonance imaging studies. *Human Brain Mapping, 33*(8), 1987-2002. https://doi.org/10.1002/hbm.21334

Hein, S., Reich, J., Thuma, P. E., & Grigorenko, E. L. (2014). Physical growth and nonverbal intelligence: Associations in Zambia. *The Journal of Pediatrics, 165*(5), 1017-1023. https://doi.org/10.1016/j.jpeds.2014.07.058

Heinz, A., Müller, D. J., Krach, S., Cabanis, M., & Kluge, U. P. (2014). The uncanny return of the race concept. *Frontiers in Human Neuroscience, 8*, 836. https://doi.org/10.3389/fnhum.2014.00836

Heller, L, & Pesmen, A. (2020, June 14). Über die rassistischen Wurzeln der Wissenschaft. Deutschlandfunk. https://www.deutschlandfunk.de/rassendenken-teil-1-ueber-die-rassistischen-wurzeln-von.740.de.html?dram:article_id=436585

*Hermann, B., Seidenberg, M., Bell, B., Rutecki, P., Sheth, R., Ruggles, K., Wendt, G., O´Leary, D., & Magnotta, V. (2002). The neurodevelopmental impact of childhood-onset temporal lobe epilepsy on brain structure and function. *Epilepsia, 43*(9), 1062-1071. https://doi.org/10.1046/j.1528-1157.2002.49901.x

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., Welch, V. A. (eds.) (2009). Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019). www.training.cochrane.org/handbook

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539-1558. https://doi.org/10.1002/sim.1186

Hill, W. D., Marioni, R. E., Maghzian, O., Ritchie, S. J., Hagenaars, S. P., McIntosh, A. M., Gale, C. R., Davies, G., & Deary, I. J. (2019). A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence. *Molecular Psychiatry, 24*(2), 169-181. https://doi.org/10.1038/s41380-017-0001-5

*Hiraiwa, A., Kawasaki, Y., Ibuki, K., Hirono, K., Matsui, M., Yoshimura, N., Origasa, H., Oishi, K., & Ichida, F. (2019, July). Brain development of children with single ventricle physiology or transposition of the great arteries: A longitudinal observation study. *Seminars in Thoracic and Cardiovascular Surgery.* WB Saunders. https://doi.org/10.1053/j.semtcvs.2019.06.013

*Hogan, M. J., Staff, R. T., Bunting, B. P., Murray, A. D., Ahearn, T. S., Deary, I. J., & Whalley, L. J. (2011). Cerebellar brain volume accounts for variance in cognitive performance in older adults. *Cortex, 47*(4), 441-450. https://doi.org/10.1016/j.cortex.2010.01.001

Hoogendam, Y. Y., Hofman, A., van der Geest, J. N., van der Lugt, A., & Ikram, M. A. (2014). Patterns of cognitive function in aging: the Rotterdam Study. *European Journal of Epidemiology, 29*(2), 133-140. https://doi.org/10.1007/s10654-014-9885-4

Hunt, E. (2010). *Human intelligence*. Cambridge University Press.

Hunter, J. E., & Schmidt, F. L. (2015). *Methods of meta-analysis*: *correcting error and bias in research findings* (3. ed.). Sage.

IntHout, J., Ioannidis, J. P., & Borm, G. F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology, 14*(1), 25. https://doi.org/10.1186/1471-2288-14-25

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials, 4*(3), 245-253. https://doi.org/10.1177%2F1740774507079441

Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Sciences, 18*(5), 235-241. https://doi.org/10.1016/j.tics.2014.02.010

*Isaacs, E. B., Fischl, B. R., Quinn, B. T., Chong, W. K., Gadian, D. G., & Lucas, A. (2010). Impact of breast milk on intelligence quotient, brain size, and white matter development. *Pediatric Research, 67*(4), 357-362. https://doi.org/10.1203/PDR.0b013e3181d026da

Ivanovic, D. M., Ibaceta, C. V., Correa, P. B., Orellana, Y. Z., Calderón, P. M., Morales, G. I., Leyton, B. D., Almagiá, A. F., Lizana, P. A., & Burrows, R. A. (2014). Brain development and scholastic achievement in the education quality measurement system tests in Chilean school-aged children. *Pediatric Research, 75*(3), 464-470. https://doi.org/10.1038/pr.2013.232

*Ivanovic, D. M., Leiva, B. P., Castro, C. G., Olivares, M. G., Jansana, J. M. M., Castro, V. G., Almagiá, A. A. F., Toro, T. D., Urrutia, M. S. C., Miller, P. T., Bosch, E. O., Larráin, C. G., & Pérez, H. T. (2004a). Brain development parameters and intelligence in Chilean high school graduates. *Intelligence, 32*(5), 461–479. https://doi.org/10.1016/j.intell.2004.07.001

*Ivanovic, D. M., Leiva, B. P., Pérez, H. T., Olivares, M. G., Dıaz, N. S., Urrutia, M. S. C., Almagiá, A. A. F., Toro, T. D., Miller, P. T., Bosch, E. O., & Larraın, C. G. (2004b). Head size and intelligence, learning, nutritional status and brain development: head, IQ, learning, nutrition and brain. *Neuropsychologia, 42*(8), 1118-1131. https://doi.org/10.1016/j.neuropsychologia.2003.11.022

Jäncke, L., Sele, S., Liem, F., Oschwald, J., & Merillat, S. (2020). Brain aging and psychometric intelligence: a longitudinal study. *Brain Structure and Function*, *225*(2), 519-536. https://doi.org/10.1007/s00429-019-02005-5

Jackson, D., Law, M., Rücker, G., & Schwarzer, G. (2017). The Hartung-Knapp modification for random-effects meta-analysis: A useful refinement but are there any residual concerns? *Statistics in Medicine, 36*(25), 3923-3934. https://doi.org/10.1002/sim.7411

Jansen, P. R., Nagel, M., Watanabe, K., Wei, Y., Savage, J. E., de Leeuw, C. A., van den Heuvel, M. P., van der Sluis, S., & Posthuma, D. (2019). GWAS of brain volume on 54,407 individuals and cross-trait analysis with intelligence identifies shared genomic loci and genes. *BioRxiv*, 613489. https://doi.org/10.1101/613489

Jaušovec, N. (2019). The neural code of intelligence: From correlation to causation. *Physics of Life Reviews, 31*, 171-187. https://doi.org/10.1016/j.plrev.2019.10.005

*Jenkins, J. V. M., Woolley, D. P., Hooper, S. R., & De Bellis, M. D. (2013). Direct and indirect effects of brain volume, socioeconomic status and family stress on child IQ. *Journal of Child and Adolescent Behavior, 1*(2). https://doi.org/10.4172/2375-4494.1000107

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Praeger.

Jensen, A. R., & Weng, L. J. (1994). What is a good g? *Intelligence, 18*(3), 231-258. https://doi.org/10.1016/0160-2896(94)90029-9

*Jensen, M. H., Bak, N., Rostrup, E., Nielsen, M. Ø., Pantelis, C., Glenthøj, B. Y., Ebdrup, B. H., & Fagerlund, B. (2019). The impact of schizophrenia and intelligence on the relationship between age and brain volume. *Schizophrenia Research: Cognition*, *15*, 1-6. https://doi.org/10.1016/j.scog.2018.09.002

Johnson, W., & Bouchard, T. J. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence, 33*(4), 393-416. https://doi.org/10.1016/j.intell.2004.12.002

*Jones, P. B., Harvey, I., Lewis, S. W., Toone, B. K., Van Os, J., Williams, M., & Murray, R. M. (1994). Cerebral ventricle dimensions as risk factors for schizophrenia and affective psychosis: an epidemiological approach to analysis. *Psychological Medicine, 24*(4), 995-1011. https://doi.org/10.1017/S0033291700029081

*Kareken, D. A., Gur, R. C., Mozley, P. D., Mozley, L. H., Saykin, A. J., Shtasel, D. L., & Gur, R. E. (1995). Cognitive functioning and neuroanatomic Vol. measures in schizophrenia. *Neuropsychology, 9*(2), 211–219. https://doi.org/10.1037/0894-4105.9.2.211

Katušić, A., Raguž, M., & Žunić Išasegi, I. (2020). Brain tissue volumes at term-equivalent age are associated with early motor behavior in very preterm infants. *International Journal of Developmental Neuroscience*, *80*(5), 409-417. https://doi.org/10.1002/jdn.10039

Kelley, T. L. (1923). *Statistical methods*. Macmillan.

*Kesler, S. R., Adams, H. F., Blasey, C. M., & Bigler, E. D. (2003). Premorbid intellectual functioning, education, and brain size in traumatic brain injury: an investigation of the cognitive reserve hypothesis. *Applied neuropsychology, 10*(3), 153-162. https://doi.org/10.1207/S15324826AN1003_04

*Kievit, R. A., Romeijn, J. W., Waldorp, L. J., Wicherts, J. M., Scholte, H. S., & Borsboom, D. (2011). Mind the gap: a psychometric approach to the reduction problem. *Psychological Inquiry, 22*(2), 67-87. https://doi.org/10.1080/1047840X.2011.550181

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine, 22*(17), 2693-2710. https://doi.org/10.1002/sim.1482

Kossmeier, M., Tran, U. S., & Voracek, M. (2020a). Metaviz: forest plots, funnel plots, and visual funnel plot inference for meta-analysis (version 0.3.1). https://cran.r-project.org/package=metaviz

Kossmeier, M., Tran, U. S., & Voracek, M. (2020b). Power-enhanced funnel plots for meta-analysis. *Zeitschrift für Psychologie, 228*, 43-49. https://doi.org/10.1027/2151-2604/a000392

*Kumra, S., Giedd, J. N., Vaituzis, A. C., Jacobsen, L. K., McKenna, K., Bedwell, J., Hamburger, S., Nelson, J. E., Lenane, M., & Rapoport, J. L. (2000). Childhood-onset psychotic disorders: magnetic resonance imaging of volumetric differences in brain structure. *American Journal of Psychiatry, 157*(9), 1467-1474. https://doi.org/10.1176/appi.ajp.157.9.1467

Kura, K., Armstrong, E. L., & Templer, D. I. (2014). Cognitive function among the Ainu people. *Intelligence*, *44*, 149-154. https://doi.org/10.1016/j.intell.2014.04.001

