Intelligence 64 (2017) 18–29

Contents lists available at ScienceDirect

Intelligence

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

Brain volume and intelligence: The moderating role of intelligence measurement quality

Gilles E. Gignac[a,*], Timothy C. Bates[b]

[a] University of Western Australia, Australia
[b] The University of Edinburgh, Scotland

ARTICLE INFO

Keywords:
Meta-analysis
Meta-regression
Brain volume
Intelligence
P-curve

ABSTRACT

A substantial amount of empirical research has estimated the association between brain volume and intelligence. The most recent meta-analysis (Pietschnig, Penke, Wicherts, Zeiler, & Voracek, 2015) reported a correlation of .24 between brain volume and intelligence – notably lower than previous meta-analytic estimates. This headline meta-analytic result was based on a mixture of samples (healthy and clinical) and sample correlations not corrected for range restriction. Additionally, the role of IQ assessment quality was not considered. Finally, evidential value of the literature was not formally evaluated. Based on the results of our meta-analysis of the Pietschnig et al.'s sample data, the corrected correlation between brain volume and intelligence in healthy adult samples was r = .31 (k = 32; N = 1758). Furthermore, the quality of intelligence measurement was found to moderate the effect between brain volume and intelligence (b = .08, p = .028). Investigations that used 'fair', 'good', and 'excellent' measures of intelligence yielded corrected brain volume and intelligence correlations of .23 (k = 9; N = 547), .32 (k = 10; N = 646), and .39 (k = 13; N = 565), respectively. The Henmi/Copas adjusted confidence intervals, the p-uniform results, and the p-curve results failed to suggest evidence of publication bias and/or p-hacking. The results were interpreted to suggest that the association between in vivo brain volume and intelligence is arguably best characterised as $r \approx .40$. Researchers are encouraged to consider intelligence measurement quality in future meta-analyses, based on the guidelines provided in this investigation.

1. Introduction

The topic of brain size and its possible association with intelligence, both within and between species, has been the subject of a substantial amount of research and debate (Mackintosh, 2011). Recently, Pietschnig et al. (2015) reported a meta-analytic observed correlation between human brain volume and intelligence of r = .24, based on 120 sample correlations (N = 6778). A limitation associated with the Pietschnig et al. (2015) investigation is that it did not provide an estimate of the association between brain volume and intelligence corrected for range restriction. Additionally, Pietschnig et al. (2015) did not explore the possibility that quality of intelligence measurement may moderate the magnitude of the association between brain volume and intelligence. Finally, Pietschnig et al. (2015) did not formally evaluate the evidential value of the reported research via a p-curve analysis.

Consequently, the purpose of this investigation was to extend the Pietschnig et al. (2015) meta-analysis in three ways. First, to estimate the correlation between in vivo human brain volume and intelligence based on correlations associated with relatively few artefacts, i.e., correlations derived from healthy adult samples and corrected for range restriction. Secondly, to develop a guide to help classify the quality of general intelligence measurement, in order to test the hypothesis that there is a positive association between intelligence test measurement quality and the magnitude of effect sizes reported across empirical investigations. Finally, to conduct a p-curve analysis to evaluate the reported brain volume and intelligence statistically significant correlations for evidential value.

1.1. Brain volume and intelligence: quantitative reviews

The association between in vivo brain volume and intelligence has been reviewed quantitatively several times over the years. More than a decade ago, Gignac, Vernon, and Wickett (2003) estimated the observed correlation between brain volume and IQ based on 14 samples (N = 858), all of which were derived from peer reviewed publications. Gignac et al. (2003) reported an N-weighted mean correlation of .37 between brain volume and intelligence. In six of the 14 investigations included in the meta-analysis, the IQ score standard deviations were

[*] Corresponding author at: School of Psychology, University of Western Australia, 35 Stirling Highway, Crawley, Western Australia 6009, Australia.
E-mail address: gilles.gignac@uwa.edu.au (G.E. Gignac).

http://dx.doi.org/10.1016/j.intell.2017.06.004
Received 6 December 2016; Received in revised form 25 May 2017; Accepted 26 June 2017

available. Consequently, Gignac et al. (2003) also reported an N-weighted mean corrected correlation of .43 between brain volume and IQ.[1]

McDaniel (2005) revisited the in vivo brain volume and intelligence association by conducting a more comprehensive meta-analysis than that of Gignac et al. (2003). McDaniel's (2005) inclusion criteria were the following: clinically healthy samples; total brain volume measurement; and well-established measures of intelligence (Wechsler scales; Raven's; but not the National Adult Reading Test, for example). Based on the samples which met those criteria ($k = 37$; $N = 1530$), McDaniel (2005) reported an observed correlation of $r = .29$ between brain volume and global intelligence. Additionally, McDaniel (2005) reported a range restricted corrected correlation of $r = .33$. Thus, the corrected correlation reported by McDaniel (2005) was smaller than the corrected correlation reported by Gignac et al. (2003; $r = .43$).

It is noteworthy that McDaniel (2005) found that the mean correlation between brain volume and intelligence was larger for adults than for children. For example, the brain volume and intelligence corrected correlation for adult males was estimated at $r = .38$, whereas the same correlation for male children was estimated at $r = .22$. McDaniel (2005) did not speculate as to why the effects may have been larger for adults in comparison to children. It is suggested here that both incomplete neurophysiological maturation and individual differences in the rate of maturation explain some of the increase in the magnitude of the brain volume and intelligence correlation from childhood to adulthood. For example, there are individual differences in the neurophysiological maturation of the frontal lobes across childhood and adolescents (Nagy, Westerberg, & Klingberg, 2004; Segalowitz & Davies, 2004). Furthermore, several of the neurophysiological characteristics of maturation may be substantially independent of brain volume (e.g., pruning, intra-cortical myelination; Paus, 2005). Thus, until such neurophysiological characteristics are largely stabilised once maturation is complete (i.e., adulthood), the correlation between brain volume and intelligence may be expected to be attenuated. Stated alternatively, the correlation between brain volume and intelligence in children may not be a fully accurate reflection of the effect.

McDaniel (2005) noted the difficulties associated with conducting a comprehensive meta-analysis, as many empirical investigations did not include standard deviation or internal consistency reliability estimates associated with the test scores. In fact, McDaniel (2005) was required to use standard deviation artefact distribution imputation for 21 of the sample correlations, as only 16 of the 37 brain volume and intelligence studies reported the standard deviation associated with intelligence test scores. Thus, the key brain volume and intelligence correlation ($r = .33$) reported by McDaniel (2005) rests upon the assumption that the imputation method worked in a valid manner.

More recently, Pietschnig et al. (2015) conducted a meta-analysis on the brain volume and intelligence empirical literature. In contrast to Gignac et al. (2003) and McDaniel (2005), Pietschnig et al. (2015) obtained a substantial number of personal communications relevant to the association between brain volume and intelligence across a variety of studies and samples. Based on 120 sample correlations derived from a mix of healthy and clinical samples ($N = 6778$), Pietschnig et al. (2015) reported a meta-analytic correlation of $r = .24$ between brain volume and global measures of intelligence (e.g., FSIQ). Thus, Pietschnig et al. (2015) reported an effect notably smaller than the meta-analytic estimates reported by McDaniel (2005; $r = .33$) and Gignac et al. (2003; $r = .43$). Pietschnig et al. (2015) suggested that the correlations reported in previous meta-analyses were likely over-estimates, as the published literature was likely affected by selective reporting (i.e., statistically non-significant effects were not reported). In support of such an argument, the meta-analytic correlation between brain volume and general intelligence based on published results was reported by Pietschnig et al. (2015) at $r = .30$ ($k = 53$; $N = 3956$). By contrast, the corresponding meta-analytic correlation in non-published work was estimated at just $r = .17$ ($k = 67$; $N = 2822$).

It should be noted, however, that both Gignac et al. (2003) and McDaniel (2005) restricted their meta-analyses to healthy samples, whereas Pietschnig et al.'s headline correlation of .24 included both healthy and clinically mixed samples. Arguably, intelligence test scores obtained from individuals suffering from various clinical conditions should not be considered optimally valid indicators of intellectual functioning. For this reason, it is commonly recommended that individuals "…should not be assessed [for intelligence] unless they appear suitably healthy and well rested." (Reschly, Myers, & Hartel, 2002, p. 101). From a statistical perspective, a correlation between intelligence and a criterion would be expected to be suppressed in clinical samples, because it is unreasonable to assume that all of the examinees suffer from the exact same condition to the same degree. Such individual differences in the clinical condition would be expected to affect the rank ordering in measurement of intelligence, in comparison to "true" intelligence, which is a threat to validity, in this context.

If we wish to estimate the population correlation accurately, sample ascertainment is critical. Whereas a sample restricted to healthy adult individuals will, allowing for sampling error, approximate the true population estimate, mixtures of samples, with non-random inclusion criteria, are likely to show considerable bias. This is true not only in extreme cases (imagine a sample of people "administered" the Raven after the consumption of 10 standard drinks of alcohol) but is likely to hold in general.

Consider, for example, the report of a relatively low correlation of $r = .07$ between brain volume and intelligence, based on a sample of 41 neurological patients (Yeo, Turkheimer, Raz, & Bigler, 1987). Nineteen of these patients presented with headache complaints, while 7 presented with symptoms of problems in concentration and memory. The validity of the brain volume and IQ correlation is not established in such a combination of groups. That is, arguably, patients suffering from concentration and memory problems will produce IQ scores which are lower than their natural maximal capacity. By contrast, migraine patients completing the IQ testing may be expected to show substantial variability, depending on, for instance, the varying level of migraine experienced during their testing, from none at all to severe. However, in both groups of cases, brain volumes likely remained stable. Consequently, rank ordering of the IQ scores in this mixed clinical sample was likely affected adversely by the heterogeneity of the clinical conditions between the patients. Such an adverse impact on rank ordering of IQ scores would also affect adversely the estimated correlation between brain volume and intelligence. In light of the above, it is our view that the best sample estimate of the true association between brain volume and intelligence, as well as tests of hypothesized moderator effects, is obtained by aggregating studies of generally healthy adult samples.

Additionally, it is important to note that Pietschnig et al. (2015) did not correct any of the correlations (published or non-published) for range restriction. By contrast, both Gignac et al. (2003) and McDaniel (2005) did take range restriction into consideration. Pietschnig et al. acknowledged the issue of range restriction in their meta-analysis, however, they did not apply a correction to their analysis, because "…a majority of the included samples' standard deviations for test performance were not reported" (p. 426–427). However, based on our review, nearly all of the studies associated with the healthy adult samples ($k = 32$) did report standard deviations for the intelligence test scores. The importance of correcting observed correlations for range restriction to obtain a more accurate estimate of the effect in the population has been well established (Le & Schmidt, 2006). For example, based on the results of a simulation investigation, Duan and Dunlap (1997) found that when the population correlation was .30 and the selection ratio was .90 (i.e., the sample standard deviation was 10% smaller than the

[1] For an introduction to the problem of range restriction and the estimation of correlations in the population, consult Wiberg and Sundström (2009). More advanced treatments can be found in Sackett and Yang (2000) and Hunter, Schmidt, and Le (2006).

population standard deviation), the observed correlation was estimated at .255, whereas the correlation corrected for range restriction was estimated at .294. Thus, to extend the findings reported in Pietschnig et al. (2015), a primary purpose of the current investigation was to estimate the correlation between brain volume and intelligence in healthy adults, corrected for range restriction in the intelligence test scores.

1.2. Measurement quality: meta-regression

In addition to range restriction, it is known that measurement quality (both reliability and validity) can attenuate the magnitude of effects estimated in a particular investigation (Furr, 2011). In the context of meta-analyses, there is some awareness of the effect of differential measurement quality on the magnitude of the effect observed between two variables. For example, commenting on a meta-analysis relevant to salt intake and the risk of stroke, Appel (2009) implicated the poor quality of dietary salt measurement in several of the empirical investigations as a key cause of significant heterogeneity in the results. In another meta-analysis relevant to the effects of parenting type on childhood depression, McLeod, Weisz, and Wood (2007) found that parental rejection was associated with childhood depression, but only when parental rejection was measured with multiple informants, in comparison to a single informant. Thus, McLeod et al. (2007) contended that measurement quality should be taken into consideration when considering the effect of one variable on another at the meta-analytic level.

With respect to the measurement of intelligence, assessments can vary from brief, group-administered, arbitrarily abbreviated, single-scale measures through to comprehensive batteries in which testing lasts over an hour. However, few, if any, meta-analyses in the area of intelligence have taken into consideration the possibility that the quality of intelligence measurement may moderate the effect between intelligence test scores and another variable. One likely reason meta-analyses do not consider the measurement quality of general intelligence (*g*) test scores is that there are no established guidelines for such a purpose. Consequently, a goal of this investigation was to test intelligence measurement quality as a moderator of the effect between brain volume and intelligence. First, however, an intelligence measurement quality classification guide needed to be developed.

In the most straightforward terms, the correlation between cognitive ability test scores and *g* would help quantify the quality of general intelligence measurement in a study. However, many combinations of cognitive ability tests have never been evaluated empirically for their association with *g*. Although a precise, non-factor analytic, algorithm for the specification of general intelligence measurement quality does not appear to have ever been published, arguably, most intelligence tests (and combination of tests) can be categorised according to their quality, particularly with respect to representations of *g*. For example, the administration of the five minute Stroop test (Golden, 1978) could not be classified justifiably as an excellent, or even a good, measure of general intellectual functioning, as it is only a single cognitive ability test which measures a single group-level dimension of intelligence. Not coincidently, the Stroop test has been found to relate to *g* only moderately at approximately .45 (Burns, Nettelbeck, & McPherson, 2009). By contrast, the FSIQ scores derived from the complete WAIS-IV would be considered an excellent measure of *g* by most clinicians and researchers (Reynolds, Floyd, & Niileksela, 2013; Sattler & Ryan, 2009). Distinguishing between the Stroop and the full WAIS-IV as indicators of general intelligence is relatively uncontentious. The challenge is to specify a more detailed guideline that may be able to accommodate all investigations which include at least one measure of cognitive ability.

As a general statement, the quality of the measurement of *g* may be determined, in part, by the number of subtests completed by the participants. Jensen (1998) recommended that a minimum of nine subtests is required to represent *g* respectably. Furthermore, the nine subtests should represent at least three group-level dimensions of cognitive ability (e.g., fluid intelligence, crystallised intelligence, processing speed). Jensen's (1998) recommendation is commonly cited (e.g., Colom, Juan-Espinosa, Abad, & García, 2000; Gignac, Shankaralingam, Walker, & Kilpatrick, 2016; Juan-Espinosa, Cuevas, Escorial, & García, 2006). Furthermore, there is empirical research which supports the notion that a stable estimate of *g* is unlikely to be achieved with fewer than 8 subtests (Major, Johnson, & Bouchard, 2011). As can be seen in Table 1, it is suggested that 1, 1–2, 2–8, and 9 + tests be classified as possibly 'poor', 'fair', 'good', and 'excellent' measures of *g*, in the absence of any other information.

Table 1
Basic guide for the categorisation of the quality of the measurement of general intelligence.

	Poor = 1	Fair = 2	Good = 3	Excellent = 4
1. Number of tests	1	1–2	2–8	9 +
2. Dimensions	1	1–2	2–3	3 +
3. Testing time	3–9 min	10–19 min	20–39 min	40 + min
4. Correlation with *g*	≤ .49	.50–.71	.72–.94	≥ .95

Note. The first three criteria can be evaluated objectively; the fourth criterion (correlation with *g*) may require some judgement on the part of the researcher, based on a combination of direct and indirect empirical evidence in the literature; in the absence of direct or indirect empirical evidence, exclusive reliance upon the first three criteria will be required.

In addition to the number of tests, the number of group-level factors of intelligence represented by the tests should also be considered. It is widely acknowledged that there are approximately 10 group-level factors of intelligence (Carroll, 2003). Commonly measured group-level factors of intelligence include crystallised intelligence (Gc), fluid intelligence (Gf), memory span (Gsm), and processing speed (Gs). Jensen (1998) recommended that a good measure of *g* be based on measures indicative of at least three group-level factors. Thus, a battery of nine short-term memory tests would not be considered an excellent measure of *g*, because all of the tests are related to a single group-level factor (Gsm). As can be seen in Table 1, it is suggested here that cognitive ability tests indicative of 1, 1–2, 2–3, and 3 + dimensions be classified as possibly 'poor', 'fair', 'good', and 'excellent' measures of *g*, in the absence of any other information. The overlap across the categories is a reflection of the fact that the various group-level factors differ in the degree to which they relate to *g*. For example, Gf and Gc are known to relate to *g* very strongly (Gignac, 2014; Kvist & Gustafsson, 2008), whereas Gsm (excluding working memory tasks) and Gs have been found to be weaker indicators of *g* (Reynolds & Keith, 2007). Thus, some consideration should be placed on the *g* saturation of the group-level factors to which the selected tests belong.

In addition to the number of tests and the amount of test diversity, the amount of time required to complete the testing should also be considered an indicator of general intelligence measurement quality. For example, a hypothetical study may administer nine tests of cognitive ability, however, due to time constraints, the investigator may choose to administer only short-forms of all of the subtests (say, even items), resulting in a testing time of only 30 min. Arguably, such an administration would not be considered as impressive as the same battery of tests which included the entire set of items and 60 min of testing time. As can be seen in Table 1, it is suggested that 3–9 min, 10–19 min, 20–39 min, and 40 + minutes be classified as 'poor', 'fair', 'good', and 'excellent' measures of *g*.

To summarize, the three key general intelligence measurement quality characteristics described above include: (1) number of tests: (2) diversity, i.e., number of group-level dimensions measured; and (3) amount of testing time. Across investigations, all three key characteristics would be expected to be correlated positively. For example, the number of tests administered would be expected to be associated with greater testing times. However, the three key characteristics would not be expected to be correlated perfectly. Consequently, all three characteristics should be considered. For example, Raven's progressive matrices take as much as 35–45 min to complete (Arthur & Day, 1994), which would suggest that it is an excellent measure of *g*. However, it is only a single test; furthermore, it measures only a single group-level dimension of intelligence. Notably, across several large, representative samples, Raven's has been found to be associated with *g* at .68 (Gignac, 2015). Thus, Raven's would be classified as a fair measure of *g*, based on the guidelines provided in Table 1.

An additional row of information has been included in Table 1 (correlation with *g*): the expected association between the test scores and *g*. It can be seen that relatively poor measures of *g* are proposed to share $\leq 24\%$ of their variance with *g* ($r \leq .45$). Fair measures are proposed to share between 25% and 50% of their variance *g* ($r = .50$ to .71). Good measures of *g* are proposed to share between 51% and 89% of their variance with *g* ($r = .51$ to .94). Finally, excellent measures of *g* are expected to be associated with *g* such that the total scores share 90% or more of their variance with *g* ($r \geq .95$).

Technically, the only information required to categorise intelligence test scores as indicators of *g* is this association with *g*. In practice, however, the three key characteristics described above are necessary because the various combinations of tests included in investigations have never been tested specifically for their association with *g*. Thus, the first three key characteristics listed in Table 1 are to be used as a necessary substitute, when the association with *g* has not been established empirically.

Once the intelligence test scores associated with the investigations included in a meta-analysis have been coded according to the guidelines reported in Table 1, intelligence test score quality can be examined as a possible moderator of the effect between an independent variable and intelligence. Such a moderator analysis can be conducted within the context of a conventional meta-regression (Huizenga, Visser, & Dolan, 2011).