*Lammers, F., Borchers, F., Feinkohl, I., Hendrikse, J., Kant, I. M., Kozma, P., Pischon, T., Slooter, A. J. C., Spies, C., van Montfort, S. J. T., Zacharias, N., Zaborszky, L., Winterer, G., & The BigCog Consortium (2018). Basal forebrain cholinergic system volume is associated with general cognitive ability in the elderly. *Neuropsychologia, 119*, 145-156. https://doi.org/10.1016/j.neuropsychologia.2018.08.005

Langan, D., Higgins, J. P., Jackson, D., Bowden, J., Veroniki, A. A., Kontopantelis, E., Viechtbauer, W., & Simmonds, M. (2018). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods, 10*(1), 83-98. https://doi.org/10.1002/jrsm.1316

*Lange, N., Froimowitz, M. P., Bigler, E. D., Lainhart, J. E., & Brain Development Cooperative Group. (2010). Associations between IQ, total and regional brain volumes, and demography in a large normative sample of healthy children and adolescents. *Developmental Neuropsychology, 35*(3), 296-317. https://doi.org/10.1080/87565641003696833

*Lawson, J. A., Vogrin, S., Bleasel, A. F., Cook, M. J., & Bye, A. M. (2000). Cerebral and cerebellar volume reduction in children with intractable epilepsy. *Epilepsia, 41*(11), 1456-1462. https://doi.org/10.1111/j.1528-1157.2000.tb00122.x

Lee, J. J., McGue, M., Iacono, W. G., Michael, A. M., & Chabris, C. F. (2019). The causal influence of brain size on human intelligence: Evidence from within-family phenotypic associations and GWAS modeling. *Intelligence, 75*, 48-58. https://doi.org/10.1016/j.intell.2019.01.011

Lenhard, W. & Lenhard, A. (2014). Signifikanztests bei Korrelationen. https://www.psychometrica.de/korrelation.html. Psychometrica. https://doi.org/10.13140/RG.2.1.2954.1367

*Leonard, C. M., Kuldau, J. M., Breier, J. I., Zuffante, P. A., Gautier, E. R., Heron, D. C., Lavery, E. M., Packing, J., Williams, S. A., & DeBose, C. A. (1999). Cumulative effect of anatomical risk factors for schizophrenia: an MRI study. *Biological Psychiatry, 46*(3), 374-382. https://doi.org/10.1016/S0006-3223(99)00052-9

Lerch, J. P., van der Kouwe, A. J., Raznahan, A., Paus, T., Johansen-Berg, H., Miller, K. L., Smith, S. M., Fischl, B., & Sotiropoulos, S. N. (2017). Studying neuroanatomy using MRI. *Nature Neuroscience, 20*(3), 314-326. https://doi.org/10.1038/nn.4501

*Lefebvre, A., Beggiato, A., Bourgeron, T., & Toro, R. (2015). Neuroanatomical diversity of corpus callosum and brain volume in autism: meta-analysis, analysis of the autism brain imaging data exchange project, and simulation. *Biological Psychiatry*, *78*(2), 126-134. https://doi.org/10.1016/j.biopsych.2015.02.010

Lin, Y. (2016). *Brain damage in chronic ketamine users: A multi-modal imaging study* (Publication Number 10632420) [doctoral dissertation, the Chinese University of Hong Kong]. ProQuest Dissertation & Thesis Global.

Liu, S., Seidlitz, J., Blumenthal, J. D., Clasen, L. S., & Raznahan, A. (2020). Integrative structural, functional, and transcriptomic analyses of sex-biased brain organization in humans. *Proceedings of the National Academy of Sciences, 117*(31), 18788-18798. https://doi.org/10.1073/pnas.1919091117

*Lodygensky, G. A., Rademaker, K., Zimine, S., Gex-Fabry, M., Lieftink, A. F., Lazeyras, F., Groenendaal, F., de Vries, L. S., & Huppi, P. S. (2005). Structural and functional brain development after hydrocortisone treatment for neonatal chronic lung disease. *Pediatrics, 116*(1), 1-7. https://doi.org/10.1542/peds.2004-1275

*Luders, E., Narr, K. L., Bilder, R. M., Thompson, P. M., Szeszko, P. R., Hamilton, L., & Toga, A. W. (2007). Positive correlations between corpus callosum thickness and intelligence. *Neuroimage, 37*(4), 1457-1464. https://doi.org/10.1016/j.neuroimage.2007.06.028

Lyden, H. (2015). Family aggression exposure and community violence exposure associated with brain volume in late adolescence: A comparison of automated versus manual segmentation (Publication No. 10799680) [Master´s thesis, University of Southern California]. ProQuest Dissertations and Theses Global.

Lynn, R. (1991). The evolution of racial differences in intelligence. *Mankind Quarterly, 32*(1), 109.

Lynn, R. (1994). Sex differences in brain size and intelligence: A paradox resolved. *Personality and Individual Differences, 17*(2), 257-271.

Lynn, R. (2017). Sex differences in intelligence: The developmental theory. *Mankind Quarterly, 58*(1), 9-42. https://doi.org/10.46469/mq.2017.58.1.2

*MacDonald, P. A., Ganjavi, H., Collins, D. L., Evans, A. C., & Karama, S. (2014). Investigating the relation between striatal volume and IQ. *Brain Imaging and Behavior, 8*(1), 52-59. https://doi.org/10.1007/s11682-013-9242-3

Maclaren, J., Han, Z., Vos, S. B., Fischbein, N., & Bammer, R. (2014). Reliability of brain volume measurements: a test-retest dataset. *Scientific data, 1*(1), 1-9. https://doi.org/10.1038/sdata.2014.37

*MacLullich, A. M. J., Ferguson, K. L., Deary, I. J., Seckl, J. R., Starr, J. M., & Wardlaw, J. M. (2002). Intracranial capacity and brain volumes are associated with cognition in elderly men. *Neurology, 59*(2), 169–174. https://doi.org/10.1212/WNL.59.2.169

Madan, C. R., & Kensinger, E. A. (2017). Test–retest reliability of brain morphology estimates. *Brain Informatics, 4*(2), 107-121. https://doi.org/10.1007/s40708-016-0060-4

*Mankovsky, B., Zherdova, N., van den Berg, E., Biessels, G. J., & de Bresser, J. (2018). Cognitive functioning and structural brain abnormalities in people with Type 2 diabetes mellitus. *Diabetic Medicine*, *35*(12), 1663-1670. https://doi.org/10.1111/dme.13800

*Martínez, K., Janssen, J., Pineda-Pardo, J. Á., Carmona, S., Román, F. J., Alemán-Gómez, Y., Garcia-Garcia, D., Escorial, S., Quiroga, M. A., Santarnecci, E., Navas-Sánchez, F. J., Desco, M., Arrango, C., & Colom, R. (2017). Individual differences in the dominance of interhemispheric connections predict cognitive ability beyond sex and brain size. *NeuroImage, 155*, 234-244. https://doi.org/10.1016/j.neuroimage.2017.04.029

*Mathias, S. R., Knowles, E. E., Mollon, J., Rodrigue, A., Koenis, M. M., Alexander-Bloch, A. F., Winkler, A. M., Olvera, R. L., Duggirala, R., Göring, H. H. H., Curran, J. E., Fox, P. T., Almasy, L., Blangero, J., & Glahn, D. C. (2020). Minimal relationship between local gyrification and general cognitive ability in humans. *Cerebral Cortex, 30*(6), 3439-3450. https://doi.org/10.1093/cercor/bhz319

Mathiesen, N. C. (2015). Multimodal brain mapping of general intelligence [Master's thesis, University of Oslo]. DUO Vitenarkiv. https://www.duo.uio.no/bitstream/handle/10852/48654/51/Mathiesen_Thesis.pdf

Mathur, M. B., & VanderWeele, T. (2019, December 18). Estimating publication bias in meta-analyses of peer-reviewed studies: A meta-meta-analysis across disciplines and journal tiers. https://doi.org/10.31219/osf.io/p3xyd

*McCoy, T. E., Conrad, A. L., Richman, L. C., Brumbaugh, J. E., Magnotta, V. A., Bell, E. F., & Nopoulos, P. C. (2014). The relationship between brain structure and cognition in transfused preterm children at school age. *Developmental Neuropsychology, 39*(3), 226-232. https://doi.org/10.1080/87565641.2013.874428

McDaniel, M. A (2005). Big-brained people are smarter: A meta-analysis of the relationship between in vivo brain volume and intelligence. *Intelligence, 33*, 337-346. https://doi.org/10.1016/j.intell.2004.11.005

McGuire, S. A., Wijtenburg, S. A., Sherman, P. M., Rowland, L. M., Ryan, M., Sladky, J. H., & Kochunov, P. V. (2017). Reproducibility of quantitative structural and physiological MRI measurements. *Brain and behavior, 7*(9), e00759. https://doi.org/10.1002/brb3.759

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*(5), 730-749. https://doi.org/10.1177/1745691616662243

Miller, G. F., & Penke, L. (2007). The evolution of human intelligence and the coefficient of additive genetic variance in human brain size. *Intelligence, 35*(2), 97-114.

*Miller, J. L., Couch, J., Schwenk, K., Long, M., Towler, S., Theriaque, D. W., He, G., Liu, Y., Driscoll, D. J., & Leonard, C. M. (2009). Early childhood obesity is associated with compromised cerebellar development. *Developmental Neuropsychology, 34*(3), 272-283. https://doi.org/10.1080/87565640802530961

*Mitchell, B. L., Cuéllar-Partida, G., Grasby, K. L., Campos, A. I., Strike, L. T., Hwang, L. D., Okbay, A., Thompson, P. M., Medland, S. E., Martin, N. G., Wright, M. J., & Rentería, M. E. (2020). Educational attainment polygenic scores are associated with cortical total surface area and regions important for language and memory. *NeuroImage, 116691*. https://doi.org/10.1016/j.neuroimage.2020.116691

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS med, 6(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

*Monson, B. B., Anderson, P. J., Matthews, L. G., Neil, J. J., Kapur, K., Cheong, J. L., Doyle, L. W., Thompson, T. K., & Inder, T. E. (2016). Examination of the pattern of growth of cerebral tissue volumes from hospital discharge to early childhood in very preterm infants. *JAMA Pediatrics, 170*(8), 772-779. https://doi.org/10.1001/jamapediatrics.2016.0781

*Mori, E., Hirono, N., Yamashita, H., Imamura, T., Ikejiri, Y., Ikeda, M., Kitagaki, H., Shimomura, T., & Yoneda, Y. (1997). Premorbid brain size as a determinant of reserve capacity against intellectual decline in Alzheimer's disease. *American Journal of Psychiatry, 154*(1), 18-24.