1.3. *p*-Curve analysis

It is known that the social sciences suffer from severe publication bias, which often distorts the literature (Franco, Malhotra, & Simonovits, 2014). For the validity of meta-analyses, then, it is critical to determine if bias affects the reviewed literature (McShane, Böckenholt, & Hansen, 2016). The results of Pietschnig et al.'s (2015) meta-analysis suggested that the brain volume and intelligence literature may have been influenced by selective reporting of significant effects, as the reported brain volume and intelligence correlations were, on average, larger than the non-reported correlations ($r = .30$ versus $r = .17$). Such differences do not, however support a formal diagnosis of bias in the literature, or, more generally of *p*-hacking (analysing data a number of different (ad hoc) ways until a statistically significant effect is observed). Simonsohn, Nelson, and Simmons (2014a) introduced the *p*-curve analysis as a method capable of formally evaluating the likelihood that published literature relevant to a particular hypothesis may be the result of *p*-hacking. The logic of the *p*-curve analysis is based principally upon the notion that *p*-hacking can be expected to yield a disproportionately large number of *p*-values just below the coveted alpha .05 threshold (i.e., $.026 < p < .050$). By contrast, when a true statistically significant effect has been reported in the literature, one should observe a significantly disproportionate number of *p*-values $< .025$ (Simonsohn, Nelson, & Simmons, 2014b). Because the analysis is based on a hypothesis about the distribution of published significant results, it does not require access to unpublished analyses.

Several *p*-curve analyses have been published recently which have called into question the evidential value of high-profile findings. For example, Vadillo, Gold, and Osman (2016) failed to observe the expected right-tailed distribution of statistically significant *p*-values in published data on the glucose model of ego depletion. In another investigation, the 33 statistically significant results supportive of the claimed effect of power-posing showed a flat distribution of *p*-values, thus supporting the alternative hypothesis that there is no power-posing effect (Simmons and Simonsohn, 2016). Additionally, Melby-Lervåg, Redick, and Hulme (2016) found that the statistically significant effects reported in the literature relevant to the generalisability of effects due to working memory training (with active control groups) were consistent with a left-skewed distribution, i.e., not supportive of a true effect in the population.

No published meta-analysis of the association of brain volume with IQ has attempted a *p*-curve analysis. Consequently, an additional purpose of this investigation was to test the possibility that statistically significant results reported in the healthy adult brain volume and intelligence published literature may have been influenced by *p*-hacking.

1.4. Summary

Although the Pietschnig et al. (2015) meta-analysis should be considered a comprehensive and competently executed meta-analysis, the reported results were limited in a number of ways. Consequently, the purpose of this investigation was to estimate the association between brain volume and intelligence, based on correlations associated with relatively few artefacts, i.e., derived from healthy adult samples and correlations corrected for range restriction. Additionally, we conducted several modern publication bias analyses, including a *p*-curve analysis, to determine whether the statistically significant results in the area support evidential value. Finally, we tested the hypothesis that the quality of measurement of intelligence, as a representation of *g*, moderated the association between brain volume and intelligence via a meta-regression.

2. Method

2.1. Dataset

In order to ensure comparability, the studies considered for inclusion in the current meta-analysis were derived from the Pietschnig et al. (2015) meta-analysis relevant to brain volume and intelligence. Specifically, the study references, study characteristics, and correlational results were drawn from the supplementary material excel file published with Pietschnig et al. (2015). Although a more extensive search could have been undertaken, we were particularly interested in comparing the results obtained from this investigation with those reported by Pietschnig et al. (2015). Consequently, we restricted our search for studies to those reported in Pietschnig et al. (2015).

2.2. Inclusion and exclusion criteria

Pietschnig et al. (2015) listed a total of 120 sample correlations between brain volume and overall intelligence derived from a total of 75 investigations. However, in order to estimate a meta-analytic derived correlation with the least number of artefacts, we excluded sample correlations based on children and/or adolescents, as well as sample correlations based on a mixture of children and adults. We also excluded samples which included participants suffering from a clinical disorder or a learning disability. Finally, we excluded a sample that had only 3 participants.[2] In some cases, Pietschnig et al. (2015) included only the correlation between brain volume and intelligence for the sexes separated into two groups. As this investigation was not particularly interested in an evaluation of sex differences, we made an effort to identify the correlation between brain volume and intelligence for the whole sample within the research papers included in the Pietschnig

[2] Pietschnig et al. (2015) included personal communication results of .00 ($N = 3$) associated with Leonard et al. (1999).

et al. meta-analysis. In some cases, the overall correlation was not obtainable, thus, some of the correlations included in the current meta-analysis were based on gender separated samples. Based on the application of the inclusion/exclusion criteria applied in this investigation, a total of 32 correlations were selected for the meta-analysis.

As mentioned in the introduction, a key purpose of the current meta-analysis was to estimate the brain volume and intelligence correlation that was not attenuated due to range restriction in intelligence test scores. The Pietschnig et al. (2015) meta-analysis did not include the standard deviations associated with the cognitive ability test scores, consequently, we searched for the standard deviations within all of the relevant empirical research papers. In cases where the standard deviation was not reported in the empirical research paper, the author(s) of the paper were contacted via email by the first author to obtain the information via personal communication.

The range restriction formula applied in this investigation requires both the sample standard deviation and the population standard deviation (Case II; Thorndike, 1949). For most of the investigations, the population standard deviation was easy to identify (e.g., Wechsler scales, $SD = 15$; Raven's, $SD = 15$; Culture Fair Intelligence Test, $SD = 16$). However, for two of the published studies that used the Standard Progressive Matrices, the raw score standard deviations were reported. Unfortunately, the Raven's technical manual (Raven, Raven, & Court, 1998a) does not report any normative sample standard deviations for the raw scores. However, the summary guide for Australian users reported a raw score standard deviation of 7.5 for Australian 17-year-olds who completed the SPM (Australian Council for Educational Research, 1991). Thus, the value of 7.5 was used in this investigation as the SPM population level standard deviation for the purposes of correcting the observed correlations which used the SPM. One study (i.e., Thoma et al., 2005) included in the current meta-analysis reported a raw score standard deviation for the Advanced Progressive Matrices. Raven, Raven, and Court (1998b) reported a normative sample standard deviation of 6.56 for the Advanced Progressive Matrices. Consequently, the value of 6.56 was used to correct the brain volume and intelligence correlation. Burgaleta et al. (2012) reported a correlation between brain volume and intelligence assessed using a combination of tests, several of which were based on only a subset (half) of the items of the full test (i.e., difficult to find norms). Fortunately, the PMA Inductive Reasoning subtest was used in its entirety in Burgaleta et al. (2012), and the standard deviation was reported at 4.54. To estimate the degree of range restriction in the data, the PMA Inductive Reasoning standard deviation reported for the Seattle Longitudinal Study (i.e., $SD = 7.4$; Schaie, 2013) was utilised to correct the correlation between brain volume and intelligence reported in Burgaleta et al. (2012). Finally, Royle et al. (2013) reported only the raw score standard deviations for the six WAIS-III subtests administered to measure intelligence. The standardized standard deviations (expected $SD = 3.0$) were obtained via personal communication (T. Booth, personal communication, October 26, 2016).

2.3. Data analysis

The core of the meta-analysis was performed via the 'metafor' (command: rma.uni) package (version 1.9–9) developed for *R*. The data were first examined for the possible effects of influential correlations and/or sample sizes (via the 'influence' command within the 'metafor' package). Specifically, the 32 study correlations and sample sizes were examined in accordance with the nine outlier evaluation statistics described by Viechtbauer and Cheung (2010). In the event that one or more studies were identified as potential outliers/influential cases, the relevant data points were Winsorized (reduced/increased in magnitude to 1 +/1- the next largest/smallest data point in the distribution), in order to reach a balance between modulating the influence of an influential study in a valid manner and maintaining statistical power (Johnson & Eagly, 2014; Lipsey & Wilson, 2001; Macnamara, Hambrick, & Oswald, 2014). Furthermore, after the application of Winsorization, a leave-one-out analysis was, nonetheless, performed to determine the degree to which one or more effect sizes may have exerted an unusually large influence on the results (Borenstein, Hedges, Higgins, & Rothstein, 2009; Kepes, McDaniel, Brannick, & Banks, 2013).

To establish a baseline to test our hypotheses, a "bare bones" meta-analysis (Hunter & Schmidt, 2004) was conducted on the uncorrected correlations and the "HS" (Hunter Schmidt) estimation method for random effects (Viechtbauer, 2010, 2016a, 2016b). As Pearson correlations are known to be biased negatively slightly, the bare bones meta-analysis was conducted on the transformed (Olkin & Pratt, 1958) correlations via the "UCOR" function with reference to the 'metafor' and 'gsl' packages. Heterogeneity was tested statistically with Cochran's *Q*. However, given Cochran's *Q* is substantially affected by statistical power (von Hippel, 2015), emphasis was placed on the interpretation of I^2, the proportion of the variance in the correlations that was due to heterogeneity.