Morton, S. G. (1849). Catalogue of skulls of man and inferior animals. Merrihew & Thompson.

Mueller, K. F., Meerpohl, J. J., Briel, M., Antes, G., von Elm, E., Lang, B., Motschall, E., Schwarzer, G., & Bassler, D. (2016). Methods for detecting, quantifying, and adjusting for dissemination bias in meta-analysis are described. *Journal of Clinical Epidemiology*, *80*, 25-33. https://doi.org/10.1016/j.jclinepi.2016.04.015

Murdoch, K., & Sullivan, L. R. (1923). A contribution to the study of mental and physical measurements in normal children. *American Physical Education Review, 28*(5), 209-215. https://doi.org/10.1080/23267224.1923.10651771

*Nakamura, M., Nestor, P. G., McCarley, R. W., Levitt, J. J., Hsu, L., Kawashima, T., Niznikiewicz, M., & Shenton, M. E. (2007). Altered orbitofrontal sulcogyral pattern in schizophrenia. *Brain, 130*(3), 693-707. https://doi.org/10.1093/brain/awm007

*Narr, K. L., Woods, R. P., Thompson, P. M., Szeszko, P., Robinson, D., Dimtcheva, T., Gurbani, M., Toga, A. W., & Bilder, R. M. (2007). Relationships between IQ and regional cortical gray matter thickness in healthy adults. *Cerebral cortex, 17*(9), 2163-2171. https://doi.org/10.1093/cercor/bhl125

Nelson, H.E. (1982). Nelson adult reading rest manual. London: The National Hospital for Nervous Diseases.

*Nikolaidis, A., Baniqued, P. L., Kranz, M. B., Scavuzzo, C. J., Barbey, A. K., Kramer, A. F., & Larsen, R. J. (2017). Multivariate associations of fluid intelligence and NAA. *Cerebral cortex, 27*(4), 2607-2616. https://doi.org/10.1093/cercor/bhw070

*Nosarti, C., Mazin, H. S., Al-Asady1, S. F., Stewart1, A. L., Rifkin, L., & Murray, R. M. (2002). Adolescents who were born very preterm have decreased brain volumes. *Brain, 125*(7), 1616–1623. https://doi.org/10.1093/brain/awf157

Nuijten, M. B., Van Assen, M. A. L. M., Augusteijn, H. E. M., Crompvoets, E. A. V., & Wicherts, J. M. (2019). Effect sizes, power, and biases in intelligence research: a meta-meta-analysis. Preprint retrieved from https://psyarxiv.com/ytsvw.

Nyborg, H. (2005). Sex-related differences in general intelligence g, brain size, and social status. *Personality and Individual Differences, 39*(3), 497-509. https://doi.org/10.1016/j.paid.2004.12.011

*Nygaard, E., Slinning, K., Moe, V., Due-Tønnessen, P., Fjell, A., & Walhovd, K. B. (2018). Neuroanatomical characteristics of youths with prenatal opioid and poly-drug exposure. *Neurotoxicology and Teratology, 68*, 13-26. https://doi.org/10.1016/j.ntt.2018.04.004

Olkin, I., Dahabreh, I. J., & Trikalinos, T. A. (2012). GOSH–a graphical display of study heterogeneity. *Research Synthesis Methods, 3*(3), 214-223. https://doi.org/10.1002/jrsm.1053

Olkin, I. & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *The Annals of Mathematical Statistics*, *29*(1), 201-211.

Oschwald, J., Guye, S., Liem, F., Rast, P., Willis, S., Röcke, C., Jäncke, L., Martin, M., & Mérillat, S. (2019). Brain structure and cognitive ability in healthy aging: a review on longitudinal correlated change. *Reviews in the Neurosciences, 31*(1), 1-57. https://doi.org/10.1515/revneuro-2018-0096

*Paradiso, S., Andreasen, N. C., O'Leary, D. S., Arndt, S., & Robinson, R. G. (1997). Cerebellar size and cognition: correlations with IQ, verbal memory and motor dexterity. *Cognitive and Behavioral Neurology, 10*(1), 1-8.

Paterson, D. G. (1930). *Physique and intellect.* Century/Random House UK.

*Paul, E. J., Larsen, R. J., Nikolaidis, A., Ward, N., Hillman, C. H., Cohen, N. J., Kramer, A. F., Barbey, A. K., & Barbey, A. K. (2016). Dissociable brain biomarkers of fluid intelligence. *NeuroImage, 137*, 201-211. https://doi.org/10.1016/j.neuroimage.2016.05.037

*Pennington, B. F., Filipek, P. A., Lefly, D., Chhabildas, N., Kennedy, D. N., Simon, J. H., Filley, C. M., Galaburda, A., & DeFries, J. C. (2000). A twin MRI study of size variations in the human brain. *Journal of Cognitive Neuroscience, 12*(1), 223–232. https://doi.org/10.1162/089892900561850

Petersson, S., Pedersen, N. L., Schalling, M., & Lavebratt, C. (1999). Primary megalencephaly at birth and low intelligence level. *Neurology, 53*(6), 1254-1254. https://doi.org/10.1212/WNL.53.6.1254

Pietschnig, J., Penke, L., Wicherts, J. M., Zeiler, M., & Voracek, M. (2015). Meta-analysis of associations between human brain volume and intelligence differences: How strong are they and what do they mean? *Neuroscience and Biobehavioral Reviews, 57*, 411-432. https://doi.org/10.1016/j.neubiorev.2015.09.017

Pietschnig, J., Siegel, M., Eder, J. S. N., & Gittler, G. (2019). Effect declines are systematic, strong and ubiquitous: a meta-meta-analysis of the decline effect in intelligence research. *Front. Psychol. 10:2874*. https://doi.org/10.3389/fpsyg.2019.02874

Pietschnig, J., Voracek, M., & Formann, A. K. (2010). Mozart effect—Shmozart effect: a meta-analysis. *Intelligence 38*, 314–323. https://doi.org/10.1016/j.intell.2010.03.001

Pigott, T.D. (2009). Handling missing data. In: H. M. Cooper, L.V. Hedges, J.C. Valentine (eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 399-416). Sage.

Pigott, T. D., & Polanin, J. R. (2020). Methodological guidance paper: high-quality meta-analysis in a systematic review. *Review of Educational Research, 90*(1), 24-46. https://doi.org/10.3102%2F0034654319877153

Protzko, J., & Schooler, J. W. (2017). Decline effects: Types, mechanisms, and personal reflections. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (p. 85–107). Wiley-Blackwell. https://doi.org/10.1002/9781119095910.ch6

Pruimboom, L., Raison, C. L., & Muskiet, F. A. (2015). Physical activity protects the human brain against metabolic stress induced by a postprandial and chronic inflammation. *Behavioural Neurology*, *2015*. https://doi.org/10.1155/2015/569869

*Qiu, A., Crocetti, D., Adler, M., Mahone, E. M., Denckla, M. B., Miller, M. I., & Mostofsky, S. H. (2009). Basal ganglia volume and shape in children with attention deficit hyperactivity disorder. *American Journal of Psychiatry, 166*(1), 74-82. https://doi.org/10.1176/appi.ajp.2008.08030426

*Raz, N., Lindenberger, U., Ghisletta, P., Rodrigue, K. M., Kennedy, K. M., & Acker, J. D. (2008). Neuroanatomical correlates of fluid intelligence in healthy adults and persons with vascular risk factors. *Cerebral cortex, 18*(3), 718-726. https://doi.org/10.1093/cercor/bhm108

*Raz, N., Torres, I. J., Briggs, S. D., Spencer, W. D., Thornton, A. E., Loken, W. J., Gunning, F. M., McQuain, D., Driesen, N. R., & Acker, J. D. (1995). Selective neuroanatornic abnormalities in Down's syndrome and their cognitive correlates: Evidence from MRI morphometry. *Neurology, 45*(2), 356-366. https://doi.org/10.1212/WNL.45.2.356

*Raz, N., Torres, I. J., Spencer, W. D., Millman, D., Baertschi, J. C., & Sarpel, G. (1993). Neuroanatomical correlates of age-sensitive and age-invariant cognitive abilities: An in vivo MRI investigation. *Intelligence, 17*(3), 407–422. https://doi.org/10.1016/0160-2896(93)90008-S

Reid, M. W., Hannemann, N. P., York, G. E., Ritter, J. L., Kini, J. A., Lewis, J. D., Sherman, P. M., Velez, C. S., Drennon, A. M., Bolzenius, J. D., & Tate, D. F. (2017). Comparing two processing pipelines to measure subcortical and cortical volumes in patients with and without mild traumatic brain injury. *Journal of Neuroimaging, 27*(4), 365-371. https://doi.org/10.1111/jon.12431

*Reiss, A. L., Abrams, M. T., Greenlaw, R., Freund, L., & Denckla, M. B. (1995). Neurodevelopmental effects of the FMR-1 full mutation in humans. *Nature medicine, 1*(2), 159-167. https://doi.org/10.1038/nm0295-159

*Reiss, A. L., Abrams, M. T., Singer, H. S., Ross, J. L., & Denckla, M. B. (1996). Brain development, gender and IQ in children. *Brain, 119*(5), 1763–1774. https://doi.org/10.1093/brain/119.5.1763

Reuter-Lorenz, P. A., & Park, D. C. (2014). How does it STAC up? Revisiting the scaffolding theory of aging and cognition. *Neuropsychology Review*, *24*(3), 355-370. https://doi.org/10.1007/s11065-014-9270-9

Ritchie, S. J., Booth, T., Hernández, M. D. C. V., Corley, J., Maniega, S. M., Gow, A. J., Royle, N. A., Pattie, A., Karama, S., Starr, J. M., Bastin, M. E, Wardlaw, J. M., & Deary, I. J. (2015). Beyond a bigger brain: Multivariable structural brain imaging and intelligence. *Intelligence, 51*, 47-56. https://doi.org/10.1016/j.intell.2015.05.001

*Ritchie, S. J., Dickie, D. A., Cox, S. R., Hernández, M. D. C. V., Sibbett, R., Pattie, A., Anblagan, D., Redmond, P., Royle, N. A., Corley, J., Maniega, S. M., Taylor, A. M., Karama, S., Booth, T., Gow, A. J., Starr, J. M., Bastin, M. E., Wardlaw, J. M., & Deary, I. J. (2018). Brain structural differences between 73-and 92-year olds matched for childhood intelligence, social background, and intracranial volume. *Neurobiology of Aging*, 62, 146-158. https://doi.org/10.1016/j.neurobiolaging.2017.10.005

Röver, C. (2020). Bayesian random-effects meta-analysis using the bayesmeta R package. *Journal of Statistical Software, 93*(6), 1-51. https://doi.org/10.18637/jss.v093.i06

*Rojas, D. C., Peterson, E., Winterrowd, E., Reite, M. L., Rogers, S. J., & Tregellas, J. R. (2006). Regional gray matter volumetric changes in autism associated with social and repetitive behavior symptoms. *BMC Psychiatry, 6*(1), 56. https://doi.org/10.1186/1471-244X-6-56

*Rojas, D. C., Smith, J. A., Benkers, T. L., Camou, S. L., Reite, M. L., & Rogers, S. J. (2004). Hippocampus and amygdala volumes in parents of children with autistic disorder. *American Journal of Psychiatry, 161*(11), 2038-2044.