In accordance with contemporary recommendations, a multi-strategy was used to evaluate the possibility of publication bias (Kepes, Banks, McDaniel, & Whetzel, 2012). Specifically, the possibility of publication bias was evaluated with a series of relatively well-known analyses: funnel plot, a contour-enhanced funnel plot, and a funnel plot with trim and fill. Additionally, Egger's regression test of funnel plot asymmetry was also performed (model: weighted regression; standard error as predictor; Egger, Smith, Schneider, & Minder, 1997). Furthermore, a series of relatively modern approaches to the evaluation of publication bias were performed. Specifically, Henmi and Copas (2010) 95% effect size confidence intervals were estimated, as they have been demonstrated to be more accurate in the presence of publication of bias. The Henmi and Copas (2010) confidence intervals were obtained from the HC function within the 'metafor' package (Viechtbauer, 2016a). Additionally, a *p*-curve analysis was performed to evaluate the possibility of *p*-hacking, in accordance with the guidelines recommended by (Simonsohn, Simmons, & Nelson, 2015). The *p*-curve results were obtained from the *p*-curve web application 4.05 (http://www.p-curve.com/app4/). Finally, the *p*-curve analysis was complimented with a *p*-uniform analysis, as the *p*-uniform analysis provides a publication bias adjusted meta-analytic effect size estimate (van Aert, Wicherts, & van Assen, 2016). The *p*-uniform analysis results were obtained from the *p*-uniform web application 1.0 (https://rvanaert.shinyapps.io/p-uniform).[3]

Next, the observed correlations were corrected for range restriction on *X* (i.e., intelligence), based on the well-known Thorndike (1949) case II formula, in order to conduct the psychometric meta-analysis (Hunter & Schmidt, 2004).[4] Although the case II formula is theoretically most appropriate for scenarios where range restriction is direct, the more advanced approaches to indirect correction (e.g., Le & Schmidt, 2006) default to the direct range restriction case, when information on the reliability of the test scores is either not available or presumed to be near 1.0 (Card, 2015). In this investigation, information on the

[3] A *p*-curve analysis requires the selection of only one effect size from a single sample (Simonsohn et al., 2015). With respect to 6 of the 32 sample correlations included in this meta-analysis, more than one cognitive ability subtest was administered, however, a brain volume and intelligence composite score correlation was not reported. The six investigations included: Hogan et al., 2011; MacLullich et al., 2002; Raz et al., 1993, 2008; Schoenemann et al., 2000. Of the six investigations in question, three of the brain volume and intelligence correlations included in the current investigation were based on the Raven's Progressive Matrices (either standard or advanced) and three of the correlations were based on the Culture Fair Intelligence Test. Raven's and the CFIT are typically considered the best single test measures of *g*. Thus, in cases where a brain volume and composite cognitive ability score was not available, the expected highest *g* loading subtest correlation was selected for inclusion in the analysis.

[4] The rma.uni command with "method = HS" estimation within the metafor package in *R* yields a type of Hunter-Schmidt psychometric meta-analysis. However, it should be noted that the HS method within the metafor package will not necessarily yield exactly the same results as other software dedicated to the Hunter-Schmidt method (Viechtbauer, 2016a, 2016b).

reliability of brain volume and intelligence scores was unavailable for almost all of the investigations that met the inclusion criteria. Thus, reliability of test scores was not considered within the context of the current psychometric meta-analysis. Duan and Dunlap (1997) found that Kelley's (1923) standard error formula was the most accurate when the corrected correlation was relatively small ($\leq .30$) and the selection ratio was relatively large ($\geq .80$), which was the circumstance for most empirical studies included in the current investigation. Thus, Kelley (1923) formula was used in the psychometric meta-analysis to estimate the range corrected correlation standard errors. Publication bias analyses were not performed on the correct correlations, as corrected correlations were not published in the literature.

In order to conduct the meta-regression, a conventional meta-regression approach was adopted (Huizenga et al., 2011). Specifically, the 'rating' variable was entered into the meta-analysis model. The 'HS' method within the 'metafor' package for *R* was applied (mixed-effects estimation). The observation of a statistically significant and positive regression coefficient was considered supportive of the hypothesis that measurement quality moderated the association between brain volume and intelligence in the hypothesized direction. Finally, a statistically significant moderator effect was followed-up with separate (basic) meta-analyses for each rating group, as recommended by Field (2013).

3. Results

3.1. Meta-analysis: bare bones

The individual study statistical results are reported in Table 2. It can be seen that the majority (59.4%; $k = 19$) of the observed correlations between brain volume and intelligence were statistically significant ($p < .05$). Prior to the meta-analysis, the data were evaluated for outliers and influential studies. The analyses identified three influential studies: study 9 (Hogan et al., 2011), study 17 (Royle et al., 2013; sample 1), and study 18 (Royle et al., 2013; sample 2). As can be seen in Fig. 1, the DFFITS, the Cook's distance, and the DFBETAS analyses identified study 9 as an outlier/influential study, whereas the hat values identified study 9, 17, and 18 as outliers/influential studies. The sample correlations associated with each of these studies were neither particularly large nor small (see Table 2). However, each of the three studies was associated with relatively large sample sizes ($N > 225$). Thus, to reduce the relatively large influence of each of the three studies on the meta-analysis, their sample sizes were Winsorized (Lipsey & Wilson, 2001). Specifically, study 9 was recoded to $N = 104$, study 17 recoded to $N = 102$, and study 18 recoded to $N = 101$. Winsorizing the three influential studies reduced the total sample size from 2305 to 1758. The meta-analyses and meta-regression results reported below were based on the Winsorized data.[5] The publication bias analyses, however, were based on the originally published (non-Winsorized) data.

The bare bones meta-analysis of the 32 correlations ($N = 1758$) was associated with a statistically significant overall effect, $r = .29$, $p < .001$ (95%CI: .24, .33; see Fig. 2 for forest plot). Furthermore, the test of heterogeneity was not statistically significant, $Q(31) = 31.70$, $p = .432$, $I^2 = 0.0\%$ (95%CI: 0% to 55.1%). A series of one-study-removed analyses found the bare bones meta-analytic estimate of $r = .29$ to be robust, as the one-study-removed correlations ranged from $r = .28$ to $r = .30$.

3.2. Publication bias: well-known tests

Next, the possibility of publication bias was evaluated with four relatively well-known analyses. As can be seen in Fig. 3 (Panel A), 91% of the correlations (29 of 32) were within the triangular area of the funnel plot, which suggested that there was only a small amount of evidence to suggest bias in the reported effects (null expectation = 95%; Sterne et al., 2011). Correspondingly, the contour-enhanced funnel plot suggested that there was only a small amount of publication bias, as a reasonably balanced number of correlations were observed in the white and grey regions of the plot (i.e., non-significant; see Panel B, Fig. 3). The trim-and-fill analysis suggested the possibility of 6 (SE = 3.77) missing studies with relatively small effect sizes (see Panel C, Fig. 3). Based on the additional 6 trim-and-fill estimated correlations included in the meta-analysis, the adjusted correlation between brain volume and intelligence was estimated at $r = .25$, $p < .001$ (95%CI: .20, .30). As the adjusted correlation of $r = .25$ was only negligibly different to the non-Winsorized meta-analytic bare bones correlation ($r = .27$; $N = 2305$; 95%CI: .23, .31),[6] publication bias was not considered a serious threat to the validity of the bare bones meta-analysis. Correspondingly, Egger's regression test of funnel plot asymmetry was not significant, $t(30) = .77$, $p = .448$, which suggested a lack of evidence in favor of publication bias in the healthy adult samples. Thus, all four well-known tests failed to suggest the presence of appreciable publication bias.

3.3. Publication bias: modern tests

First, the Henmi and Copas (2010) test of publication bias was applied, by comparing the unadjusted and adjusted DerSimonian-Laird (DL) meta-analytic correlation confidence intervals. The DL estimator yielded 95% unadjusted confidence intervals of .24 and .33. By comparison, the Henmi and Copas adjusted DL 95% confidence intervals were estimated at .26 and .34. The small difference between the unadjusted and adjusted confidence intervals suggested the absence of publication bias.

Next, the *p*-curve analysis was performed. As can be seen in Tables 2, 19 of the published correlations were statistically significant ($p < .05$). As can be seen in Fig. 4, there was a distinctly right-tailed distribution of *p*-values, which suggested evidential value for the reported effects between brain volume and intelligence in healthy adults. Furthermore, based on a binomial test, the number of statistically significant *p*-values $< .025$ was found to be statistically significantly greater than the number of *p*-values between .026 and .050 ($p = .032$). Finally, the full *p*-curve and half *p*-curve tests (i.e., combination test; Simonsohn et al., 2015) were both statistically significant ($z = -5.34$, $p < .001$; $z = -5.26$, $p < .001$, respectively).

The *p*-curve package also estimates the mean level of statistical power associated with the statistically significant correlations. For the 19 correlations in the present analysis, mean power was 69% (95%CI: 44%, 85%). This implies that if the same 19 studies were conducted again, it would be expected that approximately 69% of the studies would replicate, exceeding substantially the null expectation of 5% for alpha = .05. Thus, all of the *p*-curve related results suggested that there was evidential value in favor of a true effect between brain volume and intelligence in healthy adults.

Finally, the possibility of publication bias was evaluated with the *p*-uniform method. The mean of the statistically significant *p*-values was $< .025$ (i.e., .013), therefore, we used the default alpha setting of .05, as recommended by van Aert et al. (2016). The result the of the *p*-uniform publication bias test failed to suggest the presence of publication bias, $z = .25$, $p = .400$. Correspondingly, the fixed-effects estimation of the correlation between brain volume and intelligence ($r = .27$; 95%CI: .23, .31) was negligibly different to the *p*-uniform adjusted correlation ($r = .26$; 95%CI: .18, .33). Thus, all three modern tests failed to suggest the presence of publication bias.

[5] The sample sizes were Winsorized with reference to the other sample sizes within each rating category. The Winsorizing of the sample sizes associated with the three influential studies did not impact the key conclusions drawn in this investigation.

[6] Recall that these publication bias results are based on the published correlations and sample sizes across all 32 samples.

Table 2
Studies included in the meta-analysis: healthy adults.