Roth, G., & Dicke, U. (2005). Evolution of the brain and intelligence. *Trends in Cognitive Sciences, 9*(5), 250-257. https://doi.org/10.1016/j.tics.2005.03.005

*Royle, N. A., Booth, T., Hernández, M. C. V., Penke, L., Murray, C., Gow, A. J., Munoz Maniega, S. M., Starr, J., Bastin, M. E., Deary, I. J., & Wardlaw, J. M. (2013). Estimated maximal and current brain volume predict cognitive ability in old age. *Neurobiology of Aging, 34*(12), 2726-2733. https://doi.org/10.1016/j.neurobiolaging.2013.05.015

Ruigrok, A. N., Salimi-Khorshidi, G., Lai, M. C., Baron-Cohen, S., Lombardo, M. V., Tait, R. J., & Suckling, J. (2014). A meta-analysis of sex differences in human brain structure. *Neuroscience & Biobehavioral Reviews, 39*, 34-50. https://doi.org/10.1016/j.neubiorev.2013.12.004

Rushton, J. P., & Ankney, C. D. (1996). Brain size and cognitive ability: Correlations with age, sex, social class, and race. *Psychonomic Bulletin & Review, 3*(1), 21-36. https://doi.org/10.3758/BF03210739

Rushton, J. P., & Ankney, C. D. (2007). The evolution of brain size and intelligence. In S. M. Platek, J. P. Keenan, & T. K. Shackelford (eds.), *Evolutionary cognitive neuroscience* (pp. 121-161). MIT Press.

Rushton, J. P., & Ankney, C. D. (2009). Whole brain size and general mental ability: a review. *International Journal of Neuroscience, 119*(5), 692-732. https://doi.org/10.1080/00207450802325843

Sahin, B. (2012). Anthropometry of the intracranial volume. In: V. R Preedy (ed.), *Handbook of anthropometry* (pp. 517-530). Springer.

Saniotis, A., Grantham, J. P., Kumaratilake, J., & Henneberg, M. (2020). Neuro-hormonal regulation is a better indicator of human cognitive abilities than brain anatomy: The need for a new paradigm. *Frontiers in Neuroanatomy*, *13*, 101. https://doi.org/10.3389/fnana.2019.00101

Schild, A. H., & Voracek, M. (2015). Finding your way out of the forest without a trail of bread crumbs: development and evaluation of two novel displays of forest plots. *Research Synthesis Methods, 6*(1), 74-86. https://doi.org/10.1002/jrsm.1125

Schimmack, U. (2016). The replicability-index: Quantifying statistical research integrity. https://replicationindex.wordpress.com/2016/01/31/a-revised-introduction-to-the-r-index/

Schmidt, F. L. (2015). History and development of the Schmidt–Hunter meta-analysis methods. *Research Synthesis Methods, 6*(3), 232-239. https://doi.org/10.1002/jrsm.1134

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll Model of Intelligence. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 99–144). Guilford Press.

*Schoenemann, P. T., Budinger, T. F., Sarich, V. M., & Wang, W. S. Y. (2000). Brain size does not predict general cognitive ability within families. *Proceedings of the National Academy of Sciences, 97*(9), 4932–4937. https://doi.org/10.1073/pnas.97.9.4932

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature, 470*, 437. https://doi.org/10.1038/470437a

*Schottenbauer, M. A., Momenan, R., Kerick, M., & Hommer, D. W. (2007). Relationships among aging, IQ, and intracranial Vol. in alcoholics and control subjects. *Neuropsychology, 21*(3), 337–345. https://doi.org/10.1037/0894-4105.21.3.337

*Schumann, C. M., Hamstra, J., Goodlin-Jones, B. L., Kwon, H., Reiss, A. L., & Amaral, D. G. (2007). Hippocampal size positively correlates with verbal IQ in male children. *Hippocampus, 17*(6), 486-493. https://doi.org/10.1002/hipo.20282

Schwarzer, G., Carpenter, J. R., Rücker, G. (2020). metasens: Advanced statistical methods to model and adjust for bias in meta-analysis. R package version 0.5-0. https://CRAN.R-project.org/package=metasens

Sereno, M. I., Diedrichsen, J., Tachrount, M., Testa-Silva, G., d'Arceuil, H., & De Zeeuw, C. (2020). The human cerebellum has almost 80% of the surface area of the neocortex. *Proceedings of the National Academy of Sciences, 117*(32), 19538-19543. https://doi.org/10.1073/pnas.2002896117

*Shapleske, J., Rossell, S. L., Chitnis, X. A., Suckling, J., Simmons, A., Bullmore, E. T., Woodruff, P. W. R., & David, A. S. (2002). A computational morphometric MRI study pf schizophrenia: Effects of hallucinations. *Cerebral Cortex, 12*(12), 1331–1341. https://doi.org/10.1093/cercor/12.12.1331

*Shenkin, S. D., Rivers, C. S., Deary, I. J., Starr, J. M., & Wardlaw, J. M. (2009). Maximum (prior) brain size, not atrophy, correlates with cognition in community-dwelling older people: a cross-sectional neuroimaging study. *BMC Geriatrics, 9*(1), 12. https://doi.org/10.1186/1471-2318-9-12

Shinohara, R. T., Oh, J., Nair, G., Calabresi, P. A., Davatzikos, C., Doshi, J., Henry, R. G., Kim, G., Linn, K. A., Papinutto, N., Pelletier, D., Pham, D. L., Reich, D. S., Rooney, W., Roy, S., Stern, W., Tummala, S., Yousuf, F., Zhu, A., Sicotte, N. L., Bakshi, R., & the NAIMS Cooperative (2017). Volumetric analysis from a harmonized multisite brain MRI study of a single subject with multiple sclerosis. *American Journal of Neuroradiology, 38*(8), 1501-1509. https://doi.org/10.3174/ajnr.A5254

Sidik, K., & Jonkman, J. N. (2002). A simple confidence interval for meta-analysis. *Statistics in Medicine, 21*(21), 3153-3159. https://doi.org/10.1002/sim.1262

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366. https://doi.org/10.1177%2F0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology, 143*(2), 534–547. https://doi.org/10.1037/a0033242

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications. SSRN Scholarly Paper No. ID 2694998. Social Science Network. http://dx.doi.org/10.2139/ssrn.2694998

*Sreedharan, R. M., Sheelakumari, R., Anila, K. M., Kesavadas, C., & Thomas, S. V. (2018). Reduced brain volumes in children of women with epilepsy: A neuropsychological and voxel based morphometric analysis in pre-adolescent children. *Journal of Neuroradiology, 45*(6), 380-385. https://doi.org/10.1016/j.neurad.2018.02.005

*Staff, R. T., Murray, A. D., Deary, I. J., & Whalley, L. J. (2006). Generality and specificity in cognitive aging: a volumetric brain analysis. *NeuroImage, 30*(4), 1433-1440. https://doi.org/10.1016/j.neuroimage.2005.11.004

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702-712. https://doi.org/10.1177%2F1745691616658637

Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (eds.) *Publication bias in meta-analysis: prevention, assessment, and adjustments* (pp. 99–110). Wiley.

Strand, S., Deary, I. J., & Smith, P. (2006). Sex differences in cognitive ability test score: a UK national picture. *British Journal of Educational Psychology, 76*(3), 463–80. https://doi.org/10.1348/000709905X50906

*Takeuchi, H., Taki, Y., Nouchi, R., Yokoyama, R., Kotozaki, Y., Nakagawa, S., Sekiguchi, A., Izuka, K., Yamamoto, Y., Hanawa, S., Araki, T., Miyauchi, C. M., Shinada, T., Sakaki, K., Sassa, Y., Nozawa, T., Ikeda, S., Yokota, S., Daniele, M., & Kawashima, R. (2018). Refractive error is associated with intracranial volume. *Scientific Reports, 8*(1), 1-11. https://doi.org/10.1038/s41598-017-18669-0

*Tan, U., Tan, M., Polat, P., Ceylan, Y., Suma, S., & Okur, A. (1999). Magnetic resonance imaging brain size/IQ relations in Turkish university students. *Intelligence, 27*(1), 83–92. https://doi.org/10.1016/S0160-2896(99)00015-X

Tan, Y., Zhang, Q., Li, W., Wei, D., Qiao, L., Qiu, J., Hitchman, G., & Liu, Y. (2014). The correlation between emotional intelligence and gray matter volume in university students. *Brain and Cognition, 91*, 100-107. http://dx.doi.org/10.1016/j.bandc.2014.08.007

Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling complex meta-analytic data structures using robust variance estimation: a tutorial in R. *J Dev Life Course Criminology, 5*, 614–615. https://doi.org/10.1007/s40865-019-00127-2

*Tate, D. F., Neeley, E. S., Norton, M. C., Tschanz, J. T., Miller, M. J., Wolfson, L., Hulette, C., Leslie, C., Welsh-Bohmer, K. A., Plassman, B., & Bigler, E. D. (2011). Intracranial volume and dementia: some evidence in support of the cerebral reserve hypothesis. *Brain Research, 1385*, 151-162. https://doi.org/10.1016/j.brainres.2010.12.038

Tiedemann, F. (1836). XXIII. On the brain of the negro, compared with that of the European and the orang-outang. *Philosophical Transactions of the Royal Society of London*, (126), 497-527. https://doi.org/10.1098/rstl.1836.0025

*Thoma, R. J., Yeo, R. A., Gangestad, S. W., Halgren, E., Sanchez, N. M., & Lewine, J. D. (2005). Cortical volume and developmental instability are independent predictors of general intellectual ability. Intelligence, 33(1), 27-38. https://doi.org/10.1016/j.intell.2004.08.004

Thorndike, R. L. (1949). *Personnel selection: test and measurement techniques*. Wiley.

*Toulopoulou, T., Grech, A., Morris, R. G., Schulze, K., McDonald, C., Chapple, B., Rabe-Hesketh, S., & Murray, R. M. (2004). The relationship between volumetric brain changes and cognitive function: a family study on schizophrenia. *Biological Psychiatry, 56*(6), 447-453. https://doi.org/10.1016/j.biopsych.2004.06.026

*Tozer, D. J., Zeestraten, E., Lawrence, A. J., Barrick, T. R., & Markus, H. S. (2018). Texture analysis of T1-weighted and fluid-attenuated inversion recovery images detects abnormalities that correlate with cognitive decline in small vessel disease. *Stroke, 49*(7), 1656-1661. https://doi.org/10.1161/STROKEAHA.117.019970

*Treit, S., Zhou, D., Chudley, A. E., Andrew, G., Rasmussen, C., Nikkel, S. M., Samdup, D., Hanlon-Dearman, A., Loock, C., & Beaulieu, C. (2016). Relationships between head circumference, brain volume and cognition in children with prenatal alcohol exposure. *PloS One, 11*(2), e0150370. https://doi.org/10.1371/journal.pone.0150370

van Aert, R. C. M. (2020). puniform: meta-analysis methods correcting for publication bias. R package version 0.2.2. https://CRAN.R-project.org/package=puniform

van Aert, R. C. M., & van Assen, M. A. L. M. (2018, October 2). P-uniform*. https://doi.org/10.31222/osf.io/zqjr9

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science, 11*(5), 713-729. https://doi.org/10.1177/1745691616650874

van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods, 20*(3), 293–309. https://doi.org/10.1037/met0000025

*van der Linden, D., Dunkel, C. S., & Madison, G. (2017). Sex differences in brain size and general intelligence (g). *Intelligence*, *63*, 78-88. https://doi.org/10.1016/j.intell.2017.04.007

*van der Vlugt, J. J., Van Der Meulen, J. J., van den Braak, R. R. C., Vermeij-Keers, C., Horstman, E. G., Hovius, S. E., Verhulst, F. C., Wierdsma, A. I., Lequin, M., H., & Okkerse, J. M. (2017). Insight into the pathophysiologic mechanisms behind cognitive dysfunction in trigonocephaly. *Plastic and Reconstructive Surgery*, *139*(4), 954e-964e. https://doi.org/10.1097/PRS.0000000000003179

*van Haren, N. E. M., Setiaman, N., Koevoets, M. G. J. C., Baalbergen, H., Kahn, R. S., & Hillegers, M. H. J. (2020). Brain structure, IQ, and psychopathology in young offspring of patients with schizophrenia or bipolar disorder. *European Psychiatry*, *63*(1). https://doi.org/10.1192/j.eurpsy.2019.19

*van Leeuwen, M., Peper, J. S., van den Berg, S. M., Brouwer, R. M., Hulshoff Pol, H. E., Kahn, R. S., & Boomsma, D. I. (2009). A genetic analysis of brain volumes and IQ in children. *Intelligence, 37*(2), 181-191. https://doi.org/10.1016/j.intell.2008.10.005

van Valen, L. (1974). Brain size and intelligence in man. *American Journal of Physical Anthropology, 40*(3), 417-423. https://doi.org/10.1002/ajpa.1330400314

Van Zijl, P., & Knutsson, L. (2019). In vivo magnetic resonance imaging and spectroscopy. Technological advances and opportunities for applications continue to abound. *Journal of Magnetic Resonance*, 306, 55-65. https://doi.org/10.1016/j.jmr.2019.07.034

Vein, A. A., & Maat-Schieman, M. L. (2008). Famous Russian brains: historical attempts to understand intelligence. *Brain, 131*(2), 583-590. https://doi.org/10.1093/brain/awm326

Vernon, P. A., Wickett, J. C., Bazana, P. G., & Stelmack, R. M. (2000). The neuropsychology and psychophysiology of human intelligence. In: R. J. Sternberg, *Handbook of intelligence* (pp. 245-266). Cambridge University Press.

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. PT., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods, 7*(1), 55-79. https://doi.org/10.1002/jrsm.1164

Vevea, J. L., Coburn, K., & Sutton, A. (2019). Publication bias. In: H. Cooper, L. V. Hedges, J. C. Valentine (eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 383-430). Russel Sage Foundation.

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrica, 60*(3), 419-35. https://doi.org/10.1007/BF02294384

Viechtbauer W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics, 30*, 261–293. https://doi.org/10.3102/10769986030003261

Viechtbauer W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://www.jstatsoft.org/v36/i03/

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*(2), 112–125. https://doi.org/10.1002/jrsm.11

*Voelbel, G. T., Bates, M. E., Buckman, J. F., Pandina, G., & Hendren, R. L. (2006). Caudate nucleus volume and cognitive performance: are they related in childhood psychopathology? *Biological Psychiatry, 60*(9), 942-950. https://doi.org/10.1016/j.biopsych.2006.03.071

Voracek, M., Kossmeier, M., & Tran, U. S. (2019). Which data to meta-analyze, and how? A specification-curve and multiverse-analysis approach to meta-analysis. *Zeitschrift für Psychologie, 227(1)*, 64-82. https://doi.org/10.1027/2151-2604/a000357

*Vreeker, A., Abramovic, L., Boks, M. P., Verkooijen, S., van Bergen, A. H., Ophoff, R. A., Kahn, R. S., & van Haren, N. E. M. (2017). The relationship between brain volumes and intelligence in bipolar disorder. *Journal of Affective Disorders*, *223*, 59-64. https://doi.org/10.1016/j.jad.2017.07.009

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*(1), 35-57. https://doi.org/10.3758/s13423-017-1343-3

*Wallace, G. L., Lee, N. R., Prom-Wormley, E. C., Medland, S. E., Lenroot, R. K., Clasen, L. S., Schmitt, J. E., Neale, M. C., & Giedd, J. N. (2010). A bivariate twin study of regional brain volumes and verbal and nonverbal intellectual skills during childhood and adolescence. *Behavior Genetics*, *40*(2), 125-134. https://doi.org/10.1007/s10519-009-9329-1

*Walters, G. D., & Kiehl, K. A. (2015). Limbic correlates of fearlessness and disinhibition in incarcerated youth: Exploring the brain–behavior relationship with the Hare Psychopathy Checklist: Youth Version. *Psychiatry Research*, *230*(2), 205-210. https://doi.org/10.1016/j.psychres.2015.08.041

*Waiter, G. D., Williams, J. H., Murray, A. D., Gilchrist, A., Perrett, D. I., & Whiten, A. (2004). A voxel-based investigation of brain structure in male adolescents with autistic spectrum disorder. *Neuroimage*, *22*(2), 619-625. https://doi.org/10.1016/j.neuroimage.2004.02.029

Wang, M. C., & Bushman, B. J. (1998). Using the normal quantile plot to explore meta-analytic data sets. *Psychological Methods, 3*(1), 46–54. https://doi.org/10.1037/1082-989X.3.1.46

Warne, R. T., & Burningham, C. (2019). Spearman's *g* found in 31 non-Western nations: Strong evidence that *g* is a universal phenomenon. *Psychological bulletin, 145*(3), 237-272. http://dx.doi.org/10.1037/bul0000184

*Warwick, M. M., Doody, G. A., Lawrie, S. M., Kestelman, J. N., Best, J. J., & Johnstone, E. C. (1999). Volumetric magnetic resonance imaging study of the brain in subjects with sex chromosome aneuploidies. *Journal of Neurology, Neurosurgery & Psychiatry, 66*(5), 628-632. http://dx.doi.org/10.1136/jnnp.66.5.628

Wechsler, D. (1939). *The Measurement of Adult Intelligence*. Williams & Witkins.

Wechsler, D. (2008). Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV). *NCS Pearson, 22*(498), 1.

*Weniger, G., Lange, C., Sachsse, U., & Irle, E. (2009). Reduced amygdala and hippocampus size in trauma-exposed women with borderline personality disorder and without posttraumatic stress disorder. *Journal of Psychiatry & Neuroscience: JPN*, 34(5), 383-388.

Whipple, G. M. (1914). *Manual of mental and physical tests: Part 1: Simpler processes* (2nd ed., rev., & enlarged). Warwick & York. https://doi.org/10.1037/10822-000

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., Van Aert, R. C. M., & Van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

*Wickett, J., Vernon, P. A., & Lee, D. H. (1994). In vivo brain size, head perimeter, and intelligence in a sample of healthy adult females. *Personality and Individual Differences,16*(6), 831–838. https://doi.org/10.1016/0191-8869(94)90227-5

*Wickett, J., Vernon, P. A., & Lee, D. H. (2000). Relationships between factors of intelligence and brain volume. *Personality and Individual Differences, 29*(6), 1095–1122. https://doi.org/10.1016/S0191-8869(99)00258-5

*Willerman, L., Schulz, R., Rutledge, J. N., Bigler, E. D. (1991). In vivo brain size and intelligence. *Intelligence, 15*(2), 223-228. https://doi.org/10.1016/0160-2896(91)90031-8

Witelson, S. F., Beresh, H., & Kigar, D. L. (2006). Intelligence and brain size in 100 postmortem brains: sex, lateralization and age factors. *Brain*, *129*(2), 386-398. https://doi.org/10.1093/brain/awh696

Woodley of Menie, M. A., te Nijenhuis, J., Fernandes, H. B., & Metzen, D. (2016). Small to medium magnitude Jensen effects on brain volume: A meta-analytic test of the processing volume theory of general intelligence. *Learning and Individual Differences*, *51*, 215-219. https://doi.org/10.1016/j.lindif.2016.09.007

*Wozniak, J. R., Mueller, B. A., Chang, P. N., Muetzel, R. L., Caros, L., & Lim, K. O. (2006). Diffusion tensor imaging in children with fetal alcohol spectrum disorders. *Alcoholism: Clinical and Experimental Research*, *30*(10), 1799-1806. https://doi.org/10.1111/j.1530-0277.2006.00213.x