ID	Author	Tests	Rating	*N*	*SD*	σ	*r*	*t*	*p*	r_c
1	Raz et al. (1993)	CFIT	2	29	17.50	16	.22	1.17	.25149	.20
2	Tan et al. (1999)	CFIT	2	103	18.00	16	.40	4.39	.00003	.36
3	Schoenemann, Budinger, Sarich, and Wang (2000)	RSPM/RAPM	2	72	N/A	N/A	.22	1.89	.06332	.22
4	Garde, Mortensen, Krabbe, Rostrup, and Larsson (2000)	WAIS: DSy, BD	2	22	14.20	15	.22	1.01	.32522	.23
5	Garde et al. (2000)	WAIS: DSy, BD	2	46	14.20	15	.07	.47	.64389	.07
6	MacLullich et al. (2002)	RSPM	2	93	8.60	7.5	.39	4.04	.00011	.35
7	Shapleske et al. (2002)	Unknown (likely National Adult Reading Test)	2	23	9.20	15	.13	.60	.55438	.21
8	Raz et al. (2008)	CFIT	2	55	15.46	16	.18	1.33	.18850	.19
9	Hogan et al. (2011)	RSPM	2	234	7.74	7.5	.11	1.69	.09320	.11
10	Willerman, Schultz, Rutledge, and Bigler (1991)	WAIS-R: Voc, Sim, BD, PC	3	40	N/A	15	.35	2.30	.02683	.35
11	Egan et al. (1994)	WAIS-R: Com, Sim, Arith, BD, OA, DS, DSy	3	40	9.30	15	.32	2.08	.04412	.48
12	Gur et al. (1999)	WAIS-R: Voc, BD, CVLT, JLOT	3	80	13.21	15	.41	3.97	.00016	.45
13	Schottenbauer et al. (2007)	WAIS-R: Voc, BD	3	22	8.70	15	.60	3.35	.00316	.79
14	Schottenbauer et al. (2007)	WAIS-R: Voc, BD	3	35	10.50	15	.33	2.01	.05286	.45
15	Amat et al. (2008)	WAIS-R: BD, OA, Inf, DS, Voc	3	27	15.00	15	−.11	−.55	.58493	−.11
16	Shenkin, Rivers, Deary, Starr, and Wardlaw (2009)	MHT, RSPM, COWA, LM	3	99	11.00	11	.21	2.12	.03696	.21
17	Royle et al. (2013)	WAIS-III: BD, MR,LNS, DSB, SS, DSy	3	327	14.15	15	.27	5.06	.00001	.29
18	Royle et al. (2013)	WAIS-III: BD, MR,LNS, DSB, SS, DSy	3	293	14.03	15	.26	4.59	.00001	.30
19	Burgaleta et al. (2012)	RAPM, DAT AR, PMA IR, DAT VR, DAT NR, PMA Voc, PMA MR, DAT SR	3	100	4.54	7.40	.17	1.71	.09084	.27
20	Andreasen et al. (1993)	Complete WAIS-R	4	30	14.00	15	.44	2.59	.01497	.46
21	Andreasen et al. (1993)	Complete WAIS-R	4	37	14.00	15	.40	2.58	.01417	.42
22	Wickett et al. (1994)	Complete MAB	4	40	11.42	15	.40	2.66	.01055	.49
23	Paradiso, Andreasen, O'Leary, Arndt, and Robinson (1997)	Complete WAIS-R	4	62	12.20	15	.38	3.18	.00232	.45
24	Wickett, Vernon, and Lee (2000)	Complete MAB	4	68	10.91	15	.35	3.04	.00344	.46
25	Rojas et al. (2004)	Complete WAIS-R/WAIS-III	4	17	13.60	15	.31	1.26	.22593	.34
26	Thoma et al. (2005)	RAPM, Trails A, Trails B, Voc, BD, DS, VMR, COWA	4	19	6.36	6.56	.27	1.16	.26360	.28
27	Luders et al. (2007)	Complete WAIS-R FSIQ	4	62	12.53	15	.28	2.26	.02751	.33
28	Nakamura et al. (2007)	Complete WAIS-III FSIQ	4	44	16.10	15	.38	2.66	.01095	.36
29	Weniger, Lange, Sachsse, and Irle (2009)	Complete WAIS-R	4	25	14.50	15	.15	.73	.47420	.16
30	Hermann, Seidenberg, and Bell (2002)	Complete WAIS-III	4	67	13.39	15	.31	2.63	.01068	.34
31	Ashtari et al. (2011)	Complete WRAT-III	4	14	17.60	15	.57	2.40	.03333	.51
32	Kievit et al. (2011)	Complete WAIS-III	4	80	11.56	15	.29	2.67	.00907	.36

Note. Rating = quality of intelligence testing (2 = fair; 3 = good; 4 = excellent); CFIT = Culture Fair Intelligence Test; WAIS = Wechsler Adult Intelligence Scale; RSPM = Raven's Standard Progressive Matrices; RAPM = Raven's Advanced Progressive Matrices; DSy = Digit Symbol; BD = Block Design; Voc = Vocabulary; Sim = Similarities; PC = Picture Completion; Com = Comprehension; Arith = Arithmetic; OA = Object Assembly; DS = Digit Span; CVLT = California Verbal Learning Test; JLOT = Judgement of Line Orientation Test; MHT = Moray House Test; Inf = Information; COWA = Controlled Word Association Test; LM = Logical Memory; MR = Matrix Reasoning; LNS = Letter-Number Sequencing; DSB = Digit Span Backward; SS = Symbol Search; MAB = Multidimensional Aptitude Battery; VMR = Vandenberg Mental Rotation; the Burgaleta et al. (2012) *SD* corresponds to the complete PMA Inductive Reasoning subtest; the Willerman et al. (1991) correlation of .35 was reported by Willerman et al. as corrected (however, the *SD* was not reported in the article); σ = population standard deviation; r_c = correlation corrected for range restriction; WRAT = Wide Range Achievement Test; the re-classification of the WRAT as only a 'good' (rating = 3) measure of intelligence did not impact the results in a material way (see footnote 7).

3.4. Meta-analysis: psychometric

Next, the psychometric meta-analysis was conducted on the correlations corrected for range restriction (r_c; see Table 2). The 32 corrected correlations ($N = 1758$; sample sizes Winsorized) were associated with a statistically significant effect, $r = .31, p < .001$ (95%CI: .25, .38; see Fig. 5 for forest plot of corrected correlations). In contrast to the bare bones meta-analysis, the test of heterogeneity was statistically significant, $Q(31) = 64.77$, $p < .001$. Furthermore, the effect size measure of heterogeneity (I^2) was equal to 50.4% (95%CI: 19.2% to 68.6%), which implied a moderate amount of heterogeneity in the correlations (Higgins, Thompson, Deeks, & Altman, 2003). A series of one-study-removed analyses found the psychometric meta-analytic estimate of $r = .31$ to be robust, as the one-study-removed correlations ranged from $r = .31$ to $r = .32$.

It will be noted that the amount of heterogeneity associated with the effects increased from the barebones meta-analysis ($I^2 = 0\%$) to the psychometric analysis ($I^2 = 50.4\%$). The substantial amount of heterogeneity implied that there may have been a moderator that influenced the magnitude of the corrected sample correlations between brain volume and intelligence. Next, we tested the hypothesis that intelligence measurement quality is a moderator of the effect between brain volume and intelligence.

3.5. Meta-regression

The number and nature of the cognitive ability tests used in the investigations included in the meta-analysis are listed in Table 2 (see column labelled 'Tests'). It will be noted that nine of the intelligence measures were classified as fair (coded = 2), 10 were classified as good (coded = 3) and were 13 classified as excellent (coded = 4). Thus, none of the investigations included in the meta-analysis were considered to have used a poor measure of cognitive ability.

The meta-regression was conducted on the range corrected correlations (and Winsorized samples). The intelligence measurement quality rating moderator variable was found to be a statistically significant contributor to the model, $b = .08$ (95%CI: .01, .15), $z = 2.37$, $p = .028$. Thus, higher scores on the intelligence measurement quality scale were associated with larger brain volume and intelligence correlations. Specifically, a one unit increase in intelligence measurement quality was associated with, on average, a .08 increase in the corrected correlation between brain volume and intelligence.[7] Correspondingly,

[7] Ashtari et al.'s (2011) investigation (corrected $r = .51$, $N = 14$) used the WRAT-III

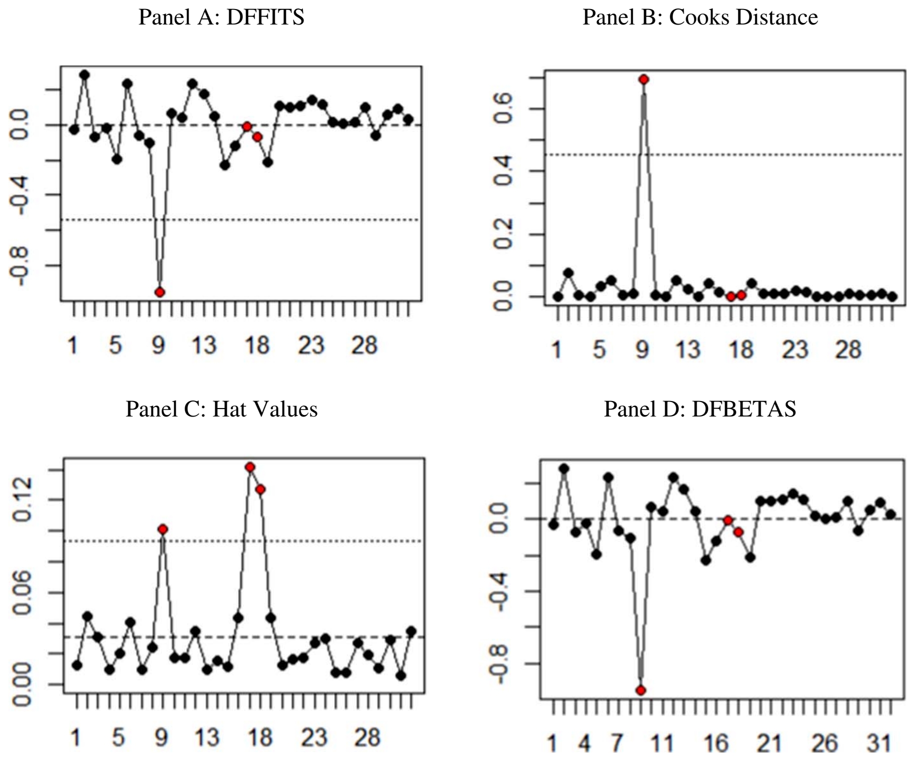

Fig. 1. Plot of four outlier diagnostic statistics; DFFITS = difference in fits; DFBETAS = difference in betas.