*Yeo, R. A., Turkheimer, E., Raz, N., & Bigler, E. D. (1987). Volumetric asymmetries of the human brain: Intellectual correlates. *Brain and Cognition, 6*(1), 15-23. https://doi.org/10.1016/0278-2626(87)90043-1

*Yurgelun-Todd, D. A., Killgore, W. D., & Cintron, C. B. (2003). Cognitive correlates of medial temporal lobe development across adolescence: a magnetic resonance imaging study. *Perceptual and Motor Skills, 96*(1), 3-17. https://doi.org/10.2466/pms.2003.96.1.3

*Zeegers, M., Pol, H. H., Durston, S., Nederveen, H., Schnack, H., van Daalen, E., Dietz, C., van Engeland, H., & Buitelaar, J. (2009). No differences in MR-based volumetry between 2-and 7-year-old children with autism spectrum disorder and developmental delay. *Brain and Development, 31*(10), 725-730. https://doi.org/10.1016/j.braindev.2008.11.002

*Zelko, F. A., Pardoe, H. R., Blackstone, S. R., Jackson, G. D., & Berg, A. T. (2014). Regional brain volumes and cognition in childhood epilepsy: Does size really matter?. *Epilepsy Research*, *108*(4), 692-700. https://doi.org/10.1016/j.eplepsyres.2014.02.003

*Zhu, B., Chen, C., Xue, G., Lei, X., Li, J., Moyzis, R. K., Dong, Q., & Lin, C. (2014). The GABRB1 gene is associated with thalamus volume and modulates the association between thalamus volume and intelligence. *Neuroimage, 102*, 756-763. https://doi.org/10.1016/j.neuroimage.2014.08.048

de Zwarte, S. M., Brouwer, R. M., Tsouli, A., Cahn, W., Hillegers, M. H., Hulshoff Pol, H. E., Kahn, R., & van Haren, N. E. M. (2019). Running in the family? Structural brain abnormalities and IQ in offspring, siblings, parents, and co-twins of patients with schizophrenia. *Schizophrenia Bulletin, 45*(6), 1209-1217. https://doi.org/10.1093/schbul/sby182

Abbreviations

Abbreviation	Meaning
CHC	Cattell-Horn-Carroll
CSF	Cerebrospinal Fluid
CT	Computed Tomography
GM	Grey Matter
GOSH	Graphical Display of Study Heterogeneity
HS	Hunter-Schmidt
ICV	Intracranial Volume
IQ	Intelligence Quotient
MRI	Magnetic Resonance Imaging
POI	Perceptual Organization Index
PM	Paule-Mandel
PSI	Processing Speed Index
QQ	Quantile-Quantile
REML	Restricted Maximum Likelihood
RVE	Robust Variance Estimation
SES	Socioeconomic Status
TBV	Total Brain Volume
VIF	Variance Inflation Factor
VCI	Verbal Comprehension Index
WM	White Matter
WMI	Working Memory Index

Appendices

Appendix A – Software Documentation

Documentation of Used Programs, Packages and R Codes.

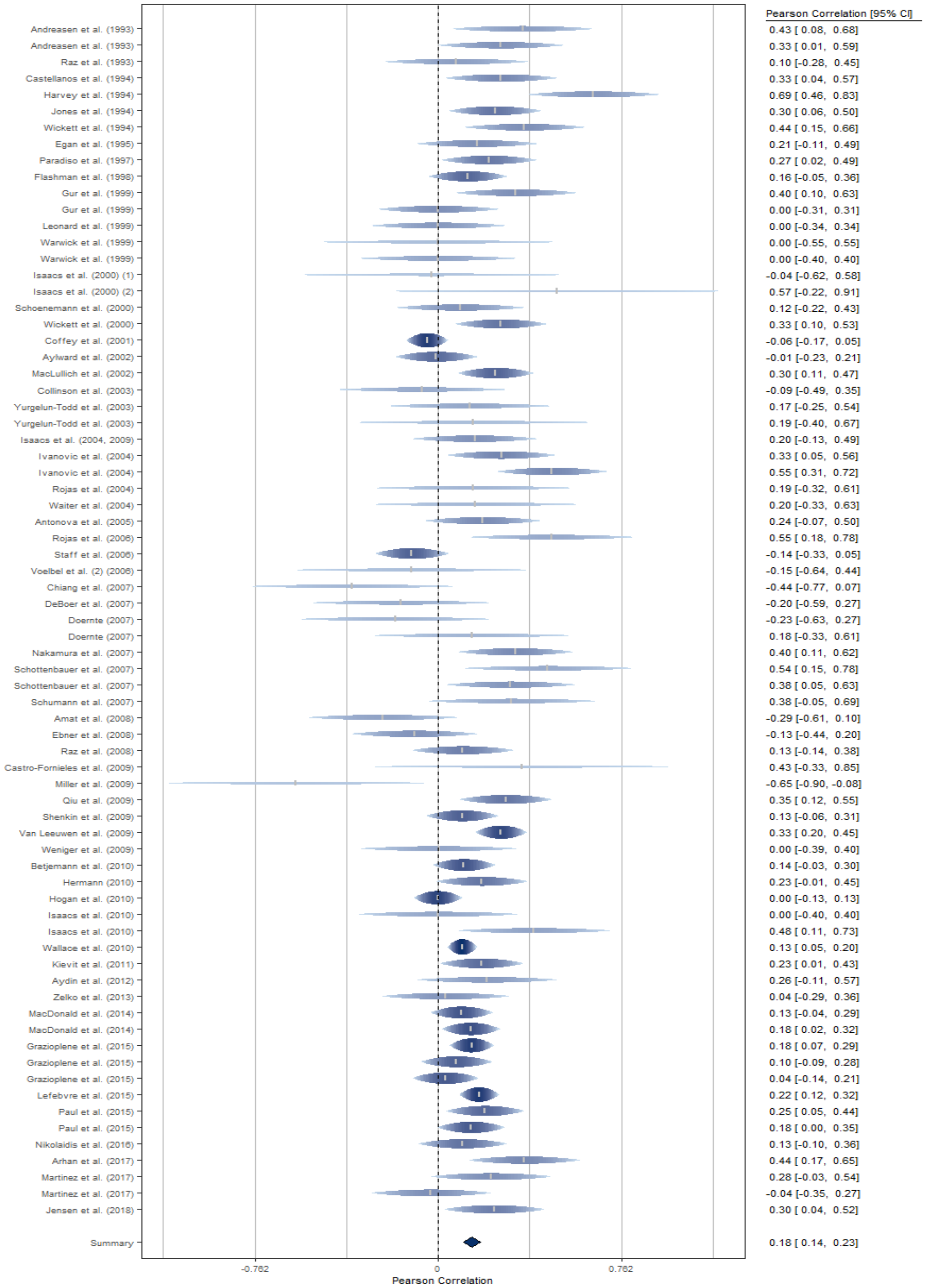
Name	Version	Platform	Reference/Source
bayesmeta	2.5	R	Röver (2020)
clubsandwich	0.5.0	R	https://cran.r-project.org/package=clubSandwich
gsl	2.1-6	R	Hankin (2006)
metafor	2.4-0	R	Viechtbauer (2010)
metasens	0.4-1	R	https://cran.r-project.org/package=metasens
metaviz	0.3.1	R	https://cran.r-project.org/package=metaviz
p-curve	4.06	website	www.p-curve.com
Prediction interval	18.12.2016	Excel sheet	https://www.meta-analysis.com/pages/prediction.php
psychmeta	2.4.2	R	Dahlke & Wiernik (2019)
psychometrica	10/2020	website	Lenhard & Lenhard (2014)
readxl	1.3.1	R	https://cran.r-project.org/package=readxl
robumeta	2.0	R	https://cran.r-project.org/package=robumeta
Rstudio	1.3.1093	R	https://rstudio.com/products/rstudio/download/
vioplot	0.3.5	R	Adler & Kelly (2020)
Voracek et al. (2019)	10.08.2018	R	Voracek et al. (2019)
weightr	2.0.2	R	https://cran.r-project.org/package=weightr

Note. The website Psychometrica was used to obtain z-to-r transformed correlation coefficients. Full documentation of used R packages and their dependencies is available at https://osf.io/47ygt/.

Appendix B – Rainforest Plots for Verbal and Performance IQ

Figure B.1

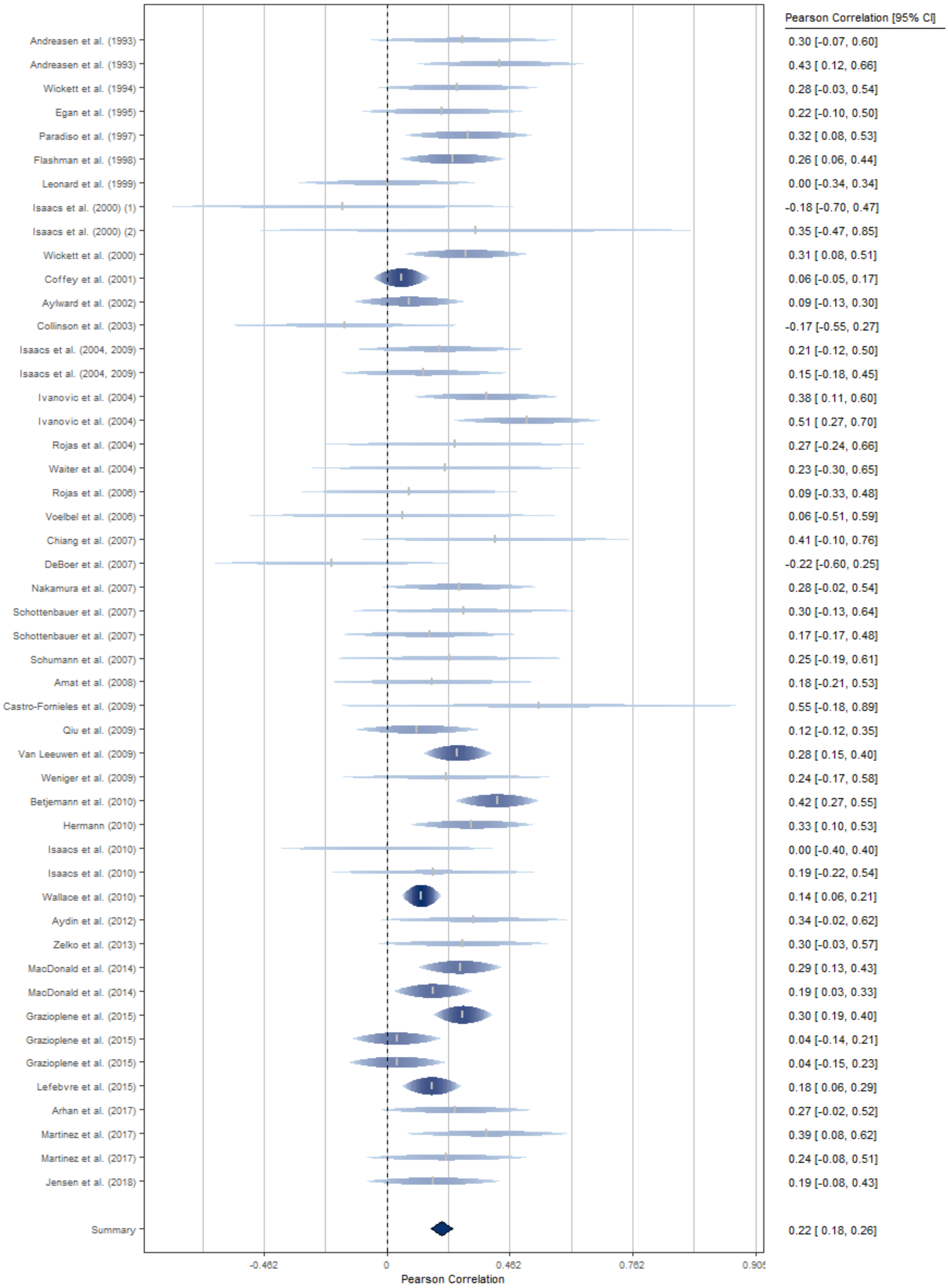
Rainforest Plot for Associations of In Vivo Brain Volume and Verbal IQ Based on Healthy Samples



Note. Summary effect is based on a random effects model and represented by the diamond; symbol size and coloring of raindrops are varied according to relative study weight within analysis.

Figure B.2

Rainforest Plot for Associations of In Vivo Brain Volume and Performance IQ Based on Healthy Samples



Note. Summary effect is based on a random effects model and represented by the diamond; symbol size and coloring of raindrops are varied according to relative study weight within analysis. Rainforest plots for verbal and performance IQ based on clinical samples are available at https://osf.io/t24wg/ and https://osf.io/qxhdu/ respectively.

Appendix C – Overview of Meta-Analytic Summary Effect

Overview of Summary Effects Based on Healthy Samples as Computed by the Different Meta-Analytical Approaches

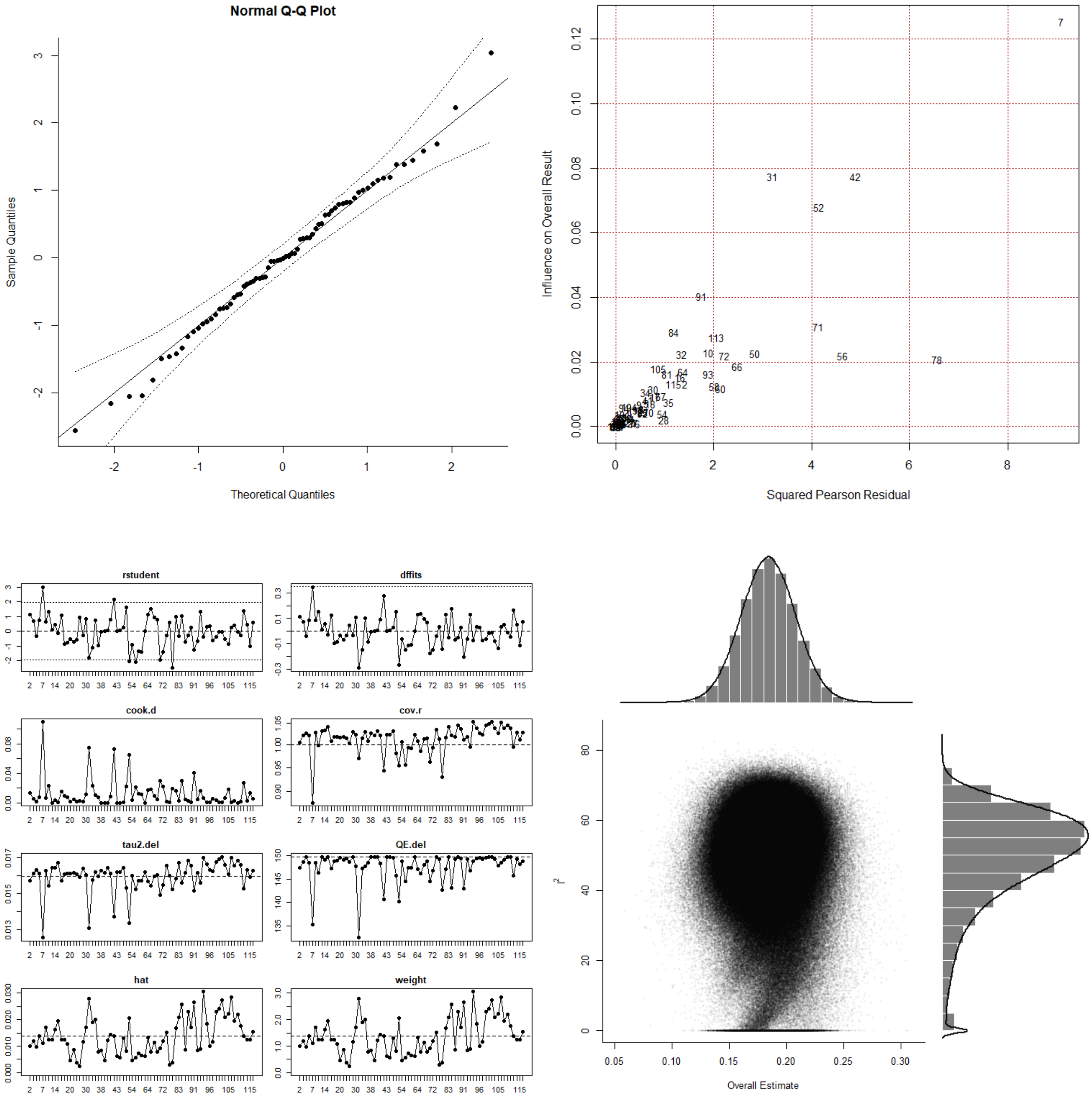
	Full-scale IQ					
	k	*n*	I^2	*r*	LCI	UCI
Hedges-Olkin	122	23359	56.35%	.24	.21	.26
"Bare-Bones"	122	23359	62.97%	.22	.15	.28
Psychometric	64	8315	36.80%	.28	.23	.33
RVE	124 (94)	23553	51.84%	.24	.21	.26
Bayesian	122	23359	-	.24	.21	.26
	Verbal IQ					
	k	*n*	I^2	*r*	LCI	UCI
Hedges-Olkin	73	5322	52.04%	.19	.14	.23
"Bare-Bones"	73	5322	66.15%	.16	.09	.23
Psychometric	31	2541	35.67%	.24	.17	.31
RVE	92 (63)	7633	57.04%	.18	.14	.23
Bayesian	73	5322	-	.18	.14	.23
	Performance IQ					
	k	*n*	I^2	*r*	LCI	UCI
Hedges-Olkin	49	3837	29.33%	.22	.18	.26
"Bare-Bones"	49	3837	29.87%	.20	.16	.25
Psychometric	28	2236	15.28%	.28	.22	.33
RVE	72 (46)	7366	29.44%	.22	.18	.26
Bayesian	49	3837	-	.22	.18	.26

Note. In the RVE approach the number of synthesized effect sizes is followed by the number of individual samples in parentheses. I^2 = percentage of variability due to variability of true effects. LCI = lower bound of 95% confidence interval. UCI = upper bound of 95% confidence interval. In the Bayesian approach these are shortest credible intervals.

Appendix D – Heterogeneity Plots for Verbal and Performance IQ Data

Figure D.1

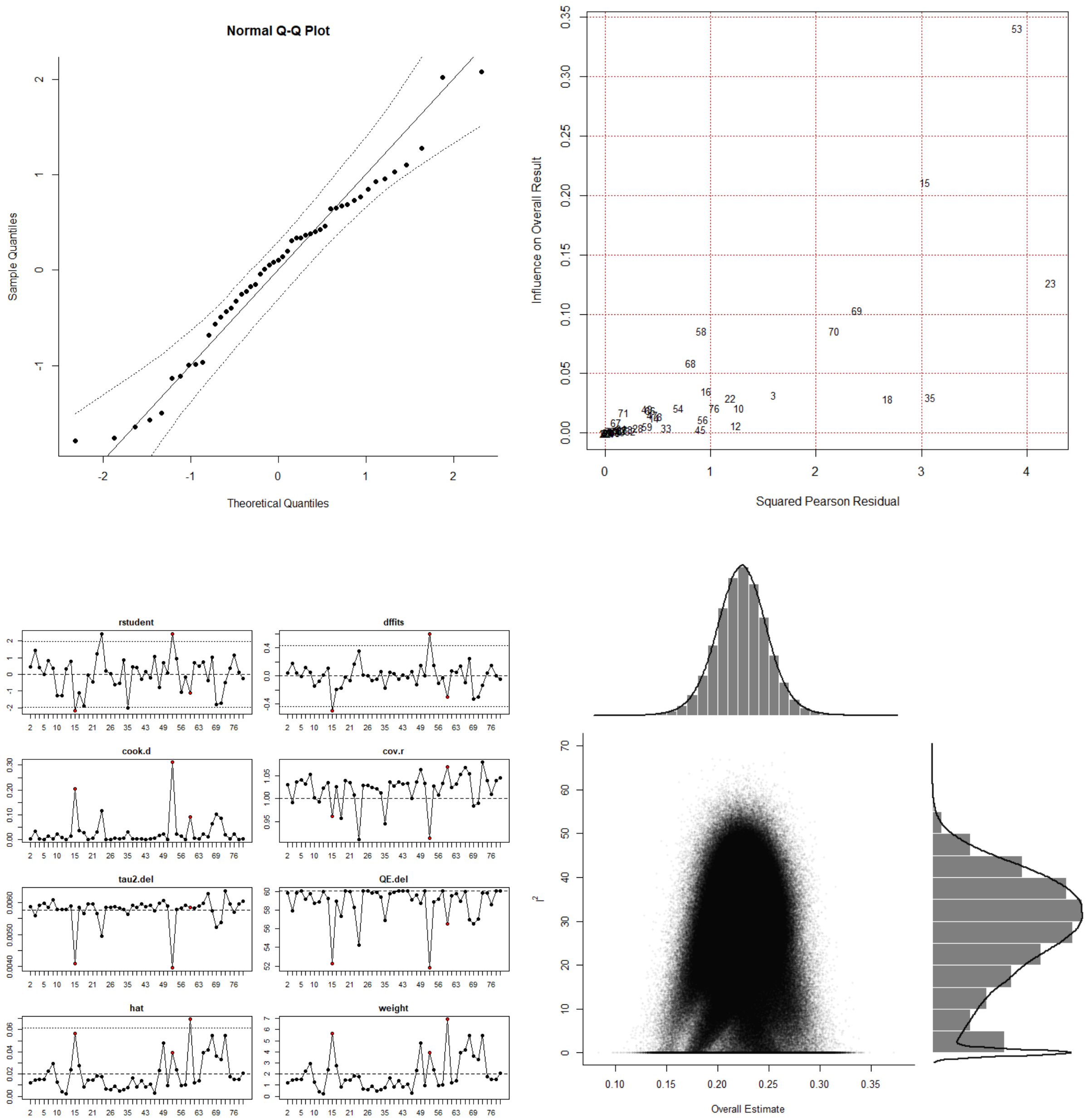
Collection of Plots Assessing Heterogeneity in Verbal IQ Data Based on Healthy Samples