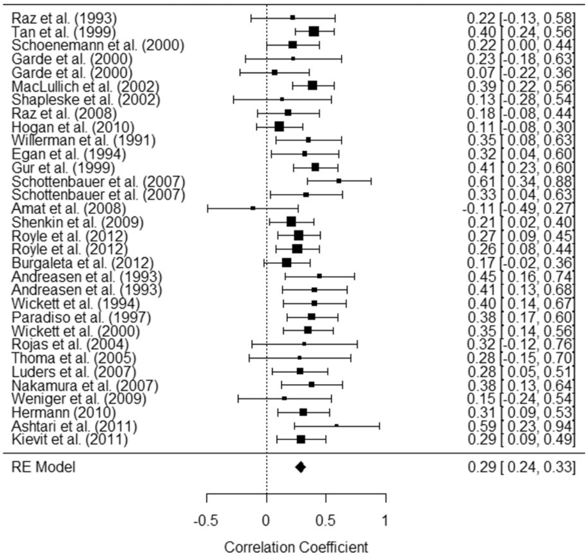

Fig. 2. Forrest plot of observed correlation coefficients; diamond represents overall effect size; square size is varied according to relative study weight within the analysis; numbers in brackets are 95% confidence intervals of point estimation; analyses were based on the Winsorized data.

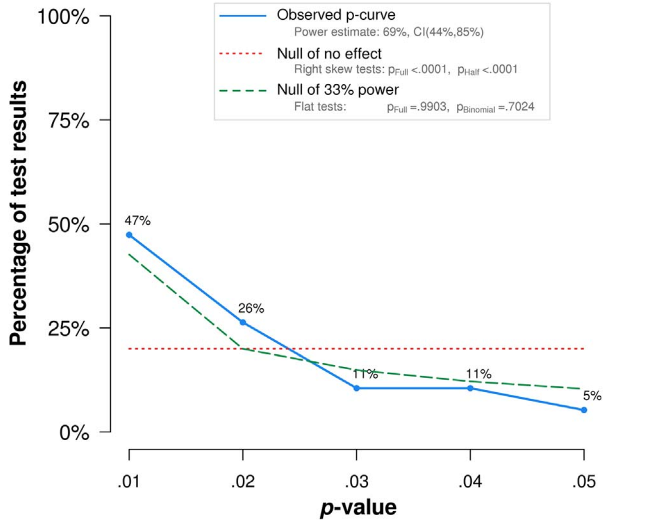

Fig. 4. Distribution of observed p-values along with the expected distribution of p-values under the null hypothesis, and if the alternative hypothesis is true but the studies are relatively underpowered (true effect, 33% power).

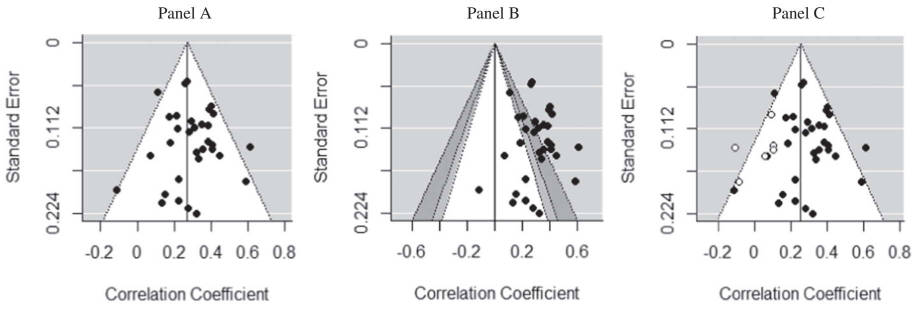

Fig. 3. Funnel plots based on observed correlations; Panel A = funnel plot; Panel B = contour-enhanced funnel plot (white region, $p > .10$; grey region, $p = .10$ to .05; dark-grey region $p = .05$ to .10; region outside funnel $p < .01$); Panel C = trim and fill funnel plot; analyses were based on the non-Winsorized data.

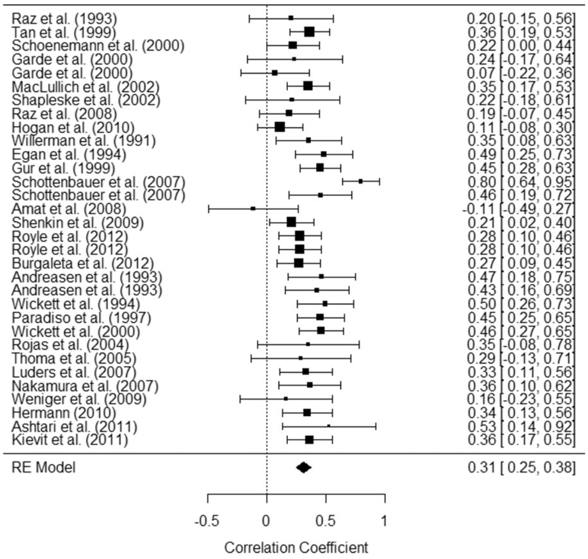

Fig. 5. Forrest plot of corrected correlation coefficients; diamond represents overall effect size; square size is varied according to relative study weight within the analysis; numbers in brackets are 95% confidence intervals of point estimation; analyses were based on the Winsorized data.

the value of I^2 was reduced to 43.4% (from 50.4%). A single study was identified as an influential case in the meta-regression (i.e., study 13, Schottenbauer, Momenan, Kerick, & Hommer, 2007, corrected $r = .79$; DFFITS = 1.00). When the meta-regression was re-run with the exclusion of the outlying correlation, the measurement quality moderator variable remained statistically significant, $b = .08$ (95%CI: .04, .14), $z = 2.94$, $p = .003$.

Separate meta-analyses were conducted to estimate the brain volume and intelligence corrected correlations across the 'fair', 'good', and 'excellent' intelligence measurement classifications. As can be seen in Table 3, the following corrected correlations were estimated: 'fair' = .21 (95%CI: .14, .28); 'good' = .32 (95%CI: .16, .46); and 'excellent' = .39 (95%CI: .32, .46).

4. Discussion

This meta-analysis indicated several findings of note regarding the association between brain volume and IQ. First, we confirmed a substantial downward bias on the effect due to sample restriction of range. Secondly, we found significant support for the influence of measurement quality on the effect sizes. Specifically, the quality of intelligence measurement was found to be a moderator of the effect between brain volume and intelligence such that investigations that used 'fair', 'good', and 'excellent' measures of intelligence yielded corrected brain volume and intelligence correlations of .23, .32, and .39, respectively. Finally, we confirmed the significant results reported in the published literature as likely the outcome of a genuine effect, as indicated in the *p*-curve analysis. These findings are discussed in more detail below.

4.1. Comparisons with previous meta-analyses

The results of this meta-analysis suggest that the association between brain volume and intelligence is at least .30, which is arguably substantially larger than the uncorrected correlation of .24 reported by Pietschnig et al. (2015). The difference in the two estimates is due, in part, to the inclusion criteria employed in this investigation: healthy adults only. Additionally, we were able to correct the included correlations for range restriction in the present investigation, whereas no corrections were applied in Pietschnig et al. (2015). The corrected $r = .31$ reported in this investigation is closely aligned with the meta-analysis reported by McDaniel (2005; $r = .33$), which included only healthy samples, in addition to corrections for range restriction in intelligence test scores.

Table 3
Key results associated with the meta-regression analyses: fair, good & excellent measures of intelligence.

	k	*N*	*M*	LB	UB
Fair	9	547	.23	.15	.31
Good	10	646	.32	.17	.46
Excellent	13	565	.39	.32	.46

Note. M = N-weighted correlations corrected for range restriction; LB = 95% confidence lower-bound; UB = 95% confidence upper-bound; analyses were based on the Winsorized data.

As contended in the introduction, people suffering from psychological and/or neurological disorders should not be expected to yield accurate estimates of intellectual functioning (Reschly et al., 2002). Additionally, there may be expected to be individual differences in the rate of developmental change across various neurophysiological characteristics, some of which may be related to be cognitive functioning (Nagy et al., 2004; Segalowitz & Davies, 2004). Unless all of those neurophysiological characteristics are correlated perfectly with brain volume, the correlations between brain volume and intelligence based on child and adolescent samples would be expected to be suppressed, if not fully, at least partly. Consequently, it is our position that the correlation of .31, based on healthy adults, is a less confounded estimate of the association between brain volume and intelligence, in comparison to the correlation of .24 reported by Pietschnig et al. (2015), which included a mixture of healthy and clinical samples, as well as children, adolescents, and adults.

Based on a quantitative review of a large number of meta-analyses in the field of differential psychology, Gignac and Szodorai (2016) found that the median observed correlation reported in the literature was .19. Thus, the observed correlation of .29 between brain volume and intelligence reported in this meta-analysis may be considered relatively large (75th percentile; Gignac et al., 2016). The corrected correlation of .31 reported in this investigation corresponds to between the 60th and 65th percentile (Gignac et al., 2016). Larger corrected meta-analytic correlations have been reported in the area of intelligence. For example, Roth et al. (2015) reported a psychometric meta-analytic correlation of .54 between intelligence and school grades. However, to-date, brain volume and intelligence appear to be the largest neurophysiological correlate of human intelligence (Ritchie et al., 2015).

4.2. Intelligence test quality as a moderator

To our knowledge, this is the first meta-analysis to use intelligence measurement quality as a moderator in a meta-analysis. The results were consistent with our hypothesis: there was a positive association between the magnitude of the association between brain volume and intelligence and the quality of general intelligence measurement. Specifically, the mean corrected correlations across the fair, good, and excellent general intelligence measurement classifications were .23, .32, and .39, respectively. In our view, the corrected .39 correlation may be the most valid representation of the association between brain volume and intelligence, as it represents the "best of" studies, at least with respect to intelligence measurement. It may be assumed that the brain volume scores obtained from the participants across the studies included in this meta-analysis were associated with less than perfect

(*footnote continued*)
(i.e., a complete multi-subtest achievement inventory). In a robustness analysis, the reclassification of the WRAT-III as only a 'good' measure of intelligence did not impact the results in a material way. For example, the moderator analysis revealed essentially the same results, $b = .08$ (95%CI: .01, .15), $z = 2.08$, $p = .037$.

levels of reliability and/or validity. Thus, the application of corrections based on the psychometrics associated with the brain volume imaging scores (e.g., reliability; range restriction), and/or the incorporation of brain imaging measurement quality in a meta-regression, may be expected to further increase the estimated correlation between brain volume and intelligence. Further research in this area is encouraged.