Note. Plots from left to right: normal QQ plot, *Baujat* plot, influence diagnostics, GOSH plot. All plots were created with the *metafor* package.

Figure D.2

Collection of Plots Assessing Heterogeneity in Performance IQ Data Based on Healthy Samples



Note. Plots from left to right: normal QQ plot, Baujat plot, influence diagnostics, GOSH plot. All plots were created with the *metafor* package. Plots based on clinical verbal and performance IQ data are available at https://osf.io/y6msp/.

Appendix E – Subgroup Results

Meta-Analytic Results for Subgroups Based on Healthy Samples

	Full-scale IQ					
	k	*n*	I^2	*r*	LCI	UCI
Healthy	122	23359	56.35%	.24	.21	.26
Clinical	66	4481	58.95%	.22	.16	.27
Reported	72	21455	64.12%	.25	.22	.28
Grey	4	66	46.41%	.13	-.49	.66
PC	46	1838	34.43%	.19	.13	.26
Children	54	4035	17.13%	.24	.20	.28
Adults	68	19324	69.36%	.23	.20	.27
TBV	74	17484	42.11%	.24	.21	.26
ICV	17	4682	79.64%	.26	.18	.34
Females	24	5994	0.02%	.26	.23	.29
Males	36	6094	10.97%	.25	.21	.29
"Fair"	12	980	71.69%	.23	.09	.37
"Elevated"	45	18174	56.70%	.20	.17	.23
"High"	50	3355	4.07%	.30	.27	.34

	Verbal IQ					
	k	*n*	I^2	*r*	LCI	UCI
Healthy	73	5322	52.04%	.19	.14	.23
Clinical	45	2263	36.30%	.22	.15	.28
Reported	42	4081	54.81%	.21	.15	.27
Grey	2	35	20.27%	-.03	-1	1
PC	29	1206	43.03%	.13	.05	.22
Children	27	2234	31.91%	.22	.15	.29
Adults	46	3088	54.59%	.17	.11	.23
TBV	47	4076	40.67%	.18	.13	.22
ICV	13	765	64.64%	.30	.15	.43
Females	15	650	0.03%	.24	.15	.32
Males	22	1038	25.72%	.23	.15	.31

	Performance IQ					
	k	*n*	I^2	*r*	LCI	UCI
Healthy	49	3837	29.33%	.22	.18	.26
Clinical	32	1840	23.82%	.19	.13	.26
Reported	28	3224	46.71%	.24	.19	.29
Grey	0	0	-	-	-	-
PC	21	613	0%	.17	.10	.23
Children	24	2108	30.03%	.23	.17	.29
Adults	25	1729	27.75%	.21	.16	.26
TBV	32	3375	40.91%	.22	.17	.26
ICV	6	180	0%	.29	.17	.40
Females	9	429	0%	.25	.17	.32
Males	16	717	17.44%	.24	.16	.31

Note. I^2 = percentage of variability due to variability of true effects. LCI = lower bound of 95% confidence interval. UCI = upper bound of 95% confidence interval. “Fair”, “elevated”, “high” refer to the rated correlation of the IQ tests with *g*.

Appendix F – RVE Meta-Regression for Verbal and Performance IQ Subdomains

A potential difference in effect size between brain volume verbal or performance IQ subdomains was examined. Verbal and performance IQ domains were divided into verbal comprehension, working memory, perceptual organization and processing speed. The subsequent coding was based on the tests used and followed the structure of Wechsler intelligence scales. Correlations where no subdomain assignment was possible were categorized as "overall". The variable was named "IQdomain2" in the data set. A comparison using RVE meta-regression between 32 verbal comprehension and 15 working memory correlations showed no statistically significant differences (*r* = .21, *r* = .19, *p* = .745). Differences between 26 perceptual organization and 12 processing speed correlations were also not statistically significant, but more pronounced (*r* = .24, *r* = .16, *p* = .241). The inconspicuous *p*-value may have been a result of the relatively low power. Since the difference in effect size was theorized before conducting the analysis, I interpreted the result as meaningful despite the *p*-value.

Appendix G – Meta-Regression of Five-Level Age Factor

A five-level coding approach was applied to test a potential moderating role of age ones more. The levels were: children (0-12yr), adolescents (13-18yr), young adults (19-34yr), adults (35-64yr), and elderly (65+yr). Coding focused on mean age under consideration of the standard deviation of mean age in a sample. Age intervals have been chosen according to stages of brain volume increases and decreases across lifetime (Hedman et al., 2012). In cases where a substantial overlap of age groups hindered categorization, samples were categorized as “mixed”. Coding choices were therefore sometimes a bit arbitrary. A random-effects meta-regression with the same specifications as the two-level moderator test was conducted. The robustness test of a moderating effect was statistically insignificant (*F*(5, 96) = 1, *p* = .422). Summary effects of age groups differed. The correlation for children was *r* = .25 (*k* = 17), for adolescents *r* = .32 (*k* = 15), for young adults *r* = .23 (*k* = 31), for adults *r* = .20 (*k* = 8), for elderly *r* = .19 (*k* = 8), and for mixed groups *r* = .24 (*k* = 23). Power was greatly reduced in comparison with other moderator analyses concerning age. Results should be interpreted in a way that age moderating effects might be sensitive (in a small degree) to variable operationalization, not as hard evidence.

Appendix H – English Abstract

Attempts to link brain volume and intelligence go back over 180 years. In numerous narrative reviews and three meta-analyses no agreement could have been reached on the effect size or their association. Although recently published meta-analyses (Pietschnig et al., 2015; Gignac & Bates, 2017) have been based on the same data pool, they have reported divergent results (*r* = .24 and *r* = .39). There was also no agreement on the influence of potential moderators and dissemination bias. In order to address these unresolved questions, a meta-analysis was conducted based on an update of the data pool by Pietschnig et al. (2015). Forty-eight new studies were included; the total number of study participants tripled. The correlation between in vivo brain volume and intelligence was estimated to be *r* = .24 based on healthy subjects using three different approaches (Hedges-Olkin, robust variance estimation, Bayesian meta-analysis). Performing a psychometric meta-analysis resulted in a higher estimate (*r* = .28). By applying meta-analytical specification analyses according to Voracek et al. (2019), the range of results under all possible specifications of previous meta-analysts was estimated to be *r* = .20 to *r* = .35. The use of range departure corrected correlations and the exclusive consideration of results based on extensive intelligence assessment had the most notable influence. The association between brain volume and intelligence generalized over age, sex, and intelligence domains. Dissemination bias was detectable in the data but had little impact on effect estimates. Nevertheless, decreasing effects were observed over the entire time span of the studies considered.

Appendix I – German Abstract

Die Geschichte von Versuchen Gehirnvolumen und Intelligenz in Zusammenhang zu bringen zählt über 180 Jahre. In zahlreichen narrativen Reviews und drei Meta-Analysen konnte keine Einigkeit über die Größe des Zusammenhangs erzielt werden. Kürzlich publizierte Meta-Analysen (Pietschnig et al., 2015; Gignac & Bates, 2017) basierten auf demselben Datenpool, berichteten aber auseinandergehende Ergebnisse (*r* = .24 und *r* = .39). Auch über den Einfluss von potenziellen Moderatoren sowie Disseminationsbias bestand keine Einigkeit. Um diese offenen Fragen aufzuklären, wurde eine neuerliche Meta-Analyse auf Grundlage eines Updates des Datenpools von Pietschnig et al. (2015) unternommen. Es wurden 48 neue Studien aufgenommen; die Gesamtzahl der Studienteilnehmer*innen konnte verdreifacht werden. Der Zusammenhang von in vivo Gehirnvolumen und Intelligenz wurde auf Grundlage von gesunden Proband*innen mit drei verschiedenen Ansätzen (Hedges-Olkin, Robust Variance Estimation, Bayesianische Meta-Analyse) auf *r* = .24 geschätzt. Die Durchführung einer psychometrischen Meta-Analyse führte zu einer höheren Schätzung (*r* = .28). Durch die Anwendung von meta-analytischen Spezifikationsanalysen nach Voracek et al. (2019) konnte der gesamte Ergebnisraum unter allen möglichen Spezifikationen vorheriger Meta-Analyst*innen auf *r* = .20 bis *r* = .35 geschätzt werden. Den größten Einfluss auf das Ergebnis hatte die Anwendung von Methoden zur Simulation von Bereichsabweichungen der Standardabweichung von Intelligenztestergebnissen sowie die Berücksichtigung ob eine umfassende Intelligenztestung stattfand oder nicht. Der Zusammenhang zwischen Gehirnvolumen und Intelligenz zeigte sich robust gegenüber Alter, Geschlecht und Intelligenz-Domänen. Disseminationsbias war in den Daten nachweisbar, hatte aber wenig Einfluss auf die Effektschätzungen. Es wurden jedoch abnehmende Effektschätzungen über die gesamte zeitliche Spanne der berücksichtigten Studien beobachtet.