The observation of a positive association between measurement quality and effect size is broadly consistent with Feinstein's (1995) view that not all empirical investigations should be considered equal in the context of a meta-analysis. That is, a meta-analysis can help overcome the problem of sampling variability, however, the inclusion of all empirical studies, without any regard for the quality of measurement, may not be the most valid approach to the estimation of the association between two theoretically linked variables. Strong inclusionist versus exclusionist stances are arguably not necessary (see Kraemer, Gardner, & Yesavage, 1998), as classifications of measurement quality can be generated and hypotheses of moderator effects tested, as conducted in this investigation. Thus, researchers in the area of intelligence are encouraged to employ the general intelligence measurement classification reported in Table 1 in future meta-analyses.

It may be presumed that researchers who administer a small number of cognitive ability tests do so because of limited amount of resources (time/money). However, the results of this investigation suggest that researchers who administer more comprehensive cognitive ability test batteries require smaller sample sizes to achieve the same level of power. For example, with respect to the uncorrected correlations, an investigator who planned to administer a single cognitive ability test, such as Raven's or the CFIT (20-minute testing time), would require a sample size of 146 to achieve power of .80, based on an expected correlation of .23. By contrast, an investigator who planned to administer 9 cognitive ability tests (40-minute testing time) would require a sample size of 49 to achieve a power of .80, based on expected correlation of .39. From this perspective, it is more efficient to administer a 40-minute comprehensive measure of intelligence across 49 participants (32.7 h of total IQ testing time), in comparison to a relatively brief 20-minute measure across 146 participants (48.7 h of total IQ testing time). Furthermore, the insights derived from an investigation which included a comprehensive measure of intelligence may be considered a more valuable contribution to the area (e.g., better scope to decompose unique effects across *g* and group-level factors).

4.3. Evidential value

The results associated with the *p*-curve analysis suggested the statistically significant correlations associated with the brain volume and intelligence literature (broadly defined) are likely not substantially the consequence of *p*-hacking. Specifically, the *p*-curve analysis found that there was a statistically significantly greater proportion of statistically significant *p*-values $< .025$, in comparison to between .026 and .049.

Despite the clear results obtained from the *p*-curve analysis in this in this investigation, it should nonetheless be acknowledged that a *p*-curve analysis has not been found necessarily to yield valid results, with respect to evidential value. For example, the validity of *p*-curve analysis results has been argued to be inversely related to the amount of heterogeneity associated with the effect sizes (van Aert et al., 2016). As the amount of heterogeneity associated with the uncorrected correlations was relatively small in this investigation ($I^2 = 12.4\%$), the *p*-curve analysis results may be considered valid. More problematic, the interpretation of the right-skewed distribution associated with a *p*-curve analysis additionally assumes the absence of parallel *p*-hacking (Ulrich & Miller, 2015) and the absence of gradual publication bias (Ulrich and Miller, 2017): assumptions which cannot be tested currently. Thus, the *p*-curve results reported in this investigation should be interpreted with caution, as there are several threats to validity which cannot be ruled out easily (van Aert et al., 2016).

In addition to the above, it should also be acknowledged that the *p*-curve approach to the evaluation of evidential value and publication bias is only one approach among several (see McShane et al., 2016, for review). Some simulation research suggests that some approaches may be expected to perform more validly under certain conditions (McShane et al., 2016). From a non-statistical perspective, we note that a large percentage of studies included in the current meta-analysis were based on FSIQ type composite scores, rather than subtest scores. The consistency in the dependent variable, as opposed to a mixture of subtest scores, is a contra-indicator of *p*-hacking (van Aert et al., 2016).

5. Limitations

Although the observed correlations included in the meta-analysis were corrected for range restriction, they were not corrected for measurement error. Thus, the current meta-analysis may not be regarded as an entirely complete psychometric meta-analysis, as a complete psychometric meta-analysis should correct the observed correlations for both range restriction and measurement error (Schmidt & Hunter, 2015). The reason the observed correlations were not corrected for measurement error is that only one investigation included in the meta-analysis reported any information about the internal consistency reliability of the intelligence test scores (i.e., Wickett, Vernon, & Lee, 1994).

It may be presumed that many researchers rely upon the very high internal consistency reliability estimates reported by test publishers in the relevant technical manuals. However, reliability is a property of test scores derived from a particular sample, rather than a property of a test (Mehrens & Lehmann, 1991). Furthermore, in practice, test score reliability tends to be lower in empirical investigations, in comparison to the estimates derived from normative samples (Vacha-Haase, Kogan, & Thompson, 2000). In light of the above, it is reasonable to suggest that the corrected brain volume and intelligence correlations reported in this investigation are underestimates of the true score effect in the population. Thus, the corrected brain volume and intelligence correlation of .39 reported for the excellent intelligence measures category is almost certainly .40 or greater at the true score level.

Although a substantial amount of the theoretical and empirical literature was taken into consideration in the development of the general intelligence measurement classification system (Table 1), it should be acknowledged that it is ultimately a subjective guide. Some may raise objections about one or more of the boundaries which demarcate one or more of the categories. Naturally, different classification systems may result in moderator effects different to those reported in this meta-analysis. Thus, the results of the meta-regressions reported in this investigation are valid to the degree that the classification system is also valid. The fact that the application of the intelligence measurement classification system yielded a statistically significant hypothesized moderator effect in the meta-regressions suggests that the classification system may be valid. Additional applications of the classification system in other meta-analyses in the area of intelligence would be valuable to further evaluate its validity (or to suggest modifications).[8]

Finally, the valid interpretation of the moderator effect obtained in this investigation assumes that the empirical investigations classified across the measurement quality categories do not differ along another

[8] We attempted to conduct additional meta-regressions on the remaining correlations within the Pietschnig et al. (2015) meta-analysis (i.e., outside the healthy adult samples). However, there were too few usable correlations within any particular category to evaluate a measurement quality moderator effect, properly. Specifically, with respect to the 31 healthy children sample correlations included in Pietschnig et al. (2015), 15 were based on a combination of different IQ tests within the same sample (e.g., some children were administered an incomplete version of the WISC-R and some were administered the complete WISC-III). Additionally, four of the healthy children studies used an 'unknown' measure of intelligence. Thus, in total, only 13 of the healthy children sample correlations were considered classifiable. For thoroughness, we note that the bare bones meta-analysis based on the 31 healthy child observed score correlations was $r = .23$ ($N = 1954$; 95%CI: .16, .31).

dimension that is related positively to the quality of general intelligence measurement classifications. For example, investigations which included a comprehensive measure of intelligence may have employed test administrators with a substantial amount of testing experience, whereas those investigations which administered a single cognitive ability test may have used test administrators with little to no psychometric experience. Such possible differences may have affected test score quality in a systematic fashion.

6. Conclusion

There is almost undoubtedly a true, positive association between brain volume and intelligence, and the magnitude of this effect is likely large, relative to typically reported correlations in the individual differences literature (Gignac & Szodorai, 2016). Researchers should now focus on why this association exists. Arguably, the best insights into the mechanisms of neurophysiology and intelligence will be achieved by investigations which include excellent neurophysiological indicators and excellent measures of intelligence.

References

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.

Andreasen, N. C., Flaum, M., Swayze, V., O'Leary, D. S., Alliger, R., Cohen, G., ... Yuh, W. T. (1993). Intelligence and brain structure in normal individuals. *American Journal of Psychiatry*, *150*, 130–134.

Ashtari, M., Avants, B., Cyckowski, L., Cervellione, K. L., Roofeh, D., Cook, P., ... Kumra, S. (2011). Medial temporal structures and memory functions in adolescents with heavy cannabis use. *Journal of Psychiatric Research*, *45*(8), 1055–1066.

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.

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.

Garde, E., Mortensen, E. L., Krabbe, K., Rostrup, E., & Larsson, H. B. (2000). Relation between age-related decline in intelligence and cerebral white-matter hyperintensities in healthy octogenarians: A longitudinal study. *The Lancet*, *356*(9230), 628–634.

Hermann, B. P., Seidenberg, M., & Bell, B. (2002). The neurodevelopmental impact of childhood onset temporal lobe epilepsy on brain structure and function and the risk of progressive cognitive effects. *Progress in Brain Research*, *135*, 429–438.

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.

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.

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.

MacLullich, A. M. J., Ferguson, K. J., Deary, I. J., Seckl, J. R., Starr, J. M., & Wardlaw, J. M. (2002). Intracranial capacity and brain volumes are associated with cognition in healthy elderly men. *Neurology*, *59*(2), 169–174.

Nakamura, M., Nestor, P. G., McCarley, R. W., Levitt, J. J., Hsu, L., Kawashima, T., ... Shenton, M. E. (2007). Altered orbitofrontal sulcogyral pattern in schizophrenia. *Brain*, *130*(3), 693–707.

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.

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.

Royle, N. A., Booth, T., Hernández, M. C. V., Penke, L., Murray, C., Gow, A. J., ... Wardlaw, J. M. (2013). Estimated maximal and current brain volume predict cognitive ability in old age. *Neurobiology of Aging*, *34*(12), 2726–2733.

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.

Schottenbauer, M. A., Momenan, R., Kerick, M., & Hommer, D. W. (2007). Relationships among aging, IQ, and intracranial volume in alcoholics and control subjects. *Neuropsychology*, *21*(3), 337–345.

Shapleske, J., Rossell, S. L., Chitnis, X. A., Suckling, J., Simmons, A., Bullmore, E. T., ... David, A. S. (2002). A computational morphometric MRI study of schizophrenia: Effects of hallucinations. *Cerebral Cortex*, *12*(12), 1331–1341.

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*, 12.

Tan, Ü., 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.

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.

Wickett, J. C., 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.

Wickett, J. C., Vernon, P. A., & Lee, D. H. (2000). Relationships between factors of intelligence and brain volume. *Personality and Individual Differences*, *29*(6), 1095–1122.

Willerman, L., Schultz, R., Rutledge, J. N., & Bigler, E. D. (1991). In vivo brain size and intelligence. *Intelligence*, *15*(2), 223–228.

van Aert, R. C., Wicherts, J. M., & van Assen, M. A. (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.

Appel, L. J. (2009). The case for population-wide salt reduction gets stronger. *BMJ*, *339*, b4980.

Arthur, W., & Day, D. V. (1994). Development of a short form for the Raven Advanced Progressive Matrices test. *Educational and Psychological Measurement*, *54*(2), 394–403.

Australian Research Council for Educational Research (1991). *Raven's progressive matrices: Summary guide for Australian users*. Hawthorn, Australia: ACER.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to metaanalysis*. London: Wiley.

Burns, N. R., Nettelbeck, T., & McPherson, J. (2009). Attention and intelligence: A factor analytic study. *Journal of Individual Differences*, *30*(1), 44–57.

Card, N. A. (2015). *Applied meta-analysis for social science research*. New York: Guilford Publications.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports *g* and about ten broad factors. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 5–21). New York: Pergamon Press.

Colom, R., Juan-Espinosa, M., Abad, F., & García, L. F. (2000). Negligible sex differences in general intelligence. *Intelligence*, *28*(1), 57–68.

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.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634.

Feinstein, A. R. (1995). Meta-analysis: Statistical alchemy for the 21st century. *Journal of Clinical Epidemiology*, *48*(1), 71–79.

Field, A. P. (2013). Meta-analysis in clinical psychology research. In J. S. Comer, & P. C. Kendall (Eds.), *The Oxford handbook of research strategies for clinical psychology* (pp. 317–335). New York, NY: Oxford University Press.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Social science. Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505.

Furr, M. (2011). *Scale construction and psychometrics for social and personality psychology*. Thousand Oaks, CA: Sage.

Gignac, G. E. (2014). Fluid intelligence shares closer to 60% of its variance with working memory capacity and is a better indicator of general intelligence. *Intelligence*, *47*, 122–133.

Gignac, G. E. (2015). Raven's is not a pure measure of general intelligence: Implications for g factor theory and the brief measurement of g. *Intelligence*, *52*, 71–79.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78.

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

Gignac, G. E., Shankaralingam, M., Walker, K., & Kilpatrick, P. (2016). Short-term memory for faces relates to general intelligence moderately. *Intelligence*, *57*, 96–104.

Golden, C. J. (1978). Stroop color and word test. *A manual for clinical and experimental uses*. Wood Dale, IL: Stoelting Company.

Gur, R. C., Turetsky, B. I., Matsui, M., Yan, 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. *The Journal of Neuroscience*, *19*(10), 4065–4072.

Henmi, M., & Copas, J. B. (2010). Confidence intervals for random effects meta-analysis and robustness to publication bias. *Statistics in Medicine*, *29*(29), 2969–2983.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*(7414), 557–560.

von Hippel, P. T. (2015). The heterogeneity statistic I^2 can be biased in small meta-analyses. *BMC Medical Research Methodology*, *15*, 35.

Huizenga, H. M., Visser, I., & Dolan, C. V. (2011). Testing overall and moderator effects in random effects meta-regression. *British Journal of Mathematical and Statistical Psychology*, *64*(1), 1–19.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Newbury Park, CA: Sage.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, *91*(3), 594–612.

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

Johnson, B. T., & Eagly, A. H. (2014). Meta-analysis of research in social and personality psychology. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 675–707). (2nd ed.). London: Cambridge University Press.

Juan-Espinosa, M., Cuevas, L., Escorial, S., & García, L. F. (2006). Testing the indifferentiation hypothesis during childhood, adolescence, and adulthood. *The Journal of Genetic Psychology*, *167*(1), 5–15.

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

Kepes, S., Banks, G. C., McDaniel, M., & Whetzel, D. L. (2012). Publication bias in the organizational sciences. *Organizational Research Methods*, *15*(4), 624–662.

Kepes, S., McDaniel, M. A., Brannick, M. T., & Banks, G. C. (2013). Meta-analytic reviews in the organizational sciences: Two meta-analytic schools on the way to MARS (the Meta-analytic Reporting Standards). *Journal of Business and Psychology*, *28*, 123–143.

Kraemer, H. C., Gardner, C., Brooks, J. O., III, & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods*, *3*(1), 23–31.

Kvist, A. V., & Gustafsson, J. E. (2008). The relation between fluid intelligence and the general factor as a function of cultural background: A test of Cattell's investment theory. *Intelligence*, *36*(5), 422–436.

Le, H., & Schmidt, F. L. (2006). Correcting for indirect range restriction in meta-analysis: Testing a new meta-analytic procedure. *Psychological Methods*, *11*(4), 416–438.

Leonard, C. M., Kuldau, J. M., Breier, J. I., Zuffante, P. A., Gautier, E. R., Heron, D. C., & DeBose, C. A. (1999). Cumulative effect of anatomical risk factors for schizophrenia: An MRI study. *Biological Psychiatry*, *46*(3), 374–382.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Mackintosh, N. J. (2011). *IQ and human intelligence*. Oxford, UK: Oxford University Press.

Macnamara, B. N., Hambrick, D. Z., & Oswald, F. L. (2014). Deliberate practice and performance in music, games, sports, education, and professions a meta-analysis. *Psychological Science*, *25*(8), 1608–1618.

Major, J. T., Johnson, W., & Bouchard, T. J. (2011). The dependability of the general factor of intelligence: Why small, single-factor models do not adequately represent g. *Intelligence*, *39*(5), 418–433.

McDaniel, M. A. (2005). Big-brained people are smarter: A meta-analysis of the relationship between in vivo brain volume and intelligence. *Intelligence*, *33*(4), 337–346.

McLeod, B. D., Weisz, J. R., & Wood, J. J. (2007). Examining the association between parenting and childhood depression: A meta-analysis. *Clinical Psychology Review*, *27*(8), 986–1003.

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.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). Fort Worth, TX: Holt, Rinehart and Winston.

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of “far transfer” evidence from a meta-analytic review. *Perspectives on Psychological Science*, *11*(4), 512–534.

Nagy, Z., Westerberg, H., & Klingberg, T. (2004). Maturation of white matter is associated with the development of cognitive functions during childhood. *Journal of Cognitive Neuroscience*, *16*(7), 1227–1233.

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

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.

Paus, T. (2005). Mapping brain maturation and cognitive development during adolescence. *Trends in Cognitive Sciences*, *9*(2), 60–68.

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 & Biobehavioral Reviews*, *57*, 411–432.

Raven, J., Raven, J. C., & Court, J. H. (1998a). *Manual for Raven's progressive matrices and vocabulary scales. Section 3, the standard progressive matrices*. (Oxford, England: Oxford Psychologists Press/San Antonio, TX: The Psychological Corporation).

Raven, J., Raven, J. C., & Court, J. H. (1998b). *Manual for Raven's progressive matrices and vocabulary scales. Section 4, the advanced progressive matrices*. (Oxford, England: Oxford Psychologists Press/San Antonio, TX: The Psychological Corporation).

Reschly, D. J., Myers, T. G., & Hartel, C. R. (2002). *Mental retardation: Determining eligibility for social security benefits*. Washington, DC: National Academy Press, 69–140.

Reynolds, M. R., & Keith, T. Z. (2007). Spearman's law of diminishing returns in hierarchical models of intelligence for children and adolescents. *Intelligence*, *35*(3), 267–281.

Reynolds, M. R., Floyd, R. G., & Niileksela, C. R. (2013). How well is psychometric g indexed by global composites? Evidence from three popular intelligence tests. *Psychological Assessment*, *25*(4), 1314–1321.

Ritchie, S. J., Booth, T., Hernández, M. D. C. V., Corley, J., Maniega, S. M., Gow, A. J., ... Bastin, M. E. (2015). Beyond a bigger brain: Multivariable structural brain imaging and intelligence. *Intelligence*, *51*, 47–56.

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, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, *53*, 118–137.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, *85*(1), 112–118.

Sattler, J. M., & Ryan, J. J. (2009). *Assessment with the WAIS–IV*. San Diego, CA: Sattler.

Schaie, K. W. (2013). *Developmental influences on adult intelligence*. New York: Oxford University Press.

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage publications.

Segalowitz, S. J., & Davies, P. L. (2004). Charting the maturation of the frontal lobe: An electrophysiological strategy. *Brain and Cognition*, *55*(1), 116–133.

Simmons, J., & Simonsohn, U. (2017). Power posing: P-curving the evidence. *Psychological Science*, *28*(5), 687–693.

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

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). p-Curve and effect size correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666–681.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). *Journal of Experimental Psychology. General*, *144*(6), 1146–1152.

Sterne, J. A., Sutton, A. J., Ioannidis, J. P., Terrin, N., Jones, D. R., Lau, J., & Tetzlaff, J. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, *343*, d4002.

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

Ulrich, R., & Miller, J. (2015). *p*-Hacking by post hoc selection with multiple opportunities: Detectability by skewness test? Comment on Simonsohn, Nelson, and Simmons (2014). *Journal of Experimental Psychology: General*, *144*, 1137–1145.

Ulrich, R., & Miller, J. Some properties of *p*-curves, with an application to gradual publication bias. Psychological Methods (in press).

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement*, *60*(4), 509–522.

Vadillo, M. A., Gold, N., & Osman, M. (2016). The bitter truth about sugar and willpower: The limited evidential value of the glucose model of ego depletion. *Psychological Science* (0956797616654911).

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Viechtbauer, W. (2016a). Package ‘metafor’ (Version 1.9-9). Retrieved from http://cran.rproject.org/web/packages/metafor/metafor.pdf.

Viechtbauer, W. (2016b). Hunter and Schmidt method. Retrieved from http://http://www.metafor-project.org/doku.php/tips:hunter_schmidt_method.

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125.

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 and Neuroscience*, *34*(5), 383–388.

Wiberg, M., & Sundström, A. (2009). A comparison of two approaches to correction of restriction of range in correlation analysis. *Practical Assessment, Research & Evaluation*, *14*(5), 2.

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.